aws

Guide du développeur

Amazon Machine Learning

Version Latest

Amazon Machine Learning: Guide du développeur

Table of Contents

Nous ne mettons plus à jour le service Amazon Machine Learning et n'acceptons plus de nouveaux utilisateurs pour celui-ci. Cette documentation est disponible pour les utilisateurs existants, mais nous ne la mettons plus à jour. Pour plus d'informations, consultez Qu'est-ce qu'Amazon Machine Learning ?

Les traductions sont fournies par des outils de traduction automatique. En cas de conflit entre le contenu d'une traduction et celui de la version originale en anglais, la version anglaise prévaudra.

Qu'est-ce qu'Amazon Machine Learning ?

Nous ne mettons plus à jour le service Amazon Machine Learning (Amazon ML) ni n'acceptons de nouveaux utilisateurs pour celui-ci. Cette documentation est disponible pour les utilisateurs existants, mais nous ne la mettons plus à jour.

AWS fournit désormais un service robuste basé sur le cloud, Amazon SageMaker AI, afin que les développeurs de tous niveaux de compétence puissent utiliser la technologie d'apprentissage automatique. SageMaker L'IA est un service d'apprentissage automatique entièrement géré qui vous aide à créer de puissants modèles d'apprentissage automatique. Grâce à l' SageMaker IA, les data scientists et les développeurs peuvent créer et entraîner des modèles d'apprentissage automatique, puis les déployer directement dans un environnement hébergé prêt pour la production.

Pour plus d'informations, consultez la documentation sur l'SageMaker IA.

Rubriques

Concepts clés d'Amazon Machine Learning

Cette section résume les concepts clés suivants et décrit plus en détail leur utilisation dans Amazon ML :

- Sources de données contiennent des métadonnées associées aux entrées de données dans Amazon ML
- Les Modèles ML génèrent des prévisions en utilisant les tendances extraites des données d'entrée
- Les Evaluations mesurent la qualité des modèles ML
- Les Prédictions par lots génèrent de façon asynchrone des prévisions pour plusieurs observations des données d'entrée
- Les Prédictions en temps réel génèrent de façon asynchrone des prévisions pour les observations de données individuelles

Sources de données

Une source de données est un objet qui contient des métadonnées relatives à vos données d'entrée. Amazon ML lit vos données d'entrée, calcule des statistiques descriptives sur ses attributs et stocke les statistiques, ainsi qu'un schéma et d'autres informations, dans le cadre de l'objet de source de données. Amazon ML utilise ensuite la source de données pour entraîner et évaluer un modèle de ML et générer des prédictions par lots.

> **Important**
>
> Une source de données ne stocke pas de copie de vos données d'entrée. Au lieu de cela, elle stocke une référence à l'emplacement Amazon S3 où résident vos données d'entrée. Si vous déplacez ou modifiez le fichier Amazon S3, Amazon ML ne peut pas y accéder ni l'utiliser pour créer un modèle de machine learning, générer des évaluations ou générer des prédictions.

Le tableau suivant définit les termes liés aux sources de données.

Durée	Définition
Attribut	Propriété nommée unique figurant dans une observation. Dans des données tabulaires, telles que des feuilles de calcul ou des fichiers CSV (de valeurs séparées par des virgules), les en-têtes des colonnes représentent les attributs et les lignes contiennent des valeurs pour chaque attribut. Synonymes : variable, nom de variable, champ, colonne
Nom de source de données	(Facultatif) Vous permet de définir un nom lisible pour une source de données. Ces noms vous permettent de rechercher et de gérer vos sources de données dans la console Amazon ML.
Données d'entrée	Nom collectif pour toutes les observations auxquelles une source de données se réfère.
Emplacement	Emplacement des données d'entrée. Actuellement, Amazon ML peut utiliser des données stockées dans des compartiments Amazon S3, des bases de

Durée	Définition
	données Amazon Redshift ou des bases de données MySQL dans Amazon Relational Database Service (RDS).
Observation	Unité individuelle de données d'entrée. Par exemple, si vous créez un modèle d'apprentissage-machine pour détecter des transactions frauduleuses, vos données d'entrée comprennent de nombreuses observations, chacune représentant une transaction individuelle. Synonymes : enregistrement, exemple, instance, ligne
ID de ligne	(Facultatif) Un indicateur qui, s'il est spécifié, identifie dans les données d'entrée un attribut à inclure dans la prédiction en sortie. Cet attribut permet d'associer plus facilement les prédictions aux observations correspondantes. Synonymes : identifiant de ligne
Schema	Les informations nécessaires pour interpréter les données d'entrée, y compris les noms d'attribut et leurs types de données attribués, et les noms des attributs spéciaux.
Statistiques	Statistiques récapitulatives pour chaque attribut dans les données d'entrée. Ces statistiques remplissent deux fonctions : La console Amazon ML les affiche sous forme de graphiques pour vous aider à comprendre vos données at-a-glance et à identifier les irrégularités ou les erreurs. Amazon ML les utilise pendant le processus de formation afin d'améliorer la qualité du modèle de ML obtenu.
Statut	Indique l'état actuel de la source de données, tel que En cours, Terminé ou Echec.

Durée	Définition
Attribut cible	Dans le contexte de l'entraînement d'un modèle de machine learning, l'attribut cible identifie le nom de l'attribut dans les données d'entrée qui contient les « bonnes » réponses. Amazon ML l'utilise pour découvrir des modèles dans les données d'entrée et générer un modèle de machine learning. Dans le contexte de l'évaluation et de la création de prédictions, l'attribut cible est l'attribut dont la valeur sera prédite par un modèle d'apprentissage-machine formé. Synonymes : cible

Modèles ML

Un modèle ML est un modèle mathématique qui génère des prédictions en trouvant des modèles dans vos données. Amazon ML prend en charge trois types de modèles de ML : classification binaire, classification multiclasse et régression.

Le tableau suivant définit les termes liés aux modèles d'apprentissage-machine.

Durée	Définition
Régression	L'objectif de la formation d'un modèle d'apprentissage-machine de régression est de prédire une valeur numérique.
Multiclasse	L'objectif de la formation d'un modèle d'apprentissage-machine multiclasse est de prédire les valeurs appartenant à un ensemble prédéfini et limité de valeurs autorisées.
Binaire	L'objectif de former un modèle d'apprentissage-machine binaire est de prédire les valeurs qui peuvent uniquement avoir deux états différents, tels que true ou false.
Taille du modèle	Les modèles d'apprentissage-machine capturent et stockent des tendances. Plus un modèle d'apprentissage-machine stocke de tendances, plus il est volumineux. La taille du modèle d'apprentissage-machine est décrite en Mo.

Durée	Définition
Nombre de passages	Lorsque vous formez un modèle d'apprentissage-machine, vous utilisez les données d'une source de données. Il est parfois avantageux d'utiliser plusieurs fois chaque enregistrement de données dans le processus d'apprent issage. Le nombre de fois que vous autorisez Amazon ML à utiliser les mêmes enregistrements de données s'appelle le nombre de passes.
Régularisation	La régularisation est une technique d'apprentissage automatique que vous pouvez utiliser pour obtenir des modèles de meilleure qualité. Amazon ML propose un paramètre par défaut qui fonctionne bien dans la plupart des cas.

Evaluations

Une évaluation mesure la qualité de votre modèle d'apprentissage-machine et détermine s'il fonctionne correctement.

Le tableau suivant définit les termes liés aux évaluations.

Durée	Définition
Analyse du modèle	Amazon ML vous fournit une métrique et un certain nombre d'informations que vous pouvez utiliser pour évaluer les performances prédictives de votre modèle.
AUC	La métrique AUC (Area Under the ROC Curve) mesure l'aptitude d'un modèle d'apprentissage-machine binaire à prédire un score plus élevé pour les exemples positifs par rapport aux exemples négatifs.
Score F1 moyenné par macro	Le score F1 moyenné par macro est utilisé pour évaluer les performances prédictives de modèles d'apprentissage-machine multiclasses.
RMSE	L'erreur quadratique moyenne (RMSE, Root Mean Square Error) est une métrique utilisée pour évaluer les performances prédictives des modèles d'apprentissage-machine de régression.

Durée	Définition
Seuil	Les modèles d'apprentissage-machine fonctionnent en générant des scores de prédiction numériques. En appliquant une valeur seuil, le système convertit ces scores en étiquettes 0 et 1.
Précision	La précision mesure le pourcentage de prédictions correctes.
Précision	La précision montre le pourcentage d'instances positives réelles (par oppositio n aux instances positives fausses) parmi les instances récupérées (celles qui devaient être positives). En d'autres termes, combien d'éléments sélectionnés sont positifs ?
Rappel	La sensibilité montre le pourcentage d'instances positives réelles parmi le nombre total d'instances pertinentes (positives réelles). En d'autres termes, combien d'éléments positifs sont sélectionnés ?

Prédictions par lots

Les prédictions par lots s'appliquent à un ensemble d'observations qui peuvent s'exécuter en même temps. Ceci est idéal pour les analyses prédictives qui ne présentent pas d'exigence en temps réel.

Le tableau suivant définit les termes liés aux prédictions par lots.

Durée	Définition
Emplacement de sortie	Les résultats d'une prédiction par lots sont stockés dans un emplacement de sortie de compartiment S3.
Fichier manifeste	Ce fichier associe chaque fichier de données d'entrée aux résultats des prédictions par lots associées. Il est stocké dans l'emplacement de sortie de compartiment S3.

Prédictions en temps réel

Les prédictions en temps réel sont appropriées pour les applications nécessitant une faible latence, telles que les applications interactives web, mobiles ou de bureau. N'importe quel modèle

d'apprentissage-machine peut être interrogé pour établir des prédictions à l'aide de l'API de prédiction en temps réel à faible latence.

Le tableau suivant définit les termes liés aux prédictions en temps réel.

Durée	Définition
API de prédiction en temps réel	L'API de prédiction en temps réel accepte une seule observation d'entrée dans la charge utile de demande et renvoie la prédiction dans la réponse.
Point de terminaison de prédiction en temps réel	Pour utiliser un modèle d'apprentissage-machine avec l'API de prédiction en temps réel, vous devez créer un point de terminaison de prédiction en temps réel. Une fois créé, ce point de terminaison contient l'URL que vous pouvez utiliser pour demander des prédictions en temps réel.

Accès à Amazon Machine Learning

Vous pouvez accéder à Amazon ML en utilisant l'une des méthodes suivantes :

Console Amazon ML

Vous pouvez accéder à la console Amazon ML en vous connectant à la console de gestion AWS et en ouvrant la console Amazon ML à l'adresse https://console.aws.amazon.com/machinelearning/.

AWS CLI

Pour plus d'informations sur l'installation et la configuration de l'interface de ligne de commande AWS, consultez la section Getting Set Up with the AWS Command Line Interface dans le guide de AWS Command Line Interface l'utilisateur.

API Amazon ML

Pour plus d'informations sur l'API Amazon ML, consultez le manuel Amazon ML API Reference.

AWS SDKs

Pour plus d'informations sur AWS SDKs, consultez la section Outils pour Amazon Web Services.

Régions et points de terminaison

Amazon Machine Learning (Amazon ML) prend en charge les points de terminaison de prédiction en temps réel dans les deux régions suivantes :

Nom de la région	Région	Point de terminaison	Protocole
USA Est (Virginie du Nord)	us-east-1	machinelearning.us-east-1.amazonaws.com	HTTPS
Europe (Irlande)	eu-west-1	machinelearning.eu-west-1.amazonaws.com	HTTPS

Vous pouvez héberger des ensembles de données, former et évaluer des modèles, et déclencher des prédictions dans n'importe quelle région.

Nous vous recommandons de conserver toutes vos ressources dans la même région. Si vos données d'entrée se trouvent dans une région différente de celle de vos ressources Amazon ML, vous devez payer des frais de transfert de données entre régions. Vous pouvez appeler un point de terminaison de prédiction en temps réel à partir de n'importe quelle région, mais l'appel d'un point de terminaison à partir d'une région qui n'a pas le point de terminaison que vous appelez peut avoir un impact sur les latences de prédiction en temps réel.

Tarification pour Amazon ML

Avec AWS les services, vous ne payez que pour ce que vous utilisez. Aucun frais minimum ni aucun engagement initial ne s'appliquent.

Amazon Machine Learning (Amazon ML) facture un taux horaire correspondant au temps de calcul utilisé pour calculer les statistiques des données et entraîner et évaluer des modèles, puis vous payez pour le nombre de prédictions générées pour votre application. Pour des prédictions en temps réel, vous payez également des frais horaires de capacité réservée en fonction de la taille de votre modèle.

Amazon ML estime les coûts des prédictions uniquement dans la console Amazon ML.

Pour plus d'informations sur la tarification d'Amazon ML, consultez la section Amazon Machine Learning Pricing.

Rubriques

Estimation du coût des prédictions par lots

Lorsque vous demandez des prédictions par lots à un modèle Amazon ML à l'aide de l'assistant Create Batch Prediction, Amazon ML estime le coût de ces prédictions. La méthode utilisée pour calculer l'estimation varie en fonction du type de données disponible.

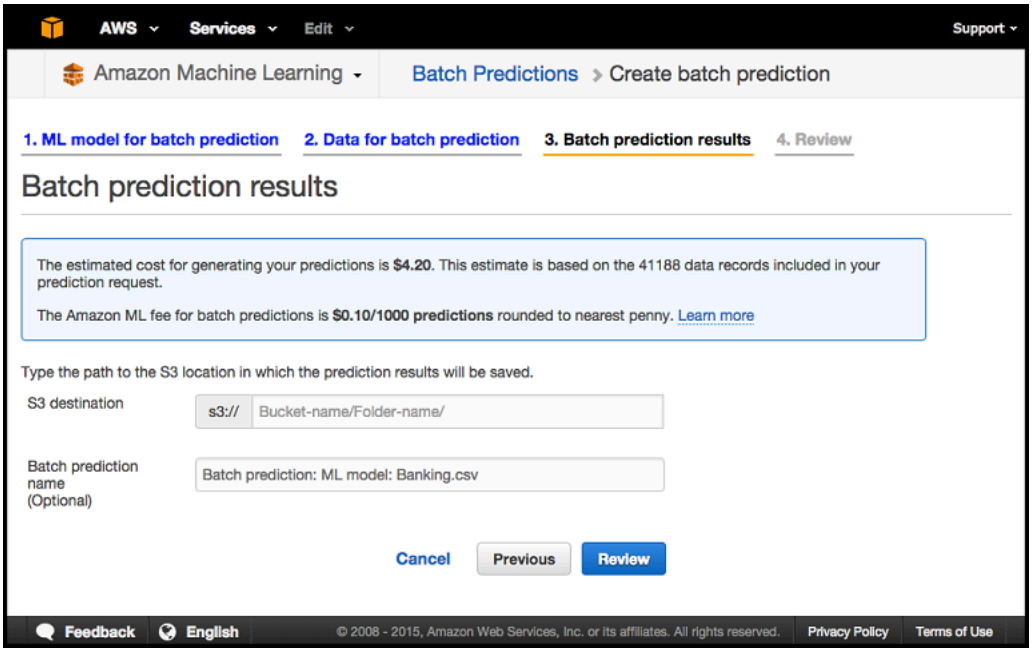
Estimation du coût des prédictions par lots lorsque les statistiques des données sont disponibles

L'estimation des coûts la plus précise est obtenue lorsqu'Amazon ML a déjà calculé des statistiques récapitulatives sur la source de données utilisée pour demander des prédictions. Ces statistiques sont toujours calculées pour les sources de données créées à l'aide de la console Amazon ML. Les utilisateurs de l'API doivent définir l'`ComputeStatistics`indicateur sur `True` lorsqu'ils créent des sources de données par programmation à l'aide du CreateDataSourceFromS3 ou du CreateDataSourceFromRedshiftRDS. CreateDataSourceFrom APIs La source de données doit être dans l'état `READY` pour que les statistiques soient disponibles.

L'une des statistiques calculées par Amazon ML est le nombre d'enregistrements de données. Lorsque le nombre d'enregistrements de données est disponible, l'assistant de prédiction Amazon ML Create Batch estime le nombre de prédictions en multipliant le nombre d'enregistrements de données par les frais associés aux prédictions par lots.

Le coût réel peut varier par rapport à cette estimation pour les raisons suivantes :

- Certains enregistrements de données peuvent ne pas être traités. Les prédictions à partir des enregistrements de données qui ont échoué ne vous sont pas facturées.
- L'estimation ne prend pas en compte les crédits préexistants ou d'autres ajustements appliqués par AWS.



Estimation du coût des prédictions par lots lorsque seul le volume des données est disponible

Lorsque vous demandez une prédiction par lots et que les statistiques de données pour la source de données demandée ne sont pas disponibles, Amazon ML estime le coût sur la base des éléments suivants :

- Le volume total des données qui est calculé et conservé pendant la validation de la source de données
- La taille moyenne des enregistrements de données, estimée par Amazon ML en lisant et en analysant les 100 premiers Mo de votre fichier de données

Pour estimer le coût de votre prédiction par lots, Amazon ML divise la taille totale des données par la taille moyenne des enregistrements de données. Cette méthode de prédiction de coût est moins précise que la méthode utilisée lorsque le nombre d'enregistrements de données est disponible, car les premiers enregistrements de votre fichier de données peut ne pas représenter avec précision la taille moyenne des enregistrements.

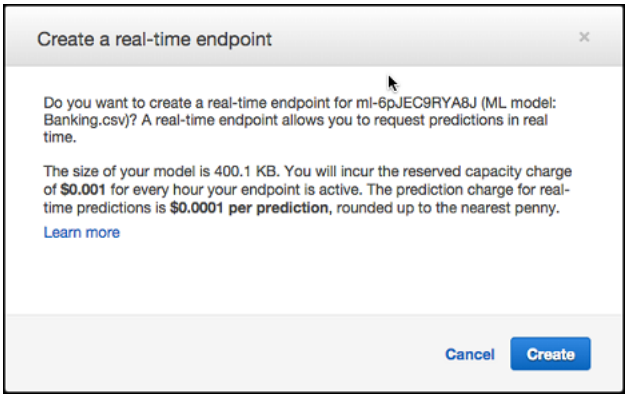
Estimation du coût des prédictions par lots lorsque ni les statistiques des données, ni le volume des données ne sont disponibles

Lorsque ni les statistiques ni la taille des données ne sont disponibles, Amazon ML ne peut pas estimer le coût de vos prévisions par lots. C'est souvent le cas lorsque la source de données que

vous utilisez pour demander des prédictions par lots n'a pas encore été validée par Amazon ML. Cela peut se produire lorsque vous avez créé une source de données basée sur une requête Amazon Redshift (Amazon Redshift) ou Amazon Relational Database Service (Amazon RDS) et que le transfert de données n'est pas encore terminé, ou lorsque la création de la source de données est mise en file d'attente après d'autres opérations sur votre compte. Dans ce cas, la console Amazon ML vous informe des frais de prédiction par lots. Vous pouvez décider de poursuivre la demande de prédiction par lots sans estimation ou d'annuler l'exécution de l'assistant et de revenir une fois que la source de données utilisée pour les prédictions est dans l'état En cours ou Prêt.

Estimation du coût des prédictions en temps réel

Lorsque vous créez un point de terminaison de prédiction en temps réel à l'aide de la console Amazon ML, les frais de capacité de réserve estimés s'affichent, qui sont des frais permanents pour la réservation du point de terminaison pour le traitement des prédictions. Ces frais varient en fonction de la taille du modèle, comme expliqué à la page de tarification du service. Vous serez également informé des frais de prédiction en temps réel standard d'Amazon ML.



Concepts d'apprentissage-machine

Le modèle d'apprentissage-machine (ML) vous permet d'utiliser les données historiques pour prendre des décisions professionnelles plus avisées. Les tendances découvertes dans les données par les algorithmes ML sont utilisées pour élaborer des modèles mathématiques. Vous pouvez ensuite utiliser ces modèles pour effectuer des prédictions sur des données à venir. Par exemple, un modèle d'apprentissage-machine pourrait permettre de prévoir la probabilité d'un achat client en fonction du comportement passé observé chez ce dernier.

Rubriques

Résolution de problèmes d'entreprise à l'aide d'Amazon Machine Learning

Vous pouvez utiliser Amazon Machine Learning pour appliquer l'apprentissage-machine dans le cadre de problèmes pour lesquels vous possédez des exemples de réponses réelles. Par exemple, pour utiliser Amazon Machine Learning afin de prédire si un e-mail correspond à du courrier indésirable, vous devez collecter des exemples d'e-mails correctement étiquetés en tant que courrier indésirable ou non. Ensuite, vous pouvez utiliser l'apprentissage-machine pour généraliser à partir de ces exemples d'e-mails, afin de prédire la probabilité selon laquelle un nouvel e-mail correspond ou non à du courrier indésirable. Cette approche d'apprentissage à partir de données ayant été étiquetées avec la réponse réelle porte le nom d'apprentissage-machine supervisé.

Vous pouvez utiliser des approches d'apprentissage-machine supervisé pour les tâches d'apprentissage-machine spécifiques suivantes : la classification binaire (prédiction du résultat entre deux résultats possibles), la classification multiclasse (prédiction du résultat entre plus de deux résultats) et la régression (prédiction d'une valeur numérique).

Exemples de problèmes de classification binaire :

- Le client achètera-t-il ce produit ou non ?
- Cet e-mail correspond-il à du courrier indésirable ou non ?

- Ce produit est-il un livre ou un animal de ferme ?
- Ce commentaire a-t-il été écrit par un client ou un robot ?

Exemples de problèmes de classification multiclasse :

- Ce produit est-il un livre, un film ou un vêtement ?
- Ce film est-il une comédie romantique, un documentaire ou un thriller ?
- Quelle catégorie de produits intéresse le plus ce client ?

Exemples de problèmes de classification de type régression :

- Quelle sera la température à Seattle demain ?
- Pour ce produit, combien d'unités se vendra-t-il ?
- Dans combien de jours ce client arrêtera-t-il d'utiliser l'application ?
- A quel prix cette maison se vendra-t-elle ?

Conditions d'utilisation de Machine Learning

Il est important de garder à l'esprit que l'apprentissage-machine n'est pas une solution pour chaque type de problème. Dans certains cas, des solutions robustes peuvent être développées sans recourir aux techniques d'apprentissage-machine. Par exemple, vous n'avez pas besoin de l'apprentissage-machine si vous pouvez déterminer une valeur cible à l'aide de règles simples, de calculs ou d'étapes prédéterminées pouvant être programmés sans apprentissage orienté données.

Utilisez l'apprentissage-machine pour les situations suivantes :

- Vous ne pouvez pas coder les règles : de nombreuses tâches humaines (telles que reconnaître si un e-mail correspond à du courrier indésirable ou non) ne peuvent pas être résolues convenablement à l'aide d'une solution simple (déterministe) basée sur des règles. Un grand nombre de facteurs peuvent influencer la réponse. Lorsque des règles dépendent d'un trop grand nombre de facteurs et qu'un grand nombre de ces règles se chevauchent ou doivent être réglées de façon très fine, il devient vite difficile pour une personne de coder précisément ces règles. Vous pouvez utiliser l'apprentissage-machine pour résoudre efficacement ce problème.
- Vous ne pouvez pas opérer à grande échelle : vous pouvez reconnaître manuellement quelques centaines d'e-mails et décider s'il s'agit de courrier indésirable ou non. Toutefois, cette tâche

devient impossible pour des millions d'e-mails. Les solutions d'apprentissage-machine sont efficaces pour le traitement des problèmes à grande échelle.

Création d'une application d'apprentissage-machine

La création d'applications d'apprentissage-machine est un processus itératif qui implique une série d'étapes. Pour créer une application d'apprentissage-machine, suivez ces étapes générales :

1. Cernez le ou les problèmes d'apprentissage-machine principaux en considérant ce qui est observé et la réponse que vous voulez que le modèle prédise.
2. Collectez, nettoyez et préparez les données pour qu'elles puissent être consommées par les algorithmes de formation de modèle d'apprentissage-machine. Visualisez et analysez les données pour exécuter des vérification d'intégrité afin de valider la qualité des données et de comprendre les données.
3. Souvent, les données brutes (variables d'entrée) et la réponse (cible) ne sont pas représentées d'une manière exploitable pour former un modèle hautement prédictif. Par conséquent, vous devez généralement essayer de construire des représentations d'entrée plus prédictives ou des entités à partir des variables brutes.
4. Fournissez les entités obtenues à l'algorithme d'apprentissage pour élaborer des modèles et évaluer la qualité de ces modèles sur les données qui ont été mises de côté lors de la création des modèles.
5. Utilisez ce modèle pour générer des prédictions de la réponse cible pour de nouvelles instances de données.

Formulation du problème

La première étape de l'apprentissage-machine consiste à décider ce que vous souhaitez prédire, également appelé « étiquette » ou « réponse cible ». Imaginez un scénario dans lequel vous souhaitez fabriquer des produits, mais votre décision concernant la fabrication de chaque produit dépend de son nombre de ventes potentielles. Dans ce scénario, vous souhaitez prédire le nombre de fois que chaque produit sera acheté (prédire le nombre de ventes). Il existe plusieurs manières de définir ce problème en utilisant l'apprentissage-machine. Le choix de la façon de définir le problème dépend de votre cas d'utilisation ou de vos besoins professionnels.

Voulez-vous prédire le nombre d'achats que vos clients effectueront pour chaque produit (auquel cas la cible est numérique et vous devez résoudre un problème de régression) ? Ou voulez-vous

prédire quels produits feront l'objet de plus de 10 achats (auquel cas la cible est binaire et vous devez résoudre un problème de classification binaire) ?

Il est important d'éviter de trop compliquer le problème et il convient de cerner la solution la plus simple qui répond à vos besoins. Toutefois, il est également important d'éviter de perdre des informations, notamment des informations dans l'historique des réponses. Ici, la conversion d'un nombre de ventes passées réelles en une variable binaire « plus de 10 » ou « moins » entraînerait la perte d'informations précieuses. Il convient d'investir du temps pour décider de la cible la plus judicieuse à prédire. Cela vous évitera d'élaborer des modèles qui ne répondent pas à votre question.

Collecte de données étiquetées

Les problèmes d'apprentissage-machine commencent avec des données, de préférence avec beaucoup de données de préférence (exemples ou observations) pour lesquelles vous connaissez déjà la réponse cible. Les données pour lesquelles vous connaissez déjà la réponse cible sont appelées données étiquetées. Dans le cadre d'un apprentissage-machine supervisé, l'algorithme enseigne à lui-même pour apprendre à partir des exemples étiquetés que nous fournissons.

Chaque exemple/observation figurant dans vos données doit contenir deux éléments :

- La cible – la réponse que vous souhaitez prédire. Vous fournissez des données qui sont étiquetées avec la cible (réponse correcte) à l'algorithme d'apprentissage-machine pour qu'il apprenne à partir d'elles. Ensuite, vous utilisez le modèle d'apprentissage-machine formé pour prédire cette réponse sur des données pour lesquelles vous ne connaissez pas la réponse cible.
- Variables/entités – ce sont des attributs de l'exemple qui peuvent être utilisés pour identifier des tendances afin de prédire la réponse cible.

Par exemple, pour le problème de classification des e-mails, la cible est une étiquette qui indique si un e-mail correspond à du courrier indésirable ou non. Comme exemples de variables, on peut citer l'expéditeur de l'e-mail, le texte dans le corps de l'e-mail, le texte dans la ligne d'objet, l'heure à laquelle l'e-mail a été envoyé et l'existence d'une correspondance antérieure entre l'expéditeur et le destinataire.

Souvent, les données ne sont pas disponibles sous une forme étiquetée. La collecte et la préparation des variables et de la cible sont souvent les étapes les plus importantes dans la résolution d'un problème d'apprentissage-machine. Les exemples de données doivent être représentatifs des données que vous aurez lorsque vous utiliserez le modèle pour établir une prédiction. Par exemple, si vous souhaitez prédire si un e-mail correspond à du courrier indésirable ou non, vous devez collecter

des positifs (courriers indésirables) et des négatifs (courriers non indésirables) pour que l'algorithme d'apprentissage-machine soit en mesure d'identifier des tendances qui permettront de distinguer les deux types d'e-mails.

Une fois que vous disposez des données étiquetées, vous pouvez être amené à les convertir dans un format acceptable par votre algorithme ou votre logiciel. Par exemple, pour utiliser Amazon ML, vous devez convertir les données au format CSV (séparé par des virgules), chaque exemple constituant une ligne du fichier CSV, chaque colonne contenant une variable d'entrée et une colonne contenant la réponse cible.

Analyse de vos données

Avant de fournir vos données étiquetées à un algorithme d'apprentissage-machine, il est recommandé d'inspecter vos données pour identifier d'éventuels problèmes et mieux connaître les données que vous utilisez. La puissance prédictive de votre modèle est proportionnelle à la qualité des données que vous lui fournissez.

Lorsque vous analysez vos données, vous devez garder à l'esprit les points suivants :

- Résumés des données variables et cibles – Il est utile de comprendre les valeurs que vos variables prennent et quelles valeurs sont dominantes dans vos données. Vous pouvez confier la réalisation de ces résumés à un spécialiste du domaine pour le problème que vous souhaitez résoudre. Demandez-vous ou demandez au spécialiste du domaine : les données correspondent-elles à vos attentes ? Avez-vous l'impression d'avoir un problème de collecte de données ? Dans votre cible, une classe est-elle plus fréquente que les autres ? Y a-t-il plus de valeurs manquantes ou non valides que ce que vous aviez prévu ?
- Corrélations variable-cible – Connaître la corrélation entre chaque variable et la classe cible est utile parce qu'une corrélation élevée implique qu'il existe une relation entre la variable et la classe cible. En général, vous voulez inclure les variables dotées d'une haute corrélation, car elles ont une puissance (signal) prédictive plus élevée, et mettre de côté les variables à faible corrélation, car elles ont peu de chances d'être pertinentes.

Dans Amazon ML, vous pouvez analyser vos données en créant une source de données et en consultant le rapport de données qui en résulte.

Traitement des entités

Une fois que vous avez appris à connaître vos données via les résumés et les visualisations, vous pouvez transformer vos variables pour les rendre plus significatives. Cela porte le nom de traitement des entités. Par exemple, supposons que vous disposez d'une variable qui capture la date et l'heure auxquelles un événement s'est produit. Cette date et cette heure ne se reproduiront jamais et ne seront donc pas utiles pour prédire votre cible. Toutefois, si cette variable est transformée en entités qui représentent l'heure de la journée, le jour de la semaine et le mois, ces variables peuvent être utiles pour savoir si l'événement a tendance à se produire à une heure particulière, un jour particulier de la semaine ou durant un mois particulier. Un tel traitement d'entités dans le but de former des points de données plus généralisables comme base d'apprentissage peut apporter des améliorations considérables aux modèles prédictifs.

Autres exemples de traitement courant d'entités :

- Remplacement des données manquantes ou non valides par des valeurs plus significatives (par exemple, si vous savez qu'une valeur manquante pour une variable de type de produit signifie en fait qu'il s'agit d'un livre, vous pouvez remplacer toutes les valeurs manquantes dans le type de produit par la valeur correspondant aux livres). Une stratégie courante utilisée pour imputer les valeurs manquantes consiste à remplacer les valeurs manquantes par la moyenne ou la valeur médiane. Il est important de comprendre vos données avant de choisir une stratégie pour remplacer les valeurs manquantes.
- Formation de produits cartésiens d'une variable avec une autre. Par exemple, si vous avez deux variables, telles que la densité de la population (urbaine, suburbaine, rurale) et l'Etat (Washington, Oregon, California), il peut y avoir des informations utiles dans les entités formées par le produit cartésien de ces deux variables, lequel se traduit par les entités (urban_Washington, suburban_Washington, rural_Washington, urban_Oregon, suburban_Oregon, rural_Oregon, urban_California, suburban_California, rural_California).
- Transformations non linéaires, telles que la discrétisation des variables numériques en catégories. Dans de nombreux cas, la relation entre une entité numérique et la cible n'est pas linéaire (la valeur de l'entité n'augmente pas et ne diminue pas de façon monotone avec la cible). Dans de tels cas, il peut être utile de discrétiser l'entité numérique en entités de catégorie représentant différentes plages de l'entité numérique. Chaque entité de catégorie (intervalle) peut ensuite être modélisée comme ayant sa propre relation linéaire avec la cible. Par exemple, supposons que vous savez que l'entité numérique continue âge n'est pas corrélée linéairement à la probabilité d'acheter un livre. Vous pouvez discrétiser l'âge en entités de catégorie susceptibles de capturer plus précisément la relation avec la cible. Le nombre optimal d'intervalles pour une variable

numérique dépend des caractéristiques de la variable et de sa relation à la cible, et la meilleure façon de le déterminer passe par l'expérimentation. Amazon ML suggère le numéro de casier optimal pour une fonctionnalité numérique sur la base des statistiques de données de la recette suggérée. Consultez le Manuel du développeur pour en savoir plus sur la recette suggérée.

- Entités spécifiques au domaine (par exemple, avec la longueur, la largeur et la hauteur comme variables séparées, vous pouvez créer une nouvelle entité volume en tant que produit de ces trois variables).
- Entités spécifiques aux variables. Certains types de variables, tels que les entités texte et les entités qui capturent la structure d'une page web ou d'une phrase, ont des méthodes génériques de traitement qui aident à extraire la structure et le contexte. Par exemple, la formation de n-grammes à partir du texte « the fox jumped over the fence » peut être représentée par des unigrammes : the, fox, jumped, over, fence, ou par des bigrammes : the fox, fox jumped, jumped over, over the, the fence.

L'insertion d'entités plus pertinentes permet d'améliorer la puissance de prédiction. De toute évidence, il n'est pas toujours possible de connaître à l'avance les entités avec un « signal » ou une influence prédictive. Il est donc judicieux d'inclure toutes les entités pouvant être associées à l'étiquette cible et de laisser l'algorithme de formation du modèle sélectionner les entités présentant les corrélations les plus fortes. Dans Amazon ML, le traitement des fonctionnalités peut être spécifié dans la recette lors de la création d'un modèle. Consultez le Manuel du développeur pour obtenir la liste des processeurs d'entités disponibles.

Fractionnement des données en données de formation et d'évaluation

L'objectif fondamental de l'apprentissage-machine est de généraliser au-delà des instances de données utilisées pour former les modèles. Nous voulons évaluer le modèle pour estimer la qualité de la généralisation des tendances pour des données avec lesquelles le modèle n'a pas été formé. Toutefois, comme les instances futures ont des valeurs cibles inconnues et que nous ne pouvons pas vérifier la précision de nos prédictions pour les instances futures, nous devons utiliser une part des données dont nous connaissons déjà la réponse comme indicateur pour les données futures. L'évaluation du modèle avec les mêmes données qui ont été utilisées pour la formation n'est pas utile. En effet, elle récompense les modèles qui peuvent « mémoriser » les données de formation, par opposition à une généralisation à partir de celles-ci.

Une stratégie courante consiste à prendre toutes les données étiquetées disponibles, et à les fractionner en sous-ensembles de formation et d'évaluation, généralement avec une proportion de 70-80 % pour la formation et de 20-30 % pour l'évaluation. Le système d'apprentissage-

machine utilise les données de formation pour former les modèles à identifier des tendances, et utilise les données d'évaluation pour évaluer la qualité prédictive du modèle formé. Le système d'apprentissage-machine évalue les performances prédictives en comparant les prédictions sur le jeu de données d'évaluation à leurs valeurs réelles (vérité de terrain) à l'aide de diverses métriques. En règle générale, vous utilisez le « meilleur » modèle sur le sous-ensemble d'évaluation pour établir des prédictions sur les instances futures pour lesquelles vous ne connaissez pas la réponse cible.

Amazon ML divise les données envoyées pour la formation d'un modèle via la console Amazon ML en 70 % pour la formation et 30 % pour l'évaluation. Par défaut, Amazon ML utilise les premiers 70 % des données d'entrée dans l'ordre dans lequel elles apparaissent dans les données source pour la source de données d'entraînement et les 30 % restants des données pour la source de données d'évaluation. Amazon ML vous permet également de sélectionner au hasard 70 % des données sources pour la formation au lieu d'utiliser les 70 % premiers et d'utiliser le complément de ce sous-ensemble aléatoire à des fins d'évaluation. Vous pouvez utiliser Amazon ML APIs pour spécifier des ratios de répartition personnalisés et pour fournir des données de formation et d'évaluation qui ont été séparées en dehors d'Amazon ML. Amazon ML propose également des stratégies pour fractionner vos données. Pour plus d'informations sur les stratégies de fractionnement, consultez Fractionnement des données.

Formation du modèle

Vous êtes maintenant prêt à fournir les données de formation à l'algorithme d'apprentissage-machine (c'est-à-dire, l'algorithme d'apprentissage). L'algorithme apprendra des données de formation les tendances mettant en correspondance les variables et la cible, et il fournira en sortie un modèle capturant ces relations. Le modèle d'apprentissage-machine peut alors être utilisé pour obtenir des prédictions sur de nouvelles données pour lesquelles vous ne connaissez pas la réponse cible.

Modèles linéaires

Il existe un grand nombre de modèles ML disponibles. Amazon ML apprend un type de modèle ML : les modèles linéaires. Le terme « modèle linéaire » implique que le modèle est spécifié sous la forme d'une combinaison linéaire d'entités. En fonction des données de formation, le processus d'apprentissage calcule un poids pour chaque entité pour former un modèle pouvant prédire ou estimer la valeur cible. Par exemple, si votre cible est le montant d'assurance qu'un client acquerra, et vos variables l'âge et le revenu, un modèle linéaire simple serait le suivant :

```
Estimated target = 0.2 + 5·age + 0.0003·income
```

Algorithme d'apprentissage

La tâche de l'algorithme d'apprentissage consiste à apprendre les poids pour le modèle. Les poids décrivent la probabilité que les tendances que le modèle apprend reflètent les relations réelles dans les données. Un algorithme d'apprentissage se compose d'une fonction de perte et d'une technique d'optimisation. La perte représente la pénalité générée lorsque l'estimation de la cible fournie par le modèle d'apprentissage-machine n'est pas parfaitement égale à la cible. Une fonction de perte quantifie cette pénalité sous la forme d'une valeur individuelle. Une technique d'optimisation cherche à minimiser la perte. Dans Amazon Machine Learning, nous utilisons trois fonctions de perte, une pour chacun des trois types de problèmes de prédiction. La technique d'optimisation utilisée dans Amazon ML est la Stochastic Gradient Descent (SGD) en ligne. La technique SGD effectue des passages séquentiels sur les données de formation et, à chaque passage, elle met à jour les poids des entités, exemple après exemple, dans le but de s'approcher des poids optimaux qui minimisent la perte.

Amazon ML utilise les algorithmes d'apprentissage suivants :

- Pour la classification binaire, Amazon ML utilise la régression logistique (fonction de perte logistique + SGD).
- Pour la classification multiclasse, Amazon ML utilise la régression logistique multinomiale (perte logistique multinomiale + SGD).
- Pour la régression, Amazon ML utilise la régression linéaire (fonction de perte au carré + SGD).

Paramètres de formation

L'algorithme d'apprentissage Amazon ML accepte des paramètres, appelés hyperparamètres ou paramètres d'entraînement, qui vous permettent de contrôler la qualité du modèle obtenu. En fonction de l'hyperparamètre, Amazon ML sélectionne automatiquement les paramètres ou fournit des valeurs statiques par défaut pour les hyperparamètres. Bien que le paramétrage par défaut des hyper-paramètres produise généralement des modèles utiles, vous pouvez améliorer les performances prédictives de vos modèles en changeant les valeurs des hyper-paramètres. Les sections suivantes décrivent les hyperparamètres courants associés aux algorithmes d'apprentissage pour les modèles linéaires, tels que ceux créés par Amazon ML.

Taux d'apprentissage

Le taux d'apprentissage est une valeur constante utilisée dans l'algorithme SGD (Stochastic Gradient Descent). Le taux d'apprentissage affecte la vitesse à laquelle l'algorithme atteint (converge vers)

les poids optimaux. L'algorithme SGD effectue des mises à jour des poids du modèle linéaire pour chaque exemple de données qu'il examine. La taille de ces mises à jour est contrôlée par le taux d'apprentissage. Un taux d'apprentissage trop élevé peut empêcher les poids de s'approcher de la solution optimale. Avec une valeur trop faible, l'algorithme nécessite de nombreux passages pour s'approcher des poids optimaux.

Dans Amazon ML, le taux d'apprentissage est sélectionné automatiquement en fonction de vos données.

Taille du modèle

Si vous disposez de nombreuses entités en entrée, un modèle de grande taille peut être généré en raison du nombre de tendances possibles dans les données. Les modèles de grande taille présentent des implications pratiques. Par exemple, plus de RAM est nécessaire pour conserver le modèle lors de sa formation et lors de la génération des prédictions. Dans Amazon ML, vous pouvez réduire la taille du modèle en utilisant la régularisation L1 ou en limitant spécifiquement la taille du modèle en spécifiant la taille maximale. Notez que si vous réduisez trop la taille du modèle, cela peut éventuellement réduire sa puissance prédictive.

Pour obtenir des informations sur la taille de modèle par défaut, consultez Paramètres de formation : Types et valeurs par défaut. Pour plus d'informations sur la régularisation, consultez Régularisation.

Nombre de passages

L'algorithme SGD effectue des passages séquentiels sur les données de formation. Le paramètre `Number of passes` contrôle le nombre de passages que l'algorithme effectue sur les données de formation. Avec davantage de passages, le modèle correspond mieux aux données (si le taux d'apprentissage n'est pas trop élevé), mais cet avantage diminue quand le nombre de passages augmente encore. Pour des jeux de données plus petits, vous pouvez augmenter significativement le nombre de passages, ce qui permet à l'algorithme d'apprentissage de correspondre effectivement plus étroitement aux données. Pour des jeux de données de très grande taille, un seul passage peut éventuellement suffire.

Pour obtenir des informations sur le nombre de passages par défaut, consultez Paramètres de formation : Types et valeurs par défaut.

Réorganisation des données

Dans Amazon ML, vous devez mélanger vos données car l'algorithme SGD est influencé par l'ordre des lignes dans les données d'apprentissage. La réorganisation de vos données de formation permet

d'améliorer les modèles d'apprentissage-machine. En effet, elle aide l'algorithme SGD à éviter des solutions qui sont optimales pour le premier type de données qu'il voit, mais pas pour la plage complète des données. La réorganisation change complètement l'ordre de vos données de manière à ce que l'algorithme SGD ne rencontre pas un seul type de données pour un trop grand nombre d'observations consécutives. S'il voit un seul type de données pour de nombreuses mises à jour de poids successives, l'algorithme peut ne pas être en mesure de corriger les poids du modèle pour un nouveau type de données parce que la mise à jour peut être trop importante. De plus, lorsque les données ne sont pas présentées de façon aléatoire, il est difficile pour l'algorithme de trouver rapidement la solution optimale pour tous les types de données ; dans certains cas, l'algorithme ne trouve jamais la solution optimale. La réorganisation des données de formation aide l'algorithme à converger plus tôt sur la solution optimale.

Par exemple, supposons que vous souhaitez former un modèle d'apprentissage-machine pour prédire un type de produit et que vos données de formation incluent les types de produit film, jouet et jeu vidéo. Si vous triez les données par colonne de type de produit avant de les télécharger sur Amazon S3, l'algorithme les voit par ordre alphabétique par type de produit. L'algorithme voit tout d'abord toutes vos données relatives aux films, et votre modèle d'apprentissage-machine commence à apprendre des tendances propres aux films. Ensuite, lorsque votre modèle aborde des données sur des jouets, chaque mise à jour que l'algorithme effectue correspond au modèle du type de produit jouet, même si ces mises à jour dégradent les tendances correspondant aux films. Ce basculement soudain du type film au type jouet peut produire un modèle qui n'apprend pas à prédire avec précision les types de produit.

Pour obtenir des informations sur le type de réorganisation par défaut, consultez Paramètres de formation : Types et valeurs par défaut.

Régularisation

La régularisation aide à empêcher les modèles linéaire de surajuster des exemples de données de formation (c'est-à-dire, de mémoriser des tendances au lieu de les généraliser) en pénalisant les valeurs pondérales extrêmes. La régularisation L1 a pour effet de réduire le nombre d'entités utilisées dans le modèle en mettant à zéro les poids d'entités qui auraient autrement de faibles poids. Par conséquent, la régularisation L1 produit des modèles dispersés et réduit la quantité de bruit dans le modèle. La régularisation L2 produit des valeurs pondérales globales plus petites, ce qui stabilise les poids lorsqu'il y a une forte corrélation entre les entités en entrée. Vous contrôlez le degré de régularisation L1 ou L2 appliqué à l'aide des paramètres `Regularization type` et `Regularization amount`. Une très grande valeur de régularisation pourrait entraîner la mise à zéro des poids de toutes les entités, ce qui empêcherait un modèle d'apprendre des tendances.

Pour obtenir des informations sur les valeurs de régularisation par défaut, consultez Paramètres de formation : Types et valeurs par défaut.

Evaluation de la précision du modèle

L'objectif du modèle d'apprentissage-machine est d'apprendre les tendances qui se prêtent bien à la généralisation pour les données inédites au lieu de simplement mémoriser les données qu'il a pu voir au cours de sa formation. Une fois que vous avez un modèle, il est important de vérifier s'il se comporte correctement sur des exemples inédits que vous n'avez pas utilisés pour la formation du modèle. Pour ce faire, vous utilisez le modèle pour prédire la réponse sur le jeu de données d'évaluation (données mises de côté), puis comparez la cible prédite à la réponse réelle (vérité du terrain).

Diverses métriques sont utilisées dans le cadre de l'apprentissage-machine pour mesurer la précision prédictive d'un modèle. Le choix de la métrique de précision dépend de la tâche d'apprentissage-machine. Il est important de passer en revue ces métriques afin de déterminer si votre modèle est efficace.

Classification binaire

La sortie réelle de nombreux algorithmes de classification binaire est un score de prédiction. Ce score indique la certitude du système que l'observation donnée appartient à la classe positive. Pour décider si l'observation doit être classée comme positive ou négative, en tant que consommateur de ce score, vous devez interpréter le score en sélectionnant une limite de classification, ou seuil, et comparer le score à ce seuil. Toute observation avec un score supérieur au seuil est alors prédite en tant que classe positive et tout score inférieur au seuil en tant que classe négative.

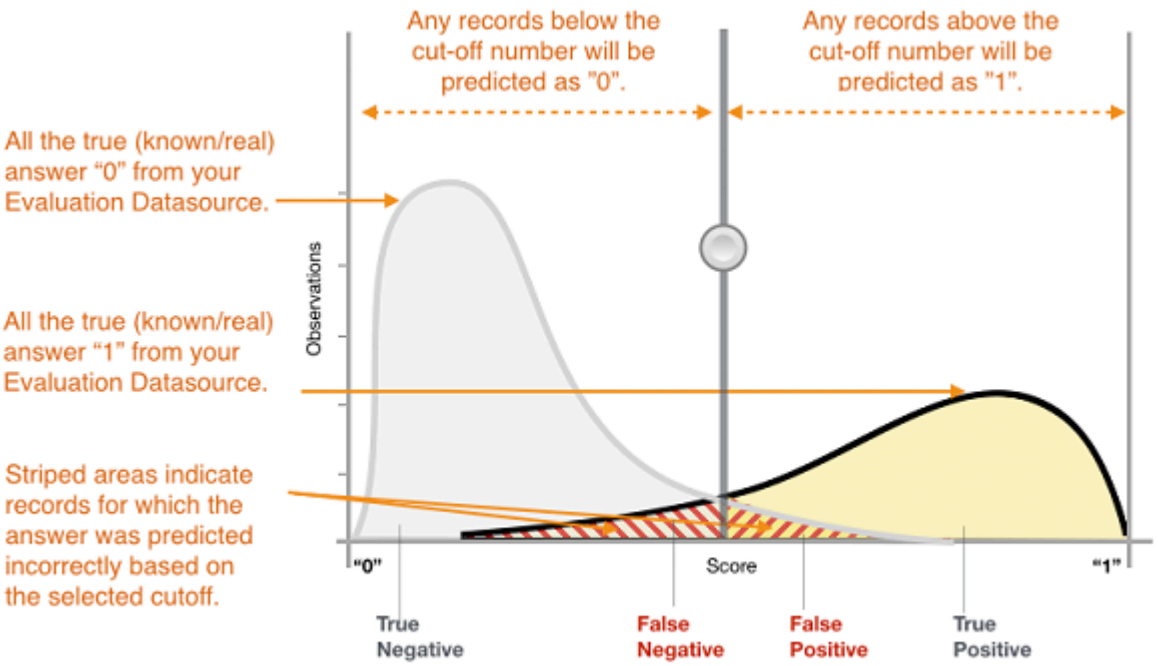


Figure 1 : Distribution des scores pour un modèle de classification binaire

Les prédictions se répartissent désormais en quatre groupes en fonction de la réponse réelle connue et de la réponse prédite : prédictions positives correctes (vrais positifs), prédictions négatives correctes (vrais négatifs), prédictions positives erronées (faux positifs) et prédictions négatives erronées (faux négatifs).

Les métriques de précision de classification binaire quantifient les deux types de prédictions correctes et les deux types d'erreurs. Les métriques standard sont la précision (ACC), le taux de positifs prédits, la sensibilité, le taux de faux négatifs, la mesure F1. Chaque métrique mesure un aspect différent du modèle prédictif. La précision (ACC) mesure la fraction de prédictions correctes. Le taux de positifs prédits mesure la fraction de positifs observés parmi les exemples prédits comme positifs. La sensibilité mesure le nombre de positifs observés qui ont été prédits comme positifs. La mesure F1 représente la moyenne harmonique entre le taux de positifs prédits et la sensibilité.

La métrique AUC est d'un autre type. Elle mesure l'aptitude du modèle à prédire un score plus élevé pour les exemples positifs par rapport aux exemples négatifs. Comme la métrique AUC est indépendante du seuil sélectionné, elle vous permet de vous faire une idée des performances de prédictions de votre modèle, sans choisir de seuil.

En fonction de votre problème, vous pouvez être plus intéressé par un modèle performant pour un sous-ensemble spécifique de ces métriques. Par exemple, deux applications métier peuvent avoir des exigences très différentes en matière de modèles d'apprentissage-machine :

- Une application peut avoir besoin d'être extrêmement certaine que les prédictions positives soient effectivement positives (taux de positifs prédits élevé) et peut se permettre de mal classer certains exemples positifs comme négatifs (sensibilité moyenne).
- Une autre application peut avoir besoin de prédire correctement autant d'exemples positifs que possible (haute sensibilité) et acceptera que certains exemples négatifs soient mal classés comme positifs (taux de positifs prédits moyen).

Dans Amazon ML, les observations obtiennent un score prévisionnel de l'ordre de [0,1]. Le seuil de score pour prendre la décision de classer les exemples comme 0 ou 1 est défini par défaut à 0,5. Amazon ML vous permet d'examiner les implications du choix de différents seuils de score et de choisir un seuil adapté aux besoins de votre entreprise.

Classification multiclasse

Contrairement au processus pour des problèmes de classification binaire, vous n'avez pas besoin de choisir un score seuil pour effectuer des prédictions. La réponse prédite est la classe (l'étiquette) avec le score prédit le plus élevé. Dans certains cas, vous pouvez utiliser la réponse prédite seulement si elle est prédite avec un score élevé. Dans ce cas, vous pouvez choisir un seuil sur les scores prédits en fonction duquel vous accepterez ou non la réponse prédite.

Les métriques standard utilisées en mode multiclasse sont les mêmes que celles utilisées dans le cas d'une classification binaire. La métrique est calculée pour chaque classe en la traitant comme un problème de classification binaire après avoir regroupé toutes les autres classes dans la seconde classe. Ensuite, la métrique binaire est moyennée sur toutes les classes pour fournir une métrique moyennée par macro (chaque classe est traitée de façon égale) ou une métrique de moyenne pondérée (pondérée par la fréquence des classes). Dans Amazon ML, la macromoyenne F1 est utilisée pour évaluer le succès prédictif d'un classificateur multiclasse.

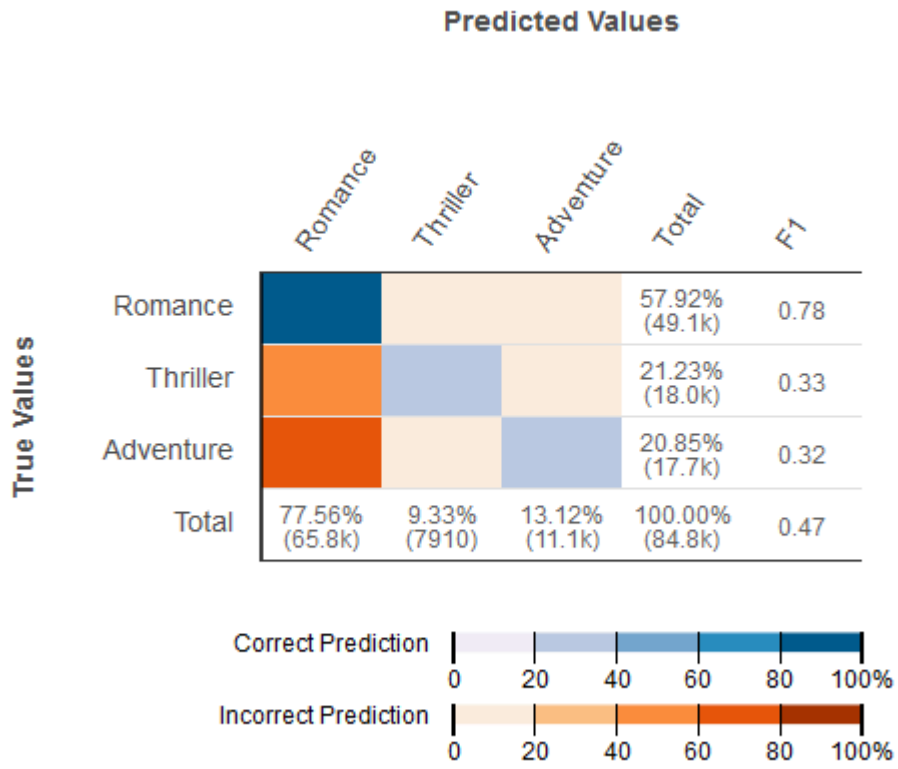


Figure 2 : Matrice de confusion pour un modèle de classification multiclasse

Il est utile de passer en revue la matrice de confusion pour rechercher d'éventuels problèmes multiclasse. La matrice de confusion est une table qui montre chaque classe dans les données d'évaluation, ainsi que le nombre ou le pourcentage de prédictions correctes et incorrectes.

Régression

Pour les tâches de régression, les métriques de précision standard sont l'erreur quadratique moyenne (RMSE, Root Mean Square Error) et l'erreur moyenne absolue en pourcentage (MAPE, Mean Absolute Percentage Error). Ces métriques mesurent la distance entre la cible numérique prédite et la réponse numérique réelle (vérité du terrain). Dans Amazon ML, la métrique RMSE est utilisée pour évaluer la précision prédictive d'un modèle de régression.

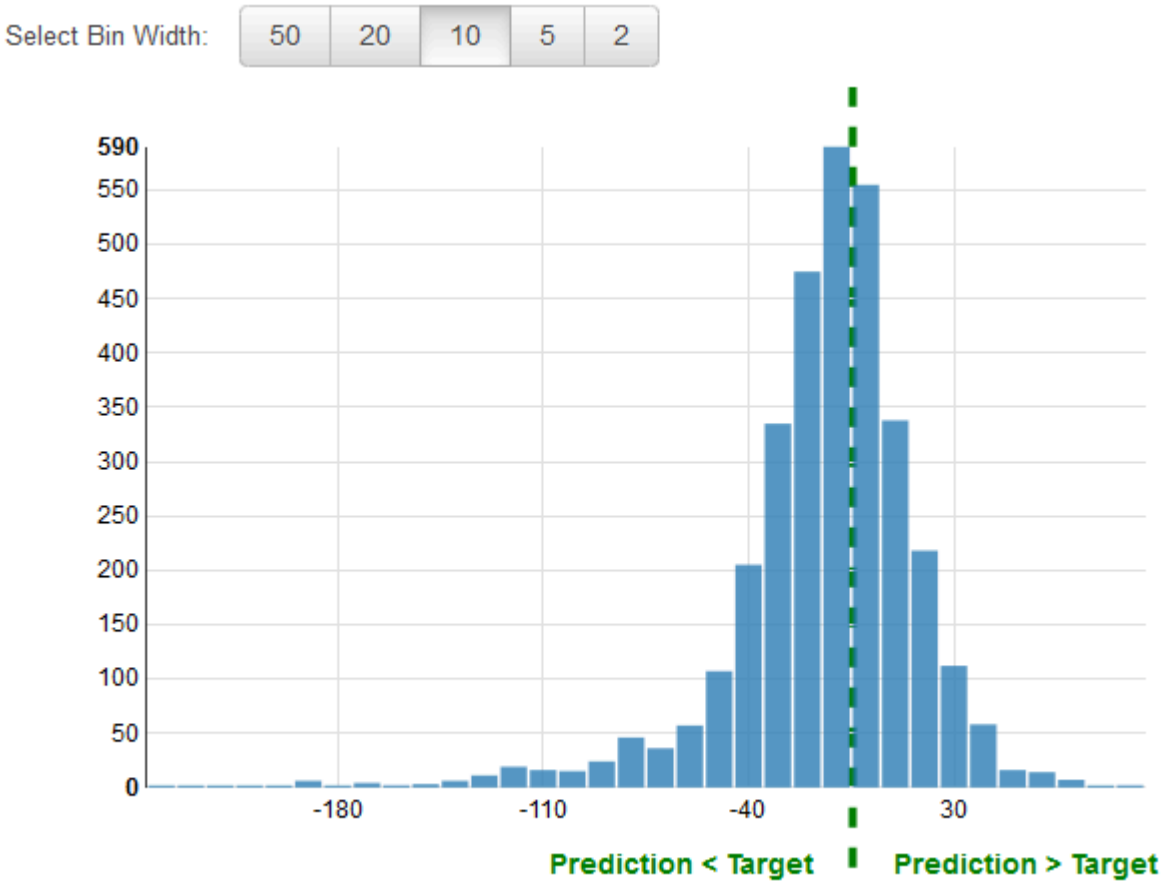


Figure 3 : Distribution des résidus pour un modèle de régression

Il est usuel de passer en revue les résidus pour identifier les problèmes de régression éventuels. Un résidu pour une observation dans les données d'évaluation représente la différence entre la cible réelle et la cible prédite. Les résidus représentent la partie de la cible que le modèle n'est pas en mesure de prédire. Un résidu positif indique que le modèle sous-estime la cible (la cible réelle est supérieure à la cible prédite). Un résidu négatif indique une surestimation (la cible réelle est inférieure à la cible prédite). L'histogramme des résidus sur les données d'évaluation lors d'une distribution en forme de cloche centrée sur zéro indique que le modèle commet des erreurs d'une manière aléatoire et qu'il ne prédit pas systématiquement trop haut ou trop bas une plage particulière de valeurs cibles. Si les résidus ne constituent pas une forme en cloche centrée sur zéro, il y a une certaine structure dans l'erreur de prédiction du modèle. L'ajout de variables supplémentaires au modèle peut aider le modèle à capturer la tendance qui n'est pas capturée par le modèle actuel.

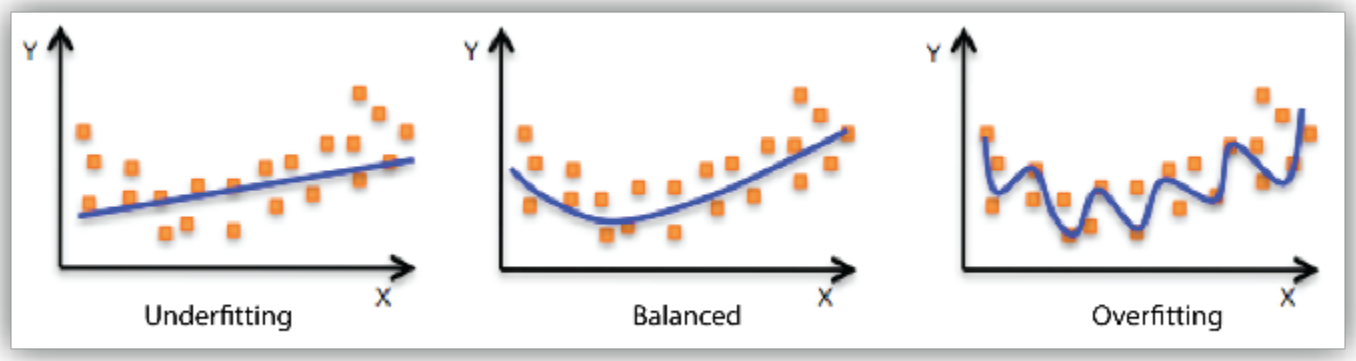
Amélioration de la précision du modèle

L'obtention d'un modèle d'apprentissage-machine qui correspond à vos besoins implique généralement d'itérer sur ce processus d'apprentissage-machine et de tester quelques variations. Vous n'obtiendrez peut-être pas un modèle très prédictif lors de la première itération ou vous souhaiterez peut-être améliorer votre modèle pour obtenir des prédictions encore meilleures. Pour améliorer les performances, vous pouvez itérer sur les étapes suivantes :

1. Collecte de données : augmentez le nombre d'exemples de formation
2. Traitement des entités : ajoutez d'autres variables et un meilleur traitement des entités
3. Réglage des paramètres du modèle : envisagez d'autres valeurs pour les paramètres de formation utilisés par votre algorithme d'apprentissage

Ajustement du modèle : sous-ajustement et surajustement

Comprendre l'ajustement du modèle est important pour comprendre la cause profonde de la mauvaise précision du modèle. Cette analyse vous guidera pour prendre des mesures correctives. Nous pouvons déterminer si un modèle prédictif constitue un sous-ajustement ou un surajustement des données de formation en examinant l'erreur de prédiction sur les données de formation et les données d'évaluation.



Votre modèle constitue un sous-ajustement des données de formation lorsque le modèle donne des résultats médiocres sur les données de formation. Cela est dû au fait que le modèle n'est pas en mesure de saisir la relation entre les exemples en entrée (souvent appelés X) et les valeurs cibles (souvent appelées Y). Votre modèle constitue un surajustement de vos données de formation lorsque vous constatez que le modèle offre de bons résultats sur les données de formation mais des résultats

médiocres sur les données d'évaluation. Cela est dû au fait que le modèle mémorise les données qu'il a vues et n'est pas en mesure de généraliser aux exemples nouveaux.

Des performances médiocres sur les données de formation peuvent être dues à un modèle trop simple (les entités en entrée ne sont pas suffisamment expressives) pour décrire correctement la cible. Il est possible d'améliorer les performances en augmentant la flexibilité du modèle. Pour augmenter la flexibilité du modèle, essayez la procédure suivante :

- Ajoutez de nouvelles entités spécifiques au domaine et d'autres produits cartésiens d'entités, puis changez le type de traitement d'entités utilisé (par exemple, en augmentant la taille des n-grammes).
- Diminuez le degré de régularisation utilisé.

Si votre modèle constitue un surajustement des données de formation, il est logique d'entreprendre des actions qui réduisent la flexibilité du modèle. Pour réduire la flexibilité du modèle, essayez la procédure suivante :

- Sélection des entités : envisagez d'utiliser moins de combinaisons d'entités, de diminuer la taille des n-grammes et de réduire le nombre d'intervalles des attributs numériques.
- Augmentez le degré de régularisation utilisé.

La mauvaise précision sur les données de formation et de test peut provenir du fait que l'algorithme d'apprentissage ne disposait pas de suffisamment de données d'apprentissage. Vous pouvez améliorer les performances en procédant comme suit :

- Augmentez le nombre d'exemples de données de formation.
- Augmentez le nombre de passages sur les données de formation existantes.

Utilisation du modèle pour effectuer des prédictions

Maintenant que vous avez un modèle d'apprentissage-machine performant, vous allez l'utiliser pour effectuer des prédictions. Dans Amazon Machine Learning, il existe deux manières d'utiliser un modèle pour effectuer des prédictions :

Prédictions par lots

La prédiction par lots est utile lorsque vous souhaitez générer des prédictions pour un ensemble d'observations à la fois, puis entreprendre une action sur un certain pourcentage ou nombre d'observations. En règle générale, une telle application ne nécessite pas une faible latence. Par exemple, lorsque vous voulez décider quels clients cibler dans le cadre d'une campagne de publicité pour un produit, vous obtenez des scores de prédiction pour tous les clients, triez les prédictions de votre modèle pour identifier les clients qui ont le plus de chances d'acheter le produit, puis ciblez éventuellement la tranche supérieure de 5 % de ces clients.

Prédictions en ligne

Les scénarios de prédiction en ligne sont destinés aux cas où vous souhaitez générer des prédictions sur one-by-one la base de chaque exemple indépendamment des autres exemples, dans un environnement à faible latence. Par exemple, vous pouvez utiliser des prédictions pour décider immédiatement si une transaction donnée correspond probablement à une transaction frauduleuse.

Reformation des modèles sur de nouvelles données

Pour qu'un modèle effectue des prédictions précises, les données sur lesquelles il se base doivent avoir une distribution similaire aux données sur lesquelles le modèle a été formé. Comme on peut s'attendre à ce que les distributions de données dérivent au fil du temps, le déploiement d'un modèle n'est pas un exercice définitif, mais plutôt un processus continu. Il est recommandé de surveiller continuellement les données entrantes et de reformer votre modèle sur des données plus récentes si vous pensez que la distribution des données s'est trop écartée de la distribution initiale des données de formation. Si la surveillance des données pour détecter une modification de la distribution des données s'avère trop complexe, une stratégie plus simple consiste à former le modèle de façon régulière, par exemple, sur une base quotidienne, hebdomadaire ou mensuelle. Afin de reformer des modèles dans Amazon ML, vous devez créer un nouveau modèle basé sur vos nouvelles données de formation.

Le processus d'Amazon Machine Learning

Le tableau suivant décrit comment utiliser la console Amazon ML pour exécuter le processus de ML décrit dans ce document.

Processus d'apprentissage-machine	Tâche Amazon ML
Analyse de vos données	Pour analyser vos données dans Amazon ML, créez une source de données et consultez la page d'informations sur les données.
Fractionnement des données en sources de données de formation et d'évaluation	Amazon ML peut diviser la source de données pour utiliser 70 % des données pour l'entraînement des modèles et 30 % pour évaluer les performances prédictives de votre modèle. Lorsque vous utilisez l'assistant Create ML Model avec les paramètres par défaut, Amazon ML divise les données pour vous. Si vous utilisez l'assistant Create ML avec les paramètres personnalisés et que vous choisissez d'évaluer le modèle ML, vous verrez une option permettant à Amazon ML de diviser les données pour vous et d'exécuter une évaluation sur 30 % des données.
Réorganisation de vos données de formation	Lorsque vous utilisez l'assistant Create ML Model avec les paramètres par défaut, Amazon ML mélange vos données pour vous. Vous pouvez également mélanger vos données avant de les importer dans Amazon ML.
Traitement des entités	Le processus de regroupement des données de formation dans un format optimal pour l'apprentissage et la généralisation est connu sous le nom de transformation d'entités. Lorsque vous utilisez l'assistant Create ML Model avec les paramètres par défaut, Amazon ML suggère des paramètres de traitement des fonctionnalités pour vos données. Pour spécifier des paramètres de traitement d'entités, utilisez l'option Personnalisé de l'assistant de création de modèle d'apprentissage-machine et fournissez une recette de traitement d'entités.
Formation du modèle	Lorsque vous utilisez l'assistant Create ML Model pour créer un modèle dans Amazon ML, Amazon ML entraîne votre modèle.
Sélection des paramètres du modèle	Dans Amazon ML, vous pouvez régler quatre paramètres qui affectent les performances prédictives de votre modèle : la taille du modèle, le nombre de passes, le type de brassage et la régularisation. Vous pouvez

Processus d'apprentissage-machine	Tâche Amazon ML
	définir ces paramètres lorsque vous utilisez l'assistant de création de modèle d'apprentissage-machine pour créer un modèle d'apprentissage-machine et choisissez l'option Personnalisé.
Evaluation des performances du modèle	Utilisez l'assistant de création d'évaluation pour évaluer les performances prédictives de votre modèle.
Sélection des entités	L'algorithme d'apprentissage Amazon ML peut supprimer des fonctionnalités qui ne contribuent pas beaucoup au processus d'apprentissage. Pour indiquer que vous souhaitez ignorer ces entités, choisissez le paramètre `L1 regularization` lorsque vous créez le modèle d'apprentissage-machine.
Définition d'un score seuil de précision des prédictions	Passez en revue les performances prédictives du modèle dans le rapport d'évaluation pour différentes valeurs de score seuil, puis définissez le score seuil en fonction de votre application métier. Le score seuil détermine comment le modèle définit une correspondance de prédictions. Ajustez ce nombre pour contrôler les faux positifs et les faux négatifs.
Utilisation du modèle	Utilisez votre modèle pour obtenir des prédictions pour un lot d'observations à l'aide de l'assistant de création de prédiction par lots. Ou, obtenez des prédictions pour des observations individuelles à la demande en permettant au modèle d'apprentissage-machine de traiter des prédictions en temps réel à l'aide de l'API `Predict`.

Configuration d'Amazon Machine Learning

Vous avez besoin d'un compte AWS pour pouvoir utiliser Amazon Machine Learning pour la première fois. Si vous n'avez pas de compte, consultez Inscrivez-vous à AWS.

Inscrivez-vous à AWS

Lorsque vous vous inscrivez à Amazon Web Services (AWS), votre compte AWS est automatiquement inscrit à tous les services d'AWS, y compris Amazon ML. Seuls les services que vous utilisez vous sont facturés. Si vous avez déjà un compte AWS, ignorez cette étape. Si tel n'est pas le AWS cas, observez la procédure suivante pour en créer un.

Pour s'inscrire à un compte AWS

1. Accédez à http://aws.amazon.com et choisissez S'inscrire.
2. Suivez les instructions à l’écran.

Dans le cadre de la procédure d'inscription, vous recevrez un appel téléphonique et vous saisirez un code PIN en utilisant le clavier numérique du téléphone.

Didacticiel : Utilisation d'Amazon ML pour prédire les réponses à une offre marketing

Avec Amazon Machine Learning (Amazon ML), vous pouvez créer et entraîner des modèles prédictifs et héberger vos applications dans une solution cloud évolutive. Dans ce didacticiel, nous vous expliquons comment utiliser la console Amazon ML pour créer une source de données, créer un modèle d'apprentissage automatique (ML) et utiliser le modèle pour générer des prédictions que vous pouvez utiliser dans vos applications.

Notre exemple d'exercice montre comment identifier des clients potentiels pour une campagne marketing ciblée, mais vous pouvez appliquer les mêmes principes pour créer et utiliser divers modèles d'apprentissage-machine. Pour réaliser cet exemple d'exercice, vous allez utiliser des jeux de données bancaires et marketing disponibles publiquement dans le Référentiel d'apprentissage-machine de l'Université de Californie à Irvine (UCI). Ces jeux de données contiennent des informations générales sur des clients, ainsi que des informations sur la façon dont ils ont répondu à des contacts marketing précédents. Vous utiliserez ces données pour identifier les clients les plus susceptibles de souscrire à votre nouveau produit, un dépôt bancaire à terme, également appelé « certificat de dépôt (CD) ».

> **Warning**
>
> Ce didacticiel n'est pas inclus dans l'offre gratuite AWS. Pour plus d'informations sur la tarification d'Amazon ML, consultez la section Amazon Machine Learning Pricing.

Prérequis

Pour effectuer ce didacticiel, vous devez disposer d'un compte AWS. Si vous n'avez pas de compte AWS, consultez Configuration d'Amazon Machine Learning.

Étapes

- Etape 1 : Préparation de vos données
- Etape 2 : Création d'une source de données de formation
- Etape 3: Création d'un modèle d'apprentissage-machine

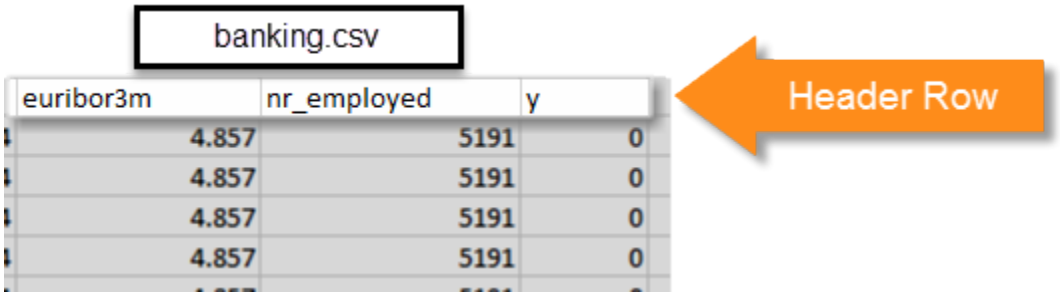
Etape 1 : Préparation de vos données

Dans le cadre de l'apprentissage-machine, vous obtenez généralement les données et veillez à ce qu'elles soient formatées convenablement avant de commencer le processus de formation. Dans le cadre de ce didacticiel, nous avons obtenu un exemple de jeu de données à partir du référentiel UCI Machine Learning, nous l'avons formaté conformément aux directives d'Amazon ML et l'avons mis à votre disposition pour téléchargement. Téléchargez le jeu de données depuis notre emplacement de stockage Amazon Simple Storage Service (Amazon S3) et chargez-le dans votre propre compartiment S3 en suivant les procédures décrites dans cette rubrique.

Pour connaître les exigences de mise en forme d'Amazon ML, consultezComprendre le format de données pour Amazon ML.

Pour télécharger les jeux de données

1. Téléchargez le fichier qui contient les données d'historique des clients qui ont acheté des produits similaires à votre dépôt bancaire à terme en cliquant sur banking.zip. Décompressez le dossier et enregistrez le fichier banking.csv sur votre ordinateur.
2. Téléchargez le fichier que vous utiliserez pour prédire si des clients potentiels vont répondre à votre offre en cliquant sur banking-batch.zip. Décompressez le dossier et enregistrez le fichier banking-batch.csv sur votre ordinateur.
3. Ouvrir `banking.csv`. Vous verrez des lignes et des colonnes de données. La ligne d'en-tête contient les noms des attributs des différentes colonnes. Un attribut est une propriété nommée unique qui décrit une caractéristique particulière de chaque client ; par exemple, nr_employed indique l'état professionnel du client. Chaque ligne représente la collection des observations relatives à un client individuel.

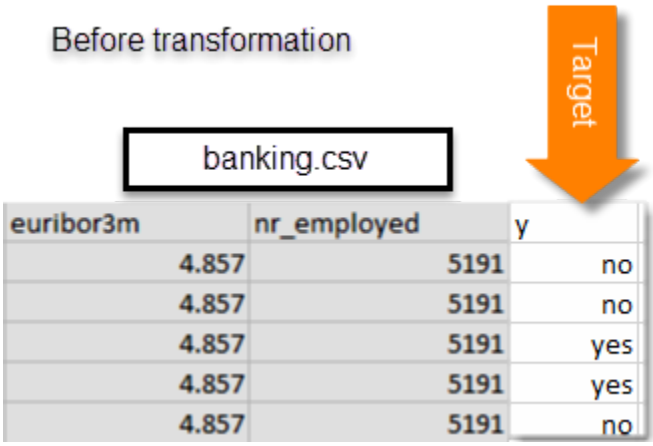


euribor3m	nr_employed	y
4.857	5191	0
4.857	5191	0
4.857	5191	0
4.857	5191	0

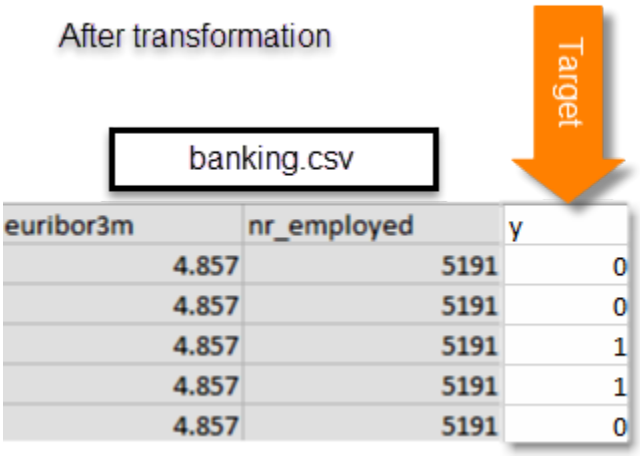
Vous souhaitez que votre modèle d'apprentissage-machine réponde à la question « Ce client optera-t-il pour mon nouveau produit ? ». Dans le jeu de données `banking.csv`, la réponse à cette question est l'attribut y, qui contient la valeur 1 (pour oui) ou 0 (pour non). L'attribut que vous souhaitez qu'Amazon ML apprenne à prévoir est appelé attribut cible.

> **Note**
>
> L'attribut y est un attribut binaire. Il peut contenir uniquement deux valeurs, dans ce cas, 0 ou 1. Dans le jeu de données UCI d'origine, l'attribut y a pour valeur Yes (Oui) ou No (Non). Nous avons modifié le jeu de données d'origine pour vous. Toutes les valeurs de l'attribut y qui signifient Oui sont désormais 1, et toutes les valeurs qui signifient Non sont désormais 0. Si vous utilisez vos propres données, vous pouvez utiliser d'autres valeurs pour un attribut binaire. Pour plus d'informations sur les valeurs valides, consultez Utilisation du AttributeType terrain.

Les exemples suivants montrent les données avant et après que nous avons remplacé les valeurs de l'attribut y par les attributs binaires 0 et 1.



euribor3m	nr_employed	y
4.857	5191	no
4.857	5191	no
4.857	5191	yes
4.857	5191	yes
4.857	5191	no



euribor3m	nr_employed	y
4.857	5191	0
4.857	5191	0
4.857	5191	1
4.857	5191	1
4.857	5191	0

Le fichier `banking-batch.csv` ne contient pas l'attribut y. Une fois que vous aurez créé un modèle d'apprentissage-machine, vous allez l'utiliser pour prédire y pour chaque enregistrement dans ce fichier.

Ensuite, téléchargez les `banking-batch.csv` fichiers `banking.csv` et sur Amazon S3.

Pour télécharger les fichiers vers un emplacement Amazon S3

1. Connectez-vous à la console Amazon S3 AWS Management Console et ouvrez-la à l'adresse https://console.aws.amazon.com/s3/.
2. Dans la liste Tous les compartiments, créez un compartiment ou choisissez l'emplacement où vous voulez charger les fichiers.
3. Dans la barre de navigation, choisissez Charger.
4. Choisissez Add Files (Ajouter des fichiers).
5. Dans la boîte de dialogue, accédez à votre bureau, choisissez `banking.csv` et `banking-batch.csv`, puis choisissez Ouvrir.

Vous êtes maintenant prêt à créer votre source de données de formation.

Etape 2 : Création d'une source de données de formation

Après avoir chargé le `banking.csv` jeu de données sur votre site Amazon Simple Storage Service (Amazon S3), vous l'utilisez pour créer une source de données de formation. Une source de données est un objet Amazon Machine Learning (Amazon ML) qui contient l'emplacement de vos données d'entrée et des métadonnées importantes relatives à ces données d'entrée. Amazon ML utilise la source de données pour des opérations telles que la formation et l'évaluation des modèles ML.

Pour créer une source de données, fournissez les éléments suivants :

- Emplacement de vos données sur Amazon S3 et autorisation d'accès aux données
- Le schéma, qui comprend les noms des attributs dans les données et le type de chaque attribut (numérique, texte, catégorie ou binaire)
- Le nom de l'attribut qui contient la réponse que vous souhaitez qu'Amazon ML apprenne à prédire, l'attribut cible

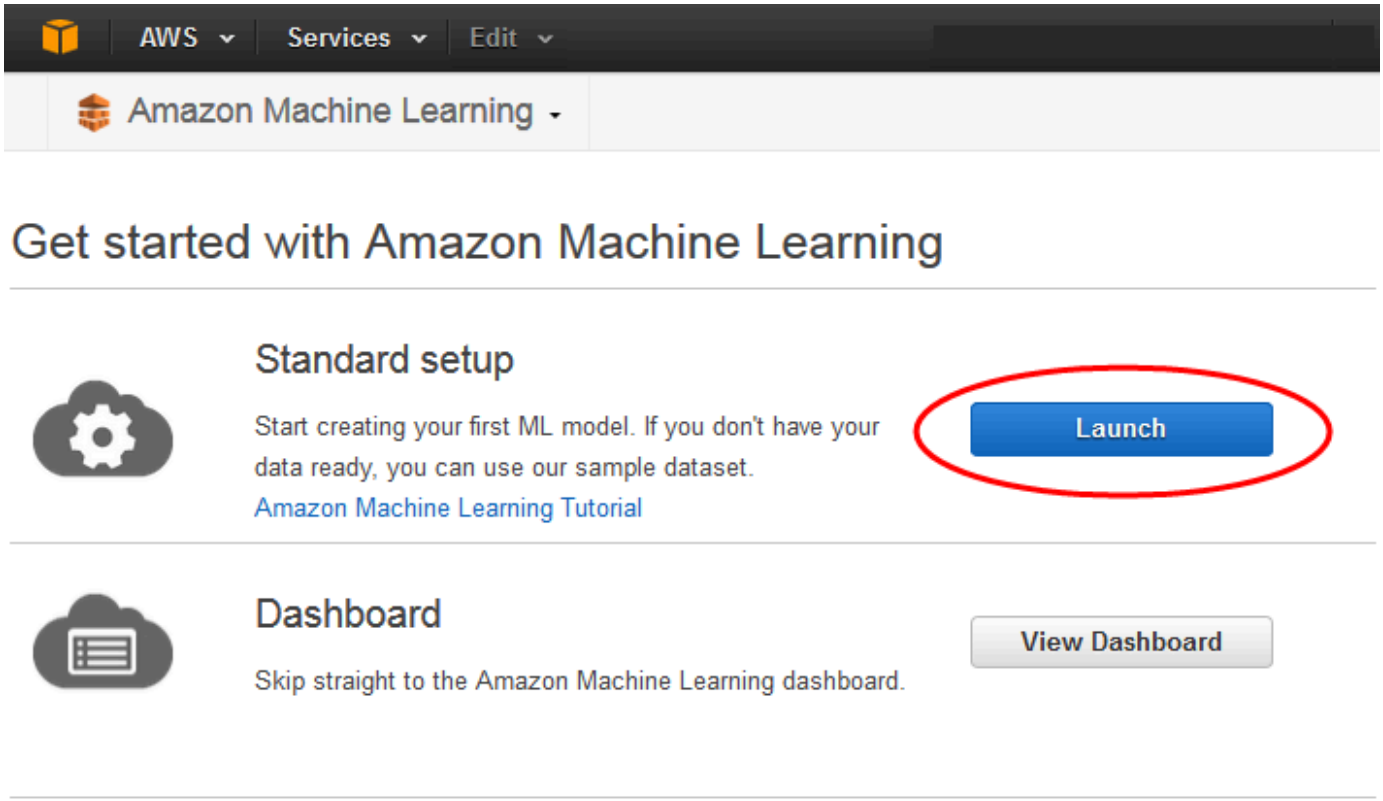
> **Note**
>
> La source de données ne stocke pas réellement vos données, mais les référence uniquement. Évitez de déplacer ou de modifier les fichiers stockés dans Amazon S3. Si vous les déplacez ou les modifiez, Amazon ML ne pourra pas y accéder pour créer un modèle de machine learning, générer des évaluations ou générer des prédictions.

Pour créer la source de données de formation

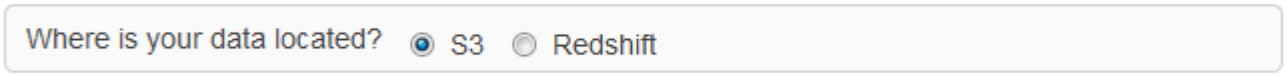
1. Ouvrez la console Amazon Machine Learning à l'adresse https://console.aws.amazon.com/machinelearning/.
2. Choisissez Démarrer.

   > **Note**
   >
   > Ce didacticiel part du principe que c'est la première fois que vous utilisez Amazon ML. Si vous avez déjà utilisé Amazon ML, vous pouvez utiliser le bouton Create new... liste déroulante sur le tableau de bord Amazon ML pour créer une nouvelle source de données.

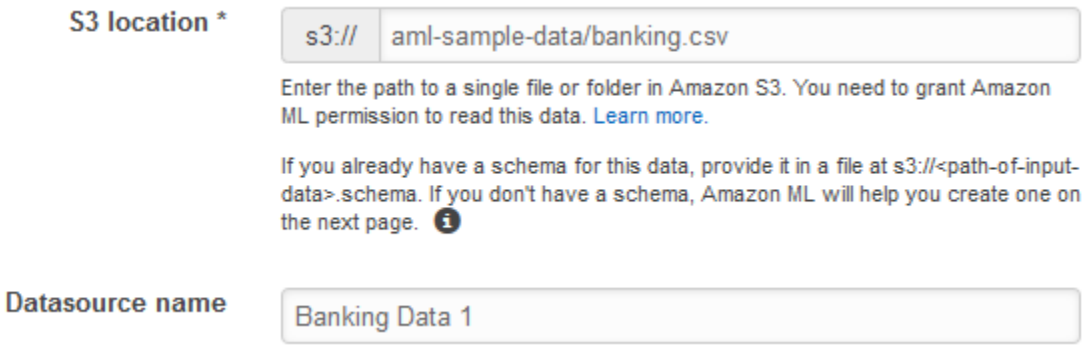
3. Sur la page Commencer avec Amazon Machine Learning, sélectionnez Launch.



4. Dans la page Input Data, pour Where is your data located?, assurez-vous que S3 est sélectionné.

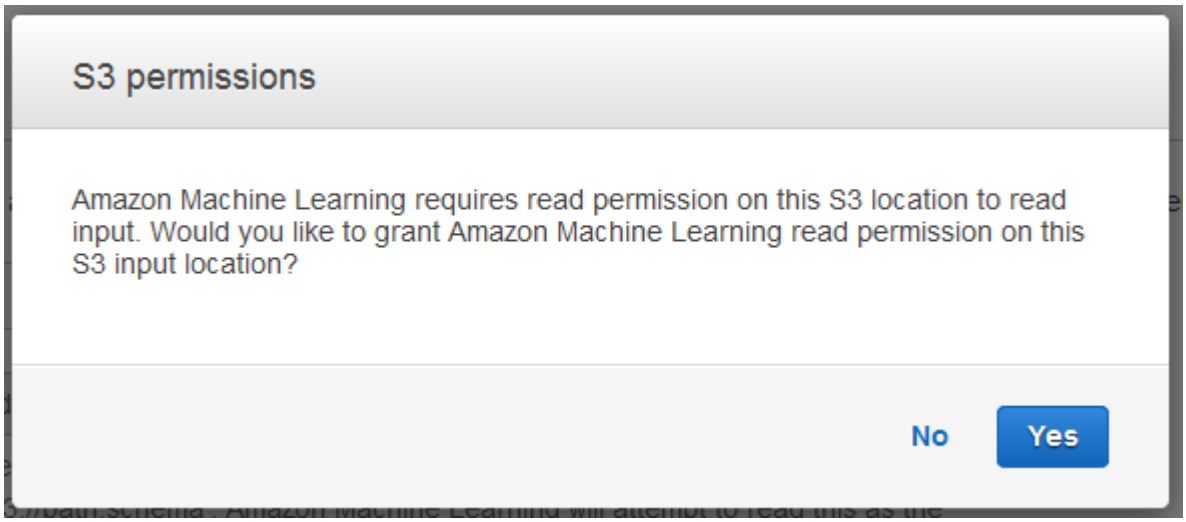


5. Pour Emplacement S3, tapez l'emplacement complet du fichier `banking.csv` de l'étape 1 : Préparation de vos données. olpPar exemple : *`your-bucket`***`/banking.csv`**. Amazon ML ajoute s3 ://au nom de votre compartiment pour vous.
6. Pour Datasource name, tapez **`Banking Data 1`**.

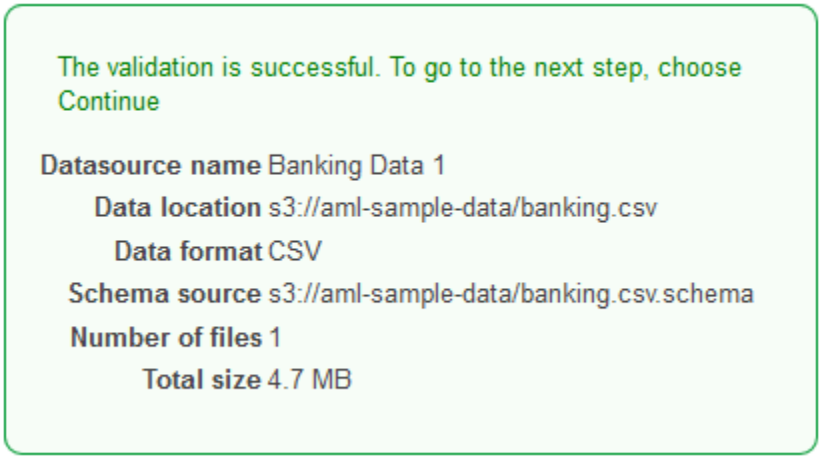


7. Choisissez Vérifier.

8. Dans la boîte de dialogue S3 permissions, choisissez Oui.



9. Si Amazon ML peut accéder au fichier de données et le lire à l'emplacement S3, vous verrez une page similaire à la suivante. Passez en revue les propriétés, puis choisissez Continuer.



Ensuite, vous devez établir un schéma. Un schéma est l'information dont Amazon ML a besoin pour interpréter les données d'entrée d'un modèle ML, y compris les noms des attributs et les types de données qui leur sont attribués, ainsi que les noms des attributs spéciaux. Il existe deux manières de fournir un schéma à Amazon ML :

- Fournissez un fichier de schéma distinct lorsque vous chargez vos données Amazon S3.
- Autorisez Amazon ML à déduire les types d'attributs et à créer un schéma pour vous.

Dans ce didacticiel, nous demanderons à Amazon ML de déduire le schéma.

Pour obtenir des informations sur la création d'un fichier de schéma distinct, consultez Création d'un schéma de données pour Amazon ML.

Pour autoriser Amazon ML à déduire le schéma

1. Sur la page Schéma, Amazon ML vous montre le schéma qu'il a déduit. Passez en revue les types de données déduits par Amazon ML pour les attributs. Il est important que le type de données approprié soit attribué aux attributs pour permettre à Amazon ML d'ingérer correctement les données et de permettre le traitement correct des fonctionnalités sur les attributs.

   - Les attributs qui ont seulement deux états possibles, tels que oui ou non, doivent être marqués comme Binary (binaire).
   - Les attributs correspondant à des chaînes ou des nombres utilisés pour indiquer une catégorie doivent être marqués comme Categorical (catégorie).
   - Les attributs correspondant à des quantités numériques dont l'ordre est important doivent être marqués comme Numeric (numérique).
   - Les attributs correspondant à des chaînes que vous souhaitez traiter comme des mots délimités par des espaces doivent être marqués comme Text (texte).

Name	Data Type	Sample Field Value 1
age	Numeric	56
campaign	Numeric	1
cons_conf_idx	Numeric	-36.4
cons_price_idx	Numeric	93.994
contact	Categorical	telephone
day_of_week	Categorical	mon
default	Categorical	no
duration	Numeric	261
education	Categorical	basic.4y
emp_var_rate	Numeric	1.1

2. Dans ce didacticiel, Amazon ML a correctement identifié les types de données pour tous les attributs. Choisissez donc Continuer.

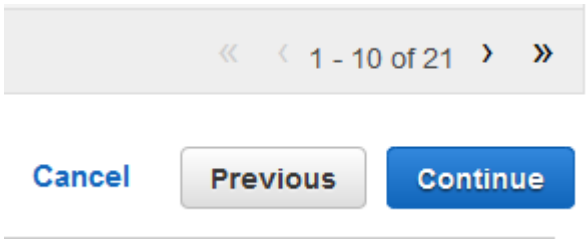
Ensuite, sélectionnez un attribut cible.

Souvenez-vous que la cible est l'attribut que le modèle d'apprentissage-machine doit apprendre à prédire. L'attribut y indique si une personne a déjà souscrit à une campagne dans le passé : 1 (oui) ou 0 (non).

> **Note**
>
> Choisissez un attribut cible seulement si vous avez l'intention d'utiliser la source de données pour la formation et l'évaluation des modèles d'apprentissage-machine.

Pour sélectionner y comme attribut cible

1. Dans la partie inférieure droite du tableau, choisissez la flèche simple pour passer à la dernière page du tableau, où figure l'attribut nommé y.



2. Dans la colonne Target, sélectionnez y.

   Amazon ML confirme que y est sélectionné comme cible.

3. Choisissez Continuer.
4. Dans la page Row ID, pour Does your data contain an identifier ? , veillez à ce que la valeur No (valeur par défaut) soit sélectionnée.
5. Choisissez Vérification, puis Continuer.

Maintenant que vous avez une source de données de formation, vous êtes prêt à créer votre modèle.

Etape 3: Création d'un modèle d'apprentissage-machine

Une fois que vous avez créé la source de données de formation, vous l'utilisez pour créer un modèle d'apprentissage-machine, former le modèle, puis évaluer les résultats. Le modèle ML est un ensemble de modèles qu'Amazon ML trouve dans vos données pendant l'entraînement. Vous utilisez le modèle pour créer des prédictions.

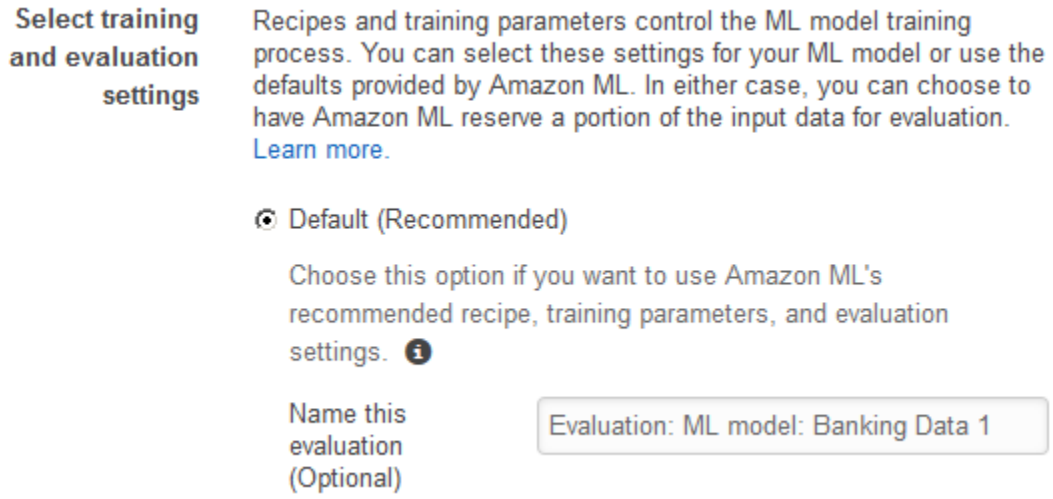
Pour créer un modèle d'apprentissage-machine

1. Dans la mesure où l'assistant de mise en route crée à la fois une source de données d'entraînement et un modèle, Amazon Machine Learning (Amazon ML) utilise automatiquement la source de données d'entraînement que vous venez de créer et vous dirige directement vers

la page des paramètres du modèle de machine learning. Dans la page ML model settings, pour ML model name, assurez-vous que la valeur par défaut **`ML model: Banking Data 1`** soit affichée.

En utilisant un nom convivial, tel que la valeur par défaut, vous pouvez facilement identifier et gérer le modèle d'apprentissage-machine.

2. Pour Training and evaluation settings, assurez-vous que la valeur Default est sélectionnée.



3. Pour Name this evaluation (Nommer cette évaluation), acceptez la valeur par défaut, **`Evaluation: ML model: Banking Data 1`**.
4. Choisissez Review, passez en revue vos paramètres, puis choisissez Finish.

Une fois que vous avez sélectionné Terminer, Amazon ML ajoute votre modèle à la file d'attente de traitement. Lorsque Amazon ML crée votre modèle, il applique les valeurs par défaut et effectue les actions suivantes :

- Il divise la source de données de formation en deux sections, l'une contenant 70 % des données et l'autre contenant les 30 % restants
- Il forme le modèle d'apprentissage-machine sur la section qui contient 70 % des données d'entrée
- Il évalue le modèle à l'aide des 30 % restants des données d'entrée

Lorsque votre modèle est dans la file d'attente, Amazon ML indique que le statut est En attente. Pendant qu'Amazon ML crée votre modèle, il indique que son statut est En cours. Lorsqu'il a terminé toutes les actions, il indique l'état Terminé. Attendez la fin de l'évaluation avant de continuer.

Maintenant, vous êtes prêt à passer en revue les performances de votre modèle et définir un score seuil.

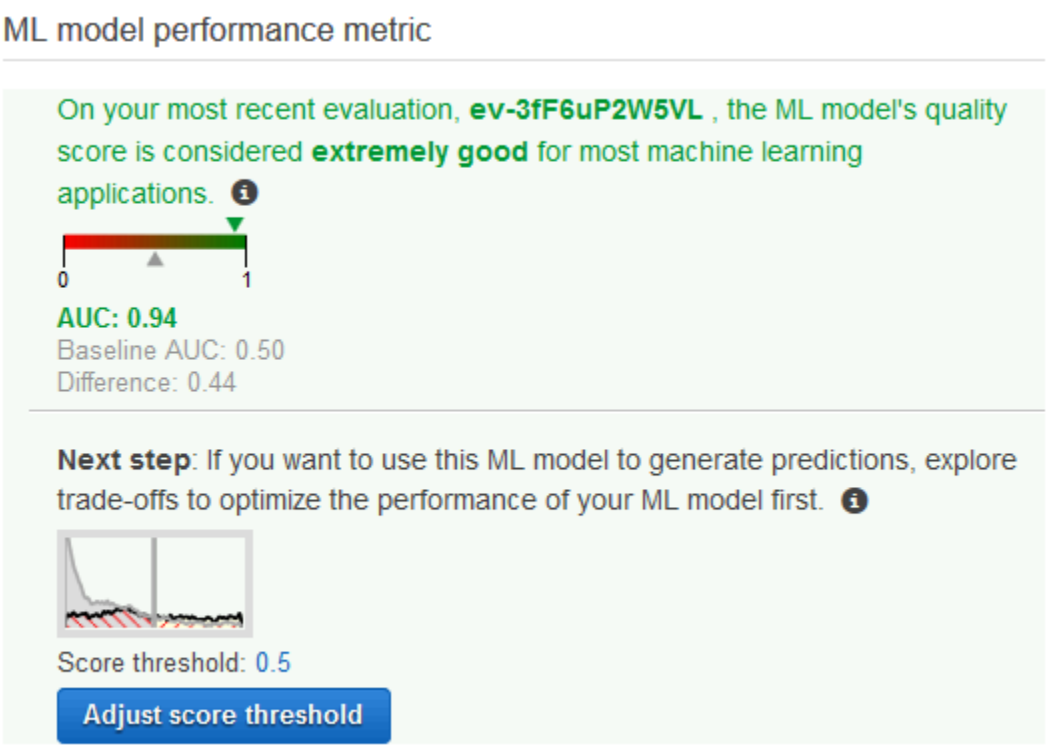
Pour plus d'informations sur la formation et l'évaluation des modèles, consultez Formation des modèles d'apprentissage-machine et evaluate an ML model.

Etape 4 : Examen des performances prédictives du modèle d'apprentissage-machine et définition d'un score seuil

Maintenant que vous avez créé votre modèle de machine learning et qu'Amazon Machine Learning (Amazon ML) l'a évalué, voyons s'il est suffisamment performant pour être utilisé. Au cours de l'évaluation, Amazon ML a calculé une métrique de qualité standard, appelée métrique Area Under a Curve (AUC), qui exprime la qualité des performances de votre modèle de ML. Amazon ML interprète également la métrique AUC pour vous indiquer si la qualité du modèle ML est adéquate pour la plupart des applications de machine learning. (Découvrez plus d'informations sur AUC dans Mesure de la précision du modèle d'apprentissage-machine.) Examinons la métrique AUC, puis ajustez le score seuil ou la limite pour optimiser les performances prédictives de votre modèle.

Pour examiner la métrique AUC de votre modèle d'apprentissage-machine

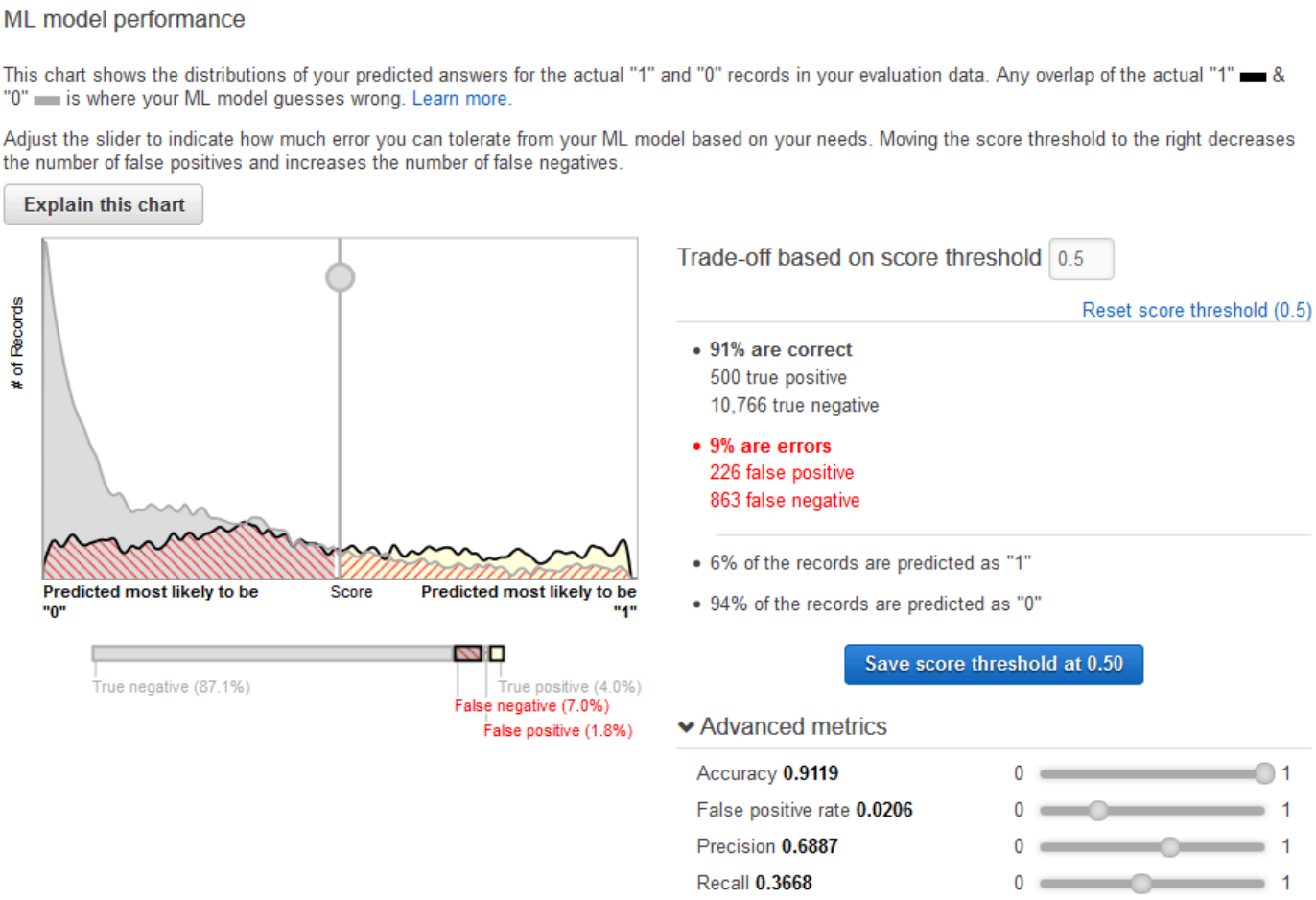
1. Dans la page ML model summary, dans le volet de navigation ML model report, choisissez Evaluations, Evaluation: ML model: Banking model 1, puis Summary.
2. Dans la page Evaluation summary, examinez le résumé d'évaluation, dont notamment la métrique de performances AUC du modèle.



Le modèle d'apprentissage-machine génère des scores de prédiction numériques pour chaque enregistrement dans une source de données de prédiction, puis applique un seuil pour convertir ces scores en étiquettes binaires 0 (pour non) ou 1 (pour oui). En changeant le score seuil, vous pouvez ajuster la manière dont le modèle d'apprentissage-machine attribue ces étiquettes. Maintenant, définissez le score seuil.

Pour définir un score seuil pour votre modèle d'apprentissage-machine

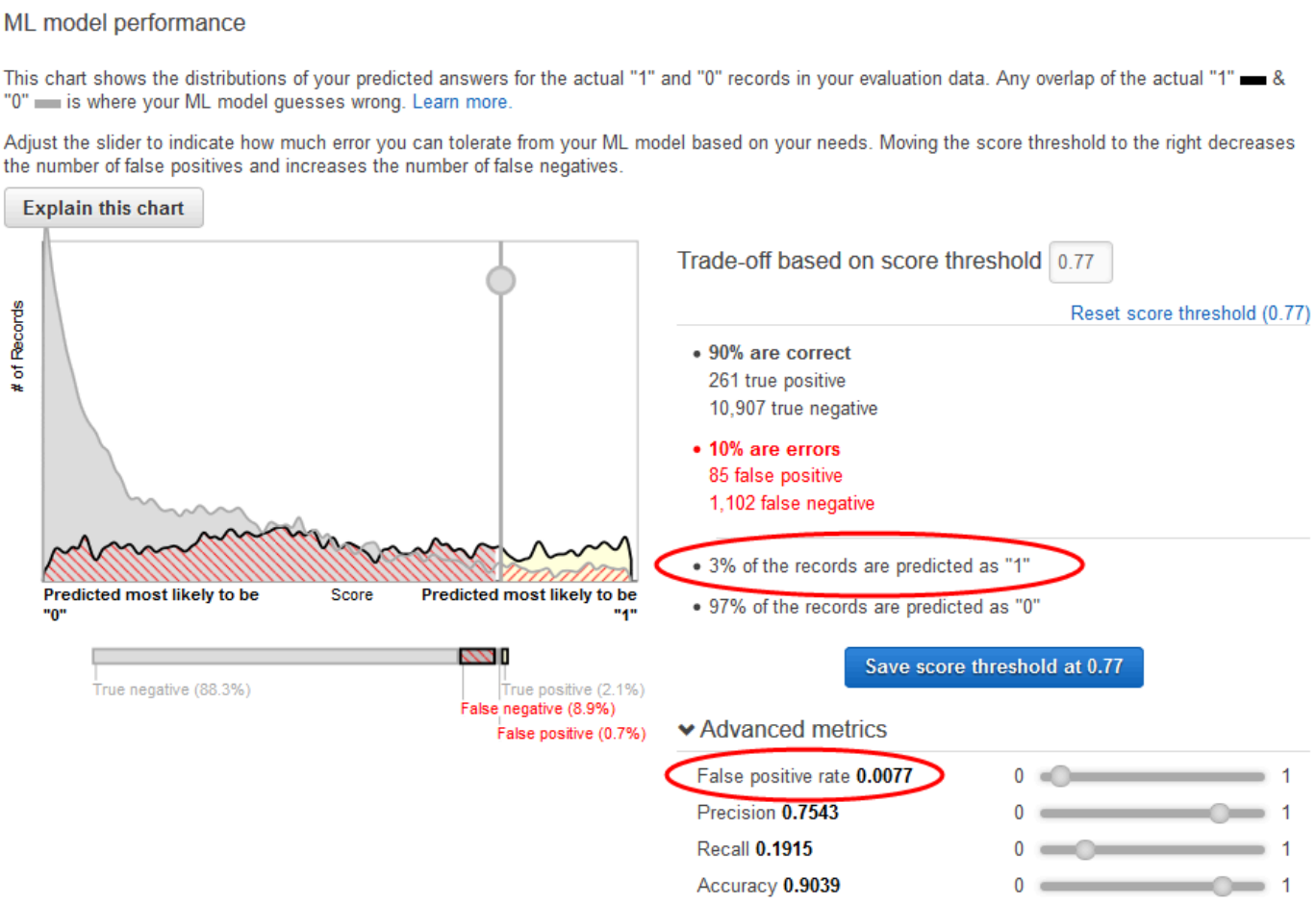
1. Dans la page Evaluation Summary, choisissez Adjust Score Threshold.



Vous pouvez affiner les métriques de performances de votre modèle d'apprentissage-machine en ajustant le score seuil. L'ajustement de cette valeur modifie le niveau de confiance que le modèle doit avoir dans une prédiction avant de considérer la prédiction comme positive. Il change également le nombre de faux négatifs et de faux positifs que vous êtes prêt à tolérer dans vos prévisions.

Vous pouvez contrôler la limite pour ce que le modèle considère comme une prédiction positive en augmentant le score seuil jusqu'à ce qu'il considère comme positives uniquement les prédictions dotées de la plus haute probabilité d'être de vrais positifs. Vous pouvez également réduire le score seuil jusqu'à ce que vous n'ayez plus aucun faux négatif. Choisissez votre limite pour refléter les besoins de votre activité. Dans le cadre de ce didacticiel, chaque faux positif a un coût financier sur la campagne, si bien que nous voulons un haut ratio de vrais positifs par rapport aux faux positifs.

2. Disons que vous voulez cibler les 3 % supérieurs des clients qui souscrivent au produit. Faites glisser le sélecteur vertical pour définir le score seuil sur une valeur qui correspond à 3% of the records are predicted as "1".



Notez l'impact de ce score seuil sur les performances du modèle d'apprentissage-machine : le taux de faux positifs est de 0,007. Supposons que ce taux de faux positifs est acceptable.

3. Choisissez Save score threshold at 0.77.

Chaque fois que vous utilisez ce modèle d'apprentissage-machine pour faire des prédictions, il prédit les enregistrements avec des scores supérieurs à 0,77 en tant que « 1 », et le reste des enregistrements en tant que « 0 ».

Pour en savoir plus sur le score seuil, consultez Classification binaire.

Maintenant, vous êtes prêt à créer des prédictions à l'aide de votre modèle.

Etape 5 : Utilisation du modèle d'apprentissage-machine pour générer des prédictions

Amazon Machine Learning (Amazon ML) peut générer deux types de prédictions : par lots et en temps réel.

Une prédiction en temps réel est une prédiction pour une seule observation générée par Amazon ML à la demande. Les prédictions en temps réel sont idéales pour les applications mobiles, les sites web et les autres applications qui doivent utiliser les résultats de façon interactive.

Une prédiction par lots est un ensemble de prédictions pour un groupe d'observations. Amazon ML traite les enregistrements d'une prédiction par lots en même temps. Le traitement peut donc prendre un certain temps. Utilisez des prédictions par lots pour les applications qui nécessitent des prédictions pour des ensembles d'observations ou des prédictions qui n'utilisent pas les résultats de façon interactive.

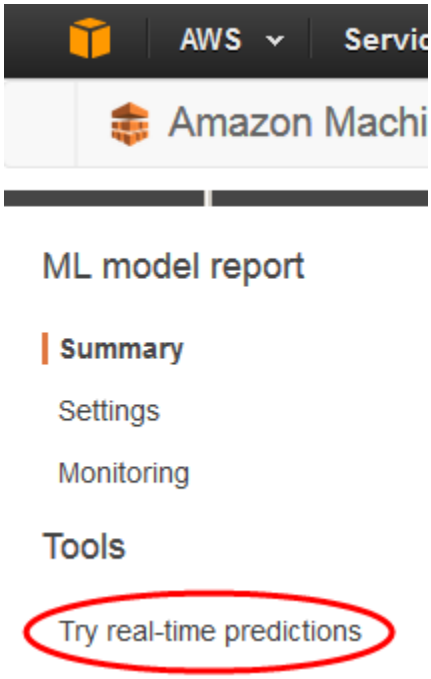
Dans le cadre de ce didacticiel, vous allez générer une prédiction en temps réel qui prédira si un client potentiel souscrira au nouveau produit. Vous allez également générer des prédictions pour un grand lot de clients potentiels. Pour la prédiction par lots, vous utiliserez le fichier `banking-batch.csv` que vous avez chargé dans Etape 1 : Préparation de vos données.

Commençons par une prédiction en temps réel.

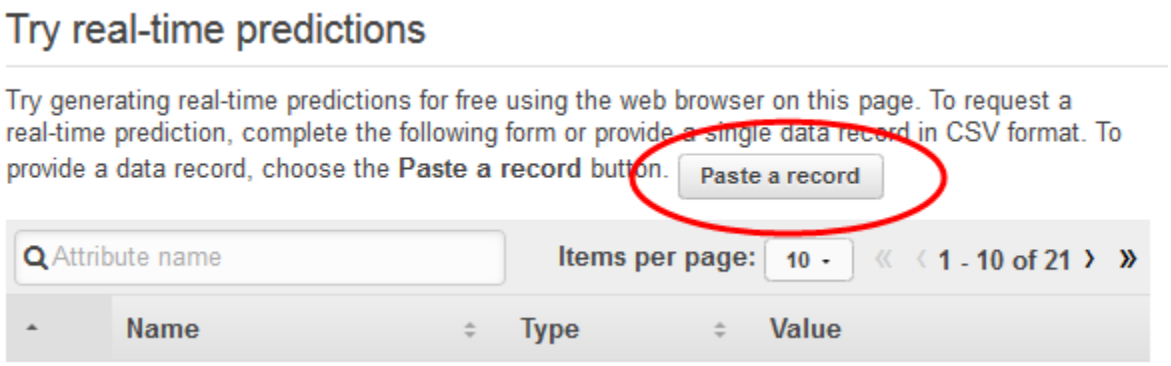
> **Note**
>
> Pour les applications qui nécessitent des prédictions en temps réel, vous devez créer un point de terminaison en temps réel pour le modèle d'apprentissage-machine. Vous accumulez des frais lorsqu'un point de terminaison en temps réel est disponible. Avant de valider l'utilisation de prédictions en temps réel et de commencer à supporter le coût qui leur est associé, vous pouvez essayer la fonctionnalité de prédiction en temps réel dans votre navigateur web, sans créer de point de terminaison en temps réel. C'est ce que nous allons faire dans le cadre de ce tutoriel.

Pour essayer d'utiliser les prédictions en temps réel

1. Dans le volet de navigation ML model report, choisissez Try real-time predictions.



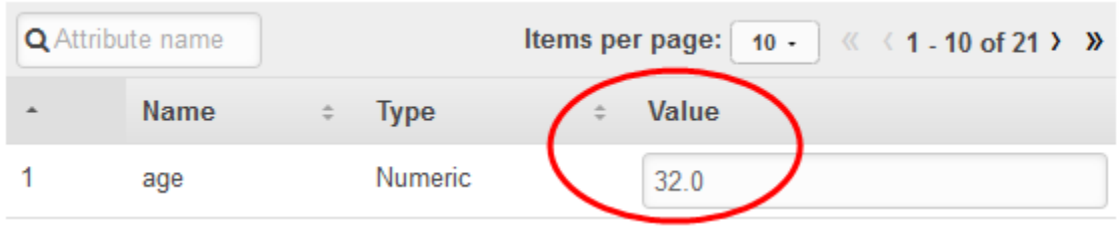
2. Choisissez Paste a record.



3. Dans la boîte de dialogue Paste a record, collez l'observation suivante :

```
32,services,divorced,basic.9y,no,unknown,yes,cellular,dec,mon,110,1,11,0,nonexistent,-1.8,9
```

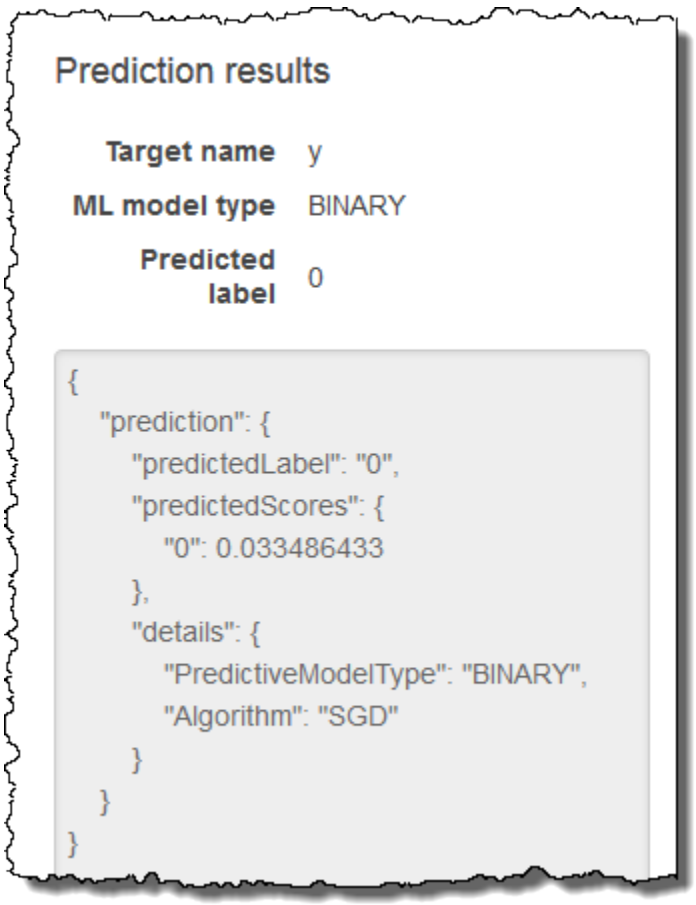
4. Dans la boîte de dialogue Coller un enregistrement, choisissez Soumettre pour confirmer que vous souhaitez générer une prédiction pour cette observation. Amazon ML renseigne les valeurs dans le formulaire de prédiction en temps réel.



> **Note**
>
> Vous pouvez également renseigner les champs Valeur en y saisissant des valeurs individuelles. Quelle que soit la méthode que vous choisissez, vous devez fournir une observation qui n'a pas été utilisée pour former le modèle.

5. En bas de la page, choisissez Create prediction.

   La prédiction apparaît dans le volet Prediction results de droite. Cette prédiction a un paramètre Predicted label égal à `0`, ce qui signifie qu'il est peu probable que ce client potentiel réponde à la campagne. Un paramètre Predicted label égal à 1 signifierait qu'il est probable que le client réponde à la campagne.

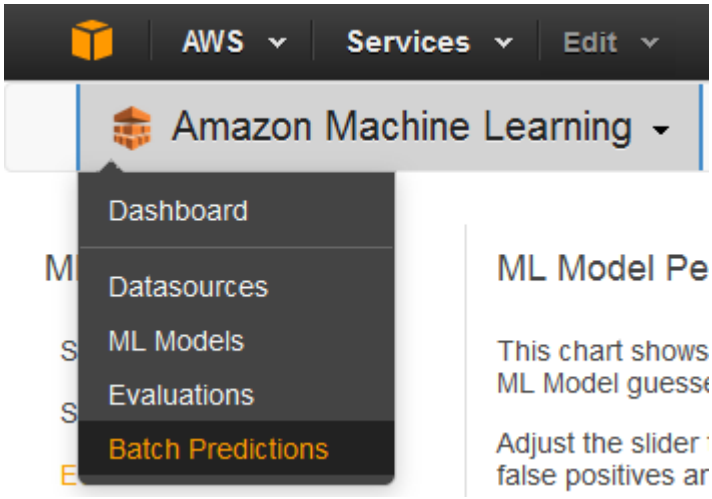


A présent, créez une prédiction par lots. Vous fournirez à Amazon ML le nom du modèle de ML que vous utilisez ; l'emplacement Amazon Simple Storage Service (Amazon S3) des données d'entrée pour lesquelles vous souhaitez générer des prédictions (Amazon ML créera une source de données

de prédiction par lots à partir de ces données) ; et l'emplacement Amazon S3 pour stocker les résultats.

Pour créer une prédiction par lots

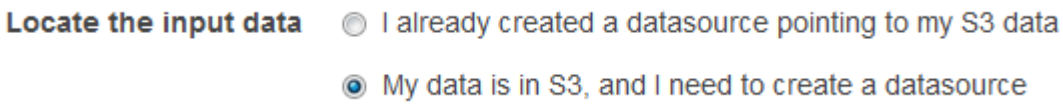
1. Choisissez Amazon Machine Learning, puis Batch Predictions.



2. Choisissez Create new batch prediction.
3. Dans la page ML model for batch predictions, choisissez ML model: Banking Data 1.

   Amazon ML affiche le nom du modèle ML, son ID, l'heure de création et l'ID de source de données associé.
4. Choisissez Continuer.
5. Pour générer des prédictions, vous devez fournir à Amazon ML les données pour lesquelles vous avez besoin de prédictions. Il s'agit des données d'entrée. Tout d'abord, placez les données d'entrée dans une source de données afin qu'Amazon ML puisse y accéder.

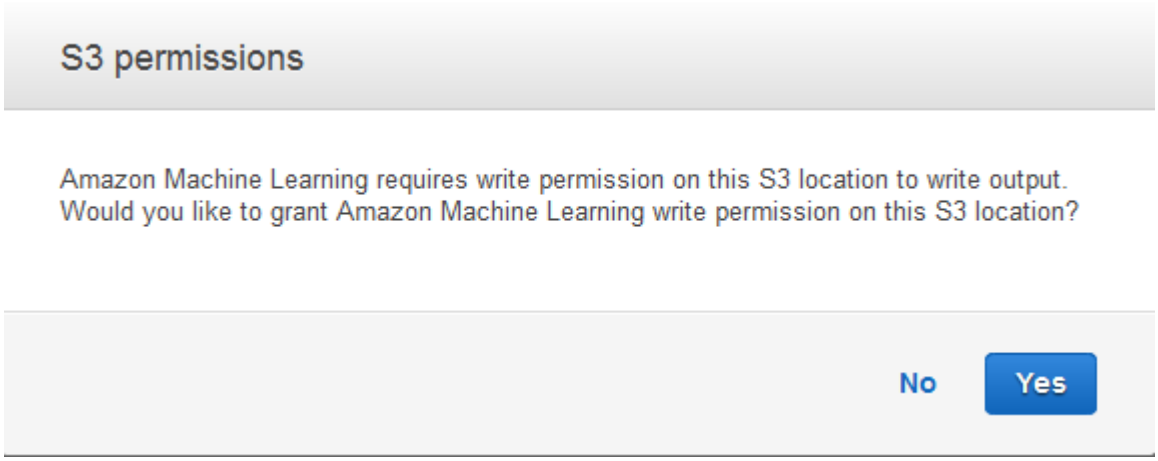
   Pour Locate the input data, choisissez My data is in S3, and I need to create a datasource.



6. Pour Datasource name, tapez **`Banking Data 2`**.
7. Pour S3 Location, saisissez l'emplacement complet du `banking-batch.csv` fichier : *`your-bucket`***`/banking-batch.csv`**.
8. Pour Does the first line in your CSV contain the column names?, choisissez Yes.
9. Choisissez Vérifier.

Amazon ML valide l'emplacement de vos données.

10. Choisissez Continuer.
11. Pour la destination S3, saisissez le nom de l'emplacement Amazon S3 où vous avez chargé les fichiers à l'étape 1 : Préparez vos données. Amazon ML y télécharge les résultats des prédictions.
12. Pour le nom de la prédiction Batch, acceptez la valeur par défaut,**`Batch prediction: ML model: Banking Data 1`**. Amazon ML choisit le nom par défaut en fonction du modèle qu'il utilisera pour créer des prédictions. Dans ce didacticiel, le modèle et les prédictions sont nommés d'après la source de données de formation, `Banking Data 1`.
13. Choisissez Examiner.
14. Dans la boîte de dialogue S3 permissions, choisissez Oui.

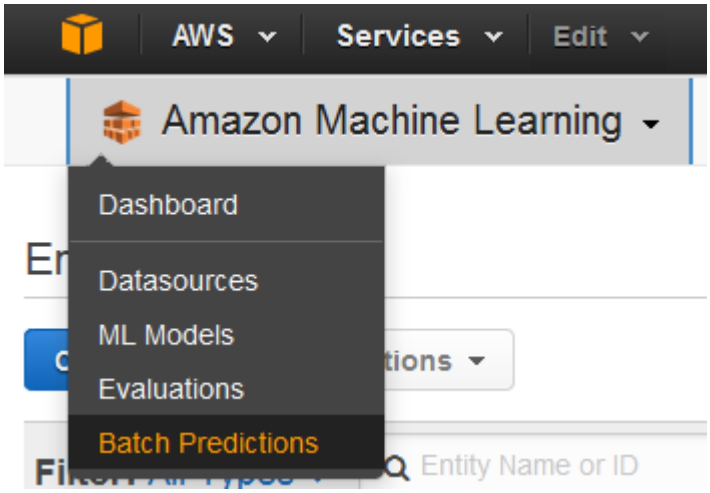


15. Dans la page Vérification, choisissez Terminer.

    La demande de prédiction par lots est envoyée à Amazon ML et entrée dans une file d'attente. Le temps nécessaire à Amazon ML pour traiter une prédiction par lots dépend de la taille de votre source de données et de la complexité de votre modèle de machine machine learning. Pendant qu'Amazon ML traite la demande, il indique le statut En cours. Une fois que la prédiction par lots est terminée, l'état de la demande devient Terminé. Maintenant, vous pouvez afficher les résultats.

Pour afficher les prédictions

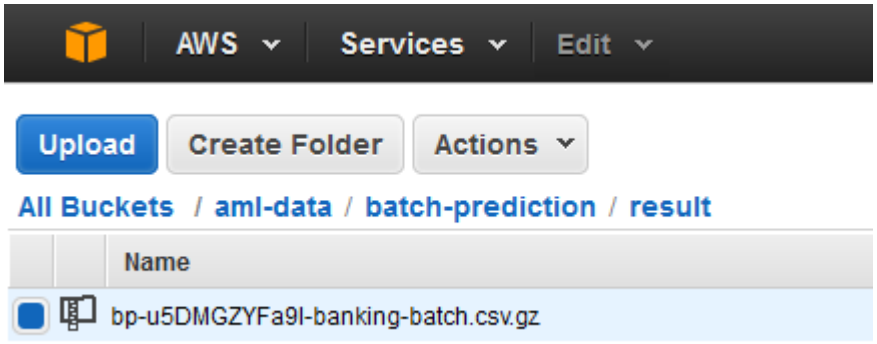
1. Choisissez Amazon Machine Learning, puis Batch Predictions.



2. Dans la liste des prédictions, choisissez la prédiction Batch prediction: ML model: Banking Data
   1. La page Batch prediction info apparaît.

Name	Subscription propensity Predictions
ID	bp-u5DMGZYFa9l
Creation Time	Mar 5, 2015 3:28:33 PM
Status	Completed
Log	Download Log
Datasource ID	ds-33Rqgz9w3ee
ML Model ID	ml-u7ljoShX2kX
Input S3 URL	s3://aml-data/banking-batch.csv
Output S3 URL	s3://aml-data/

3. Pour consulter les résultats de la prédiction par lots, accédez à la console Amazon S3 à l'adresse https://console.aws.amazon.com/s3/et accédez à l'emplacement Amazon S3 référencé dans le champ URL de sortie S3. De là, accédez au dossier des résultats, qui aura un nom similaire à `s3://aml-data/batch-prediction/result`.



La prédiction est stockée dans un fichier .gzip compressé avec l'extension .gz.

4. Téléchargez le fichier de prédiction sur votre bureau, décompressez-le et ouvrez-le.

bestAnswer	score
0	0.06046
0	0.00507
0	0.01410
0	0.00170
0	0.00184
0	0.07133
0	0.30811

Le fichier possède deux colonnes, bestAnswer et score, et une ligne pour chaque observation de votre source de données. Les résultats de la colonne bestAnswer sont basés sur le score seuil de 0,77 que vous avez défini dans Etape 4 : Examen des performances prédictives du modèle d'apprentissage-machine et définition d'un score seuil. Un score supérieur à 0,77 génère une valeur bestAnswer de 1, ce qui représente une prédiction ou une réponse positive, et un score inférieur à 0,77 génère une valeur bestAnswer de 0, ce qui représente une prédiction ou une réponse négative.

Les exemples suivants présentent des prédictions positives ou négatives en fonction du score seuil de 0,77.

Prédiction positive :

bestAnswer	score
1	0.8228876

Dans cet exemple, la valeur de bestAnswer est 1 et la valeur de score est 0,8228876. La valeur de bestAnswer est 1 parce que le score est supérieur au score seuil de 0,77. Une valeur bestAnswer égale à 1 indique qu'il est probable que le client achète votre produit, ce qui est donc considéré comme une prédiction positive.

Prédiction négative :

bestAnswer	score
0	0.7695356

Dans cet exemple, la valeur de bestAnswer est 0, car la valeur de score est 0,7695356, qui est inférieure au score seuil de 0,77. La valeur bestAnswer de 0 indique qu'il est improbable que le client achète votre produit, ce qui est donc considéré comme une prédiction négative.

Chaque ligne du résultat correspond à une ligne dans votre lot d'entrée (une observation dans votre source de données).

Après l'analyse des prédictions, vous pouvez exécuter votre campagne marketing ciblée ; par exemple, en envoyant des dépliants à toutes les personnes dotées d'un score prédit de 1.

Maintenant que vous avez créé, revu et utilisé votre modèle, nettoyez les données et les ressources AWS que vous avez créées pour éviter d'accumuler des frais inutiles et pour maintenir dégagé votre espace de travail.

Étape 6 : nettoyer

Pour éviter d'avoir à payer des frais supplémentaires liés à Amazon Simple Storage Service (Amazon S3), supprimez les données stockées dans Amazon S3. Les autres ressources Amazon ML non utilisées ne vous sont pas facturées, mais nous vous recommandons de les supprimer pour préserver la propreté de votre espace de travail.

Pour supprimer les données d'entrée stockées dans Amazon S3

1. Ouvrez la console Amazon S3 à l'adresse https://console.aws.amazon.com/s3/.
2. Accédez à l'emplacement Amazon S3 où vous avez stocké les `banking-batch.csv` fichiers `banking.csv` et.
3. Sélectionnez les fichiers `banking.csv`, `banking-batch.csv` et `.writePermissionCheck.tmp`.
4. Choisissez Actions, puis Supprimer.
5. Lorsque vous êtes invité à confirmer l'opération, choisissez OK.

Bien que la conservation de la prédiction par lots exécutée par Amazon ML ou des sources de données, du modèle et de l'évaluation que vous avez créés pendant le didacticiel ne vous soit pas facturée, nous vous recommandons de les supprimer pour éviter d'encombrer votre espace de travail.

Pour supprimer les prédictions par lots

1. Accédez à l'emplacement Amazon S3 où vous avez stocké le résultat de la prédiction par lots.

2. Choisissez le dossier `batch-prediction`.
3. Choisissez Actions, puis Supprimer.
4. Lorsque vous êtes invité à confirmer l'opération, choisissez OK.

Pour supprimer les ressources Amazon ML

1. Sur le tableau de bord Amazon ML, sélectionnez les ressources suivantes.
   - La source de données `Banking Data 1`
   - La source de données `Banking Data 1_[percentBegin=0, percentEnd=70, strategy=sequential]`
   - La source de données `Banking Data 1_[percentBegin=70, percentEnd=100, strategy=sequential]`
   - La source de données `Banking Data 2`
   - Le modèle d'apprentissage-machine `ML model: Banking Data 1`
   - L'évaluation `Evaluation: ML model: Banking Data 1`
2. Choisissez Actions, puis Supprimer.
3. Dans la boîte de dialogue, choisissez Supprimer pour supprimer toutes les ressources sélectionnées.

Vous avez maintenant terminé le didacticiel. Pour continuer à utiliser la console afin de créer des sources de données, des modèles et des prédictions, consultez le manuel Amazon Machine Learning Developer Guide. Pour apprendre à utiliser l'API, consultez la Référence d'API Amazon Machine Learning.

Création et utilisation des sources de données

Vous pouvez utiliser les sources de données Amazon ML pour entraîner un modèle ML, évaluer un modèle ML et générer des prédictions par lots à l'aide d'un modèle ML. Les objets source de données contiennent des métadonnées relatives à vos données d'entrée. Lorsque vous créez une source de données, Amazon ML lit vos données d'entrée, calcule des statistiques descriptives sur ses attributs et stocke les statistiques, un schéma et d'autres informations dans le cadre de l'objet de source de données. Après avoir créé une source de données, vous pouvez utiliser les informations de données Amazon ML pour explorer les propriétés statistiques de vos données d'entrée, et vous pouvez utiliser la source de données pour entraîner un modèle de machine machine learning.

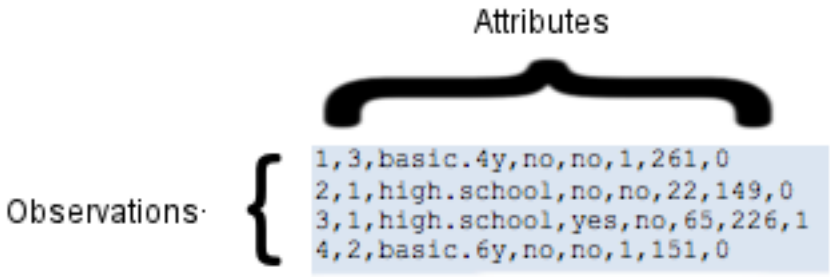
> **Note**
>
> Cette section part du principe que vous connaissez les concepts d'Amazon Machine Learning.

Rubriques

Comprendre le format de données pour Amazon ML

Les données d'entrée sont les données que vous utilisez pour créer une source de données. Vous devez enregistrer vos données d'entrée au format CSV (valeurs séparées par des virgules). Chaque ligne du fichier .csv est un enregistrement de données ou une observation unique. Chaque colonne du fichier .csv contient un attribut de l'observation. Par exemple, la figure suivante illustre le contenu d'un fichier .csv qui compte quatre observations, chacune sur sa propre ligne. Chaque observation contient huit attributs, séparés par des virgules. Les attributs représentent les informations suivantes

concernant chaque individu représenté par une observation : CustomerID, JoBid, éducation, logement, prêt, campagne, durée, campagne. willRespondTo



Attributs

Amazon ML nécessite un nom pour chaque attribut. Vous pouvez spécifier les noms des attributs en :

- incluant les noms des attributs dans la première ligne (également connue sous le nom de ligne d'en-tête) du fichier .csv que vous utilisez en tant que données d'entrée ;
- incluant les noms des attributs dans un fichier de schéma distinct qui est situé dans le même compartiment S3 que vos données d'entrée.

Pour plus d'informations sur l'utilisation des fichiers de schéma, consultez Création d'un schéma de données.

L'exemple suivant d'un fichier .csv comprend les noms des attributs dans la ligne d'en-tête.

```
customerId,jobId,education,housing,loan,campaign,duration,willRespondToCampaign

1,3,basic.4y,no,no,1,261,0

2,1,high.school,no,no,22,149,0

3,1,high.school,yes,no,65,226,1

4,2,basic.6y,no,no,1,151,0
```

Exigences en matière de format du fichier d'entrée

Le fichier .csv qui contient vos données d'entrée doit répondre aux exigences suivantes :

- Il doit être en texte brut et utiliser un jeu de caractères tel qu'ASCII, Unicode ou EBCDIC.
- Il est composé d'observations, une observation par ligne.
- Pour chaque observation, les valeurs d'attribut doivent être séparées par des virgules.
- Si une valeur d'attribut contient une virgule (délimiteur), la totalité de la valeur d'attribut doit être entre guillemets doubles.
- Chaque observation doit être terminée par un end-of-line caractère spécial ou une séquence de caractères indiquant la fin d'une ligne.
- Les valeurs d'attribut ne peuvent pas inclure de end-of-line caractères, même si la valeur d'attribut est placée entre guillemets.
- Chaque observation doit avoir le même nombre d'attributs et la même séquence d'attributs.
- Chaque observation ne doit pas dépasser 100 Ko. Amazon ML rejette toute observation supérieure à 100 Ko pendant le traitement. Si Amazon ML rejette plus de 10 000 observations, il rejette l'intégralité du fichier .csv.

Utilisation de plusieurs fichiers comme entrée de données dans Amazon ML

Vous pouvez fournir vos données à Amazon ML sous forme de fichier unique ou de collection de fichiers. Les collections doivent satisfaire les conditions suivantes :

- Tous les fichiers doivent avoir le même schéma de données.
- Tous les fichiers doivent résider dans le même préfixe Amazon Simple Storage Service (Amazon S3), et le chemin que vous fournissez pour la collection doit se terminer par une barre oblique («/»).

Par exemple, si vos fichiers de données sont nommés input1.csv, input2.csv et input3.csv, et que le nom de votre compartiment S3 est s3://examplebucket, les chemins de vos fichiers peuvent ressembler à :

s3 ://examplebucket/path/to/data/input1.csv

s3 ://examplebucket/path/to/data/input2.csv

s3 ://examplebucket/path/to/data/input3.csv

Vous devez fournir l'emplacement S3 suivant en entrée à Amazon ML :

's3 :///' examplebucket/path/to/data

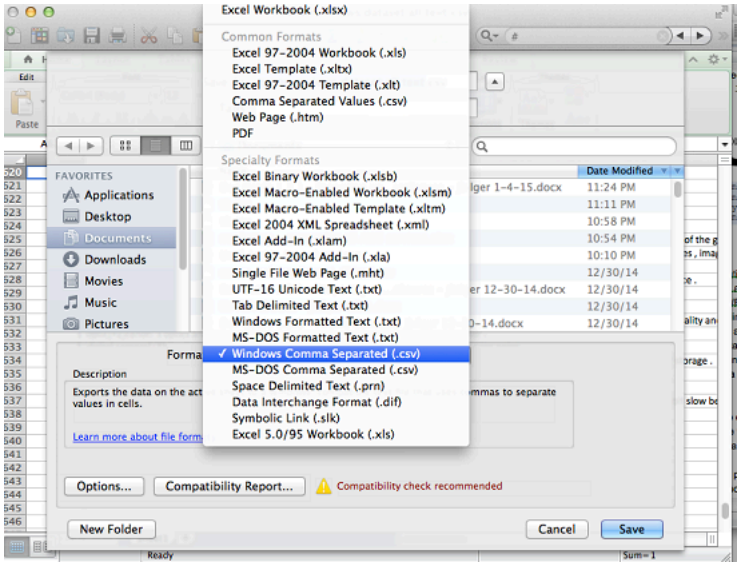
End-of-Line Caractères au format CSV

Lorsque vous créez votre fichier .csv, chaque observation sera terminée par un end-of-line caractère spécial. Ce caractère n'est pas visible, mais il est automatiquement inclus à la fin de chaque observation lorsque vous appuyez sur votre touche Entrée ou Return. Le caractère spécial qui représente le end-of-line varie en fonction de votre système d'exploitation. Les systèmes Unix, tels que Linux ou OS X, utilisent un caractère de saut de ligne indiqué par « \n » (code ASCII 10 en notation décimale ou 0x0a en notation hexadécimale). Microsoft Windows utilise deux caractères appelés retour chariot et saut de ligne qui sont indiqués par « \r\n » (codes ASCII 13 et 10 en notation décimale, ou 0x0d et 0x0a en notation hexadécimale).

Si vous souhaitez utiliser OS X et Microsoft Excel pour créer votre fichier .csv, effectuez la procédure suivante. Veillez à choisir le format correct.

Pour enregistrer un fichier .csv si vous utilisez OS X et Excel

1. Quand vous enregistrez le fichier .csv, choisissez Format, puis choisissez CSV (Windows) (séparateur : point-virgule) (.csv).
2. Choisissez Save (Enregistrer).



> **Important**
>
> N'enregistrez pas le fichier .csv en utilisant les formats valeurs séparées par des virgules (.csv) ou MS-DOS séparés par des virgules (.csv) car Amazon ML ne peut pas les lire.

Création d'un schéma de données pour Amazon ML

Un schéma est composé de tous les attributs figurant dans les données d'entrée et de leurs types de données correspondants. Cela permet à Amazon ML de comprendre les données de la source de données. Amazon ML utilise les informations du schéma pour lire et interpréter les données d'entrée, calculer les statistiques, appliquer les transformations d'attributs correctes et affiner ses algorithmes d'apprentissage. Si vous ne fournissez pas de schéma, Amazon ML en déduit un à partir des données.

Exemple de schéma

Pour qu'Amazon ML puisse lire correctement les données d'entrée et produire des prédictions précises, le type de données approprié doit être attribué à chaque attribut. Passons en revue un exemple pour voir comment les types de données sont attribués aux attributs et comment les attributs et les types de données sont inclus dans un schéma. Nous appellerons notre exemple « Campagne clients », car nous voulons prédire quels clients répondront à notre campagne d'e-mailing. Notre fichier d'entrée est un fichier .csv contenant neuf colonnes :

```
1,3,web developer,basic.4y,no,no,1,261,0

2,1,car repair,high.school,no,no,22,149,0

3,1,car mechanic,high.school,yes,no,65,226,1

4,2,software developer,basic.6y,no,no,1,151,0
```

Le schéma de ces données est le suivant :

```
{
    "version": "1.0",
    "rowId": "customerId",
    "targetAttributeName": "willRespondToCampaign",
    "dataFormat": "CSV",
    "dataFileContainsHeader": false,
    "attributes": [
        {
            "attributeName": "customerId",
            "attributeType": "CATEGORICAL"
        },
        {
```

```
            "attributeName": "jobId",
            "attributeType": "CATEGORICAL"
        },
        {
            "attributeName": "jobDescription",
            "attributeType": "TEXT"
        },
        {
            "attributeName": "education",
            "attributeType": "CATEGORICAL"
        },
        {
            "attributeName": "housing",
            "attributeType": "CATEGORICAL"

        },
        {
            "attributeName": "loan",
            "attributeType": "CATEGORICAL"

        },
        {
            "attributeName": "campaign",
            "attributeType": "NUMERIC"
        },
        {
            "attributeName": "duration",
            "attributeType": "NUMERIC"
        },

        {
            "attributeName": "willRespondToCampaign",
            "attributeType": "BINARY"
        }

    ]
}
```

Dans le fichier de schéma de cet exemple, la valeur de `rowId` est `customerId` :

```
"rowId": "customerId",
```

L'attribut `willRespondToCampaign` est défini en tant qu'attribut cible :

```
"targetAttributeName": "willRespondToCampaign ",
```

L'attribut `customerId` et le type de données `CATEGORICAL` sont associés à la première colonne, l'attribut `jobId` et le type de données `CATEGORICAL` sont associés à la deuxième colonne, l'attribut `jobDescription` et le type de données `TEXT` sont associés à la troisième colonne, l'attribut `education` et le type de données `CATEGORICAL` sont associés à la quatrième colonne, etc. La neuvième colonne est associée à l'attribut `willRespondToCampaign` avec un type de données `BINARY`, et cet attribut est également défini en tant qu'attribut cible.

Utilisation du targetAttributeName terrain

La valeur `targetAttributeName` est le nom de l'attribut que vous voulez prédire. Vous devez attribuer un `targetAttributeName` lors de la création ou de l'évaluation d'un modèle.

Lorsque vous entraînez ou évaluez un modèle de machine learning, `targetAttributeName` identifie le nom de l'attribut dans les données d'entrée qui contient les « bonnes » réponses pour l'attribut cible. Amazon ML utilise la cible, qui inclut les bonnes réponses, pour découvrir des modèles et générer un modèle de machine learning.

Lorsque vous évaluez votre modèle, Amazon ML utilise la cible pour vérifier l'exactitude de vos prévisions. Une fois que vous avez créé et évalué le modèle d'apprentissage-machine, vous pouvez utiliser les données avec un champ `targetAttributeName` non attribué pour générer des prédictions avec votre modèle d'apprentissage-machine.

Vous définissez l'attribut cible dans la console Amazon ML lorsque vous créez une source de données ou dans un fichier de schéma. Si vous créez votre propre fichier de schéma, utilisez la syntaxe suivante pour définir l'attribut cible :

```
"targetAttributeName": "exampleAttributeTarget",
```

Dans cet exemple, `exampleAttributeTarget` est le nom de l'attribut dans votre fichier d'entrée qui est l'attribut cible.

Utilisation du champ rowID

Le champ `row ID` est un indicateur optionnel associé à un attribut dans les données d'entrée. S'il est spécifié, l'attribut marqué en tant que `row ID` est inclus dans la prédiction fournie en sortie. Cet attribut permet d'associer plus facilement les prédictions aux observations correspondantes. Par exemple, un ID client ou un attribut unique similaire constitue un champ `row ID` efficace.

> **Note**
>
> L'ID de ligne est fourni à titre de référence uniquement. Amazon ML ne l'utilise pas lors de la formation d'un modèle de ML. Le fait de sélectionner un attribut en tant qu'ID de ligne exclut l'éventualité d'utiliser l'attribut pour former un modèle d'apprentissage-machine.

Vous le définissez `row ID` dans la console Amazon ML lorsque vous créez une source de données ou dans un fichier de schéma. Si vous créez votre propre fichier de schéma, utilisez la syntaxe suivante pour définir le champ `row ID` :

```
"rowId": "exampleRow",
```

Dans l'exemple précédent, `exampleRow` est le nom de l'attribut figurant dans votre fichier d'entrée qui est défini comme ID de ligne.

Lors de la génération de prédictions par lots, vous pouvez obtenir le résultat suivant :

```
tag,bestAnswer,score
55,0,0.46317
102,1,0.89625
```

Dans cet exemple, `RowID` représente l'attribut `customerId`. Par exemple, il est prédit que le client customerId `55` répondra à notre campagne d'e-mailing avec un faible degré de confiance (0,46317), tandis qu'il est prédit que le client `customerId` `102` répondra à notre campagne d'e-mailing avec un niveau de confiance élevé (0,89625).

Utilisation du AttributeType terrain

Dans Amazon ML, il existe quatre types de données pour les attributs :

Binaire

Choisissez `BINARY` pour un attribut qui a seulement deux états possibles, tels que `yes` et `no`.

Par exemple, l'attribut `isNew`, permettant de vérifier si une personne est un nouveau client, a une valeur `true` si la personne est un nouveau client ou une valeur `false` si elle n'est pas un nouveau client.

Les valeurs négatives valides sont `0`, `n`, `no`, `f` et `false`.

Les valeurs positives valides sont `1`, `y`, `yes`, `t` et `true`.

Amazon ML ignore le cas des entrées binaires et supprime l'espace blanc environnant. Par exemple, `" FaLSe "` est une valeur binaire valide. Vous pouvez mélanger les valeurs binaires que vous utilisez dans la même source de données, par exemple en utilisant `trueno`, et`1`. Amazon ML produit des sorties uniquement `0` et `1` pour les attributs binaires.

Categorical (catégorie)

Choisissez `CATEGORICAL` pour un attribut qui accepte un nombre limité de valeurs de chaîne uniques. Par exemple, un ID d'utilisateur, le mois et un code postal sont des valeurs de catégorie. Les attributs de catégorie sont traités comme une chaîne individuelle et ne sont pas tokenisés davantage.

Numérique

Choisissez `NUMERIC` pour un attribut qui accepte une quantité en tant que valeur.

Par exemple, une température, un poids et un taux de clics sont des valeurs numériques.

Tous les attributs qui contiennent des nombres ne sont pas numériques. Les attributs catégoriels, tels que les jours du mois et IDs, sont souvent représentés par des nombres. Pour être considéré comme numérique, un nombre doit être comparable à un autre nombre. Par exemple, l'ID client `664727` ne vous dit rien sur l'ID client `124552`, alors qu'un poids de `10` indique que l'attribut est plus lourd qu'un attribut avec un poids de `5`. Les jours du mois ne sont pas numériques, car le premier d'un mois peut se produire avant ou après le second d'un autre mois.

> **Note**
>
> Lorsque vous utilisez Amazon ML pour créer votre schéma, il attribue le type de `Numeric` données à tous les attributs utilisant des nombres. Si Amazon ML crée votre schéma, vérifiez qu'il n'y a pas d'assignations incorrectes et définissez ces attributs sur`CATEGORICAL`.

Text

Choisissez `TEXT` pour un attribut qui est une chaîne de mots. Lors de la lecture d'attributs de texte, Amazon ML les convertit en jetons, délimités par des espaces blancs.

Par exemple, `email subject` devient `email` et `subject`, et `email-subject here` devient `email-subject` et `here`.

Si le type de données d'une variable du schéma d'entraînement ne correspond pas au type de données de cette variable dans le schéma d'évaluation, Amazon ML modifie le type de données d'évaluation pour qu'il corresponde au type de données d'entraînement. Par exemple, si le schéma de données de formation attribue un type de données de `TEXT` à la variable`age`, mais que le schéma d'évaluation attribue un type de données de `NUMERIC` à`age`, Amazon ML traite les âges des données d'évaluation comme des `TEXT` variables plutôt que comme des variables. `NUMERIC`

Pour obtenir des informations sur les statistiques associées à chaque type de données, consultez Statistiques descriptives.

Fourniture d'un schéma à Amazon ML

Chaque source de données a besoin d'un schéma. Vous pouvez choisir entre deux méthodes pour fournir un schéma à Amazon ML :

- Autorisez Amazon ML à déduire les types de données de chaque attribut dans le fichier de données d'entrée et à créer automatiquement un schéma pour vous.
- Fournissez un fichier de schéma lorsque vous chargez vos données Amazon Simple Storage Service (Amazon S3).

Autoriser Amazon ML à créer votre schéma

Lorsque vous utilisez la console Amazon ML pour créer une source de données, Amazon ML utilise des règles simples, basées sur les valeurs de vos variables, pour créer votre schéma. Nous vous recommandons vivement de consulter le schéma créé par Amazon ML et de corriger les types de données s'ils ne sont pas exacts.

Fourniture d'un schéma

Après avoir créé votre fichier de schéma, vous devez le mettre à la disposition d'Amazon ML. Vous avez deux options :

1. Fournissez le schéma à l'aide de la console Amazon ML.

   Utilisez la console pour créer votre source de données et incluez le fichier de schéma en ajoutant l'extension .schema au nom de fichier de votre fichier de données d'entrée. Par exemple, si l'URI

Amazon Simple Storage Service (Amazon S3) vers vos données d'entrée est s3 :my-bucket-name///data/input.csv, the URI to your schema will be s3://my-bucket-name/data/input.csv.schema. Amazon ML localise automatiquement le fichier de schéma que vous fournissez au lieu d'essayer de déduire le schéma à partir de vos données.

Pour utiliser un répertoire de fichiers comme entrée de données dans Amazon ML, ajoutez l'extension .schema au chemin de votre répertoire. Par exemple, si vos fichiers de données se trouvent à l'emplacement s3 ://examplebucket/path/to/data/, the URI to your schema will be s3:// examplebucket/path/to/data/.schema.

2. Fournissez le schéma à l'aide de l'API Amazon ML.

   Si vous envisagez d'appeler l'API Amazon ML pour créer votre source de données, vous pouvez télécharger le fichier de schéma dans Amazon S3, puis fournir l'URI de ce fichier dans l'`DataSchemaLocationS3`attribut de l'`CreateDataSourceFromS3`API. Pour plus d'informations, consultez CreateDataSourceFromS3.

   Vous pouvez fournir le schéma directement dans la charge utile `CreateDataSource` de* `APIs` au lieu de l'enregistrer d'abord sur Amazon S3. Pour ce faire, placez la chaîne de schéma complète dans l'`DataSchema`attribut de `CreateDataSourceFromS3CreateDataSourceFromRDS`, ou `CreateDataSourceFromRedshift` APIs. Pour plus d'informations, consultez la Référence d'API Amazon Machine Learning.

Fractionnement des données

L'objectif fondamental d'un modèle d'apprentissage-machine est d'effectuer des prédictions précises sur des instances de données futures au-delà de celles utilisées pour former les modèles. Avant d'utiliser un modèle d'apprentissage-machine pour effectuer des prédictions, nous devons évaluer les performances prédictives du modèle. Pour estimer la qualité des prédictions d'un modèle d'apprentissage-machine avec des données qu'il ne connaît pas, nous pouvons réserver, ou fractionner, une partie des données pour lesquelles nous connaissons déjà la réponse comme indicateur pour les données futures, et évaluer la qualité avec laquelle le modèle d'apprentissage-machine prédit les réponses correctes pour ces données. Vous fractionnez la source de données de manière à obtenir une partie pour la source de données de formation et une partie pour la source de données d'évaluation.

Amazon ML propose trois options pour fractionner vos données :

- Pré-fractionner les données : vous pouvez diviser les données en deux emplacements d'entrée de données, avant de les télécharger sur Amazon Simple Storage Service (Amazon S3) et de créer deux sources de données distinctes avec elles.
- Fractionnement séquentiel Amazon ML : vous pouvez demander à Amazon ML de diviser vos données de manière séquentielle lors de la création des sources de données de formation et d'évaluation.
- Fractionnement aléatoire Amazon ML : vous pouvez demander à Amazon ML de fractionner vos données à l'aide d'une méthode aléatoire prédéfinie lors de la création des sources de données de formation et d'évaluation.

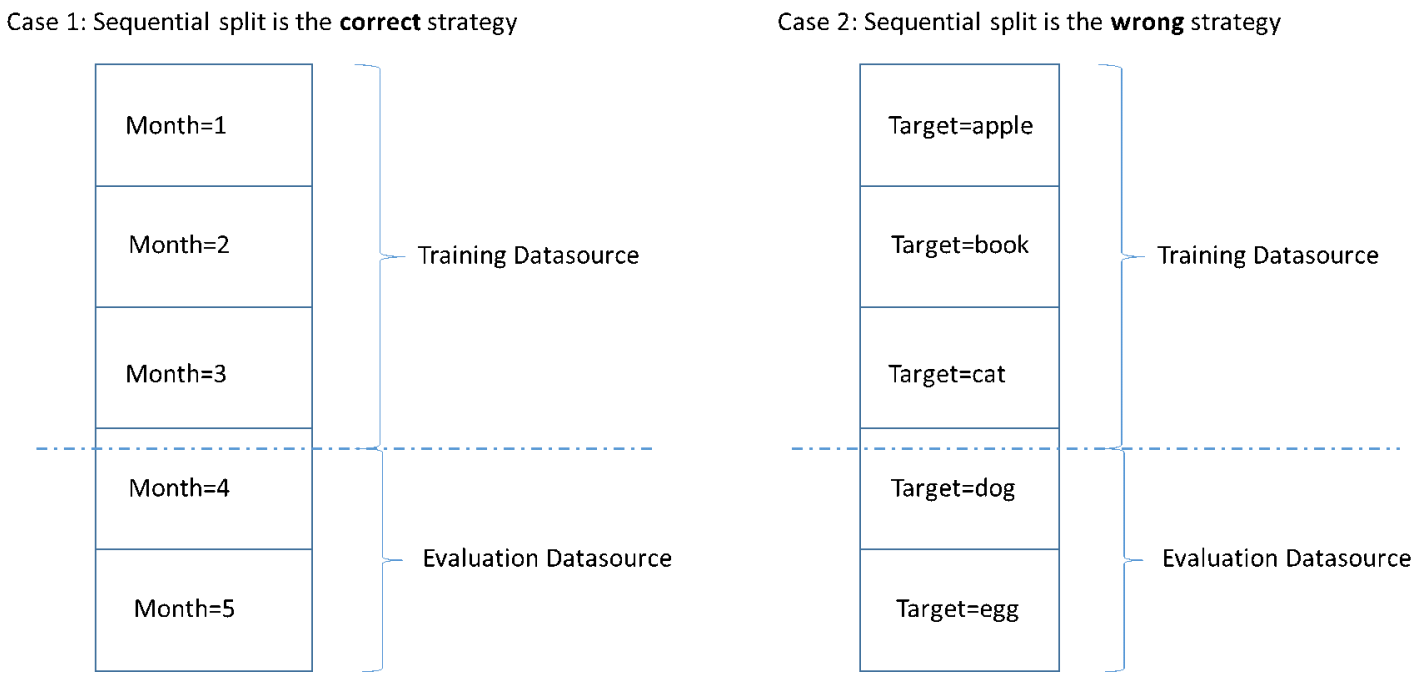
Pré-fractionnement des données

Si vous souhaitez un contrôle explicite des données dans vos sources de données de formation et d'évaluation, fractionnez les données en emplacements de données séparés et créez des sources de données distinctes pour les emplacements d'entrée et d'évaluation.

Fractionnement séquentiel des données

Un moyen simple de fractionner vos données d'entrée pour la formation et l'évaluation consiste à sélectionner des sous-ensembles sans chevauchement de vos données tout en conservant l'ordre des enregistrements de données. Cette approche est utile si vous souhaitez évaluer vos modèles d'apprentissage-machine sur les données pour une date donnée ou au sein d'une période donnée. Par exemple, imaginons que vous disposez des données d'engagement client des cinq derniers mois et que vous souhaitez utiliser ces données historiques pour prédire l'engagement des clients au mois suivant. L'utilisation du début de la période pour la formation et des données de la fin de la période pour l'évaluation peut générer une estimation plus précise de la qualité du modèle que l'utilisation des données d'enregistrement extraites de la plage de données complète.

La figure suivante montre des exemples illustrant quand vous devez utiliser une stratégie de fractionnement séquentiel et quand vous devez utiliser une stratégie aléatoire.



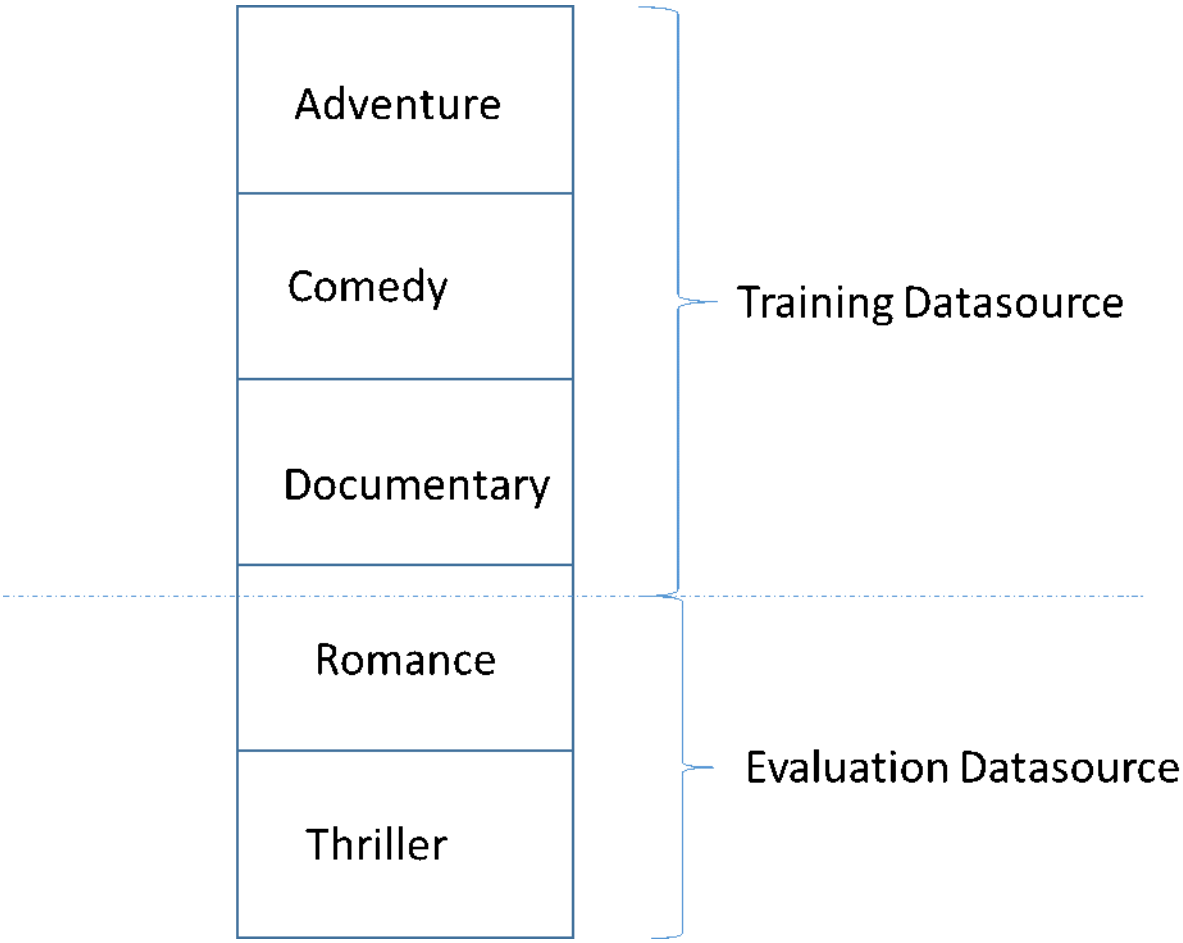
Lorsque vous créez une source de données, vous pouvez choisir de la diviser de manière séquentielle, et Amazon ML utilise les premiers 70 % de vos données pour la formation et les 30 % restants pour l'évaluation. Il s'agit de l'approche par défaut lorsque vous utilisez la console Amazon ML pour diviser vos données.

Fractionnement aléatoire des données

Le fractionnement aléatoire des données d'entrée dans les sources de données de formation et d'évaluation garantit une distribution des données similaire dans les sources de données de formation et d'évaluation. Choisissez cette option lorsque vous n'avez pas besoin de préserver l'ordre de vos données d'entrée.

Amazon ML utilise une méthode de génération de nombres pseudo-aléatoires prédéfinis pour diviser vos données. La valeur initiale est basée en partie sur une valeur de chaîne en entrée et en partie sur le contenu des données lui-même. Par défaut, la console Amazon ML utilise l'emplacement S3 des données d'entrée comme chaîne. Les utilisateurs d'API peuvent fournir une chaîne personnalisée. Cela signifie qu'avec le même compartiment S3 et les mêmes données, Amazon ML divise les données de la même manière à chaque fois. Pour modifier la façon dont Amazon ML divise les données, vous pouvez utiliser l'`CreateDatasourceFromRDS`API `CreateDatasourceFromS3``CreateDatasourceFromRedshift`, ou et fournir une valeur pour la chaîne de départ. Lorsque vous les utilisez APIs pour créer des sources de données distinctes pour la formation et l'évaluation, il est important d'utiliser la même valeur de chaîne de départ pour les

deux sources de données et l'indicateur de complément pour une source de données, afin de garantir qu'il n'y a pas de chevauchement entre les données d'entraînement et d'évaluation.



Lors du développement d'un modèle d'apprentissage-machine de haute qualité, un piège courant consiste à évaluer le modèle d'apprentissage-machine sur des données qui ne sont pas similaires à celles utilisées pour la formation. Par exemple, imaginons que vous utilisez l'apprentissage-machine pour prédire le genre de films et que vos données de formation contiennent des films des genres Aventure, Comédie et Documentaire. Toutefois, vos données d'évaluation contiennent uniquement des données des genres Film romantique et Thriller. Dans ce cas, le modèle d'apprentissage-machine n'a appris aucune information sur les genres Film romantique et Thriller, et l'évaluation n'a pas évalué la manière dont le modèle a appris les tendances pour les genres Aventure, Comédie et Documentaire. En conséquence, les informations de genre sont inutiles et la qualité des prédictions du modèle d'apprentissage-machine est compromise pour tous les genres. Le modèle et l'évaluation sont trop dissemblables (ont des statistiques descriptives extrêmement différentes) pour être utiles. Cela peut se produire lorsque les données d'entrée sont triées selon l'une des colonnes du jeu de données, puis fractionnées de manière séquentielle.

Si vos sources de données de formation et d'évaluation ont des distributions de données différentes, vous voyez une alerte d'évaluation dans votre évaluation de modèle. Pour plus d'informations sur les alertes d'évaluation, consultez Alertes d'évaluation.

Il n'est pas nécessaire d'utiliser le découpage aléatoire dans Amazon ML si vous avez déjà randomisé vos données d'entrée, par exemple en les mélangeant de manière aléatoire dans Amazon S3 ou en utilisant une fonction de requête SQL Amazon Redshift `random()` ou une fonction de requête SQL MySQL lors de la création des sources de `rand()` données. Dans ces cas, vous pouvez vous appuyer sur l'option de fractionnement séquentiel pour créer des sources de données de formation et d'évaluation avec des distributions similaires.

Analyse des données

Amazon ML calcule des statistiques descriptives sur vos données d'entrée que vous pouvez utiliser pour comprendre vos données.

Statistiques descriptives

Amazon ML calcule les statistiques descriptives suivantes pour différents types d'attributs :

Numérique:

- Histogrammes de distribution
- Nombre de valeurs non valides
- Valeurs minimale, médiane, moyenne et maximale

Binary (binaire) et Categorical (catégorie) :

- Nombre (de valeurs distinctes par catégorie)
- Histogramme de distribution de valeurs
- Valeurs les plus fréquentes
- Nombres de valeurs uniques
- Pourcentage de valeur vraie (binaire uniquement)
- Mots les plus visibles
- Mots les plus fréquents

Text:

- Nom de l'attribut
- Corrélation avec la cible (si une cible est définie)
- Total de mots
- Mots uniques
- Plage du nombre de mots dans une ligne
- Plage des longueurs de mot
- Mots les plus visibles

Accès à Data Insights sur la console Amazon ML

Sur la console Amazon ML, vous pouvez choisir le nom ou l'ID de n'importe quelle source de données pour afficher sa page Data Insights. Cette page fournit des métriques et des visualisations qui vous permettent d'en savoir plus sur les données d'entrée associées à la source de données, y compris les informations suivantes :

- Récapitulatif des données
- Distributions cibles
- Valeurs manquantes
- Valeurs non valides
- Statistiques récapitulatives des variables par type de données
- Distributions des variables par type de données

Les sections suivantes décrivent les métriques et les visualisations de manière plus détaillée.

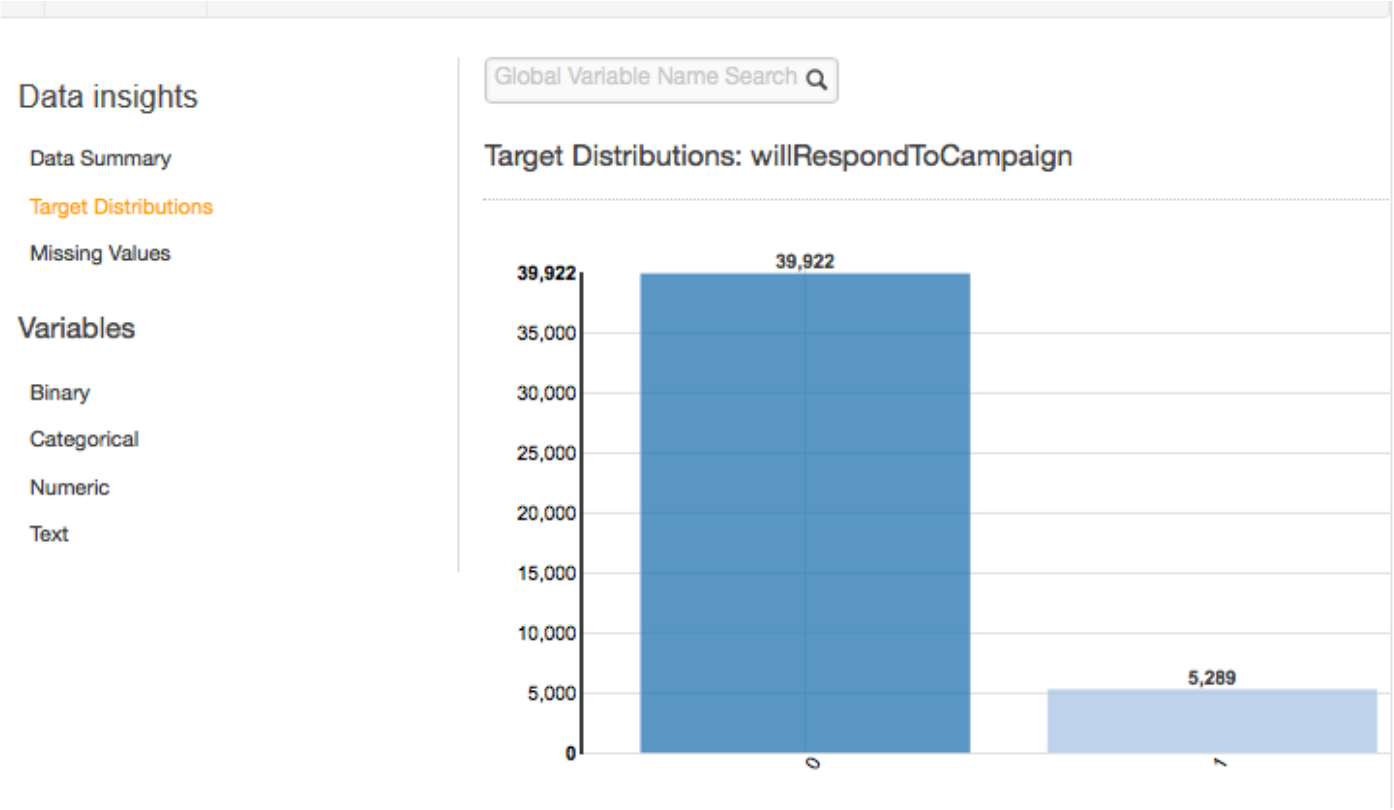
Récapitulatif des données

Le rapport récapitulatif des données d'une source de données affiche des informations récapitulatives, y compris l'ID de la source de données, son nom, l'emplacement où elle a été élaborée, l'état actuel, l'attribut cible, les informations de saisie de données (emplacement du compartiment S3, format des données, nombre d'enregistrements traités et nombre d'enregistrements incorrects rencontrés lors du traitement) ainsi que le nombre de variables par type de données.

Distributions cibles

Le rapport des distributions cibles montre la distribution de l'attribut cible de la source de données. Dans l'exemple suivant, il existe 39 922 observations pour lesquelles l'attribut cible de

la willRespondTo campagne est égal à 0. Il s'agit du nombre de clients qui n'ont pas répondu à la campagne d'e-mail. Il y a 5 289 observations pour lesquelles willRespondTo Campaign est égal à 1. Il s'agit du nombre de clients qui ont répondu à la campagne d'e-mail.



Valeurs manquantes

Le rapport des valeurs manquantes répertorie les attributs dans les données d'entrée pour lesquels des valeurs sont manquantes. Seuls les attributs d'un type de données numérique peuvent avoir des valeurs manquantes. Etant donné que des valeurs manquantes peuvent affecter la qualité de formation d'un modèle d'apprentissage-machine, nous vous recommandons de fournir les valeurs manquantes, si possible.

Pendant l'entraînement du modèle ML, si l'attribut cible est absent, Amazon ML rejette l'enregistrement correspondant. Si l'attribut cible est présent dans l'enregistrement, mais qu'une valeur pour un autre attribut numérique est manquante, Amazon ML ignore la valeur manquante. Dans ce cas, Amazon ML crée un attribut de remplacement et le définit sur 1 pour indiquer que cet attribut est manquant. Cela permet à Amazon ML d'apprendre des modèles à partir de l'occurrence de valeurs manquantes.

Valeurs non valides

Des valeurs non valides peuvent survenir uniquement avec les types de données Numeric (numérique) et Binary (binaire). Vous pouvez relever des valeurs non valides en affichant les statistiques récapitulatives des variables dans les rapports par type de données. Dans les exemples suivants, il y a une valeur non valide dans l'attribut numérique de durée et deux valeurs non valides de type de données binaire (une dans l'attribut de logement et l'autre dans l'attribut de prêt).

Numeric Variables

Variables	Correlations to Target	Missing Values	Invalid Values	Range	Mean	Median	Preview
duration	0.05165	2 (0%)	1 (0%)	0 - 4918	258.1618	180	

Binary Variables

Variables	Correlations to Target	Percent True	Invalid Values	Preview
campaign	NA	100%	27667 (61%)	
housing	0.01842	56%	1 (0%)	
loan	0.00656	16%	1 (0%)	
willRespondToCampaign	NA	12%	0 (0%)	

Corrélation Variable - Cible

Après avoir créé une source de données, Amazon ML peut évaluer la source de données et identifier la corrélation, ou l'impact, entre les variables et la cible. Par exemple, le prix d'un produit peut avoir un impact significatif sur le fait qu'il compte ou non parmi les meilleures ventes de l'année, tandis que les dimensions du produit peuvent n'avoir qu'une faible valeur prédictive.

Une bonne pratique consiste généralement à inclure autant de variables que possible dans les données de formation. Toutefois, le bruit introduit en incluant de nombreuses variables à faible valeur prédictive peut avoir une incidence négative sur la qualité et l'exactitude de votre modèle d'apprentissage-machine.

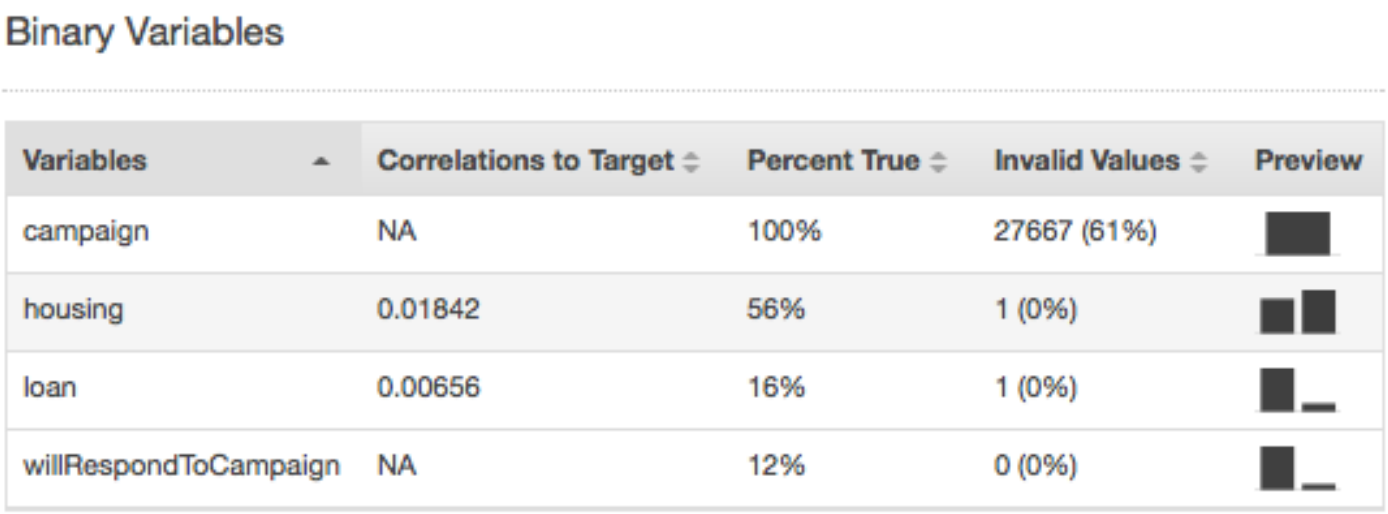
Vous pouvez améliorer les performances prédictives de votre modèle en supprimant des variables qui n'ont que peu d'impact lorsque vous formez votre modèle. Vous pouvez définir les variables mises à disposition du processus d'apprentissage automatique dans une recette, qui est un mécanisme de transformation d'Amazon ML. Pour en savoir plus sur les recettes, consultez Transformation des données pour l'apprentissage-machine.

Statistiques récapitulatives des attributs par type de données

Dans le rapport d'analyse des données, vous pouvez visualiser des statistiques récapitulatives d'attribut des types de données suivants :

- Binaire
- Categorical (catégorie)
- Numérique
- Texte

Les statistiques récapitulatives pour le type de données binaire (Binary) montrent tous les attributs binaires. La colonne Correlations to target (Corrélations avec la cible) montre les informations partagées entre la colonne cible et la colonne d'attribut. La colonne Percent true (Pourcentage true) indique le pourcentage d'observations de valeur 1. La colonne Invalid values (Valeurs non valides) indique le nombre de valeurs non valides, ainsi que le pourcentage de valeurs non valides pour chaque attribut. La colonne Aperçu fournit un lien vers un graphique de distribution pour chaque attribut.



Les statistiques récapitulatives pour le type de données Categorical (catégorie) affichent tous les attributs de catégorie avec le nombre de valeurs uniques, la valeur la plus fréquente et la valeur la moins fréquente. La colonne Aperçu fournit un lien vers un graphique de distribution pour chaque attribut.

Categorical Variables

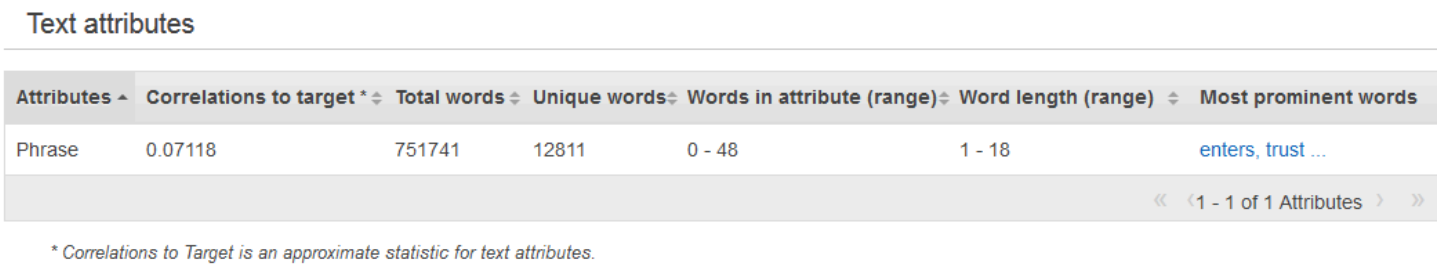
Variables	Correlations to Target	Unique Values	Most Frequent	Least Frequent	Preview
campaign	0.00433	49	1	39	
customerId	NA	45211	45211	1	
education	0.00355	5	secondary		
housing	0.01846	4	1		
jobId	0.00671	13	blue-collar		
willRespondToCampaign	NA	3	0		

Les statistiques récapitulatives pour le type de données Numeric (numérique) montrent tous les attributs numériques avec le nombre de valeurs manquantes, les valeurs non valides, la plage de valeurs, la moyenne et la valeur médiane. La colonne Aperçu fournit un lien vers un graphique de distribution pour chaque attribut.

Numeric Variables

Variables	Correlations to Target	Missing Values	Invalid Values	Range	Mean	Median	Preview
duration	0.05165	2 (0%)	1 (0%)	0 - 4918	258.1618	180	

Les statistiques récapitulatives pour le type de données Text (texte) montrent tous les attributs texte, le nombre total de mots dans cet attribut, le nombre de mots uniques dans cet attribut, la plage de mots dans un attribut, la plage des longueurs de mot et les mots les plus visibles. La colonne Aperçu fournit un lien vers un graphique de distribution pour chaque attribut.



Text attributes

Attributes	Correlations to target *	Total words	Unique words	Words in attribute (range)	Word length (range)	Most prominent words
Phrase	0.07118	751741	12811	0 - 48	1 - 18	enters, trust ...

« ‹ 1 - 1 of 1 Attributes › »

* Correlations to Target is an approximate statistic for text attributes.

L'exemple suivant montre les statistiques du type de données Text pour une variable texte appelée revue, avec quatre enregistrements.

```
1. The fox jumped over the fence.
2. This movie is intriguing.
3.
4. Fascinating movie.
```

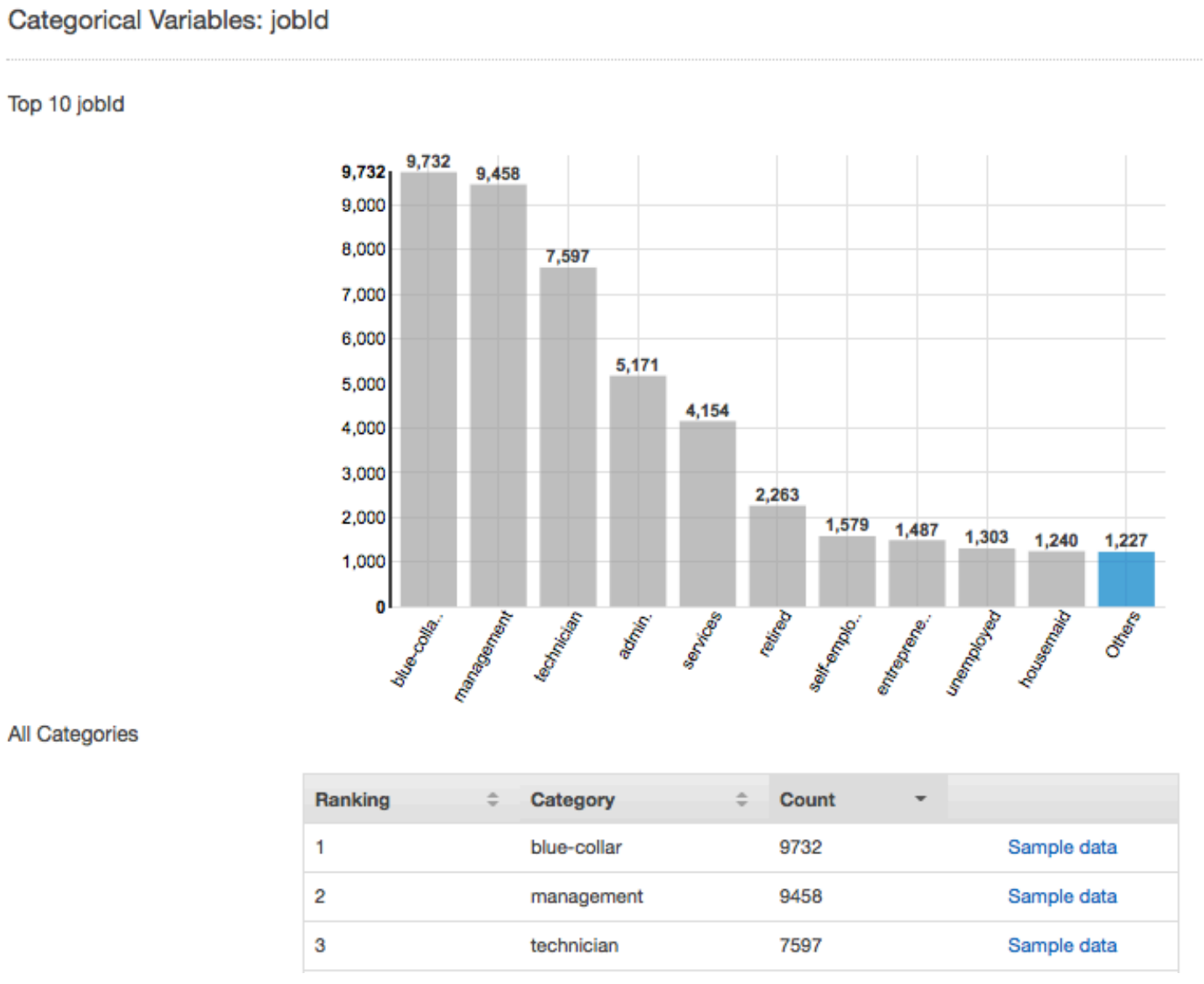
Pour cet exemple, les colonnes indiquent les informations suivantes.

- La colonne Attributes indique le nom de la variable. Dans cet exemple, cette colonne indique « revue ».
- La colonne Correlations to target existe uniquement si une cible est spécifiée. La corrélation mesure la quantité d'informations que cet attribut fournit sur la cible. Plus la corrélation est élevée, plus cet attribut vous donne d'informations sur la cible. La corrélation est mesurée en termes d'informations mutuelles entre une représentation simplifiée de l'attribut texte et la cible.
- La colonne Total words indique le nombre de mots générés lors de la segmentation (génération de jetons) de chaque enregistrement, les mots étant délimités par des espaces. Dans cet exemple, cette colonne indique « 12 ».
- La colonne Unique words indique le nombre de mots uniques pour un attribut. Dans cet exemple, cette colonne indique « 10 ».
- La colonne Words in attribute (range) indique le nombre de mots dans les lignes individuelles de l'attribut. Dans cet exemple, cette colonne indique « 0-6 ».
- La colonne Word length (range) indique la plage des nombres de caractères dans les mots. Dans cet exemple, cette colonne indique « 2-11 ».
- La colonne Most prominent words affiche une liste classant les mots qui figurent dans l'attribut. S'il y a un attribut cible, les mots sont classés en fonction de leur corrélation avec la cible, ce qui signifie que les mots de plus forte corrélation sont les premiers répertoriés. Si aucune cible n'est présente dans les données, les mots sont classés en fonction de leur entropie.

Présentation de la distribution des attributs binaires et de catégorie

En cliquant sur le lien Aperçu associé à un attribut binaire ou de catégorie, vous pouvez afficher la distribution de cet attribut, ainsi que les exemples de données issus du fichier d'entrée pour chaque valeur de catégorie de l'attribut.

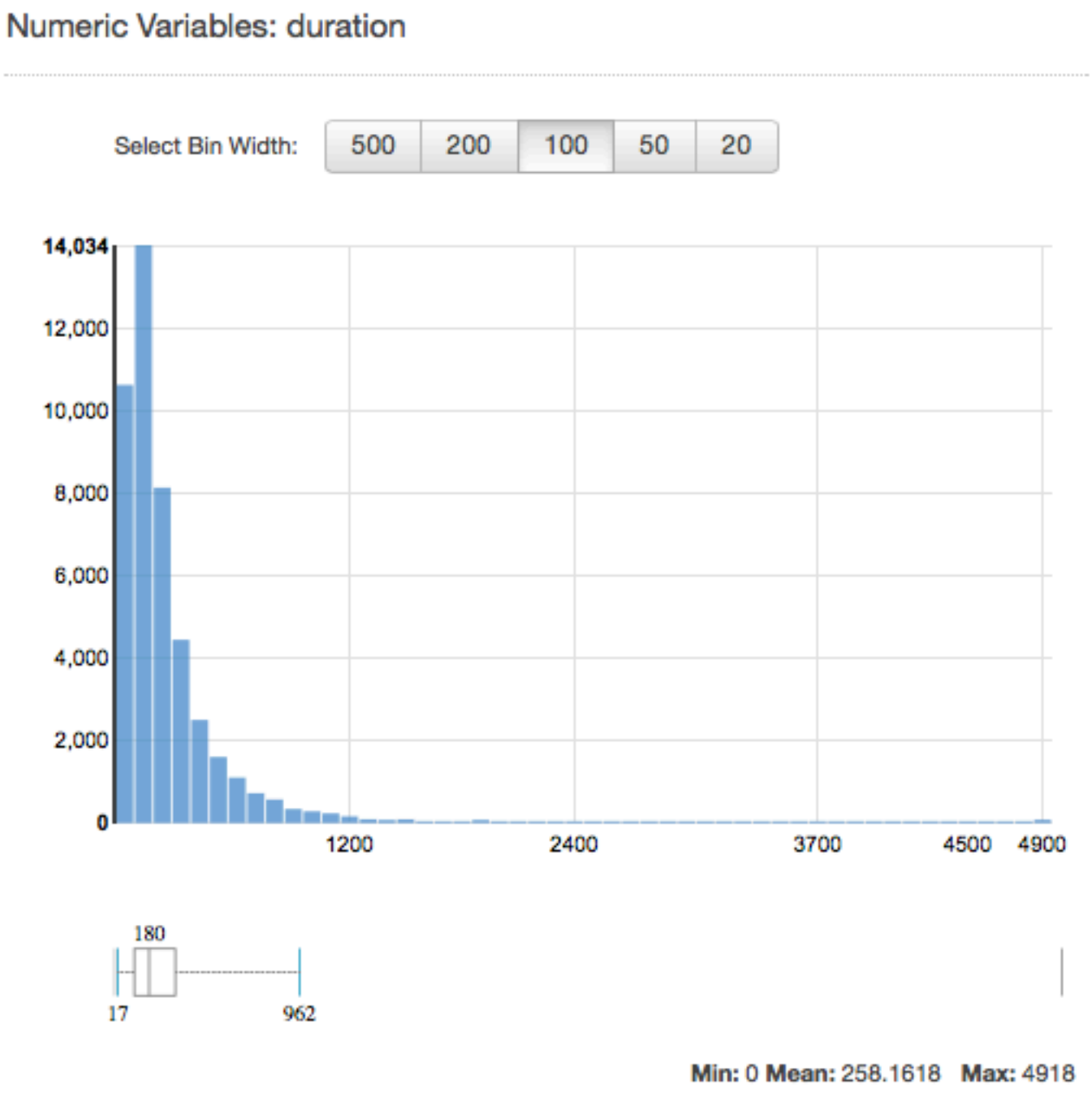
Par exemple, la capture d'écran suivante illustre la distribution de l'attribut de catégorie JobId. Cette distribution affiche les 10 valeurs de catégorie les plus élevées, avec toutes les autres valeurs regroupées dans « Autres ». Elle classe chacune des 10 valeurs de catégorie les plus élevées avec le nombre d'observations dans le fichier d'entrée qui contiennent cette valeur, ainsi qu'un lien pour afficher les exemples d'observations à partir du fichier de données d'entrée.



Présentation de la distribution des attributs numériques

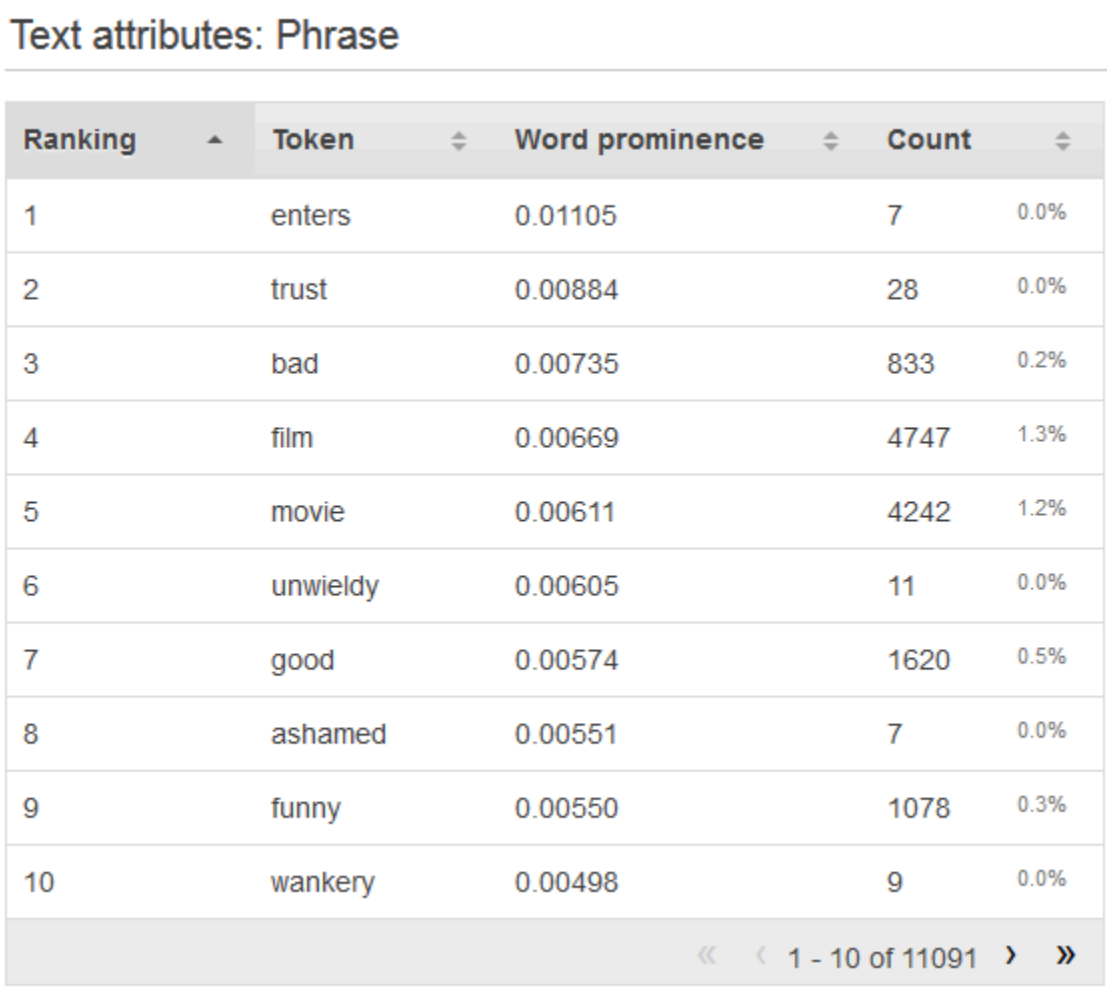
Pour afficher la distribution d'un attribut numérique, cliquez sur le lien Aperçu de cet attribut. Lorsque vous consultez la distribution d'un attribut numérique, vous pouvez choisir une taille de compartiment de 500, 200, 100, 50 ou 20. Plus la taille de compartiment est élevée, plus le nombre de barres de graphique affichées sera bas. De plus, la résolution de la distribution est grossière pour de grandes tailles de compartiment. Au contraire, la configuration d'une taille de compartiment de 20 augmente la résolution de la distribution affichée.

Les valeurs minimale, moyenne et maximale sont également affichées, comme illustré dans la capture d'écran ci-dessous.



Présentation de la distribution des attributs texte

Pour afficher la distribution d'un attribut texte, cliquez sur le lien Aperçu de cet attribut. Lorsque vous consultez la distribution d'un attribut texte, vous voyez les informations suivantes.



Ranking (classement)

Les jetons de texte sont classés en fonction de la quantité d'informations qu'ils convoient, des plus informatifs aux moins informatifs.

Jeton

La colonne Token indique le mot issu du texte d'entrée auquel se rapporte la ligne de statistiques.

Word prominence (Prédominance du mot)

S'il y a un attribut cible, les mots sont classés en fonction de leur corrélation avec la cible, de sorte que les mots de plus forte corrélation sont les premiers répertoriés. Si les données ne contiennent pas de cible, les mots sont classés en fonction de leur entropie, c'est-à-dire de la quantité d'informations qu'ils peuvent communiquer.

Count (décompte)

Ce nombre indique le nombre d'enregistrements d'entrée dans lesquels le jeton apparaît.

Pourcentage

Ce pourcentage indique le pourcentage des lignes de données d'entrée où le jeton apparaît.

Utilisation d'Amazon S3 avec Amazon ML

Amazon Simple Storage Service (Amazon S3) est une solution de stockage sur Internet. Vous pouvez utiliser Amazon S3 pour stocker et récupérer n'importe quelle quantité de données, n'importe quand et depuis n'importe quel emplacement sur le Web. Amazon ML utilise Amazon S3 comme référentiel de données principal pour les tâches suivantes :

- Pour accéder à vos fichiers d'entrée afin de créer des objets source de données pour la formation et l'évaluation de vos modèles d'apprentissage-machine.
- Pour accéder à vos fichiers d'entrée afin de générer des prédictions par lots.
- Lorsque vous générez des prédictions par lots à l'aide de vos modèles d'apprentissage-machine, pour fournir en sortie le fichier de prédiction dans un compartiment S3 que vous spécifiez.
- Pour copier les données que vous avez stockées dans Amazon Redshift ou Amazon Relational Database Service (Amazon RDS) dans un fichier .csv et chargez-le sur Amazon S3.

Pour permettre à Amazon ML d'effectuer ces tâches, vous devez autoriser Amazon ML à accéder à vos données Amazon S3.

> **Note**
>
> Vous ne pouvez pas fournir en sortie les fichiers de prédiction par lots dans un compartiment S3 qui accepte uniquement des fichiers chiffrés côté serveur. Assurez-vous que votre stratégie de compartiment permet le chargement de fichiers non chiffrés, en confirmant que cette stratégie n'inclut pas d'effet `Deny` pour l'action `s3:PutObject`, lorsqu'il n'y a pas d'en-tête `s3:x-amz-server-side-encryption` dans la demande. Pour plus d'informations sur les politiques relatives aux compartiments de chiffrement côté serveur S3, consultez la section Protection des données à l'aide du chiffrement côté serveur dans le guide de l'utilisateur d'Amazon Simple Storage Service.

Chargement de vos données sur Amazon S3

Vous devez télécharger vos données d'entrée sur Amazon Simple Storage Service (Amazon S3) car Amazon ML lit les données depuis les sites Amazon S3. Vous pouvez télécharger vos données directement sur Amazon S3 (par exemple, depuis votre ordinateur), ou Amazon ML peut copier les données que vous avez stockées dans Amazon Redshift ou Amazon Relational Database Service (RDS) dans un fichier .csv et les charger sur Amazon S3.

Pour plus d'informations sur la copie de vos données depuis Amazon Redshift ou Amazon RDS, consultez Utilisation d'Amazon Redshift avec Amazon ML ou Utilisation d'Amazon RDS avec Amazon ML, respectivement.

Le reste de cette section décrit comment télécharger vos données d'entrée directement depuis votre ordinateur vers Amazon S3. Avant de commencer les procédures de cette section, veillez à ce que vos données figurent dans un fichier .csv. Pour savoir comment formater correctement votre fichier .csv afin qu'Amazon ML puisse l'utiliser, consultez Comprendre le format de données pour Amazon ML.

Pour télécharger vos données depuis votre ordinateur vers Amazon S3

1. Connectez-vous à la console de gestion AWS et ouvrez la console Amazon S3 à l'adresse https://console.aws.amazon.com/s3.
2. Créez un compartiment ou sélectionnez un compartiment existant.

    a. Pour créer un compartiment, choisissez Créer un compartiment. Nommez votre compartiment, choisissez une région (vous pouvez choisir n'importe quelle région disponible), puis choisissez Créer. Pour plus d'informations, consultez Création d'un compartiment dans le Guide de mise en route Amazon Simple Storage Service.

    b. Pour utiliser un compartiment existant, recherchez le compartiment en le choisissant dans la liste Tous les compartiments. Lorsque le nom du compartiment apparaît, sélectionnez-le, puis choisissez Charger.
3. Dans la boîte de dialogue Charger, choisissez Ajouter des fichiers.
4. Accédez au dossier qui contient le fichier .csv de vos données d'entrée, puis choisissez Ouvrir.

Autorisations

Pour autoriser Amazon ML à accéder à l'un de vos compartiments S3, vous devez modifier la politique relative aux compartiments.

Pour plus d'informations sur l'octroi à Amazon ML de l'autorisation de lire les données de votre compartiment dans Amazon S3, consultez la section Octroi d'autorisations à Amazon ML pour lire vos données depuis Amazon S3.

Pour plus d'informations sur l'octroi à Amazon ML de l'autorisation de transmettre les résultats des prédictions par lots à votre compartiment dans Amazon S3, consultez la section Octroi à Amazon ML des autorisations de sortie de prédictions vers Amazon S3.

Pour plus d'informations sur la gestion des autorisations d'accès aux ressources Amazon S3, consultez le manuel du développeur Amazon S3.

Création d'une source de données Amazon ML à partir des données d'Amazon Redshift

Si vous avez des données stockées dans Amazon Redshift, vous pouvez utiliser l'assistant Create Datasource de la console Amazon Machine Learning (Amazon ML) pour créer un objet de source de données. Lorsque vous créez une source de données à partir des données Amazon Redshift, vous spécifiez le cluster qui contient vos données et la requête SQL permettant de récupérer vos données. Amazon ML exécute la requête en appelant la commande Amazon `Unload` Redshift sur le cluster. Amazon ML stocke les résultats dans l'emplacement Amazon Simple Storage Service (Amazon S3) de votre choix, puis utilise les données stockées dans Amazon S3 pour créer la source de données. La source de données, le cluster Amazon Redshift et le compartiment S3 doivent tous se trouver dans la même région.

> **Note**
>
> Amazon ML ne prend pas en charge la création de sources de données à partir de clusters Amazon Redshift en privé. VPCs Le cluster doit avoir une adresse IP publique.

Rubriques

Paramètres obligatoires pour l'assistant de création de sources de données

Pour permettre à Amazon ML de se connecter à votre base de données Amazon Redshift et de lire les données en votre nom, vous devez fournir les informations suivantes :

- L'Amazon Redshift `ClusterIdentifier`
- Le nom de la base de données Amazon Redshift
- Les informations d'identification de la base de données Amazon Redshift (nom d'utilisateur et mot de passe)
- Le rôle Amazon ML Amazon Redshift AWS Identity and Access Management (IAM)
- La requête SQL Amazon Redshift
- (Facultatif) L'emplacement du schéma Amazon ML
- L'emplacement intermédiaire d'Amazon S3 (où Amazon ML place les données avant de créer la source de données)

En outre, vous devez vous assurer que les utilisateurs ou les rôles IAM qui créent les sources de données Amazon Redshift (que ce soit par le biais de la console ou de l'action) disposent de `CreateDatasourceFromRedshift` l'autorisation. `iam:PassRole`

Amazon Redshift **`ClusterIdentifier`**

Utilisez ce paramètre distinguant majuscules et minuscules pour permettre à Amazon ML de trouver votre cluster et de s'y connecter. Vous pouvez obtenir l'identifiant du cluster (nom) sur la console Amazon Redshift. Pour plus d'informations sur les clusters, consultez Amazon Redshift Clusters.

Nom de la base de données Amazon Redshift

Utilisez ce paramètre pour indiquer à Amazon ML quelle base de données du cluster Amazon Redshift contient les données que vous souhaitez utiliser comme source de données.

Informations d'identification de la base de données Amazon Redshift

Utilisez ces paramètres pour spécifier le nom d'utilisateur et le mot de passe de l'utilisateur de base de données Amazon Redshift dans le contexte duquel la requête de sécurité sera exécutée.

> **Note**
>
> Amazon ML a besoin d'un nom d'utilisateur et d'un mot de passe Amazon Redshift pour se connecter à votre base de données Amazon Redshift. Une fois les données déchargées sur Amazon S3, Amazon ML ne réutilise jamais votre mot de passe et ne le stocke jamais.

Amazon ML (rôle Amazon Redshift)

Utilisez ce paramètre pour spécifier le nom du rôle IAM qu'Amazon ML doit utiliser pour configurer les groupes de sécurité pour le cluster Amazon Redshift et la politique de compartiment pour le site de transit Amazon S3.

Si vous ne disposez pas d'un rôle IAM pouvant accéder à Amazon Redshift, Amazon ML peut créer un rôle pour vous. Lorsqu'Amazon ML crée un rôle, il crée et associe une politique gérée par le client à un rôle IAM. La politique créée par Amazon ML accorde à Amazon ML l'autorisation d'accéder uniquement au cluster que vous spécifiez.

Si vous disposez déjà d'un rôle IAM pour accéder à Amazon Redshift, vous pouvez saisir l'ARN du rôle ou choisir le rôle dans la liste déroulante. Les rôles IAM avec accès à Amazon Redshift sont répertoriés en haut de la liste déroulante.

Le rôle IAM doit avoir le contenu suivant :

```
{
    "Version": "2012-10-17",
    "Statement": [
    {
        "Effect": "Allow",
        "Principal": {
            "Service": "machinelearning.amazonaws.com"
        },
        "Action": "sts:AssumeRole",
        "Condition": {
            "StringEquals": { "aws:SourceAccount": "123456789012" },
           "ArnLike": { "aws:SourceArn": "arn:aws:machinelearning:us-
east-1:123456789012:datasource/*" }
        }
    }]
}
```

Pour plus d'informations sur les politiques gérées par le client, consultez la section Politiques gérées par le client dans le guide de l'utilisateur IAM.

Requête SQL Amazon Redshift

Utilisez ce paramètre pour spécifier la requête SQL SELECT qu'Amazon ML exécute sur votre base de données Amazon Redshift afin de sélectionner vos données. Amazon ML utilise l'action Amazon Redshift UNLOAD pour copier en toute sécurité les résultats de votre requête vers un emplacement Amazon S3.

> **Note**
>
> Amazon ML fonctionne mieux lorsque les enregistrements d'entrée sont classés dans un ordre aléatoire (mélangés). Vous pouvez facilement mélanger les résultats de votre requête SQL Amazon Redshift à l'aide de la fonction Amazon Redshift random (). Par exemple, supposons que la requête d'origine est :
>
> ```
> "SELECT col1, col2, … FROM training_table"
> ```
>
> Vous pouvez intégrer une réorganisation aléatoire en mettant à jour la requête comme ceci :
>
> ```
> "SELECT col1, col2, … FROM training_table ORDER BY random()"
> ```

Emplacement du schéma (facultatif)

Utilisez ce paramètre pour spécifier le chemin Amazon S3 vers votre schéma pour les données Amazon Redshift qu'Amazon ML exportera.

Si vous ne fournissez pas de schéma pour votre source de données, la console Amazon ML crée automatiquement un schéma Amazon ML basé sur le schéma de données de la requête SQL Amazon Redshift. Les schémas Amazon ML contiennent moins de types de données que les schémas Amazon Redshift. Il ne s'agit donc pas d'une conversion. one-to-one La console Amazon ML convertit les types de données Amazon Redshift en types de données Amazon ML en utilisant le schéma de conversion suivant.

Types de données Amazon Redshift	Alias Amazon Redshift	Type de données Amazon ML
SMALLINT	INT2	NUMERIC
INTEGER	ENTIER, INT4	NUMERIC
BIGINT	INT8	NUMERIC
DECIMAL	NUMERIC	NUMERIC
REAL	FLOAT4	NUMERIC
DOUBLE PRECISION	FLOAT8, FLOTTEUR	NUMERIC
BOOLEAN	BOOL	BINAIRE
CHAR	CHARACTER, NCHAR, BPCHAR	CATEGORICAL (catégorie)
VARCHAR	CHARACTER VARYING, NVARCHAR, TEXT	TEXT
DATE		TEXT
TIMESTAMP	TIMESTAMP WITHOUT TIME ZONE	TEXT

Pour être converties en types de `Binary` données Amazon ML, les valeurs des booléens Amazon Redshift présentes dans vos données doivent être compatibles avec les valeurs binaires Amazon ML. Si votre type de données booléen comporte des valeurs non prises en charge, Amazon ML les convertit dans le type de données le plus spécifique possible. Par exemple, si un booléen Amazon Redshift contient les valeurs`0`, 1 et qu'2Amazon ML convertit le booléen en type de données. `Numeric` Pour plus d'informations sur les valeurs binaires prises en charge, consultez Utilisation du AttributeType terrain.

Si Amazon ML ne parvient pas à identifier un type de données, sa valeur par défaut est. `Text`

Une fois qu'Amazon ML a converti le schéma, vous pouvez consulter et corriger les types de données Amazon ML attribués dans l'assistant de création de source de données, et réviser le schéma avant qu'Amazon ML ne crée la source de données.

Emplacement de transit d'Amazon S3

Utilisez ce paramètre pour spécifier le nom de l'emplacement intermédiaire Amazon S3 où Amazon ML stocke les résultats de la requête SQL Amazon Redshift. Après avoir créé la source de données, Amazon ML utilise les données sur le site de transit au lieu de les renvoyer vers Amazon Redshift.

> **Note**
>
> Amazon ML assumant le rôle IAM défini par le rôle Amazon Redshift d'Amazon ML, Amazon ML est autorisé à accéder à tous les objets se trouvant dans l'emplacement de transit Amazon S3 spécifié. C'est pourquoi nous vous recommandons de ne stocker que les fichiers ne contenant pas d'informations sensibles dans l'emplacement intermédiaire Amazon S3. Par exemple, si votre compartiment racine l'est`s3://mybucket/`, nous vous suggérons de créer un emplacement pour stocker uniquement les fichiers auxquels vous souhaitez qu'Amazon ML accède, tels que`s3://mybucket/AmazonMLInput/`.

Création d'une source de données avec Amazon Redshift Data (console)

La console Amazon ML propose deux méthodes pour créer une source de données à l'aide des données Amazon Redshift. Vous pouvez créer une source de données en suivant l'assistant de création de source de données ou, si vous avez déjà créé une source de données à partir des données Amazon Redshift, vous pouvez copier la source de données d'origine et modifier ses paramètres. La copie d'une source de données vous permet de facilement créer plusieurs sources de données similaires.

Pour plus d'informations sur la création d'une source de données à l'aide de l'API, consultez CreateDataSourceFromRedshift.

Pour plus d'informations sur les paramètres utilisés dans les procédures suivantes, consultez Paramètres obligatoires pour l'assistant de création de sources de données.

Rubriques

- Création d'une source de données (console)

Création d'une source de données (console)

Pour décharger des données d'Amazon Redshift vers une source de données Amazon ML, utilisez l'assistant de création de source de données.

Pour créer une source de données à partir de données dans Amazon Redshift

1. Ouvrez la console Amazon Machine Learning à l'adresse https://console.aws.amazon.com/machinelearning/.
2. Sur le tableau de bord Amazon ML, sous Entités, choisissez Create new... , puis choisissez Datasource.
3. Sur la page des données d'entrée, sélectionnez Amazon Redshift.
4. Dans l'assistant de création de sources de données, pour Cluster identifier, tapez le nom de votre cluster.
5. Dans Nom de la base de données, saisissez le nom de la base de données Amazon Redshift.
6. Pour Database user name, tapez votre nom d'utilisateur de base de données.
7. Pour Database password, tapez votre mot de passe de base de données.
8. Pour Rôle IAM, choisissez votre rôle IAM. Si vous n'en avez pas déjà un, choisissez Créer un nouveau rôle. Amazon ML crée un rôle IAM Amazon Redshift pour vous.
9. Pour tester vos paramètres Amazon Redshift, choisissez Test Access (à côté du rôle IAM). Si Amazon ML ne parvient pas à se connecter à Amazon Redshift avec les paramètres fournis, vous ne pouvez pas continuer à créer une source de données. Pour bénéficier d'une aide à la résolution des problèmes, consultez Dépannage des erreurs.
10. Pour SQL query, tapez votre requête SQL.
11. Pour l'emplacement du schéma, indiquez si vous souhaitez qu'Amazon ML crée un schéma pour vous. Si vous avez créé un schéma vous-même, saisissez le chemin Amazon S3 vers votre fichier de schéma.
12. Pour l'emplacement intermédiaire d'Amazon S3, saisissez le chemin Amazon S3 vers le compartiment dans lequel vous souhaitez qu'Amazon ML place les données qu'il décharge depuis Amazon Redshift.
13. (Facultatif) Pour Datasource name, tapez un nom pour votre source de données.
14. Choisissez Vérifier. Amazon ML vérifie qu'il peut se connecter à votre base de données Amazon Redshift.

15. Dans la page Schema, passez en revue les types de données pour tous les attributs et corrigez-les, si nécessaire.
16. Choisissez Continuer.
17. Si vous voulez utiliser cette source de données pour créer ou évaluer un modèle d'apprentissage-machine, pour Do you plan to use this dataset to create or evaluate an ML model?, choisissez Yes. Si vous choisissez Yes, choisissez votre ligne cible. Pour en savoir plus sur les cibles, consultez Utilisation du targetAttributeName terrain.

    Si vous voulez utiliser cette source de données avec un modèle que vous avez déjà créé, afin de créer des prédictions, choisissez No.
18. Choisissez Continuer.
19. Pour Does your data contain an identifier ?, si vos données ne contiennent pas d'identifiant de ligne, choisissez No.

    Si vos données contiennent un identifiant de ligne, choisissez Yes. Pour obtenir des informations sur les identifiants de ligne, consultez Utilisation du champ rowID.
20. Choisissez Examiner.
21. Dans la page Révision, passez en revue vos paramètres, puis choisissez Terminer.

Une fois que vous avez créé une source de données, vous pouvez l'utiliser pour create an ML model. Si vous avez déjà créé un modèle, vous pouvez utiliser la source de données pour evaluate an ML model ou generate predictions.

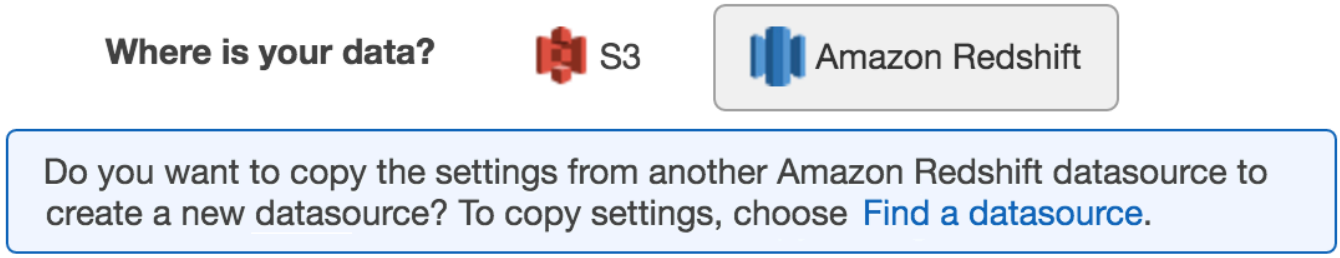
Copie d'une source de données (console)

Lorsque vous souhaitez créer une source de données similaire à une source de données existante, vous pouvez utiliser la console Amazon ML pour copier la source de données d'origine et modifier ses paramètres. Par exemple, vous pouvez choisir de commencer par une source de données existante, puis de modifier le schéma de données pour qu'il corresponde mieux à vos données, de modifier la requête SQL utilisée pour décharger les données d'Amazon Redshift ou de spécifier un AWS Identity and Access Management autre utilisateur (IAM) pour accéder au cluster Amazon Redshift.

Pour copier et modifier une source de données Amazon Redshift

1. Ouvrez la console Amazon Machine Learning à l'adresse https://console.aws.amazon.com/machinelearning/.

2. Sur le tableau de bord Amazon ML, sous Entités, choisissez Create new... , puis choisissez Datasource.
3. Sur la page Données d'entrée, pour Où sont vos données ? , choisissez Amazon Redshift. Si vous avez déjà créé une source de données à partir des données Amazon Redshift, vous avez la possibilité de copier les paramètres d'une autre source de données.



   Si vous n'avez pas encore créé de source de données à partir des données Amazon Redshift, cette option n'apparaît pas.
4. Choisissez Find a datasource.
5. Sélectionnez la source de données que vous souhaitez copier, puis choisissez Copier les paramètres. Amazon ML remplit automatiquement la plupart des paramètres de la source de données avec les paramètres de la source de données d'origine. Il ne copie pas le mot de passe de la base de données, l'emplacement du schéma ni le nom de la source de données à partir de la source de données d'origine.
6. Modifiez tous les paramètres renseignés automatiquement que vous souhaitez. Par exemple, si vous souhaitez modifier les données déchargées par Amazon ML depuis Amazon Redshift, modifiez la requête SQL.
7. Pour Database password, tapez votre mot de passe de base de données. Amazon ML ne stocke ni ne réutilise votre mot de passe. Vous devez donc toujours le fournir.
8. (Facultatif) Pour l'emplacement du schéma, Amazon ML présélectionne Je veux qu'Amazon ML génère un schéma recommandé pour vous. Si vous avez déjà créé un schéma, choisissez Je souhaite utiliser le schéma que j'ai créé et stocké dans Amazon S3 et saisissez le chemin d'accès à votre fichier de schéma dans Amazon S3.
9. (Facultatif) Pour Datasource name, tapez un nom pour votre source de données. Dans le cas contraire, Amazon ML génère un nouveau nom de source de données pour vous.
10. Choisissez Vérifier. Amazon ML vérifie qu'il peut se connecter à votre base de données Amazon Redshift.

11. (Facultatif) Si Amazon ML a déduit le schéma pour vous, sur la page Schéma, passez en revue les types de données pour tous les attributs et corrigez-les si nécessaire.
12. Choisissez Continuer.
13. Si vous voulez utiliser cette source de données pour créer ou évaluer un modèle d'apprentissage-machine, pour Do you plan to use this dataset to create or evaluate an ML model?, choisissez Yes. Si vous choisissez Yes, choisissez votre ligne cible. Pour en savoir plus sur les cibles, consultez Utilisation du targetAttributeName terrain.

    Si vous voulez utiliser cette source de données avec un modèle que vous avez déjà créé, afin de créer des prédictions, choisissez No.
14. Choisissez Continuer.
15. Pour Does your data contain an identifier ?, si vos données ne contiennent pas d'identifiant de ligne, choisissez No.

    Si vos données contiennent un identifiant de ligne, choisissez Oui et sélectionnez la ligne que vous souhaitez utiliser comme identifiant. Pour obtenir des informations sur les identifiants de ligne, consultez Utilisation du champ rowID.
16. Choisissez Examiner.
17. Passez en revue vos paramètres, puis choisissez Terminer.

Une fois que vous avez créé une source de données, vous pouvez l'utiliser pour create an ML model. Si vous avez déjà créé un modèle, vous pouvez utiliser la source de données pour evaluate an ML model ou generate predictions.

Résolution des problèmes liés à Amazon Redshift

Lorsque vous créez votre source de données Amazon Redshift, vos modèles de machine learning et votre évaluation, Amazon Machine Learning (Amazon ML) indique le statut de vos objets Amazon ML dans la console Amazon ML. Si Amazon ML renvoie des messages d'erreur, utilisez les informations et ressources suivantes pour résoudre les problèmes.

Pour obtenir des réponses aux questions générales concernant Amazon ML, consultez le site Amazon Machine Learning FAQs. Vous pouvez également rechercher des réponses et publier des questions sur le forum Amazon Machine Learning.

Rubriques

- Dépannage des erreurs
- Contacter AWS Support

Dépannage des erreurs

Le format du rôle n'est pas valide. Fournissez un rôle IAM valide. Par exemple, arn:aws:iam : : ID:Role/. YourAccount YourRedshiftRole

Cause

Le format de l'Amazon Resource Name (ARN) de votre rôle IAM est incorrect.

Solution

Dans l'assistant de création de sources de données, corrigez l'ARN de votre rôle. Pour plus d'informations sur le rôle de formatage ARNs, voir IAM ARNs dans le guide de l'utilisateur d'IAM. La région est facultative pour le rôle ARNs IAM.

Le rôle n'est pas valide. Amazon ML ne peut pas assumer le <role ARN>rôle IAM. Fournissez un rôle IAM valide et rendez-le accessible à Amazon ML.

Cause

Votre rôle n'est pas configuré pour permettre à Amazon ML de l'assumer.

Solution

Dans la console IAM, modifiez votre rôle afin qu'il dispose d'une politique de confiance permettant à Amazon ML d'assumer le rôle qui lui est associé.

Cet utilisateur < ARN d'utilisateur> n'est pas autorisé à fournir le rôle IAM <ARN du rôle>.

Cause

Votre utilisateur IAM ne dispose d'aucune politique d'autorisation lui permettant de transmettre un rôle à Amazon ML.

Solution

Associez une politique d'autorisation à votre utilisateur IAM qui vous permet de transmettre des rôles à Amazon ML. Vous pouvez attacher une stratégie d'autorisations à votre utilisateur IAM dans la console IAM.

La transmission d'un rôle IAM entre des comptes n'est pas autorisée. Le rôle IAM doit appartenir à ce compte.

Cause

Vous ne pouvez pas transmettre un rôle qui appartient à un autre compte IAM.

Solution

Connectez-vous au compte AWS que vous avez utilisé pour créer le rôle. Vous pouvez voir vos rôles IAM dans votre console IAM.

Le rôle spécifié n'a pas les autorisations nécessaires pour effectuer cette opération. Fournissez un rôle doté d'une politique qui fournit à Amazon ML les autorisations requises.

Cause

Votre rôle IAM n'a pas les autorisations nécessaires pour effectuer l'opération demandée.

Solution

Modifiez la stratégie d'autorisation attachée à votre rôle dans la console IAM pour fournir les autorisations requises.

Amazon ML ne peut pas configurer de groupe de sécurité sur ce cluster Amazon Redshift avec le rôle IAM spécifié.

Cause

Votre rôle IAM ne dispose pas des autorisations requises pour configurer un cluster de sécurité Amazon Redshift.

Solution

Modifiez la stratégie d'autorisation attachée à votre rôle dans la console IAM pour fournir les autorisations requises.

Une erreur s'est produite lorsqu'Amazon ML a tenté de configurer un groupe de sécurité sur votre cluster. Réessayez ultérieurement.

Cause

Lorsqu'Amazon ML a essayé de se connecter à votre cluster Amazon Redshift, il a rencontré un problème.

Solution

Vérifiez que le rôle IAM que vous avez fourni dans l'assistant de création de sources de données possède toutes les autorisations requises.

Le format de l'ID du cluster n'est pas valide. IDs Le cluster doit commencer par une lettre et ne doit contenir que des caractères alphanumériques et des traits d'union. Ils ne peuvent pas contenir deux traits d'union consécutifs ou se terminer par un trait d'union.

Cause

Le format de votre identifiant de cluster Amazon Redshift est incorrect.

Solution

Dans l'assistant de création de sources de données, corrigez votre ID de cluster pour qu'il contienne uniquement des caractères alphanumériques et des tirets, et qu'il ne contienne pas deux traits d'union consécutifs ni ne se termine par un trait d'union.

Il n'existe aucun <Amazon Redshift cluster name>cluster, ou le cluster ne se trouve pas dans la même région que votre service Amazon ML. Spécifiez un cluster dans la même région que cet Amazon ML.

Cause

Amazon ML ne trouve pas votre cluster Amazon Redshift car il ne se trouve pas dans la région où vous créez une source de données Amazon ML.

Solution

Vérifiez que votre cluster existe sur la page Clusters de la console Amazon Redshift, que vous créez une source de données dans la même région que celle où se trouve votre cluster Amazon Redshift et que l'ID de cluster spécifié dans l'assistant de création de source de données est correct.

Amazon ML ne peut pas lire les données de votre cluster Amazon Redshift. Fournissez l'ID de cluster Amazon Redshift correct.

Cause

Amazon ML ne peut pas lire les données du cluster Amazon Redshift que vous avez spécifié.

Solution

Dans l'assistant de création de source de données, spécifiez l'ID de cluster Amazon Redshift correct, vérifiez que vous créez une source de données dans la même région que votre cluster Amazon Redshift et que votre cluster est répertorié sur la page Amazon Redshift Clusters.

Le <Amazon Redshift cluster name>cluster n'est pas accessible au public.

Cause

Amazon ML ne peut pas accéder à votre cluster car celui-ci n'est pas accessible au public et ne possède pas d'adresse IP publique.

Solution

Rendez le cluster publiquement accessible et attribuez-lui une adresse IP publique. Pour plus d'informations sur la façon de rendre les clusters accessibles au public, consultez la section Modification d'un cluster dans le guide de gestion Amazon Redshift.

L'<Redshift>état du cluster n'est pas disponible pour Amazon ML. Utilisez la console Amazon Redshift pour visualiser et résoudre ce problème d'état du cluster. L'état du cluster doit être « disponible ».

Cause

Amazon ML ne peut pas voir l'état du cluster.

Solution

Assurez-vous que votre cluster est disponible. Pour plus d'informations sur la vérification de l'état de votre cluster, consultez Getting an Overview of Cluster Status dans le guide de gestion Amazon Redshift. Pour plus d'informations sur le redémarrage du cluster afin qu'il soit disponible, consultez la section Redémarrage d'un cluster dans le guide de gestion Amazon Redshift.

Il n'y a pas de base de données <nom de base de données> dans ce cluster. Vérifiez que le nom de la base de données est correct ou spécifiez un autre cluster ou une autre base de données.

Cause

Amazon ML ne trouve pas la base de données spécifiée dans le cluster spécifié.

Solution

Vérifiez que le nom de base de données saisi dans l'assistant de création de sources de données est correct, ou spécifiez les noms corrects de cluster et de base de données.

Amazon ML n'a pas pu accéder à votre base de données. Fournissez un mot de passe valide pour l'utilisateur de base de données <nom d'utilisateur>.

Cause

Le mot de passe que vous avez fourni dans l'assistant de création de source de données pour autoriser Amazon ML à accéder à votre base de données Amazon Redshift est incorrect.

Solution

Entrez le mot de passe correct pour l'utilisateur de votre base de données Amazon Redshift.

Une erreur s'est produite lorsqu'Amazon ML a tenté de valider la requête.

Cause

Il y a un problème avec votre requête SQL.

Solution

Vérifiez que votre requête est une requête SQL valide.

Une erreur s'est produite lors de l'exécution de votre requête SQL. Vérifiez le nom de base de données et la requête fournie. Cause racine : {serverMessage}.

Cause

Amazon Redshift n'a pas pu exécuter votre requête.

Solution

Vérifiez que vous avez spécifié le nom de base de données correct dans l'assistant de création de sources de données, et que votre requête est une requête SQL valide.

Une erreur s'est produite lors de l'exécution de votre requête SQL. Cause racine : {serverMessage}.

Cause

Amazon Redshift n'a pas pu trouver la table spécifiée.

Solution

Vérifiez que la table que vous avez spécifiée dans l'assistant de création de source de données est présente dans la base de données de votre cluster Amazon Redshift et que vous avez saisi l'ID de cluster, le nom de base de données et la requête SQL corrects.

Contacter AWS Support

Si vous avez souscrit un plan AWS Premium Support, vous pouvez créer une demande d'assistance technique dans le Centre AWS Support.

Utilisation des données d'une base de données Amazon RDS pour créer une source de données Amazon ML

Amazon ML vous permet de créer un objet de source de données à partir de données stockées dans une base de données MySQL dans Amazon Relational Database Service (Amazon RDS). Lorsque vous effectuez cette action, Amazon ML crée un objet AWS Data Pipeline qui exécute la requête SQL que vous spécifiez et place la sortie dans un compartiment S3 de votre choix. Amazon ML utilise ces données pour créer la source de données.

> **Note**
>
> Amazon ML prend uniquement en charge les bases de données MySQL dans VPCs.

Avant qu'Amazon ML puisse lire vos données d'entrée, vous devez exporter ces données vers Amazon Simple Storage Service (Amazon S3). Vous pouvez configurer Amazon ML pour effectuer l'exportation à votre place à l'aide de l'API. (RDS est limité à cette API et n'est pas disponible à partir de la console.)

Pour qu'Amazon ML puisse se connecter à votre base de données MySQL dans Amazon RDS et lire les données en votre nom, vous devez fournir les informations suivantes :

- L'identifiant d'instance de base de données RDS
- Le nom de la base de données MySQL
- Rôle AWS Identity and Access Management (IAM) utilisé pour créer, activer et exécuter le pipeline de données

- Les informations d'identification de l'utilisateur de base de données :
  - Nom utilisateur
  - Mot de passe
- Les informations de sécurité d'AWS Data Pipeline :
  - Le rôle de ressource IAM
  - Le rôle de service IAM
- Informations de sécurité Amazon RDS :
  - L'ID de sous-réseau
  - Le groupe de sécurité IDs
- La requête SQL qui spécifie les données que vous souhaitez utiliser pour créer la source de données
- L'emplacement (compartiment) de sortie S3 utilisé pour stocker les résultats de la requête
- (Facultatif) L'emplacement du fichier de schéma de données

En outre, vous devez vous assurer que les utilisateurs ou rôles IAM qui créent des sources de données Amazon RDS à l'aide de l'opération CreateDataSourceFromRDS sont autorisés. `iam:PassRole` Pour de plus amples informations, veuillez consulter Contrôle de l'accès aux ressources Amazon ML à l'aide d'IAM.

Rubriques

- Identifiant d'instance de base de données RDS
- Nom de la base de données MySQL
- Informations d'identification de l'utilisateur de base de données
- Informations de sécurité d'AWS Data Pipeline
- Informations de sécurité Amazon RDS
- Requête SQL MySQL
- Emplacement de sortie S3

Identifiant d'instance de base de données RDS

L'identifiant d'instance de base de données RDS est un nom unique que vous fournissez et qui identifie l'instance de base de données qu'Amazon ML doit utiliser lors de l'interaction avec Amazon RDS. Vous pouvez trouver l'identifiant de l'instance de base de données RDS dans la console Amazon RDS.

Nom de la base de données MySQL

Le nom de base de données MySQL spécifie le nom de la base de données MySQL dans l'instance de base de données RDS.

Informations d'identification de l'utilisateur de base de données

Pour vous connecter à l'instance de base de données RDS, vous devez fournir le nom d'utilisateur et le mot de passe de l'utilisateur de la base de données qui dispose d'autorisations suffisantes pour exécuter la requête SQL que vous fournissez.

Informations de sécurité d'AWS Data Pipeline

Pour activer l'accès sécurisé à AWS Data Pipeline, vous devez fournir les noms du rôle de ressource IAM et du rôle de service IAM.

Une EC2 instance joue le rôle de ressource pour copier les données d'Amazon RDS vers Amazon S3. La manière la plus simple de créer ce rôle de ressource consiste à utiliser le modèle `DataPipelineDefaultResourceRole` et à répertorier **`machinelearning.aws.com`** comme service approuvé. Pour plus d'informations sur ce modèle, consultez Configuration de rôles IAM dans le Manuel du développeur AWS Data Pipeline.

Si vous créez votre propre rôle, celui-ci doit comporter le contenu suivant :

```
{
    "Version": "2012-10-17",
    "Statement": [
    {
        "Effect": "Allow",
        "Principal": {
            "Service": "machinelearning.amazonaws.com"
        },
        "Action": "sts:AssumeRole",
        "Condition": {
            "StringEquals": { "aws:SourceAccount": "123456789012" },
            "ArnLike": { "aws:SourceArn": "arn:aws:machinelearning:us-
east-1:123456789012:datasource/*" }
        }
    }]
}
```

AWS Data Pipeline assume le rôle de service chargé de surveiller la progression de la copie des données d'Amazon RDS vers Amazon S3. La manière la plus simple de créer ce rôle de ressource consiste à utiliser le modèle `DataPipelineDefaultRole` et à répertorier `machinelearning.aws.com` comme service approuvé. Pour plus d'informations sur ce modèle, consultez Configuration de rôles IAM dans le Manuel du développeur AWS Data Pipeline.

Informations de sécurité Amazon RDS

Pour activer l'accès sécurisé à Amazon RDS, vous devez fournir le `VPC Subnet ID` et`RDS Security Group IDs`. Vous devez également configurer des règles de trafic entrant appropriées pour le sous-réseau VPC sur lequel pointe le paramètre `Subnet ID`, et fournir l'ID du groupe de sécurité qui possède cette autorisation.

Requête SQL MySQL

Le paramètre `MySQL SQL Query` spécifie la requête SQL SELECT que vous voulez exécuter sur votre base de données MySQL. Les résultats de cette requête sont copiés dans l'emplacement (compartiment) de sortie S3 que vous spécifiez.

> **Note**
>
> La technologie d'apprentissage-machine fonctionne le mieux lorsque les enregistrements d'entrée sont présentés dans un ordre aléatoire (réorganisé aléatoirement). Vous pouvez facilement réorganiser aléatoirement les résultats de votre requête SQL MySQL à l'aide de la fonction `rand()`. Par exemple, supposons que la requête d'origine est :
> « SELECT col1, col2, … FROM training_table »
> Vous pouvez ajouter une réorganisation aléatoire en mettant à jour la requête comme ceci :
> « SELECT col1, col2, … FROM training_table ORDER BY rand() »

Emplacement de sortie S3

Le `S3 Output Location` paramètre spécifie le nom de l'emplacement « intermédiaire » Amazon S3 où les résultats de la requête SQL MySQL sont générés.

> **Note**
>
> Vous devez vous assurer qu'Amazon ML est autorisé à lire les données depuis cet emplacement une fois les données exportées depuis Amazon RDS. Pour plus d'informations

sur la définition de ces autorisations, consultez Octroi à Amazon ML des autorisations nécessaires pour lire vos données depuis Amazon S3.

Formation des modèles d'apprentissage-machine

Le processus de formation d'un modèle d'apprentissage-machine implique la fourniture d'un algorithme d'apprentissage-machine (c'est-à-dire, l'algorithme d'apprentissage) avec des données de formation qui serviront à l'apprentissage. Le terme modèle d'apprentissage-machine fait référence à l'artefact de modèle qui est créé par le processus de formation.

Les données de formation doivent contenir la réponse correcte, qui porte le nom de cible ou d'attribut cible. L'algorithme d'apprentissage identifie des tendances dans les données de formation, qui mettent en correspondance les attributs des données d'entrée avec la cible (la réponse que vous souhaitez prédire), et il fournit en sortie un modèle d'apprentissage-machine qui capture ces tendances.

Vous pouvez utiliser le modèle d'apprentissage-machine pour obtenir des prédictions sur de nouvelles données pour lesquelles vous ne connaissez pas la cible. Par exemple, supposons que vous voulez former un modèle d'apprentissage-machine pour prédire si un e-mail correspond à du courrier indésirable ou non. Vous devez fournir à Amazon ML des données de formation contenant des e-mails dont vous connaissez la cible (c'est-à-dire une étiquette indiquant si un e-mail est un spam ou non). Amazon ML entraînerait un modèle de ML en utilisant ces données, pour obtenir un modèle qui tente de prédire si les nouveaux e-mails seront considérés comme du spam ou non.

Pour obtenir des informations générales sur les modèles d'apprentissage-machine et les algorithmes d'apprentissage-machine, consultez Concepts d'apprentissage-machine.

Rubriques

Types de modèles d'apprentissage-machine

Amazon ML prend en charge trois types de modèles de ML : classification binaire, classification multiclasse et régression. Le type de modèle que vous devez choisir dépend du type de cible que vous voulez prédire.

Modèle de classification binaire

Les modèles d'apprentissage-machine pour les problèmes de classification binaire prédisent un résultat binaire (une classe entre deux classes possibles). Pour entraîner des modèles de classification binaire, Amazon ML utilise l'algorithme d'apprentissage standard connu sous le nom de régression logistique.

Exemples de problèmes de classification binaire

- « Cet e-mail correspond-il à du courrier indésirable ou non ? »
- « Le client achètera-t-il ce produit ? »
- « Ce produit est-il un livre ou un animal de ferme ? »
- « Ce commentaire a-t-il été écrit par un client ou un robot ? »

Modèle de classification multiclasse

Les modèles d'apprentissage-machine pour les problèmes de classification multiclasse vous permettent de générer des prédictions pour plusieurs classes (prédire l'un d'au moins trois résultats). Pour la formation de modèles multiclasses, Amazon ML utilise l'algorithme d'apprentissage standard connu sous le nom de régression logistique multinomiale.

Exemples de problèmes multiclasse

- « Ce produit est-il un livre, un film ou un vêtement ? »
- « Ce film est-il une comédie romantique, un documentaire ou un thriller ? »
- « Quelle catégorie de produits intéresse le plus ce client ? »

Modèle de régression

Les modèles d'apprentissage-machine pour les problèmes de régression prédisent une valeur numérique. Pour l'entraînement des modèles de régression, Amazon ML utilise l'algorithme d'apprentissage standard connu sous le nom de régression linéaire.

Exemples de problèmes de régression

- « Quelle sera la température à Seattle demain ? »
- « Pour ce produit, combien d'unités se vendra-t-il ? »
- « A quel prix cette maison se vendra-t-elle ? »

Processus de formation

Pour former un modèle d'apprentissage-machine, vous devez spécifier ce qui suit :

- La source de données de formation en entrée
- Le nom de l'attribut de données qui contient la cible à prédire
- Les instructions de transformation de données requises
- Les paramètres de formation pour contrôler l'algorithme d'apprentissage

Pendant le processus de formation, Amazon ML sélectionne automatiquement le bon algorithme d'apprentissage à votre place, en fonction du type de cible que vous avez spécifié dans la source de données de formation.

Paramètres de formation

En règle générale, les algorithmes d'apprentissage-machine acceptent les paramètres pouvant servir à contrôler certaines propriétés du processus de formation et du modèle d'apprentissage-machine résultant. Dans Amazon Machine Learning, ces paramètres sont appelés paramètres d'entraînement. Vous pouvez définir ces paramètres à l'aide de la console Amazon ML, de l'API ou de l'interface de ligne de commande (CLI). Si vous ne définissez aucun paramètre, Amazon ML utilisera des valeurs par défaut reconnues pour leur efficacité dans le cadre d'un large éventail de tâches d'apprentissage automatique.

Vous pouvez spécifier des valeurs pour les paramètres de formation suivants :

- Taille maximale du modèle
- Nombre maximal de passages sur les données de formation
- Type de réorganisation
- Type de régularisation
- Degré de régularisation

Dans la console Amazon ML, les paramètres d'entraînement sont définis par défaut. Les paramètres par défaut sont appropriés pour la plupart des problèmes d'apprentissage-machine, mais vous pouvez choisir d'autres valeurs pour optimiser les performances. Certains autres paramètres de formation, tels que le taux d'apprentissage, sont configurés pour vous en fonction de vos données.

Les sections suivantes fournissent plus d'informations sur les paramètres de formation.

Taille maximale du modèle

La taille maximale du modèle est la taille totale, en octets, des modèles créés par Amazon ML lors de l'entraînement d'un modèle ML.

Par défaut, Amazon ML crée un modèle de 100 Mo. Vous pouvez demander à Amazon ML de créer un modèle plus petit ou plus grand en spécifiant une taille différente. Pour connaître la plage des tailles disponibles, consultez Types de modèles d'apprentissage-machine

Si Amazon ML ne trouve pas suffisamment de modèles pour remplir la taille du modèle, il crée un modèle plus petit. Par exemple, si vous spécifiez une taille de modèle maximale de 100 Mo, mais qu'Amazon ML trouve des modèles d'une taille totale de 50 Mo, le modèle obtenu sera de 50 Mo. Si Amazon ML trouve plus de modèles que ce qui correspond à la taille spécifiée, il applique une limite maximale en coupant les modèles qui affectent le moins la qualité du modèle appris.

Le choix de la taille du modèle vous permet de contrôler le compromis entre la qualité prédictive d'un modèle et le coût de son utilisation. Les modèles plus petits peuvent amener Amazon ML à supprimer de nombreux modèles pour respecter la limite de taille maximale, ce qui affecte la qualité des prédictions. D'un autre côté, des modèles plus grands sont plus coûteux dans le cadre des requêtes de prédictions en temps réel.

> **Note**
>
> Si vous utilisez un modèle d'apprentissage-machine pour générer des prédictions en temps réel, vous devrez vous acquitter de faibles frais de réservation de capacités, déterminés par la taille du modèle. Pour de plus amples informations, veuillez consulter Tarification pour Amazon ML.

Des ensembles de données d'entrée plus grands n'entraînent pas nécessairement la création de plus grands modèles, parce que les modèles stockent les tendances et non pas les données d'entrée. Si les tendances sont peu nombreuses et simples, le modèle résultant est de petite taille. Les données d'entrée comportant un grand nombre d'attributs bruts (colonnes d'entrée) ou de caractéristiques dérivées (sorties des transformations de données Amazon ML) comporteront probablement davantage de modèles découverts et stockés pendant le processus de formation. Pour sélectionner la bonne taille de modèle pour vos données et votre problème, la meilleure approche met en jeu quelques expériences. Le journal d'entraînement du modèle Amazon ML (que vous pouvez télécharger depuis la console ou via l'API) contient des messages indiquant dans quelle

mesure le modèle a été ajusté (le cas échéant) pendant le processus de formation, ce qui vous permet d'estimer la hit-to-prediction qualité potentielle.

Nombre maximal de passages sur les données

Pour de meilleurs résultats, Amazon ML peut avoir besoin de passer plusieurs fois sur vos données pour découvrir des modèles. Par défaut, Amazon ML effectue 10 passes, mais vous pouvez modifier la valeur par défaut en définissant un nombre allant jusqu'à 100. Amazon ML assure le suivi de la qualité des modèles (convergence des modèles) au fur et à mesure et arrête automatiquement l'entraînement lorsqu'il n'y a plus de points de données ou de modèles à découvrir. Par exemple, si vous définissez le nombre de passes à 20, mais qu'Amazon ML découvre qu'aucun nouveau modèle ne peut être trouvé au bout de 15 passes, la formation s'arrêtera à 15 passes.

En général, les ensembles de données contenant seulement quelques observations nécessitent plus de passages sur les données pour obtenir une meilleure qualité de modèle. Les ensembles de données plus grands contiennent de nombreux points de données similaires, ce qui élimine la nécessité d'avoir un grand nombre de passages. L'impact du choix d'un plus grand nombre de passages sur vos données est double : la formation du modèle est plus longue et a un coût plus élevé.

Type de réorganisation des données de formation

Dans Amazon ML, vous devez mélanger vos données d'entraînement. La réorganisation change complètement l'ordre de vos données de manière à ce que l'algorithme SGD ne rencontre pas un seul type de données pour un trop grand nombre d'observations consécutives. Par exemple, si vous formez un modèle d'apprentissage-machine pour prédire un type de produit et que vos données de formation incluent les types de produits film, jouet et jeu vidéo, si vous avez trié vos données selon la colonne de type de produit avant de les charger, l'algorithme voit les données dans l'ordre alphabétique, par type de produit. L'algorithme voit tout d'abord toutes vos données relatives aux films, et votre modèle d'apprentissage-machine commence à apprendre des tendances propres aux films. Ensuite, lorsque votre modèle aborde des données sur des jouets, chaque mise à jour que l'algorithme effectue correspond au modèle du type de produit jouet, même si ces mises à jour dégradent les tendances correspondant aux films. Ce basculement soudain du type film au type jouet peut produire un modèle qui n'apprend pas à prédire avec précision les types de produit.

Vous devez réorganiser aléatoirement vos données de formation, même si vous avez choisi l'option de fractionnement aléatoire lorsque vous avez divisé la source de données d'entrée en deux parties pour la formation et l'évaluation. La stratégie de fractionnement aléatoire choisit un sous-ensemble

aléatoire des données pour chaque source de données, mais elle ne modifie pas l'ordre des lignes dans la source de données. Pour plus d'informations sur le fractionnement de vos données, consultez Fractionnement des données.

Lorsque vous créez un modèle ML à l'aide de la console, Amazon ML mélange par défaut les données à l'aide d'une technique de brassage pseudo-aléatoire. Quel que soit le nombre de passes demandées, Amazon ML ne mélange les données qu'une seule fois avant d'entraîner le modèle de ML. Si vous avez mélangé vos données avant de les fournir à Amazon ML et que vous ne souhaitez pas qu'Amazon ML les mélange à nouveau, vous pouvez définir le type Shuffle sur. `none` Par exemple, si vous avez mélangé de manière aléatoire les enregistrements de votre fichier .csv avant de le télécharger sur Amazon S3, si vous avez utilisé la fonction `rand()` dans votre requête SQL MySQL lors de la création de votre source de données à partir d'Amazon RDS, ou si vous avez utilisé la fonction dans `random()` votre requête SQL Amazon Redshift lors de la création de votre source de données à partir d'Amazon Redshift, le fait de définir le type Shuffle sur n'aura aucun impact sur la précision prédictive de votre modèle ML. `none` Réorganiser vos données une seule fois réduit le temps et le coût d'exécution pour la création d'un modèle d'apprentissage-machine.

> **Important**
>
> Lorsque vous créez un modèle de machine learning à l'aide de l'API Amazon ML, Amazon ML ne mélange pas vos données par défaut. Si vous utilisez l'API à la place de la console pour créer votre modèle d'apprentissage-machine, nous vous recommandons fortement de réorganiser vos données en définissant le paramètre `sgd.shuffleType` sur `auto`.

Type et degré de régularisation

Les performances prédictives de modèles d'apprentissage-machine complexes (dotés de nombreux attributs d'entrée) pâtissent lorsque les données contiennent un trop grand nombre de tendances. Plus le nombre de tendances est élevé, plus le modèle a de chances d'apprendre des artefacts non intentionnels plutôt que des tendances réelles. Dans ce cas, le modèle se comporte très bien sur les données de formation, mais ne peut pas généraliser correctement aux données nouvelles. Ce phénomène est appelé surajustement des données de formation.

La régularisation permet d'empêcher le surajustement des exemples de données de formation dans le cas de modèles linéaires, en pénalisant les valeurs d'un poids extrême. La régularisation L1 réduit le nombre d'entités utilisées dans le modèle en mettant à zéro le poids d'entités qui auraient autrement un très faible poids. La régularisation L1 produit des modèles dispersés et réduit la

quantité de bruit dans le modèle. La régularisation L2 produit des valeurs pondérales globales plus petites, ce qui stabilise les poids lorsqu'il y a une forte corrélation entre les entités. Vous pouvez contrôler le degré de régularisation L1 ou L2 à l'aide du paramètre `Regularization amount`. La spécification d'une très grande valeur `Regularization amount` peut entraîner la mise à zéro du poids de toutes les entités.

La sélection et l'ajustement de la valeur de régularisation optimale est un thème actuel de recherche dans le domaine de l'apprentissage-machine. Vous aurez probablement intérêt à sélectionner un niveau modéré de régularisation L2, qui est la valeur par défaut dans la console Amazon ML. Les utilisateurs avancés peuvent choisir entre trois types de régularisation (aucune, L1 et L2) et le degré. Pour plus d'informations sur la régularisation, consultez Régularisation (mathématiques).

Paramètres de formation : Types et valeurs par défaut

Le tableau suivant répertorie les paramètres d'entraînement Amazon ML, ainsi que les valeurs par défaut et la plage autorisée pour chacun d'entre eux.

Paramètre de formation	Type	Valeur par défaut	Description
maximum MLModel SizeInBytes	Entier	100 000 000 octets (100 Mio)	Plage autorisée : de 100 000 (100 Kio) à 2 147 483 648 (2 Gio) En fonction des données d'entrée, la taille du modèle peut avoir un impact sur les performances.
sgd.maxPasses	Entier	10	Plage autorisée : 1-100
sgd.shuffleType	Chaîne	auto	Valeurs autorisées : `auto` ou `none`
sgd.l1 RegularizationAmount	Double	0 (par défaut, L1 n'est pas utilisée)	Plage autorisée : de 0 à MAX_DOUBLE Les valeurs L1 comprises entre 1E-4 et 1E-8 sont connues pour donner de bons résultats. Des valeurs supérieures sont susceptibles de produire des modèles qui ne sont pas très utiles.

Paramètre de formation	Type	Valeur par défaut	Description
			Vous ne pouvez pas définir à la fois L1 et L2. Vous devez choisir l'une ou l'autre.
sgd.l2 RegularizationAmount	Double	1E-6 (par défaut, L2 est utilisée avec ce degré de régularisation)	Plage autorisée : de 0 à MAX_DOUBLE Les valeurs L2 comprises entre 1E-2 et 1E-6 sont connues pour donner de bons résultats. Des valeurs supérieures sont susceptibles de produire des modèles qui ne sont pas très utiles. Vous ne pouvez pas définir à la fois L1 et L2. Vous devez choisir l'une ou l'autre.

Création d'un modèle d'apprentissage-machine

Une fois que vous avez créé une source de données, vous êtes prêt à créer un modèle d'apprentissage-machine. Si vous utilisez la console Amazon Machine Learning pour créer un modèle, vous pouvez choisir d'utiliser les paramètres par défaut ou de personnaliser votre modèle en appliquant des options personnalisées.

Les options personnalisées sont les suivantes :

- Paramètres d'évaluation : vous pouvez demander à Amazon ML de réserver une partie des données d'entrée afin d'évaluer la qualité prédictive du modèle de ML. Pour obtenir des informations sur les évaluations, consultez Evaluation des modèles d'apprentissage-machine.
- Une recette : une recette indique à Amazon ML quels attributs et transformations d'attributs sont disponibles pour l'entraînement des modèles. Pour plus d'informations sur les recettes Amazon ML, consultez la section Transformations de fonctionnalités avec des recettes de données.

- Paramètres de formation : les paramètres contrôlent certaines propriétés du processus de formation et du modèle d'apprentissage-machine qui en résulte. Pour plus d'informations sur les paramètres de formation, consultez Paramètres de formation.

Pour sélectionner ou spécifier des valeurs pour ces paramètres, choisissez l'option Personnalisé lorsque vous utilisez l'assistant de création de modèle d'apprentissage-machine. Si vous souhaitez qu'Amazon ML applique les paramètres par défaut, choisissez Default.

Lorsque vous créez un modèle ML, Amazon ML sélectionne le type d'algorithme d'apprentissage qu'il utilisera en fonction du type d'attribut de votre attribut cible. (L'attribut cible est l'attribut qui contient les réponses « correctes ».) Si votre attribut cible est binaire, Amazon ML crée un modèle de classification binaire qui utilise l'algorithme de régression logistique. Si votre attribut cible est catégorique, Amazon ML crée un modèle multiclasse qui utilise un algorithme de régression logistique multinomial. Si votre attribut cible est numérique, Amazon ML crée un modèle de régression qui utilise un algorithme de régression linéaire.

Rubriques

Prérequis

Avant d'utiliser la console Amazon ML pour créer un modèle de ML, vous devez créer deux sources de données, l'une pour entraîner le modèle et l'autre pour évaluer le modèle. Si vous n'avez pas créé deux sources de données, consultez Etape 2 : Création d'une source de données de formation dans ce didacticiel.

Création d'un modèle d'apprentissage-machine avec les options par défaut

Choisissez les options par défaut si vous souhaitez qu'Amazon ML :

- fractionne les données d'entrée pour en utiliser 70 % pour la formation et les 30 % restants pour l'évaluation ;
- suggère une recette basée sur les statistiques collectées sur la source de données de formation, qui représente 70 % de la source de données d'entrée ;
- choisisse les paramètres de formation par défaut.

Pour choisir les options par défaut

1. Dans la console Amazon ML, choisissez Amazon Machine Learning, puis choisissez ML models.
2. Dans la page récapitulative ML models, choisissez Create a new ML model.
3. Dans la page Input data, assurez-vous que l'option I already created a datasource pointing to my S3 data est sélectionnée.
4. Dans la table, choisissez votre source de données, puis choisissez Continue.
5. Dans la page ML model settings, pour ML model name, tapez un nom pour votre modèle d'apprentissage-machine.
6. Pour Training and evaluation settings, assurez-vous que la valeur Par défaut est sélectionnée.
7. Pour Nommer cette évaluation, tapez le nom de l'évaluation, puis choisissez Réviser. Amazon ML contourne le reste de l'assistant et vous dirige vers la page de révision.
8. Passez en revue vos données, supprimez les balises copiées depuis la source de données que vous ne voulez pas appliquer à votre modèle et à vos évaluations, puis choisissez Terminer.

Création d'un modèle d'apprentissage-machine avec des options personnalisées

La personnalisation de votre modèle d'apprentissage-machine vous offre les possibilités suivantes :

- Fournir votre propre recette. Pour obtenir des informations sur la façon de fournir votre propre recette, consultez Référence de format des recettes.
- Choisir les paramètres de formation. Pour plus d'informations sur les paramètres de formation, consultez Paramètres de formation.
- Choisir un rapport de fractionnement pour formation/évaluation autre que le rapport 70/30 par défaut ou fournir une autre source de données que vous avez déjà préparée pour l'évaluation. Pour obtenir des informations sur les stratégies de fractionnement, consultez Fractionnement des données.

Vous pouvez également choisir les valeurs par défaut d'un ou plusieurs de ces paramètres.

Si vous avez déjà créé un modèle à l'aide des options par défaut et que vous souhaitez améliorer les performances prédictives de votre modèle, utilisez l'option Personnalisé pour créer un nouveau modèle avec certains paramètres personnalisés. Par exemple, vous pouvez ajouter des

transformations d'entité supplémentaires à la recette ou augmenter le nombre de passages dans le paramètre de formation.

Pour créer un modèle avec des options personnalisées

1. Dans la console Amazon ML, choisissez Amazon Machine Learning, puis choisissez ML models.
2. Dans la page récapitulative ML models, choisissez Create a new ML model.
3. Si vous avez déjà créé une source de données, dans la page Input data, choisissez I already created a datasource pointing to my S3 data. Dans la table, choisissez votre source de données, puis choisissez Continue.

   Si vous avez besoin de créer une source de données, choisissez My data is in S3, and I need to create a datasource, puis Continue. Vous êtes redirigé vers l'assistant Create a Datasource. Spécifiez si vos données se trouvent dans S3 ou Redshift, puis choisissez Vérifier. Terminez la procédure de création d'une source de données.

   Une fois que vous avez créé une source de données, vous êtes redirigé vers l'étape suivante dans l'assistant Create ML Model.
4. Dans la page ML model settings, pour ML model name, tapez un nom pour votre modèle d'apprentissage-machine.
5. Dans Select training and evaluation settings, choisissez Custom, puis choisissez Continue.
6. Dans la page Recette, vous pouvez customize a recipe. Si vous ne souhaitez pas personnaliser de recette, Amazon ML vous en suggère une. Choisissez Continuer.
7. Dans la page Advanced settings (Paramètres avancés), spécifiez les valeurs Maximum ML model Size (Taille maximum de modèle d'apprentissage automatique), Maximum number of data passes (Nombre maximum de passes de données), Shuffle type for training data (Type de réorganisation pour les données de formation), Regularization type (Type de réorganisation) et Regularization amount (Montant de régularisation). Si vous ne les spécifiez pas, Amazon ML utilise les paramètres d'entraînement par défaut.

   Pour plus d'informations sur ces paramètres et leurs valeurs par défaut, consultez Paramètres de formation.

   Choisissez Continuer.
8. Dans la page Evaluation, spécifiez si vous souhaitez évaluer le modèle d'apprentissage-machine immédiatement. Si vous ne voulez pas évaluer le modèle d'apprentissage-machine maintenant, choisissez Review.

Si vous souhaitez évaluer le modèle d'apprentissage-machine maintenant :

a. Pour Name this evaluation (Nommer cette évaluation), tapez un nom pour l'évaluation.
b. Pour Select evaluation data, indiquez si vous souhaitez qu'Amazon ML réserve une partie des données d'entrée à des fins d'évaluation et, le cas échéant, comment vous souhaitez diviser la source de données, ou choisissez de fournir une autre source de données pour l'évaluation.
c. Choisissez Examiner.

9. Dans la page Review, modifiez vos sélections, supprimez les balises copiées depuis la source de données que vous ne voulez pas appliquer à votre modèle et à vos évaluations, puis choisissez Finish.

Une fois que vous avez créé le modèle, consultez Etape 4 : Examen des performances prédictives du modèle d'apprentissage-machine et définition d'un score seuil.

Transformations de données pour l'apprentissage-machine

Les modèles d'apprentissage-machine ne sont pas meilleurs que les données qui ont servi à les former. Une caractéristique clé de bonnes données de formation est qu'elles sont fournies d'une manière optimisée pour l'apprentissage et la généralisation. Le processus visant à rassembler les données dans ce format optimal est connu dans le secteur sous le nom de transformation de fonctionnalité.

Rubriques

Importance de la transformation des entités

Considérez un modèle d'apprentissage-machine dont la tâche consiste à déterminer si une transaction de carte de crédit est frauduleuse ou non. Sur la base de la connaissance du contexte de votre application et de l'analyse des données, vous pouvez décider des champs de données (ou entités) qu'il est important d'inclure dans les données d'entrée. Par exemple, le montant de la transaction, le nom du vendeur, l'adresse et l'adresse du propriétaire de la carte de crédit sont importants à fournir au processus d'apprentissage. D'un autre côté, un ID de transaction généré de façon aléatoire ne porte aucune information (si nous savons qu'il est vraiment aléatoire) et n'est pas utile.

Une fois que vous avez décidé quels champs inclure, vous transformez ces entités pour faciliter le processus d'apprentissage. Les transformations ajoutent une expérience en arrière-plan aux données d'entrée, ce qui permet au modèle d'apprentissage-machine de bénéficier de cette expérience. Par exemple, l'adresse de vendeur suivante est représentée sous la forme d'une chaîne :

« 123 Main Street, Seattle, WA 98101 »

Par elle-même, l'adresse a un pouvoir expressif limité : elle est utile uniquement pour apprendre les tendances associées à l'adresse exacte. La décomposer en éléments constitutifs, toutefois,

peut créer des entités supplémentaires telles que « Address » (123 Main Street), « City » (Seattle), « State » (WA) et « Zip » (98101). Maintenant, l'algorithme d'apprentissage peut regrouper plus de transactions disparates et découvrir des tendances plus larges. Certains codes zip de vendeurs connaissent éventuellement plus d'activités frauduleuses que d'autres.

Pour plus d'informations sur l'approche et le processus de transformation d'entités, consultez Concepts d'apprentissage-machine.

Transformations d'entités à l'aide de recettes de données

Il existe deux façons de transformer des entités avant de créer des modèles d'apprentissage-machine à l'aide d'Amazon ML : vous pouvez transformer vos données d'entrée directement avant de les montrer à Amazon ML, ou vous pouvez utiliser les transformations de données intégrées d'Amazon ML. Vous pouvez utiliser les recettes Amazon ML, qui sont des instructions préformatées pour les transformations courantes. Les recettes vous permettent d'effectuer les actions suivantes :

- Choisir des transformations dans une liste de transformations d'apprentissage-machine courantes intégrées, et les appliquer à des variables individuelles ou à des groupes de variables
- Sélectionner les transformations et les variables d'entrée à mettre à la disposition du processus d'apprentissage-machine

L'utilisation des recettes Amazon ML offre plusieurs avantages. Amazon ML effectue les transformations de données pour vous, de sorte que vous n'avez pas besoin de les mettre en œuvre vous-même. En outre, les transformations sont rapides, car Amazon ML les applique lors de la lecture des données d'entrée, et fournit les résultats au processus d'apprentissage sans l'étape intermédiaire d'enregistrement de ces résultats sur le disque.

Référence de format des recettes

Les recettes Amazon ML contiennent des instructions pour transformer vos données dans le cadre du processus d'apprentissage automatique. Les recettes sont définies à l'aide d'une syntaxe de type JSON, mais elles présentent des restrictions supplémentaires au-delà des restrictions JSON normales. Les recettes possèdent les sections suivantes, qui doivent apparaître dans l'ordre indiqué ici :

- La section groups (groupes) permet le regroupement de plusieurs variables, simplifiant ainsi l'application de transformations. Par exemple, vous pouvez créer un groupe de toutes les variables

ayant à faire avec les parties en texte libre d'une page web (titre, corps), puis effectuer une transformation globale sur l'ensemble de ces parties.
- La section assignments (affectations) permet la création de variables nommées intermédiaires qui peuvent être réutilisées dans le traitement.
- La section outputs (sorties) définit les variables qui seront utilisées dans le processus d'apprentissage, ainsi que les transformations (le cas échéant) qui s'appliquent à ces variables.

Groups

Vous pouvez définir des groupes de variables afin de transformer collectivement toutes les variables au sein des groupes ou d'utiliser ces variables pour l'apprentissage-machine sans les transformer. Par défaut, Amazon ML crée les groupes suivants pour vous :

ALL_TEXT, ALL_NUMERIC, ALL_CATEGORICAL, ALL_BINARY : groupes spécifiques aux types, basés sur les variables définies dans le schéma de source de données.

> **Note**
>
> Vous ne pouvez pas créer de groupe avec `ALL_INPUTS`.

Ces variables peuvent être utilisées dans la section outputs de votre recette sans être définies. Vous pouvez également créer des groupes personnalisés en ajoutant ou en soustrayant des variables dans les groupes existants, ou directement dans une collection de variables. Dans l'exemple suivant, nous illustrons ces trois approches, ainsi que la syntaxe pour l'affectation des groupements :

```
"groups": {

"Custom_Group": "group(var1, var2)",
"All_Categorical_plus_one_other": "group(ALL_CATEGORICAL, var2)"

}
```

Les noms des groupes doivent commencer par un caractère alphabétique et peuvent comporter entre 1 et 64 caractères. Si le nom de groupe ne commence pas par un caractère alphabétique ou s'il contient des caractères spéciaux (, ' " \t \r \n ( ) \), le nom doit être placé entre guillemets pour être inclus dans la recette.

Assignments (affectations)

Vous pouvez affecter une ou plusieurs transformations à une variable intermédiaire, pour des raisons de commodité et de lisibilité. Par exemple, si vous avez une variable texte nommée email_subject et que vous lui appliquez la transformation en minuscules, vous pouvez nommer la variable résultante email_subject_lowercase, ce qui facilite son suivi ailleurs dans la recette. Les affectations peuvent également être chaînées, ce qui vous permet d'appliquer plusieurs transformations dans un ordre donné. L'exemple suivant montre des affectations individuelles et chaînées dans la syntaxe d'une recette :

```
"assignments": {

"email_subject_lowercase": "lowercase(email_subject)",

"email_subject_lowercase_ngram":"ngram(lowercase(email_subject), 2)"

}
```

Les noms des variables intermédiaires doivent commencer par un caractère alphabétique et peuvent comporter entre 1 et 64 caractères. Si le nom ne commence pas par un caractère alphabétique ou s'il contient des caractères spéciaux (, ' " \t \r \n ( ) \), le nom doit être placé entre guillemets pour être inclus dans la recette.

Outputs

La section outputs contrôle quelles variables d'entrée seront utilisées pour le processus d'apprentissage et quelles transformations leur sont applicables. Une section outputs vide ou inexistante est une erreur, car aucune donnée ne sera transmise au processus d'apprentissage.

La section outputs la plus simple comprend simplement le groupe ALL_INPUTS prédéfini, ce qui indique à Amazon ML d'utiliser toutes les variables définies dans la source de données pour l'apprentissage :

```
"outputs": [

"ALL_INPUTS"
```

```
]
```

La section outputs peut également faire référence aux autres groupes prédéfinis en indiquant à Amazon ML d'utiliser toutes les variables de ces groupes :

```
"outputs": [

"ALL_NUMERIC",

"ALL_CATEGORICAL"

]
```

La section outputs peut également faire référence à des groupes personnalisés. Dans l'exemple suivant, un seul des groupes personnalisés définis dans la section des affectations de groupement de l'exemple précédent sera utilisé pour l'apprentissage-machine. Toutes les autres variables seront ignorées :

```
"outputs": [

"All_Categorical_plus_one_other"

]
```

La section outputs peut également faire référence aux affectations de variables définies dans la section assignments :

```
"outputs": [

"email_subject_lowercase"

]
```

De plus, les transformations et les variables d'entrée peuvent être définies directement dans la section outputs :

```
"outputs": [

"var1",
```

```
"lowercase(var2)"

]
```

La sortie doit spécifier explicitement toutes les variables et variables transformées qui sont censées être disponibles pour le processus d'apprentissage. Supposons, par exemple, que vous incluez dans la section outputs le produit cartésien de var1 et var2. Si vous souhaitez inclure à la fois les variables brutes var1 et var2, vous devez ajouter les variables brutes dans la section outputs :

```
"outputs": [

"cartesian(var1,var2)",

"var1",

"var2"

]
```

La section outputs peut inclure des commentaires pour des raisons de lisibilité, le texte de commentaire étant ajouté à côté de la variable :

```
"outputs": [

"quantile_bin(age, 10) //quantile bin age",

"age // explicitly include the original numeric variable along with the
binned version"

]
```

Vous pouvez combiner toutes ces approches dans la section outputs.

> **Note**
>
> Les commentaires ne sont pas autorisés dans la console Amazon ML lors de l'ajout d'une recette.

Exemple de recette complète

L'exemple suivant fait référence à plusieurs processeurs de données intégrés, qui ont été présentés dans les exemples précédents :

```
{

"groups": {

"LONGTEXT": "group_remove(ALL_TEXT, title, subject)",

"SPECIALTEXT": "group(title, subject)",

"BINCAT": "group(ALL_CATEGORICAL, ALL_BINARY)"

},

"assignments": {

"binned_age" : "quantile_bin(age,30)",

"country_gender_interaction" : "cartesian(country, gender)"

},

"outputs": [

"lowercase(no_punct(LONGTEXT))",

"ngram(lowercase(no_punct(SPECIALTEXT)),3)",

"quantile_bin(hours-per-week, 10)",

"hours-per-week // explicitly include the original numeric variable
along with the binned version",

"cartesian(binned_age, quantile_bin(hours-per-week,10)) // this one is
critical",

"country_gender_interaction",

"BINCAT"
```

```
]

}
```

Recettes suggérées

Lorsque vous créez une nouvelle source de données dans Amazon ML et que les statistiques sont calculées pour cette source de données, Amazon ML crée également une recette suggérée qui peut être utilisée pour créer un nouveau modèle d'apprentissage-machine à partir de la source de données. La source de données suggérée est basée sur les données et sur l'attribut cible présent dans les données, et fournit un point de départ utile pour créer et affiner vos modèles d'apprentissage-machine.

Pour utiliser la recette suggérée dans la console Amazon ML, choisissez Datasource ou Datasource and ML model dans la liste déroulante Create new. Pour les paramètres du modèle d'apprentissage-machine, vous avez le choix entre les paramètres Default et Custom Training and Evaluation dans l'étape de paramétrage du modèle d'apprentissage-machine de l'assistant Create ML Model. Si vous choisissez l'option Par défaut, Amazon ML utilise automatiquement la recette suggérée. Si vous choisissez l'option personnalisée, l'éditeur de recettes à l'étape suivante affichera la recette suggérée, et vous pourrez la vérifier ou la modifier si nécessaire.

> **Note**
>
> Amazon ML vous permet de créer une source de données, puis de l'utiliser immédiatement pour créer un modèle d'apprentissage-machine, avant la fin du calcul des statistiques. Dans ce cas, vous ne pourrez pas consulter la recette suggérée dans l'option personnalisée, mais vous pourrez toujours passer outre cette étape et indiquer à Amazon ML d'utiliser la recette par défaut pour la formation du modèle.

Pour utiliser la recette suggérée avec l'API Amazon ML, vous pouvez transmettre une chaîne vide dans les paramètres de la recette et de RecipeUri l'API. Il n'est pas possible de récupérer la recette suggérée à l'aide de l'API Amazon ML.

Référence des transformations de données

Rubriques

Transformation n-gramme

La transformation n-gramme accepte une variable texte comme entrée et génère des chaînes correspondant au glissement d'une fenêtre de n mots (configurable par l'utilisateur), en générant des résultats au cours de ce processus. Par exemple, considérez la chaîne de texte « I really enjoyed reading this book » (J'ai vraiment apprécié la lecture de ce livre).

La transformation n-gramme paramétrée avec une taille de fenêtre de 1 vous donne simplement tous les mots individuels figurant dans cette chaîne :

```
{"I", "really", "enjoyed", "reading", "this", "book"}
```

La transformation n-gramme paramétrée avec une taille de fenêtre de 2 vous donne toutes les combinaisons de deux mots, ainsi que les mots individuels :

```
{"I really", "really enjoyed", "enjoyed reading", "reading this", "this
book", "I", "really", "enjoyed", "reading", "this", "book"}
```

La transformation n-gramme paramétrée avec une taille de fenêtre de 3 ajoute les combinaisons à 3 mots à cette liste, et donne le résultat suivant :

```
{"I really enjoyed", "really enjoyed reading", "enjoyed reading this",
"reading this book", "I really", "really enjoyed", "enjoyed reading",
"reading this", "this book", "I", "really", "enjoyed", "reading",
"this", "book"}
```

Vous pouvez demander des n-grammes avec une taille allant de 2 à 10 mots. Les n-grammes de taille 1 sont générés implicitement pour toutes les entrées dont le type est marqué en tant que texte dans le schéma de données, de sorte que vous n'avez pas à les demander. Enfin, souvenez-vous que les n-grammes sont générés en divisant les données d'entrée au niveau des espaces. Cela signifie que, par exemple, les caractères de ponctuation sont considérés comme faisant partie des jetons : la création de n-grammes avec une fenêtre de taille 2 pour la chaîne « rouge, vert, bleu » donne {"rouge,", "vert,", "bleu,", "rouge, vert", "vert, bleu"}. Vous pouvez utiliser le processeur de suppression de ponctuation (décrit plus loin dans ce document) pour supprimer les symboles de ponctuation si vous n'en voulez pas.

Pour calculer les n-grammes de la variable var1 pour une taille de fenêtre de 3 :

```
"ngram(var1, 3)"
```

Transformation bigramme d'analyse orthogonale (OSB, Orthogonal Sparse Bigram)

La transformation OSB est destinée à faciliter l'analyse des chaînes de texte et constitue une alternative à la transformation en deux grammes (n-gramme avec une taille de fenêtre de 2). OSBs sont générés en faisant glisser la fenêtre de taille n sur le texte et en sortant chaque paire de mots qui inclut le premier mot de la fenêtre.

Pour élaborer un OSB, les mots le constituant sont joints par le caractère « _ » (trait de soulignement) et chaque jeton ignoré est indiqué par l'ajout d'un autre trait de soulignement dans l'OSB. Ainsi, l'OSB ne code pas seulement les jetons visibles au sein d'une fenêtre, mais fournit aussi une indication du nombre de jetons ignorés dans cette même fenêtre.

À titre d'illustration, considérez la ficelle « Le rapide renard brun saute par-dessus le chien paresseux », et elle est OSBs de taille 4. Les six fenêtres de quatre mots et les deux dernières fenêtres plus courtes à partir de la fin de la chaîne sont illustrées dans l'exemple suivant, ainsi que OSBs générées à partir de chacune d'elles :

Fenêtre, {OSBs générée}

```
"The quick brown fox", {The_quick, The__brown, The___fox}

"quick brown fox jumps", {quick_brown, quick__fox, quick___jumps}
```

```
"brown fox jumps over", {brown_fox, brown__jumps, brown___over}

"fox jumps over the", {fox_jumps, fox__over, fox___the}

"jumps over the lazy", {jumps_over, jumps__the, jumps___lazy}

"over the lazy dog", {over_the, over__lazy, over___dog}

"the lazy dog", {the_lazy, the__dog}

"lazy dog", {lazy_dog}
```

Les bigrammes d'analyse orthogonale constituent une alternative aux n-grammes et sont susceptibles de mieux fonctionner dans certaines situations. Si vos données contiennent de grands champs textuels (10 mots ou plus), faites des expériences pour voir ce qui convient le mieux. Notez que la qualification « grand champ textuel » peut varier en fonction de la situation. Cependant, dans le cas de champs de texte plus grands, OSBs il a été démontré empiriquement qu'ils représentent le texte de manière unique en raison du symbole spécial de saut (le trait de soulignement).

Vous pouvez demander une taille de fenêtre de 2 à 10 pour les transformations OSB sur des variables de texte en entrée.

Pour calculer OSBs avec une taille de fenêtre 5 pour la variable var1 :

"osb(var1, 5)"

Transformation en minuscules

Le processeur de transformation en minuscules convertit les entrées de texte en minuscules. Par exemple, avec les données d'entrée « The Quick Brown Fox Jumps Over the Lazy Dog », le processeur fournit en sortie « the quick brown fox jumps over the lazy dog ».

Pour appliquer une transformation en minuscules à la variable var1 :

"lowercase(var1)"

Transformation de suppression de la ponctuation

Amazon ML fractionne implicitement les entrées marquées en tant que texte dans le schéma de données au niveau des espaces. La ponctuation figurant dans la chaîne finit soit scindée aux jetons, soit considérée comme des jetons à part entière, en fonction des espaces qui l'entourent. Si ce n'est

pas ce que vous souhaitez, vous pouvez utiliser la transformation de suppression de la ponctuation pour supprimer les symboles de ponctuation des entités générées. Par exemple, dans le cas de la chaîne « Welcome to AML - please fasten your seat-belts! », l'ensemble de jetons suivant est généré implicitement :

```
{"Welcome", "to", "Amazon", "ML", "-", "please", "fasten", "your", "seat-belts!"}
```

L'application du processeur de suppression de la ponctuation à cette chaîne fournit l'ensemble suivant :

```
{"Welcome", "to", "Amazon", "ML", "please", "fasten", "your", "seat-belts"}
```

Notez que seuls les signes de ponctuation en préfixe ou en suffixe sont supprimés. La ponctuation qui figure au milieu d'un jeton, par exemple, le trait d'union dans « seat-belts », n'est pas supprimée.

Pour appliquer la suppression de la ponctuation à la variable var1 :

"no_punct(var1)"

Transformation de discrétisation par quantiles

Le processeur de discrétisation par quantiles accepte deux entrées, une variable numérique et un paramètre appelé nombre d'intervalles, et génère en sortie une variable de catégorie. L'objectif est de découvrir une non-linéarité dans la distribution de la variable en regroupant les valeurs observées.

Dans de nombreux cas, la relation entre une variable numérique et la cible n'est pas linéaire (la valeur de la variable numérique n'augmente pas et ne diminue pas de façon monotone avec la cible). Dans de tels cas, il peut être utile de discrétiser l'entité numérique en une entité de catégorie représentant différentes plages de l'entité numérique. Chaque valeur d'entité de catégorie (intervalle) peut ensuite être modélisée comme ayant sa propre relation linéaire avec la cible. Par exemple, supposons que vous savez que l'entité numérique continue account_age (âge du compte) n'est pas corrélée linéairement à la probabilité d'acheter un livre. Vous pouvez discrétiser l'âge en entités de catégorie susceptibles de capturer plus précisément la relation avec la cible.

Le processeur de discrétisation par quantiles peut être utilisé pour demander à Amazon ML d'établir n intervalles de taille égale sur la distribution de toutes les valeurs d'entrée de la variable d'âge, puis de remplacer chaque nombre par un jeton textuel contenant l'intervalle. Le nombre optimal d'intervalles pour une variable numérique dépend des caractéristiques de la variable et de sa relation à la cible, et la meilleure façon de le déterminer passe par l'expérimentation. Amazon ML suggère

le nombre optimal d'intervalles pour une entité numérique en fonction des statistiques des données figurant dans la recette suggérée.

Vous pouvez demander le calcul d'entre 5 et 1 000 intervalles pour n'importe quelle variable numérique en entrée.

L'exemple suivant montre comment calculer et utiliser 50 intervalles à la place de la variable numérique var1 :

"quantile_bin(var1, 50)"

Transformation de normalisation

Le transformateur de normalisation normalise les variables numériques pour obtenir une moyenne de 0 et une variance égale à 1. La normalisation des variables numériques peut aider le processus d'apprentissage s'il existe de très grandes différences de plages entre les variables numériques, car les variables de grandeur la plus élevée pourraient dominer le modèle d'apprentissage-machine, que l'entité apporte ou non des informations sur la cible.

Pour appliquer cette transformation à la variable numérique var1, ajoutez ce qui suit à la recette :

normalize(var1)

Ce transformateur peut également accepter un groupe défini par l'utilisateur de variables numériques ou le groupe prédéfini pour toutes les variables numériques (ALL_NUMERIC) comme entrée :

normalize(ALL_NUMERIC)

Remarque

Il n'est pas obligatoire d'utiliser le processeur de normalisation pour les variables numériques.

Transformation par produit cartésien

La transformation cartésienne génère des permutations d'au moins deux variables d'entrée de texte ou de catégorie. Cette transformation est utilisée lorsqu'une interaction entre les variables est suspectée. Par exemple, considérez le jeu de données marketing bancaire utilisé dans Didacticiel : Utilisation d'Amazon ML pour prédire les réponses à une offre marketing. Avec ce jeu de données, nous aimerions prédire si une personne répondra positivement à une promotion de sa banque, en fonction des informations économiques et démographiques. Nous soupçonnons éventuellement que le type de travail de la personne joue également un rôle (il existe peut-être une corrélation entre

le fait d'être employé dans certains domaines et d'avoir des économies disponibles), et le niveau le plus élevé de formation atteint est également important. Nous pouvons aussi avoir une intuition plus profonde qu'il existe un signal fort dans l'interaction de ces deux variables ; par exemple, que la promotion est particulièrement adaptée aux clients qui sont des entrepreneurs ayant obtenu un diplôme universitaire.

La transformation par produit cartésien accepte des variables de catégorie ou du texte en entrée, et génère de nouvelles entités qui capturent l'interaction entre ces variables d'entrée. En particulier, pour chaque exemple de formation, elle crée une combinaison d'entités et les ajoute en tant qu'entité autonome. Par exemple, supposons que nos lignes d'entrée simplifiées ressemblent à ceci :

target, education, job

0, university.degree, technician

0, high.school, services

1, university.degree, admin

Si nous spécifions que la transformation cartésienne doit être appliquée aux champs education et job des variables de catégorie, l'entité obtenue education_job_interaction ressemblera à ceci :

target, education_job_interaction

0, university.degree_technician

0, high.school_services

1, university.degree_admin

La transformation cartésienne est encore plus puissante quand il s'agit de travailler sur des séquences de jetons, comme c'est le cas lorsque l'un de ses arguments est une variable de texte fractionnée implicitement ou explicitement en jetons. Par exemple, considérez la tâche consistant à classer un livre comme manuel ou non. Intuitivement, nous pouvons penser que des éléments du titre peuvent nous indiquer s'il s'agit d'un manuel (certains mots peuvent apparaître plus fréquemment dans les titres des manuels), et nous pouvons également croire que certaines caractéristiques de la reliure du livre peuvent être prédictives (les manuels ont plus de chances d'être reliés [Hardcover] que d'avoir une couverture souple [Softcover]), mais c'est vraiment la combinaison de certains mots du titre et de la reliure qui est la plus prédictive. Comme exemple concret, le tableau suivant montre les résultats de l'application du processeur cartésien aux variables d'entrée Binding (reliure) et Title (titre) :

Manu	Title	Binding	Produit cartésien de no_punct(Title) et Bindind
1	Economics : Principles, Problems, Policies	Hardcover	{"Economics_Hardcover", "Principles_Hardcover", "Problems_Hardcover", "Policies_Hardcover"}
0	The Invisible Heart: An Economics Romance	Softcover	{"The_Softcover", "Invisible_Softcover", "Heart_Softcover", "An_Softcover", "Economics_Softcover", "Romance_ Softcover"}
0	Fun With Problems	Softcover	{"Fun_Softcover", "With_Softcover", "Problems_Softcover"}

L'exemple suivant montre comment appliquer le transformateur cartésien à var1 et var2 :

cartesian(var1, var2)

Réorganisation des données

La fonctionnalité de réorganisation de données vous permet de créer une source de données basée uniquement sur une partie des données d'entrée sur laquelle elle pointe. Par exemple, lorsque vous créez un modèle ML à l'aide de l'assistant Create ML de la console Amazon ML et que vous choisissez l'option d'évaluation par défaut, Amazon ML réserve automatiquement 30 % de vos données pour l'évaluation du modèle ML et utilise les 70 % restants pour la formation. Cette fonctionnalité est activée par la fonctionnalité de réarrangement des données d'Amazon ML.

Si vous utilisez l'API Amazon ML pour créer des sources de données, vous pouvez spécifier la partie des données d'entrée qui sera basée sur une nouvelle source de données. Pour ce faire, vous devez transmettre les instructions du `DataRearrangement` paramètre au`CreateDataSourceFromS3`, `CreateDataSourceFromRedshift` ou `CreateDataSourceFromRDS` APIs. Le contenu de la DataRearrangement chaîne est une chaîne JSON contenant les emplacements de début et de fin de vos données, exprimés sous forme de pourcentages, d'un indicateur de complément et d'une stratégie de division. Par exemple, la DataRearrangement chaîne suivante indique que les 70 % premiers des données seront utilisés pour créer la source de données :

```
{
    "splitting": {
```

```
        "percentBegin": 0,
        "percentEnd": 70,
        "complement": false,
        "strategy": "sequential"
    }
}
```

DataRearrangement Paramètres

Pour modifier la façon dont Amazon ML crée une source de données, utilisez les paramètres suivants.

PercentBegin (Facultatif)

Utilisez `percentBegin` pour indiquer où les données pour la source de données commencent. Si vous n'incluez pas `percentBegin` et`percentEnd`, Amazon ML inclut toutes les données lors de la création de la source de données.

Les valeurs valides vont de `0` à `100`, bornes incluses.

PercentEnd (Facultatif)

Utilisez `percentEnd` pour indiquer où les données pour la source de données finissent. Si vous n'incluez pas `percentBegin` et`percentEnd`, Amazon ML inclut toutes les données lors de la création de la source de données.

Les valeurs valides vont de `0` à `100`, bornes incluses.

Complement (facultatif)

Le `complement` paramètre indique à Amazon ML d'utiliser les données qui ne sont pas incluses dans la plage de `percentBegin` `percentEnd` to pour créer une source de données. Le paramètre `complement` est utile si vous avez besoin de créer des sources de données complémentaires pour la formation et l'évaluation. Pour créer une source de données complémentaire, utilisez les mêmes valeurs pour `percentBegin` et `percentEnd`, ainsi que le paramètre `complement`.

Par exemple, les deux sources de données suivantes ne partagent aucune donnée, et peuvent être utilisées pour former et évaluer un modèle. La première source de données comporte 25 % des données, et la seconde 75 % des données.

Source de données pour l'évaluation :

```
{
    "splitting":{
        "percentBegin":0,
        "percentEnd":25
    }
}
```

Source de données pour la formation :

```
{
    "splitting":{
        "percentBegin":0,
        "percentEnd":25,
        "complement":"true"
    }
}
```

Les valeurs valides sont `true` et `false`.

Strategy (facultatif)

Pour modifier la façon dont Amazon ML divise les données d'une source de données, utilisez le `strategy` paramètre.

La valeur par défaut du `strategy` paramètre est`sequential`, ce qui signifie qu'Amazon ML prend tous les enregistrements de données compris entre les `percentEnd` paramètres `percentBegin` et de la source de données, dans l'ordre dans lequel les enregistrements apparaissent dans les données d'entrée

Les deux lignes `DataRearrangement` suivantes sont des exemples de sources de données de formation et d'évaluation ordonnées de manière séquentielle :

Source de données pour l'évaluation : `{"splitting":{"percentBegin":70, "percentEnd":100, "strategy":"sequential"}}`

Source de données pour la formation : `{"splitting":{"percentBegin":70, "percentEnd":100, "strategy":"sequential", "complement":"true"}}`

Pour créer une source de données à partir d'une sélection aléatoire des données, affectez au paramètre `strategy` la valeur `random` et fournissez une chaîne qui est utilisée comme valeur d'amorçage pour le fractionnement des données aléatoires (par exemple, vous pouvez utiliser le chemin d'accès S3 à vos données comme chaîne d'amorçage aléatoire). Si vous choisissez

la stratégie de répartition aléatoire, Amazon ML attribue à chaque ligne de données un nombre pseudo-aléatoire, puis sélectionne les lignes dont le numéro est compris entre `percentBegin` et. `percentEnd` Les nombres pseudo-aléatoires sont attribués à l'aide du décalage d'octets en tant qu'amorçage, si bien que la modification des données entraîne un fractionnement différent. Tout ordre préexistant est préservé. La stratégie de fractionnement aléatoire garantit que les variables figurant dans les données de formation et d'évaluation seront distribuées de façon similaire. Elle est utile dans les cas où les données d'entrée peuvent avoir un ordre de tri implicite, qui conduirait autrement à ce que les sources de données de formation et d'évaluation contiennent des enregistrements de données non similaires.

Les deux lignes `DataRearrangement` suivantes sont des exemples de sources de données de formation et d'évaluation ordonnées de manière non séquentielle :

Source de données pour l'évaluation :

```
{
    "splitting":{
        "percentBegin":70,
        "percentEnd":100,
        "strategy":"random",
        "strategyParams": {
            "randomSeed":"RANDOMSEED"
        }
    }
}
```

Source de données pour la formation :

```
{
    "splitting":{
        "percentBegin":70,
        "percentEnd":100,
        "strategy":"random",
        "strategyParams": {
            "randomSeed":"RANDOMSEED"
        }
        "complement":"true"
    }
}
```

Les valeurs valides sont `sequential` et `random`.

(Facultatif) Stratégie : RandomSeed

Amazon ML utilise le RandomSeed pour diviser les données. L'amorce par défaut pour l'API est une chaîne vide. Pour spécifier une amorce pour la stratégie de fractionnement aléatoire, fournissez une chaîne. Pour plus d'informations sur les valeurs de départ aléatoires, consultez Fractionnement aléatoire des données le manuel Amazon Machine Learning Developer Guide.

Pour obtenir un exemple de code expliquant comment utiliser la validation croisée avec Amazon ML, rendez-vous sur Github Machine Learning Samples.

Evaluation des modèles d'apprentissage-machine

Vous devez toujours évaluer un modèle pour déterminer s'il contribuera à prédire correctement la cible dans le cadre de nouvelles données à venir. Comme les instances futures ont des valeurs cibles inconnues, vous devez vérifier la métrique de précision du modèle d'apprentissage-machine sur des données dont vous connaissez déjà la réponse cible, puis utiliser cette évaluation comme indicateur de la précision prédictive des données futures.

Pour évaluer correctement un modèle, vous disposez d'un échantillon des données qui ont été étiquetées avec la cible (vérité du terrain) à partir de la source de données de formation. L'évaluation de la précision prédictive d'un modèle d'apprentissage-machine avec les mêmes données qui ont été utilisées pour la formation n'est pas utile. En effet, elle récompense les modèles qui peuvent « mémoriser » les données de formation, par opposition à une généralisation à partir de celles-ci. Une fois que vous avez terminé la formation du modèle d'apprentissage-machine, vous envoyez à ce modèle les observations mises de côté dont vous connaissez les valeurs cibles. Vous comparez alors les prédictions renvoyées par le modèle d'apprentissage-machine aux valeurs cibles connues. Enfin, vous calculez une métrique récapitulative indiquant la qualité de correspondance entre les valeurs prévues et les valeurs réelles.

Dans Amazon ML, vous évaluez un modèle de ML en créant une évaluation. Pour créer une évaluation pour un modèle d'apprentissage-machine, vous avez besoin d'un modèle d'apprentissage-machine à évaluer et de données étiquetées qui n'ont pas été utilisées pour la formation. Tout d'abord, créez une source de données à des fins d'évaluation en créant une source de données Amazon ML avec les données conservées. Les données utilisées dans l'évaluation doivent avoir le même schéma que les données utilisées dans la formation et inclure des valeurs réelles pour la variable cible.

Si toutes vos données se trouvent dans un seul fichier ou répertoire, vous pouvez utiliser la console Amazon ML pour les diviser. Le chemin par défaut dans l'assistant de création de modèle d'apprentissage-machine fractionne la source de données d'entrée et utilise les premiers 70 % comme source de données de formation et les autres 30 % comme source de données d'évaluation. Vous pouvez également personnaliser le rapport de fractionnement en utilisant l'option Personnalisé dans l'assistant de création de modèle d'apprentissage-machine. Vous pouvez alors choisir de sélectionner un échantillon aléatoire de 70 % pour la formation et d'utiliser les 30 % restants pour l'évaluation. Pour continuer à spécifier des rapports de fractionnement personnalisés, utilisez la chaîne de réorganisation des données dans l'API de création d'une source de données. Lorsque vous

disposez d'une source d'évaluation et d'un modèle d'apprentissage-machine, vous pouvez créer une évaluation et passer en revue les résultats de cette évaluation.

Rubriques

Analyse du modèle d'apprentissage-machine

Lorsque vous évaluez un modèle d'apprentissage-machine, Amazon ML fournit une métrique conforme aux normes du secteur et un certain nombre d'informations pour vérifier la précision prédictive de votre modèle. Dans Amazon ML, le résultat d'une évaluation contient les éléments suivants :

- Une métrique de précision des prédictions pour établir un rapport sur la réussite globale du modèle
- Des visualisations pour vous aider à étudier la précision de votre modèle au-delà de la métrique de précision de prédiction
- La possibilité de passer en revue l'impact de la configuration d'un score seuil (uniquement pour la classification binaire)
- Des alertes sur les critères permettant de vérifier la validité de l'évaluation

Le choix de la métrique et de la visualisation dépend du type de modèle d'apprentissage-machine que vous évaluez. Il est important de passer en revue ces visualisations pour déterminer si votre modèle est suffisamment performant pour répondre aux besoins de votre entreprise.

Analyse du modèle binaire

Interprétation des prédictions

La sortie réelle de nombreux algorithmes de classification binaire est un score de prédiction. Ce score indique la certitude du système que l'observation donnée appartient à la classe des positifs (la valeur cible réelle est 1). Les modèles de classification binaire d'Amazon ML génèrent un score compris entre 0 et 1. En tant que consommateur de ce score, pour décider si l'observation doit être classée comme 1 ou 0, vous interprétez le score en sélectionnant une limite de classification, ou seuil, et comparez le score à ce seuil. Toutes les observations avec des scores supérieurs au seuil sont prédites comme cible = 1, et les scores inférieurs au seuil sont prédits comme cible = 0.

Dans Amazon ML, le seuil de score par défaut est de 0,5. Vous pouvez choisir de mettre à jour cette limite en fonction des besoins de votre entreprise. Vous pouvez utiliser les visualisations disponibles dans la console pour comprendre comment le choix du seuil affectera votre application.

Mesure de la précision du modèle d'apprentissage-machine

Amazon ML fournit une métrique de précision standard pour les modèles de classification binaire appelée Area Under the (Receiver Operating Characteristic) Curve (AUC). La métrique AUC mesure l'aptitude du modèle à prédire un score plus élevé pour les exemples positifs par rapport aux exemples négatifs. Comme cela est indépendant du score seuil, vous pouvez vous faire une idée de la précision des prédictions de votre modèle à partir de la métrique AUC, sans choisir de seuil.

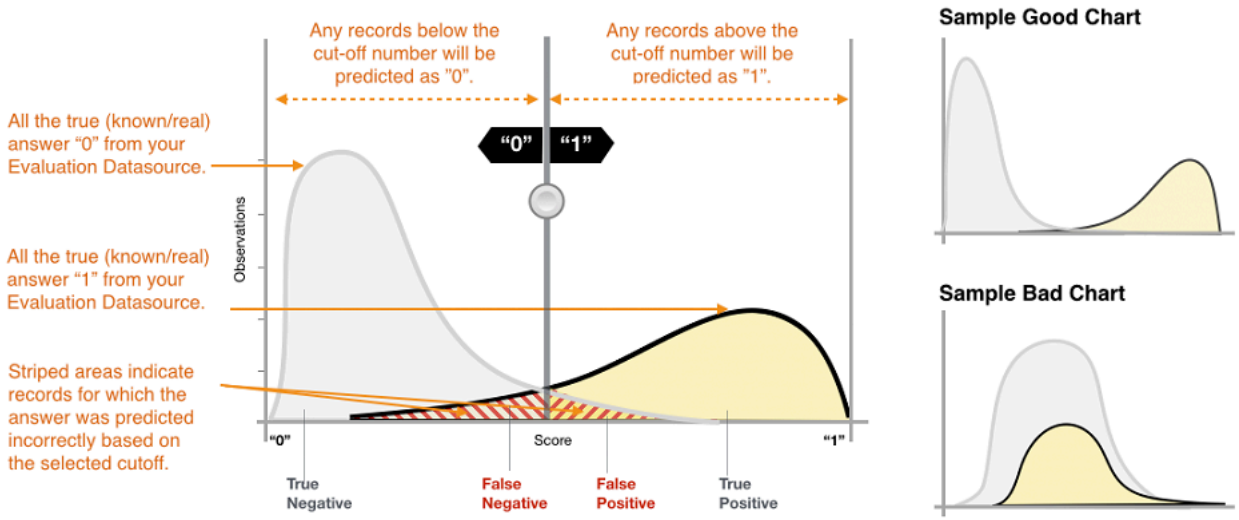
Elle renvoie une valeur décimale comprise entre 0 et 1. Les valeurs AUC proches de 1 indiquent un modèle ML qui est très précis. Les valeurs proches de 0,5 indiquent un modèle d'apprentissage-machine qui n'est pas meilleur que de deviner au hasard. Les valeurs proches de 0 sont inhabituelles et indiquent généralement un problème avec les données. Fondamentalement, une valeur AUC proche de 0 indique que le modèle d'apprentissage-machine a appris les bonnes tendances, mais les utilise pour effectuer des prédictions inversées par rapport à la réalité (les « 0 » sont prédits comme des « 1 » et vice versa). Pour plus d'informations sur la métrique AUC, accédez à la page Courbe ROC sur Wikipédia.

La métrique AUC de base pour un modèle binaire est 0,5. C'est la valeur correspondant à un modèle d'apprentissage-machine hypothétique qui prédit de façon aléatoire une réponse égale à 1 ou à 0. Votre modèle d'apprentissage-machine binaire doit avoir de meilleurs résultats que cette valeur pour commencer à être intéressant.

Utilisation de la visualisation des performances

Pour découvrir la précision du modèle ML, vous pouvez consulter les graphiques sur la page d'évaluation de la console Amazon ML. Cette page montre deux histogrammes : a) un histogramme des scores pour les positifs observés (la cible est 1) et b) un histogramme des scores pour les négatifs observés (la cible est 0) dans les données d'évaluation.

Un modèle d'apprentissage-machine qui a une bonne précision prédictive prédira de meilleurs scores pour les « 1 » observés et des scores inférieurs pour les « 0 » observés. Un modèle parfait aura ces deux histogrammes aux deux extrémités de l'axe des x, montrant que les positifs observés ont tous eu des scores élevés et que les négatifs observés ont tous eu des scores bas. Cependant, les modèles d'apprentissage-machine commettent des erreurs et un graphique typique indiquera que les deux histogrammes se chevauchent pour certains scores. Un modèle extrêmement inefficace ne pourra pas distinguer entre les classes des positifs et des négatifs, et les deux classes auront des histogrammes majoritairement superposés.



Grâce aux visualisations, vous pouvez identifier le nombre de prédictions qui appartiennent aux deux types de prédictions correctes et aux deux types de prédictions erronées.

Prédictions correctes

- Vrai positif (TP) : Amazon ML a prédit la valeur comme 1, et la vraie valeur est 1.
- Vrai négatif (TN) : Amazon ML a prédit la valeur comme 0, et la vraie valeur est 0.

Prédictions erronées

- Faux positif (FP) : Amazon ML a prédit la valeur comme 1, mais la vraie valeur est 0.

- Faux négatif (FN) : Amazon ML a prédit la valeur comme 0, mais la vraie valeur est 1.

> **Note**
>
> Le nombre de VP, VN, FP et FN dépend du score seuil sélectionné, et l'optimisation de l'un quelconque de ces nombres s'effectue aux dépens des autres. Un nombre élevé de TPs se traduit généralement par un nombre élevé FPs et un faible nombre de TNs.

Ajustement du score seuil

Les modèles d'apprentissage-machine fonctionnent en générant des scores de prédiction numériques, puis en appliquant un seuil pour convertir ces scores en étiquettes 0/1 binaires. En changeant le score seuil, vous pouvez ajuster le comportement du modèle lorsqu'il effectue une erreur. Sur la page d'évaluation de la console Amazon ML, vous pouvez examiner l'impact des différents seuils de score, et vous pouvez enregistrer le seuil de score que vous souhaitez utiliser pour votre modèle.

Lorsque vous ajustez le score seuil, observez le compromis que cela représente entre les deux types d'erreurs. Déplacer le seuil vers la gauche capture plus de vrais positifs, mais la contre-partie est une augmentation du nombre d'erreurs de faux positifs. Le déplacer vers la droite capture moins d'erreurs de faux positifs, mais la contre-partie est de manquer certains vrais positifs. Pour votre application prédictive, vous décidez du type d'erreur qui est plus tolérable en sélectionnant un score seuil approprié.

Revue des métriques avancées

Amazon ML fournit les mesures supplémentaires suivantes pour mesurer la précision prédictive du modèle de ML : exactitude, précision, rappel et taux de faux positifs.

Précision

La précision (ACC) mesure la fraction de prédictions correctes. La plage est comprise entre 0 et 1. Plus la valeur est grande et meilleure est la précision prédictive :

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Précision

Le taux de positifs prédits mesure la fraction de positifs observés parmi les exemples prédits comme positifs. La plage est comprise entre 0 et 1. Plus la valeur est grande et meilleure est la précision prédictive :

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Rappel

La sensibilité mesure la fraction de positifs observés qui sont prédits comme positifs. La plage est comprise entre 0 et 1. Plus la valeur est grande et meilleure est la précision prédictive :

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Taux de faux positifs

Le taux de faux positifs (TFP) mesure le taux de fausses alertes ou la fraction de négatifs observés qui sont prédits comme positifs. La plage est comprise entre 0 et 1. Plus la valeur est petite et meilleure est la précision prédictive :

$$FPR = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

En fonction de votre problème, vous pouvez être plus intéressé par un modèle performant pour un sous-ensemble spécifique de ces métriques. Par exemple, deux applications métier peuvent avoir des exigences très différentes en matière de modèle d'apprentissage-machine :

- Une application peut avoir besoin d'être extrêmement certaine que les prédictions positives soient effectivement positives (taux de positifs prédits élevé) et peut se permettre de mal classer certains exemples positifs comme négatifs (sensibilité moyenne).
- Une autre application peut avoir besoin de prédire correctement autant d'exemples positifs que possible (haute sensibilité) et acceptera que certains exemples négatifs soient mal classés comme positifs (taux de positifs prédits moyen).

Amazon ML vous permet de choisir un seuil de score correspondant à une valeur particulière de l'un des indicateurs avancés précédents. Il montre également les compromis induits par l'optimisation d'une métrique quelconque. Par exemple, si vous sélectionnez un seuil qui correspond à un taux de positifs prédits élevé, vous devez généralement accepter une baisse de la sensibilité en contre-partie.

> **Note**
>
> Vous devez enregistrer le score seuil pour qu'il prenne effet lors de la classification de futures prédictions par votre modèle d'apprentissage-machine.

Analyse du modèle multiclasse

Interprétation des prédictions

La sortie réelle d'un algorithme de classification multiclasse est un ensemble de scores de prédiction. Les scores indiquent la certitude du modèle que l'observation donnée appartient à chacune des classes. Contrairement aux problèmes de classification binaire, vous n'avez pas besoin de choisir un score seuil pour effectuer des prédictions. La réponse prédite est la classe (par exemple, une étiquette) avec le score prédit le plus élevé.

Mesure de la précision du modèle d'apprentissage-machine

Les métriques standard utilisées en mode multiclasse sont les mêmes que celles utilisées dans le cas d'une classification binaire après avoir calculé leur moyenne sur l'ensemble des classes. Dans Amazon ML, le score F1 macromoyen est utilisé pour évaluer la précision prédictive d'une métrique multiclasse.

Score F1 moyenné par macro

Le score F1 est une métrique de classification binaire qui prend en compte à la fois la sensibilité et le taux de positifs prédits des métriques binaires. Il s'agit de la moyenne harmonique entre la sensibilité et le taux de positifs prédits. La plage est comprise entre 0 et 1. Plus la valeur est grande et meilleure est la précision prédictive :

$$F1\ score = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

Le score F1 moyenné par macro correspond à la moyenne non pondérée du score F1 sur l'ensemble des classes du cas multiclasse. Il ne prend pas en compte la fréquence d'apparition des classes dans le jeu de données d'évaluation. Plus la valeur est grande et meilleure est la précision prédictive. L'exemple suivant montre K classes dans la source de données d'évaluation :

$$Macro\ average\ F1\ score = \frac{1}{K}\sum_{k=1}^{K} F1\ score\ for\ class\ k$$

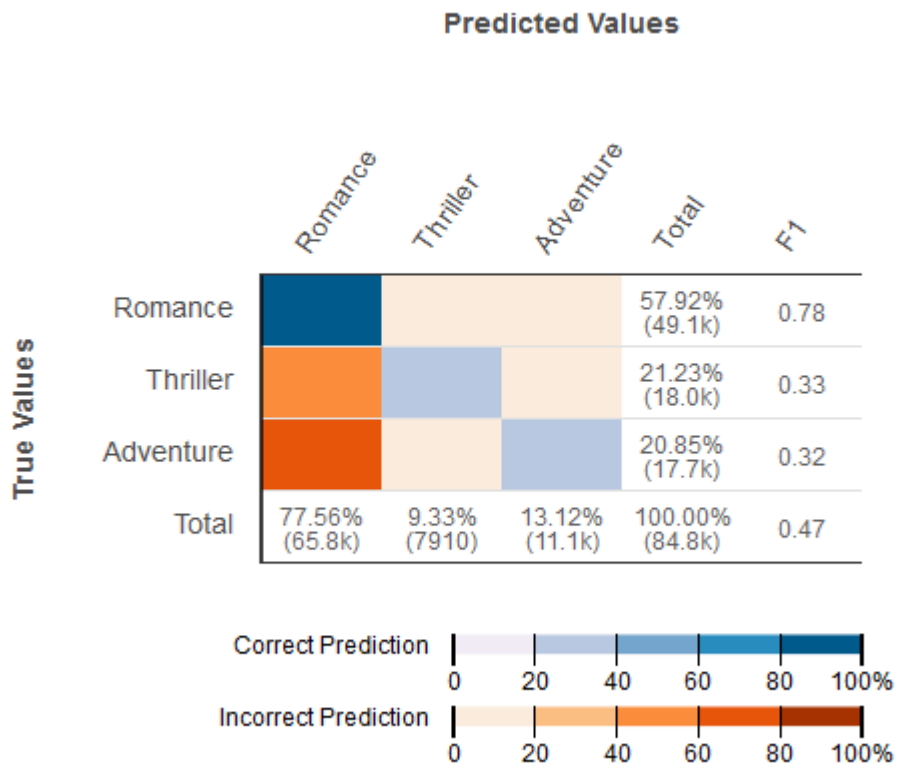
Score F1 moyenné par macro de référence

Amazon ML fournit une métrique de référence pour les modèles multiclasses. Il s'agit du score F1 moyenné par macro pour un modèle multiclasse hypothétique qui prédirait toujours la classe la plus fréquente comme réponse. Par exemple, dans le cadre de la prédiction du genre d'un film, si le genre le plus courant figurant dans vos données de formation était Film romantique, alors le modèle de référence prédirait toujours le genre comme Film romantique. Vous pouvez comparer votre modèle d'apprentissage-machine à cette référence afin de valider si votre modèle d'apprentissage-machine est meilleur qu'un modèle d'apprentissage-machine qui prédit cette réponse invariable.

Utilisation de la visualisation des performances

Amazon ML fournit une matrice de confusion afin de visualiser la précision des modèles prédictifs de classification multiclasses. La matrice de confusion illustre dans une table le nombre ou le pourcentage de prédictions correctes et incorrectes pour chaque classe en comparant la classe prédite d'une observation à sa véritable classe.

Par exemple, si vous essayez de classer un film par genre, le modèle prédictif peut prédire que son genre (sa classe) est Film romantique. Toutefois, son véritable genre pourrait en fait être Thriller. Lorsque vous évaluez la précision d'un modèle ML de classification multiclasse, Amazon ML identifie ces erreurs de classification et affiche les résultats dans la matrice de confusion, comme indiqué dans l'illustration suivante.



Les informations suivantes sont affichées dans une matrice de confusion :

- Nombre de prédictions correctes et incorrectes pour chaque classe : chaque ligne de la matrice de confusion correspond aux métriques pour l'une des véritables classes. Par exemple, la première ligne indique que pour les films qui appartiennent réellement au genre Film romantique, le modèle d'apprentissage-machine multiclasse aboutit à des prédictions exactes dans plus de 80 % des cas. Il prédit de façon erronée le genre comme Thriller dans moins de 20 % des cas, et Aventure dans moins de 20 % des cas.
- Score F1 au niveau de la classe : la dernière colonne indique le score F1 pour chacune des classes.
- Fréquences de classe véritables dans les données d'évaluation : l'avant-dernière colonne montre cela dans le jeu de données d'évaluation, 57,92 % des observations dans les données d'évaluation correspondent à Film romantique, 21,23 % à Thriller et 20,85 % à Aventure.
- Fréquences de classe prévues pour les données d'évaluation : La dernière ligne indique la fréquence de chaque classe dans les prédictions. 77,56 % des observations sont prédites comme Romance, 9,33 % sont prédites comme Thriller et 13,12 % sont prédites comme Adventure.

La console Amazon ML fournit un affichage visuel qui prend en charge jusqu'à 10 classes dans la matrice de confusion, répertoriées par ordre de classe la plus fréquente à la moins fréquente dans les données d'évaluation. Si vos données d'évaluation comportent plus de 10 classes, les 9 classes les plus fréquentes apparaîtront dans la matrice de confusion, et toutes les autres classes seront regroupées dans une classe appelée « autres ». Amazon ML permet également de télécharger la matrice de confusion complète via un lien sur la page des visualisations multiclasses.

Analyse du modèle de régression

Interprétation des prédictions

La sortie d'un modèle d'apprentissage-machine de régression est une valeur numérique pour la prédiction de la cible du modèle. Par exemple, si vous effectuez une prédiction des prix de logements, la prédiction du modèle peut être une valeur telle que 254 013.

> **Note**
>
> La plage des prédictions peut différer de la plage de la cible dans les données de formation. Par exemple, supposons que vous effectuez des prédictions de prix de logements et que la cible dans les données de formation avait des valeurs comprises dans une plage de 0 à 450 000. La cible prédite n'est pas nécessairement dans la même plage et peut prendre n'importe quelle valeur positive (supérieure à 450 000) ou négative (inférieure à zéro). Il est important de planifier comment gérer les valeurs de prédiction qui sortent d'une plage acceptable pour votre application.

Mesure de la précision du modèle d'apprentissage-machine

Pour les tâches de régression, Amazon ML utilise la métrique d'erreur quadratique moyenne (RMSE, Root Mean Square Error) qui est un standard de l'industrie. Il s'agit d'une mesure de distance entre la cible numérique prédite et la réponse numérique réelle (vérité du terrain). Plus la valeur de la métrique RMSE est petite, meilleure est la précision du modèle. Un modèle avec des prédictions parfaitement correctes aurait une métrique RMSE de 0. L'exemple suivant montre des données d'évaluation qui contiennent N enregistrements :

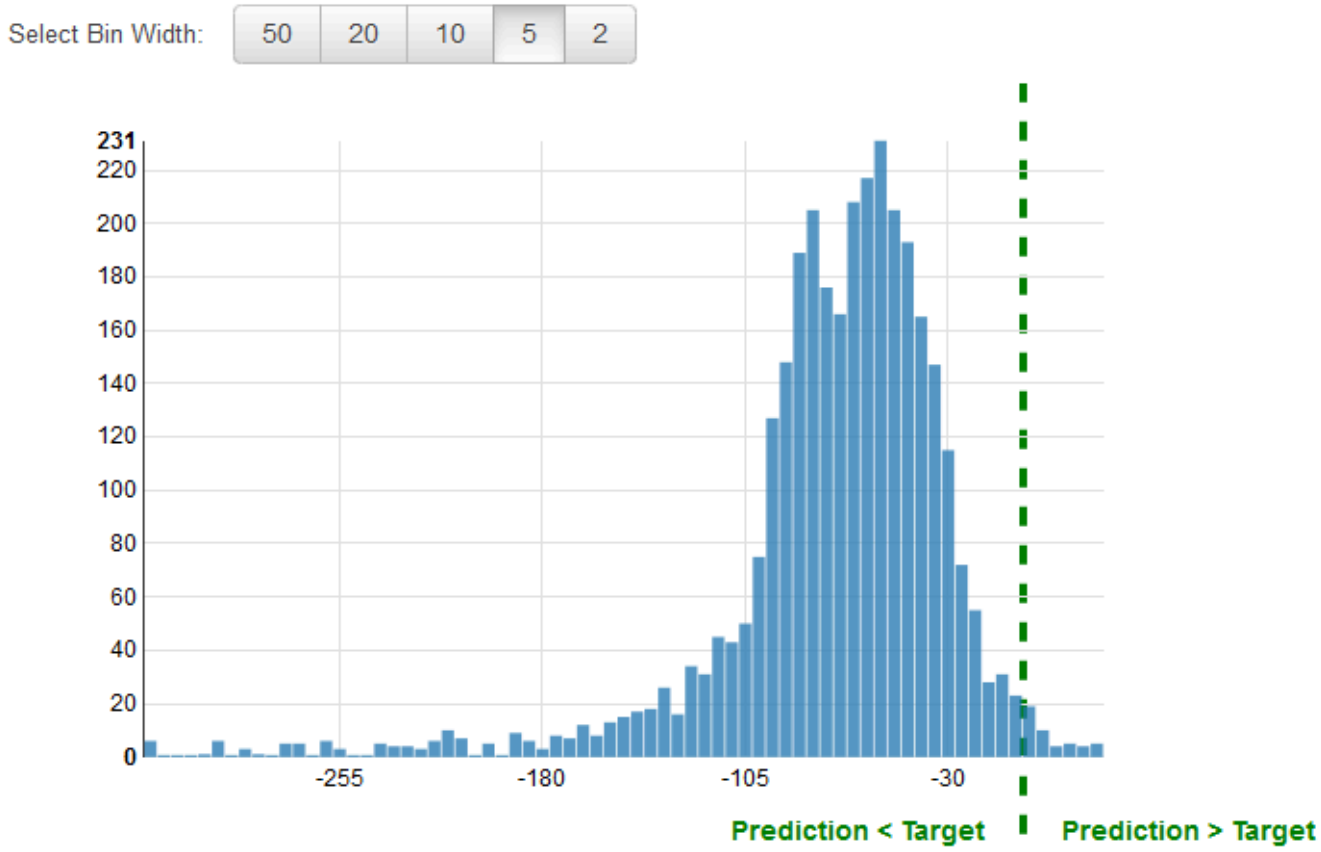
$$RMSE = \sqrt{1/N \sum_{i=1}^{N} (actual\ target - predicted\ target)^2}$$

RMSE de référence

Amazon ML fournit une métrique de référence pour les modèles de régression. Il s'agit de la métrique RMSE pour un modèle de régression hypothétique qui prédirait toujours la moyenne de la cible comme réponse. Par exemple, si vous essayez de prédire l'âge d'un acheteur de maison et que l'âge moyen pour les observations dans vos données de formation est 35, le modèle de référence prédira toujours la réponse 35. Vous pouvez comparer votre modèle d'apprentissage-machine à cette référence afin de valider si votre modèle d'apprentissage-machine est meilleur qu'un modèle d'apprentissage-machine qui prédit cette réponse invariable.

Utilisation de la visualisation des performances

Il est usuel de passer en revue les résidus pour identifier les problèmes de régression éventuels. Un résidu pour une observation dans les données d'évaluation représente la différence entre la cible réelle et la cible prédite. Les résidus représentent la partie de la cible que le modèle n'est pas en mesure de prédire. Un résidu positif indique que le modèle sous-estime la cible (la cible réelle est supérieure à la cible prédite). Un résidu négatif indique une surestimation (la cible réelle est inférieure à la cible prédite). L'histogramme des résidus sur les données d'évaluation lors d'une distribution en forme de cloche centrée sur zéro indique que le modèle commet des erreurs d'une manière aléatoire et qu'il ne prédit pas systématiquement trop haut ou trop bas une plage particulière de valeurs cibles. Si les résidus ne constituent pas une forme en cloche centrée sur zéro, il y a une certaine structure dans l'erreur de prédiction du modèle. L'ajout de variables supplémentaires au modèle peut aider le modèle à capturer la tendance qui n'est pas capturée par le modèle actuel. L'illustration suivante montre des résidus qui ne sont pas centrés sur zéro.



Prévention d'un surajustement

Lors de la création et de la formation d'un modèle d'apprentissage-machine, l'objectif est de sélectionner le modèle qui réalise les meilleures prédictions, ce qui revient à sélectionner le modèle doté des meilleurs paramètres (paramètres ou hyper-paramètres de modèle d'apprentissage-machine). Dans Amazon Machine Learning, vous pouvez définir quatre hyperparamètres : le nombre de passes, la régularisation, la taille du modèle et le type de shuffle. Toutefois, si vous sélectionnez les paramètres de modèle qui produisent les « meilleures » performances prédictives sur les données d'évaluation, vous pouvez surajuster votre modèle. Un surajustement se produit lorsqu'un modèle a mémorisé les tendances qui apparaissent dans les sources de données de formation et d'évaluation, mais n'a pas réussi à généraliser ces tendances dans les données. Il se produit souvent lorsque les données de formation incluent toutes les données utilisées dans l'évaluation. Un modèle surajusté présente de bons résultats pendant les évaluations, mais ne permet pas des prédictions précises sur des données inconnues.

Pour éviter de sélectionner un modèle surajusté comme meilleur modèle, vous pouvez réserver des données supplémentaires pour valider les performances du modèle d'apprentissage-machine. Par

exemple, vous pouvez séparer vos données en en dédiant 60 % à la formation, 20 % à l'évaluation et encore 20 % à la validation. Après avoir sélectionné les paramètres de modèle qui fonctionnent bien pour les données d'évaluation, vous effectuez une seconde évaluation avec les données de validation pour voir l'efficacité du modèle d'apprentissage-machine sur les données de validation. Si le modèle répond à vos attentes sur les données de validation, cela signifie que le modèle ne surajuste pas les données.

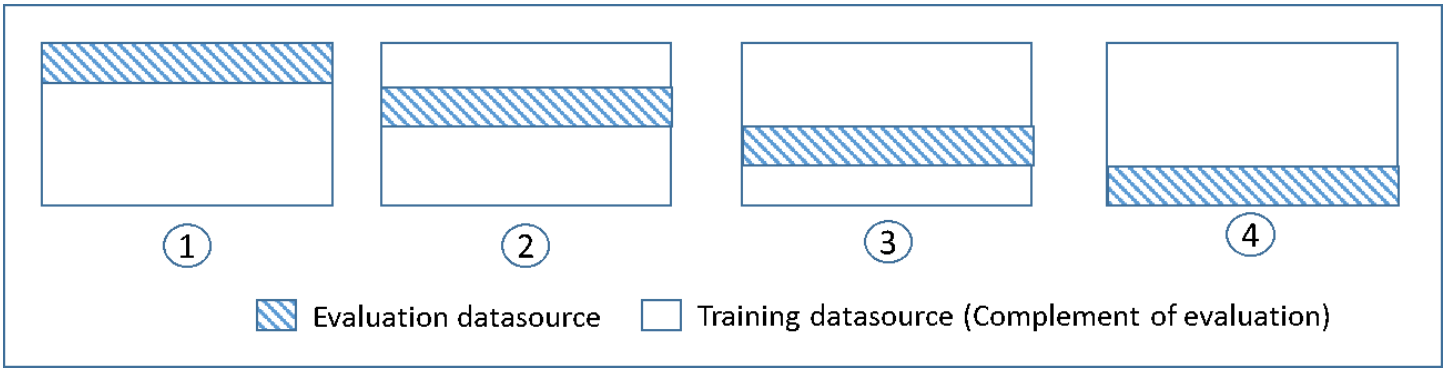
L'utilisation d'un troisième ensemble de données pour la validation vous aide à sélectionner les paramètres de modèle d'apprentissage-machine appropriés pour empêcher tout surajustement. Toutefois, mettre de côté des données du processus de formation pour effectuer l'évaluation et la validation réduit la quantité de données disponibles pour la formation. Ceci peut s'avérer particulièrement problématique avec de petits ensembles de données, car il est toujours préférable d'utiliser autant de données que possible pour la formation. Pour résoudre ce problème, vous pouvez effectuer une validation croisée. Pour plus d'informations sur la validation croisée, consultez Validation croisée.

Validation croisée

La validation croisée est une technique d'évaluation des modèles d'apprentissage-machine via la formation de plusieurs modèles d'apprentissage-machine sur des sous-ensembles des données d'entrée disponibles et via leur évaluation sur le sous-ensemble complémentaire des données. Utilisez la validation croisée pour détecter un surajustement, par exemple, l'échec de la généralisation d'une tendance.

Dans Amazon ML, vous pouvez utiliser la méthode de validation croisée k-fold pour effectuer une validation croisée. Dans le cadre de la validation croisée à k volets, vous divisez les données d'entrée en k sous-ensembles de données (également appelés plis). Vous entraînez un modèle ML sur tous les sous-ensembles sauf un (k-1), puis vous évaluez le modèle sur le sous-ensemble qui n'a pas été utilisé pour l'entraînement. Ce processus est répété k fois, avec un sous-ensemble différent réservé à l'évaluation (et exclu de la formation) à chaque fois.

Le diagramme suivant illustre un exemple des sous-ensembles de formation et des sous-ensembles d'évaluation complémentaire générés pour chacun des quatre modèles qui sont créés et formés au cours d'une validation croisée à 4 échantillons. Le modèle 1 utilise 25 % des données pour l'évaluation et les 75 % restants pour la formation. Le modèle 2 utilise le deuxième sous-ensemble de 25 % (de 25 % à 50 %) pour l'évaluation et les trois sous-ensembles restants



Chaque modèle est formé et évalué à l'aide de sources de données complémentaires - les données figurant dans la source de données d'évaluation incluent toutes les données qui ne se trouvent pas dans la source de données de formation, et sont limitées à cela. Vous créez des sources de données pour chacun de ces sous-ensembles avec le `DataRearrangement` paramètre dans le `createDatasourceFromS3` `createDatasourceFromRedShift`, et. `createDatasourceFromRDS` APIs Dans le paramètre `DataRearrangement`, spécifiez le sous-ensemble de données à inclure dans une source de données en spécifiant où commencer et finir chaque segment. Pour créer les sources de données complémentaires requises pour une validation croisée à 4 000 échantillons, spécifiez le paramètre `DataRearrangement` comme indiqué dans l'exemple suivant :

Modèle 1 :

Source de données pour l'évaluation :

```
{"splitting":{"percentBegin":0, "percentEnd":25}}
```

Source de données pour la formation :

```
{"splitting":{"percentBegin":0, "percentEnd":25, "complement":"true"}}
```

Modèle 2 :

Source de données pour l'évaluation :

```
{"splitting":{"percentBegin":25, "percentEnd":50}}
```

Source de données pour la formation :

```
{"splitting":{"percentBegin":25, "percentEnd":50, "complement":"true"}}
```

Modèle 3 :

Source de données pour l'évaluation :

```
{"splitting":{"percentBegin":50, "percentEnd":75}}
```

Source de données pour la formation :

```
{"splitting":{"percentBegin":50, "percentEnd":75, "complement":"true"}}
```

Modèle 4 :

Source de données pour l'évaluation :

```
{"splitting":{"percentBegin":75, "percentEnd":100}}
```

Source de données pour la formation :

```
{"splitting":{"percentBegin":75, "percentEnd":100, "complement":"true"}}
```

L'exécution d'une validation croisée en 4 étapes génère quatre modèles, quatre sources de données pour entraîner les modèles, quatre sources de données pour évaluer les modèles et quatre évaluations, une pour chaque modèle. Amazon ML génère une métrique de performance du modèle pour chaque évaluation. Par exemple, dans une validation croisée à 4 échantillons pour un problème de classification binaire, chacune des évaluations signale une métrique Aire sous une courbe (AUC, Area Under a Curve). Vous pouvez obtenir la mesure des performances globales en calculant la moyenne de ces quatre métriques AUC. Pour obtenir des informations sur la métrique AUC, consultez Mesure de la précision du modèle d'apprentissage-machine.

Pour un exemple de code expliquant comment créer une validation croisée et calculer la moyenne des scores du modèle, consultez l'exemple de code Amazon ML.

Ajustement de vos modèles

Après avoir effectué la validation croisée des modèles, vous pouvez ajuster les paramètres pour le modèle suivant si votre modèle ne fonctionne pas selon vos normes. Pour plus d'informations sur le surajustement, consultez Ajustement du modèle : sous-ajustement et surajustement. Pour

plus d'informations sur la régularisation, consultez Régularisation. Pour plus d'informations sur la modification des paramètres de régularisation, consultez Création d'un modèle d'apprentissage-machine avec des options personnalisées.

Alertes d'évaluation

Amazon ML fournit des informations qui vous aident à vérifier si vous avez correctement évalué le modèle. Si l'évaluation ne répond pas à l'un des critères de validation, la console Amazon ML vous avertit en affichant le critère de validation qui n'a pas été respecté, comme suit.

- L'évaluation du modèle d'apprentissage-machine s'effectue sur des données mises de côté

  Amazon ML vous avertit si vous utilisez la même source de données pour la formation et l'évaluation. Si vous utilisez Amazon ML pour fractionner vos données, vous répondrez à ce critère de validité. Si vous n'utilisez pas Amazon ML pour diviser vos données, assurez-vous d'évaluer votre modèle ML avec une source de données autre que la source de données d'entraînement.
- Des données suffisantes ont été utilisées pour l'évaluation du modèle prédictif

  Amazon ML vous avertit si le nombre d'observations/d'enregistrements dans vos données d'évaluation est inférieur à 10 % du nombre d'observations que vous avez dans votre source de données d'entraînement. Pour évaluer correctement votre modèle, il est important de fournir un échantillon de données suffisamment grand. Ce critère vous permet de savoir si vous utilisez trop peu de données. La quantité de données requise pour évaluer votre modèle de machine learning est subjective. 10 % sont sélectionnés ici comme solution provisoire en l'absence d'une meilleure mesure.
- Correspondance des schémas

  Amazon ML vous avertit si le schéma de la source de données de formation et d'évaluation n'est pas le même. Si certains attributs n'existent pas dans la source de données d'évaluation ou si vous avez des attributs supplémentaires, Amazon ML affiche cette alerte.
- Tous les enregistrements issus des fichiers d'évaluation ont été utilisés pour l'évaluation des performances du modèle prédictif

  Il est important de savoir si tous les enregistrements fournis pour l'évaluation ont réellement été utilisés pour évaluer le modèle. Amazon ML vous avertit si certains enregistrements de la source de données d'évaluation n'étaient pas valides et n'ont pas été inclus dans le calcul des mesures de précision. Par exemple, si la variable cible est absente pour certaines observations de la source de données d'évaluation, Amazon ML n'est pas en mesure de vérifier si les prédictions du modèle

ML pour ces observations sont correctes. Dans ce cas, les enregistrements avec des valeurs cibles manquantes sont considérés comme non valides.

- Distribution de la variable cible

  Amazon ML vous indique la distribution de l'attribut cible à partir des sources de données de formation et d'évaluation afin que vous puissiez vérifier si la cible est distribuée de la même manière dans les deux sources de données. Si le modèle a été formé sur des données de formation avec une distribution de la cible différente de la distribution de la cible sur les données d'évaluation, la qualité de l'évaluation pourrait en pâtir, car elle serait calculée sur des données dotées de statistiques très différentes. Il est recommandé d'avoir des données distribuées de manière similaire dans les données de formation et d'évaluation, et de faire en sorte que ces ensembles de données imitent autant que possible les données que le modèle rencontrera lorsqu'il réalisera des prédictions.

  Si cette alerte se déclenche, essayez d'utiliser la stratégie de fractionnement aléatoire pour diviser les données en sources de données de formation et d'évaluation. Dans de rares cas, cette alerte peut vous avertir par erreur des différences de distribution cible, même si vous répartissez vos données de manière aléatoire. Amazon ML utilise des statistiques de données approximatives pour évaluer les distributions de données, déclenchant parfois cette alerte par erreur.

Génération et interprétation des prédictions

Amazon ML fournit deux mécanismes pour générer des prédictions : asynchrone (par lots) et synchrone (). one-at-a-time

Utilisez les prédictions asynchrones, ou prédictions par lots, lorsque vous avez un certain nombre d'observations et que vous souhaitez obtenir des prédictions pour toutes les observations à la fois. Le processus utilise une source de données en entrée et fournit des prédictions en sortie, dans un fichier .csv stocké dans un compartiment S3 de votre choix. Vous devez attendre que le processus de prédiction par lots se termine avant de pouvoir accéder aux résultats de prédiction. La taille maximale d'une source de données qu'Amazon ML peut traiter dans un fichier batch est de 1 To (environ 100 millions d'enregistrements). Si votre source de données est supérieure à 1 To, votre tâche échouera et Amazon ML renverra un code d'erreur. Pour éviter cela, fractionnez vos données en plusieurs lots. Si vos enregistrements sont globalement plus longs, vous atteindrez la limite de 1 To avant de traiter 100 millions d'enregistrements. Dans ce cas, nous vous recommandons de contacter AWS support pour augmenter la taille de la tâche pour votre prédiction par lots.

Utilisez des prédictions en temps réel, synchrones, pour obtenir des prédictions à faible latence. L'API de prédiction en temps réel accepte une seule observation en entrée, sérialisée sous la forme d'une chaîne JSON, et renvoie de façon synchrone la prédiction et les métadonnées associées dans le cadre de la réponse de l'API. Vous pouvez appeler simultanément l'API plusieurs fois pour obtenir des prédictions synchrones en parallèle. Pour plus d'informations sur les limites de débit de l'API de prédiction en temps réel, reportez-vous aux limites de prédiction en temps réel dans Référence d'API Amazon ML.

Rubriques

Création d'une prédiction par lots

Pour créer une prédiction par lots, vous devez créer un `BatchPrediction` objet à l'aide de la console ou de l'API Amazon Machine Learning (Amazon ML). Un `BatchPrediction` objet décrit

un ensemble de prédictions générées par Amazon ML à l'aide de votre modèle ML et d'un ensemble d'observations d'entrée. Lorsque vous créez un `BatchPrediction` objet, Amazon ML lance un flux de travail asynchrone qui calcule les prédictions.

Vous devez utiliser le même schéma pour la source de données que vous utilisez pour obtenir des prédictions par lots et la source de données que vous avez utilisée pour former le modèle d'apprentissage-machine que vous interrogez pour obtenir des prédictions. La seule exception est que la source de données pour une prédiction par lots n'a pas besoin d'inclure l'attribut cible, car Amazon ML prédit la cible. Si vous fournissez l'attribut cible, Amazon ML ignore sa valeur.

Création d'une prédiction par lots (console)

Pour créer une prédiction par lots à l'aide de la console Amazon ML, utilisez l'assistant Create Batch Prediction.

Pour créer une prédiction par lots (console)

1. Connectez-vous à la console Amazon Machine Learning AWS Management Console et ouvrez-la à l'adresse https://console.aws.amazon.com/machinelearning/.
2. Sur le tableau de bord Amazon ML, sous Objets, choisissez Create new... , puis choisissez Prédiction par lots.
3. Choisissez le modèle Amazon ML que vous souhaitez utiliser pour créer la prédiction par lots.
4. Pour confirmer que vous souhaitez utiliser ce modèle, choisissez Continue.
5. Choisissez la source de données pour laquelle vous voulez créer les prédictions. La source de données doit avoir le même schéma que votre modèle, même si elle n'est pas tenue d'inclure l'attribut cible.
6. Choisissez Continuer.
7. Pour S3 destination, tapez le nom de votre compartiment S3.
8. Choisissez Examiner.
9. Passez en revue vos paramètres et choisissez Create batch prediction.

Création d'une prédiction par lots (API)

Pour créer un `BatchPrediction` objet à l'aide de l'API Amazon ML, vous devez fournir les paramètres suivants :

ID de source de données

ID de la source de données qui pointe sur les observations pour lesquelles vous souhaitez obtenir des prédictions. Par exemple, si vous voulez des prédictions pour les données d'un fichier nommé `s3://examplebucket/input.csv`, vous devez créer un objet source de données qui pointe sur le fichier de données, puis transmettre l'ID de cette source de données avec ce paramètre.

BatchPrediction ID

ID à attribuer à la prédiction par lots.

ID du modèle d'apprentissage-machine

L'ID du modèle ML qu'Amazon ML doit interroger pour les prédictions.

URI de sortie

L'URI du compartiment S3 dans lequel stocker le résultat de la prédiction. Amazon ML doit être autorisé à écrire des données dans ce compartiment.

Le paramètre `OutputUri` doit faire référence à un chemin S3 qui se termine par une barre oblique (« / »), comme indiqué dans l'exemple suivant :

s3://examplebucket/examplepath/

Pour obtenir des informations sur la configuration des autorisations S3, consultez Octroi d'autorisations à Amazon ML pour fournir en sortie des prédictions dans Amazon S3.

(Facultatif) BatchPrediction Nom

(Facultatif) Nom contrôlable de visu pour votre prédiction par lots.

Examen des métriques de prédiction par lots

Une fois qu'Amazon Machine Learning (Amazon ML) a créé une prédiction par lots, il fournit deux métriques `Records seen` : `Records failed to process` et. `Records seen`vous indique le nombre d'enregistrements examinés par Amazon ML lors de l'exécution de votre prédiction par lots. `Records failed to process`indique le nombre d'enregistrements qu'Amazon ML n'a pas pu traiter.

Pour permettre à Amazon ML de traiter les enregistrements ayant échoué, vérifiez le formatage des enregistrements dans les données utilisées pour créer votre source de données, et assurez-vous que

tous les attributs requis sont présents et que toutes les données sont correctes. Après avoir corrigé vos données, vous pouvez recréer votre prédiction par lots ou créer une nouvelle source de données avec les enregistrements ayant échoué, puis créer une nouvelle prédiction par lots à l'aide de la nouvelle source de données.

Examen des métriques de prédiction par lots (console)

Pour consulter les métriques dans la console Amazon ML, ouvrez la page de résumé des prédictions Batch et consultez la section Informations traitées.

Examen des détails et des métriques de prédiction par lots (API)

Vous pouvez utiliser Amazon ML APIs pour récupérer des informations sur les `BatchPrediction` objets, y compris les métriques d'enregistrement. Amazon ML fournit les appels d'API de prédiction par lots suivants :

- `CreateBatchPrediction`
- `UpdateBatchPrediction`
- `DeleteBatchPrediction`
- `GetBatchPrediction`
- `DescribeBatchPredictions`

Pour plus d'informations, consultez le manuel Amazon ML API Reference.

Lecture des fichiers de sortie de prédiction par lots

Effectuez les opérations suivantes pour récupérer les fichiers de sortie de prédiction par lots :

1. Localisez le fichier manifeste de prédiction par lots.
2. Lisez le fichier manifeste pour déterminer les emplacements des fichiers de sortie.
3. Récupérez les fichiers de sortie qui contiennent les prédictions.
4. Interprétez le contenu des fichiers de sortie. Le contenu varie en fonction du type de modèle d'apprentissage-machine qui a été utilisé pour générer les prédictions.

Les sections suivantes décrivent ces étapes de façon plus détaillée.

Localisation du fichier manifeste de prédiction par lots

Les fichiers manifestes de prédiction par lots contiennent les informations qui mettent en correspondance vos fichiers d'entrée avec les fichiers de sortie de prédiction.

Pour localiser le fichier manifeste, commencez avec l'emplacement de sortie que vous avez spécifié en créant l'objet de prédiction par lots. Vous pouvez interroger un objet de prédiction par lots terminé pour récupérer l'emplacement S3 de ce fichier à l'aide de l'API Amazon ML ou du https://console.aws.amazon.com/machinelearning/.

Le fichier manifeste est situé dans l'emplacement de sortie, avec un chemin constitué de la chaîne statique `/batch-prediction/` ajoutée à l'emplacement de sortie et du nom du fichier manifeste, qui correspond à l'ID de la prédiction par lots, avec l'extension `.manifest` ajoutée à cela.

Par exemple, si vous créez un objet de prédiction par lots avec l'ID `bp-example`, et que vous spécifiez l'emplacement S3 `s3://examplebucket/output/` comme emplacement de sortie, vous trouverez votre fichier manifeste ici :

`s3://examplebucket/output/batch-prediction/bp-example.manifest`

Lecture du fichier manifeste

Le contenu du fichier .manifest est codé sous la forme d'un mappage JSON, dans lequel la clé est une chaîne du nom d'un fichier de données d'entrée S3, et la valeur est une chaîne du fichier de résultats de prédiction par lots associé. Il y a une ligne de mappage pour chaque paire de fichiers d'entrée/sortie. En poursuivant notre exemple, si l'entrée pour la création de l'objet `BatchPrediction` se compose d'un fichier individuel nommé data.csv et situé dans `s3://examplebucket/input/`, vous pouvez voir une chaîne de mappage similaire à :

```
{"s3://examplebucket/input/data.csv":"
s3://examplebucket/output/batch-prediction/result/bp-example-data.csv.gz"}
```

Si l'entrée pour la création de l'objet `BatchPrediction` se compose de trois fichiers appelés data1.csv, data2.csv et data3.csv, et qu'ils sont tous stockés dans l'emplacement S3 `s3://examplebucket/input/`, vous pouvez voir une chaîne de mappage similaire à :

```
{"s3://examplebucket/input/data1.csv":"s3://examplebucket/output/batch-prediction/
result/bp-example-data1.csv.gz",
```

```
"s3://examplebucket/input/data2.csv":"
s3://examplebucket/output/batch-prediction/result/bp-example-data2.csv.gz",

"s3://examplebucket/input/data3.csv":"
s3://examplebucket/output/batch-prediction/result/bp-example-data3.csv.gz"}
```

Récupération des fichiers de sortie de prédiction par lots

Vous pouvez télécharger chaque fichier de prédiction par lots obtenu à partir du mappage de manifeste et le traiter localement. Le format de fichier est CSV, compressé avec l'algorithme gzip. Ce fichier comprend une ligne pour chaque observation en entrée dans le fichier d'entrée correspondant.

Pour joindre les prédictions au fichier d'entrée de la prédiction par lots, vous pouvez effectuer une simple record-by-record fusion des deux fichiers. Le fichier de sortie de la prédiction par lots contient toujours le même nombre d'enregistrements que le fichier d'entrée de prédiction, dans le même ordre. Si une observation en entrée échoue au cours du traitement et qu'aucune prédiction ne peut être générée, le fichier de sortie de la prédiction par lots aura une ligne vide dans l'emplacement correspondant.

Interprétation du contenu des fichiers de prédiction par lots pour un modèle d'apprentissage-machine de classification binaire

Les colonnes du fichier de prédiction par lots pour un modèle de classification binaire sont nommées bestAnswer et score.

La colonne bestAnswer contient l'étiquette de prédiction (« 1 » ou « 0 ») qui est obtenue en évaluant le score de prédiction par rapport au score seuil. Pour plus d'informations sur les scores limites, consultez Ajustement du score seuil. Vous définissez un score limite pour le modèle ML en utilisant l'API Amazon ML ou la fonctionnalité d'évaluation du modèle sur la console Amazon ML. Si vous ne définissez pas de score limite, Amazon ML utilise la valeur par défaut de 0,5.

La colonne de score contient le score de prédiction brut attribué par le modèle ML pour cette prédiction. Amazon ML utilise des modèles de régression logistique. Ce score tente donc de modéliser la probabilité que l'observation corresponde à une valeur vraie (« 1 »). Notez que le score est indiqué en notation scientifique ; dans la première ligne de l'exemple suivant, la valeur 8,7642E-3 est donc égale à 0,0087642.

Par exemple, si le score seuil pour le modèle d'apprentissage-machine est 0,75, le contenu du fichier de sortie de prédiction par lots pour un modèle de classification binaire peut ressembler à ce qui suit :

```
bestAnswer,score

0,8.7642E-3

1,7.899012E-1


0,6.323061E-3

0,2.143189E-2


1,8.944209E-1
```

Les deuxième et cinquième observations dans le fichier d'entrée ont reçu des scores de prédiction supérieurs à 0,75, de sorte que la colonne bestAnswer pour ces observations indique la valeur « 1 », tandis que les autres observations ont la valeur « 0 ».

Interprétation du contenu des fichiers de prédiction par lots pour un modèle d'apprentissage-machine de classification multiclasse

Le fichier de prédiction par lots pour un modèle multiclasse contient une colonne pour chaque classe figurant dans les données de formation. Les noms des colonnes apparaissent dans la ligne d'en-tête du fichier de prédiction par lots.

Lorsque vous demandez des prédictions à partir d'un modèle multiclasse, Amazon ML calcule plusieurs scores de prédiction pour chaque observation du fichier d'entrée, un pour chacune des classes définies dans le jeu de données d'entrée. Cela équivaut à demander : « Quelle est la probabilité (mesurée entre 0 et 1) pour que cette observation appartienne à cette classe, plutôt qu'à l'une des autres classes ? ». Chaque score peut être interprété comme une « probabilité pour que l'observation appartienne à cette classe ». Dans la mesure où les scores de prédiction modélisent les probabilités sous-jacentes pour que l'observation appartienne à une classe ou à une autre, la somme de tous les scores de prédiction sur une ligne est égale à 1. Vous devez choisir une classe en tant que classe prévue pour le modèle. Le plus souvent, vous devez choisir comme meilleure réponse la classe qui présente la plus haute probabilité.

Considérons par exemple une tentative visant à prédire l'évaluation d'un produit par un client sur une échelle de 1 à 5 étoiles. Si les classes sont nommées `1_star`, `2_stars`, `3_stars`, `4_stars` et `5_stars`, le fichier de sortie de prédiction multiclasse peut ressembler à ceci :

```
1_star, 2_stars, 3_stars, 4_stars, 5_stars

8.7642E-3, 2.7195E-1, 4.77781E-1, 1.75411E-1, 6.6094E-2

5.59931E-1, 3.10E-4, 2.48E-4, 1.99871E-1, 2.39640E-1

7.19022E-1, 7.366E-3, 1.95411E-1, 8.78E-4, 7.7323E-2

1.89813E-1, 2.18956E-1, 2.48910E-1, 2.26103E-1, 1.16218E-1

3.129E-3, 8.944209E-1, 3.902E-3, 7.2191E-2, 2.6357E-2
```

Dans cet exemple, la première observation possède le score de prédiction le plus élevé pour la classe `3_stars` (score de prédiction = 4,77781E-1) et vous interpréteriez les résultats comme indiquant que la classe `3_stars` est la meilleure réponse pour cette observation. Notez que les scores de prédiction sont indiqués en notation scientifique et qu'un score de prédiction de 4,77781E-1 est égal à 0,477781.

Dans certaines circonstances, il est possible que vous ne souhaitiez pas choisir la classe dotée de la plus haute probabilité. Par exemple, vous pouvez établir un seuil minimum sous lequel vous ne considérerez pas une classe comme étant la meilleure réponse, même si elle a le meilleur score de prédiction. Supposons que vous classez des films par genre et que vous voulez que le score de prédiction soit au moins 5E-1 avant de déclarer le genre comme votre meilleure réponse. Vous obtenez un score de prédiction de 3E-1 pour les comédies, de 2,5E-1 pour les drames, de 2,5E-1 pour les documentaires et de 2E-1 pour les films d'action. Dans ce cas, le modèle d'apprentissage-machine prédit que la comédie est votre choix le plus probable, mais vous décidez de ne pas la choisir comme meilleure réponse. Dans la mesure où aucun des scores de prédiction n'a dépassé votre score de prédiction de base de 5E-1, vous décidez que la prédiction est insuffisante pour prédire de manière fiable le genre et vous décidez de choisir quelque chose d'autre. Votre application peut alors traiter le champ genre de ce film comme « inconnu ».

Interprétation du contenu des fichiers de prédiction par lots pour un modèle d'apprentissage-machine de régression

Le fichier de prédiction par lots d'un modèle de régression contient une seule colonne nommée score. Cette colonne contient les prédictions numériques brutes de toutes les observations figurant dans les données d'entrée. Les valeurs sont indiquées en notation scientifique ; la valeur score de -1,526385E1 est donc égale à -15,26835 dans la première ligne de l'exemple suivant.

Cet exemple montre un fichier de sortie pour une prédiction par lots effectuée sur un modèle de régression :

```
score

-1.526385E1

-6.188034E0

-1.271108E1

-2.200578E1

8.359159E0
```

Demande de prédiction en temps réel

Une prédiction en temps réel est un appel synchrone à Amazon Machine Learning (Amazon ML). La prédiction est effectuée lorsque Amazon ML reçoit la demande, et la réponse est renvoyée immédiatement. Les prédictions en temps réel sont couramment utilisées pour fournir des fonctionnalités prédictives au sein d'applications interactives web, mobiles ou de bureau. Vous pouvez interroger un modèle ML créé avec Amazon ML pour obtenir des prédictions en temps réel à l'aide de l'`Predict`API à faible latence. L'opération `Predict` accepte une seule observation d'entrée dans la charge utile de demande et renvoie de façon synchrone la prédiction dans la réponse. Cela le distingue de l'API de prédiction par lots, qui est invoquée avec l'ID d'un objet de source de données Amazon ML qui pointe vers l'emplacement des observations en entrée, et renvoie de manière asynchrone un URI vers un fichier contenant des prédictions pour toutes ces observations. Amazon ML répond à la plupart des demandes de prédiction en temps réel dans un délai de 100 millisecondes.

Vous pouvez essayer des prédictions en temps réel sans frais dans la console Amazon ML. Si vous décidez ensuite d'utiliser les prédictions en temps réel, vous devez commencer par créer un point de terminaison pour la génération des prédictions en temps réel. Vous pouvez le faire dans la console Amazon ML ou à l'aide de l'`CreateRealtimeEndpoint`API. Une fois que vous avez un point de terminaison, utilisez l'API de prédiction en temps réel pour générer des prédictions en temps réel.

> **Note**
>
> Une fois que vous aurez créé un point de terminaison en temps réel pour votre modèle, vous commencerez à payer des frais de réservation de capacité en fonction de la taille du modèle. Pour plus d'informations, consultez Pricing (Tarification CTlong). Si vous créez le point de terminaison en temps réel dans la console, celle-ci affiche un détail des frais estimés que le point de terminaison comptabilisera de façon continue. Pour cesser d'engager des frais lorsque vous n'avez plus besoin d'obtenir de prédictions en temps réel à partir de ce modèle, supprimez le point de terminaison en temps réel à l'aide de la console ou de l'opération `DeleteRealtimeEndpoint`.

Pour des exemples de `Predict` demandes et de réponses, consultez Predict dans le manuel Amazon Machine Learning API Reference. Pour voir un exemple du format de réponse exact que votre modèle utilise, consultez Essai d'utilisation des prédictions en temps réel.

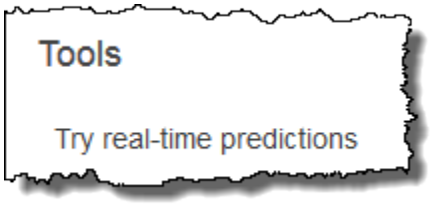
Rubriques

Essai d'utilisation des prédictions en temps réel

Pour vous aider à décider d'activer ou non les prédictions en temps réel, Amazon ML vous permet d'essayer de générer des prédictions sur des enregistrements de données uniques sans encourir les frais supplémentaires associés à la configuration d'un point de terminaison de prédiction en temps réel. Pour essayer les prédictions en temps réel, vous devez disposer d'un modèle d'apprentissage-machine. Pour créer des prédictions en temps réel à plus grande échelle, utilisez l'API Predict dans le manuel Amazon Machine Learning API Reference.

Pour essayer d'utiliser les prédictions en temps réel

1. Connectez-vous à la console Amazon Machine Learning AWS Management Console et ouvrez-la à l'adresse https://console.aws.amazon.com/machinelearning/.
2. Dans la barre de navigation, dans le menu déroulant Amazon Machine Learning, choisissez ML models.
3. Choisissez le modèle que vous souhaitez utiliser pour essayer les prédictions en temps réel, tel que le modèle `Subscription propensity model` issu du didacticiel.
4. Dans la page de rapport du modèle d'apprentissage-machine, sous Predictions, choisissez Summary, puis choisissez Try real-time predictions.



   Amazon ML affiche une liste des variables qui constituaient les enregistrements de données utilisés par Amazon ML pour entraîner votre modèle.
5. Vous pouvez continuer en entrant des données dans tous les champs du formulaire ou en collant un enregistrement de données unique, au format CSV, dans la zone de texte.

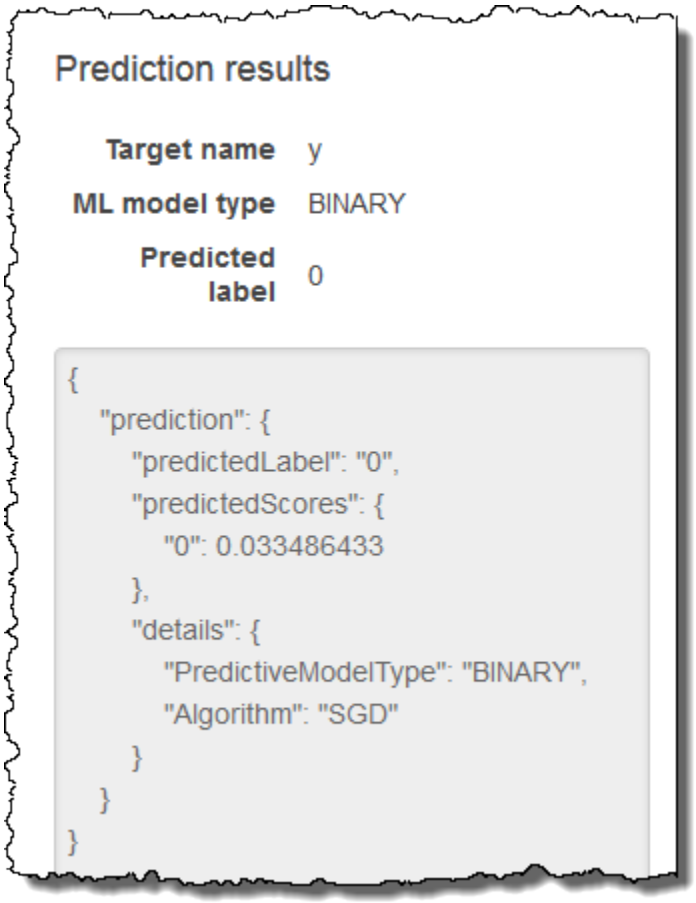
   Pour utiliser le formulaire, pour chaque champ Valeur, entrez les données que vous souhaitez utiliser pour tester vos prédictions en temps réel. Si l'enregistrement de données que vous entrez ne contient pas de valeurs pour un ou plusieurs attributs de données, laissez les champs d'entrée vides.

   Pour fournir un enregistrement de données, choisissez Paste a record. Collez une seule ligne de données au format CSV dans le champ de texte, puis choisissez Soumettre. Amazon ML remplit automatiquement les champs de valeur pour vous.

   > **Note**
   >
   > Les données de l'enregistrement de données doivent avoir le même nombre de colonnes que les données de formation, et être organisées dans le même ordre. La seule exception est que vous devez omettre la valeur cible. Si vous incluez une valeur cible, Amazon ML l'ignore.

6. En bas de la page, choisissez Create prediction. Amazon ML renvoie la prédiction immédiatement.

   Dans le volet Prediction results, vous voyez l'objet prédiction que l'appel à l'API `Predict` renvoie, ainsi que le type de modèle d'apprentissage-machine, le nom de la variable cible et la classe ou la valeur prédite. Pour obtenir des informations sur l'interprétation des résultats, consultez Interprétation du contenu des fichiers de prédiction par lots pour un modèle d'apprentissage-machine de classification binaire.



Création d'un point de terminaison en temps réel

Pour générer des prédictions en temps réel, vous devez créer un point de terminaison en temps réel. Pour créer un point de terminaison en temps réel, vous devez déjà avoir un modèle d'apprentissage-machine pour lequel vous souhaitez générer des prédictions en temps réel. Vous pouvez créer un point de terminaison en temps réel à l'aide de la console Amazon ML ou en appelant l'`CreateRealtimeEndpoint`API. Pour plus d'informations sur l'utilisation de l'`CreateRealtimeEndpoint`API, consultez https://docs.aws.amazon.com/machine-learning/

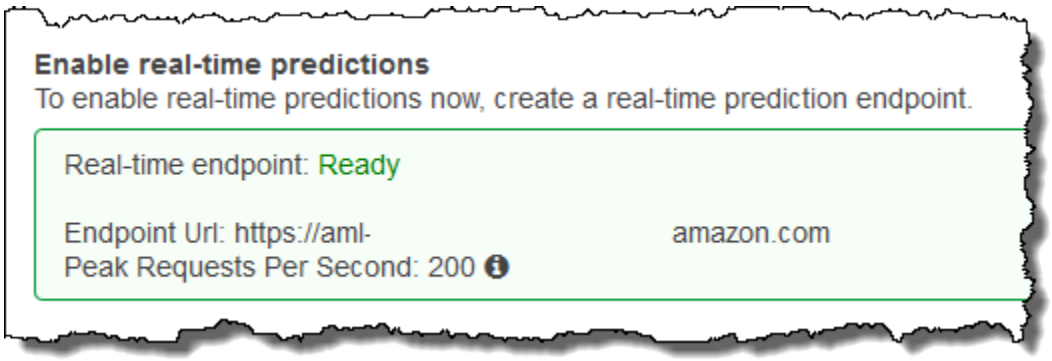
latest/APIReference/API_CreateRealtimeEndpoint.html le manuel Amazon Machine Learning API Reference.

Pour créer un point de terminaison en temps réel

1. Connectez-vous à la console Amazon Machine Learning AWS Management Console et ouvrez-la à l'adresse https://console.aws.amazon.com/machinelearning/.
2. Dans la barre de navigation, dans le menu déroulant Amazon Machine Learning, choisissez ML models.
3. Choisissez le modèle pour lequel vous souhaitez générer des prédictions en temps réel.
4. Dans la page ML model summary, sous Predictions, choisissez Create real-time endpoint.

   Une boîte de dialogue qui explique comment les prédictions en temps réel sont facturées apparaît.
5. Sélectionnez Create (Créer). La demande de point de terminaison en temps réel est envoyée à Amazon ML et entrée dans une file d'attente. L'état du point de terminaison en temps réel est Mise à jour en cours.

6. Lorsque le point de terminaison en temps réel est prêt, le statut passe à Prêt et Amazon ML affiche l'URL du point de terminaison. Utilisez l'URL du point de terminaison pour créer des demandes de prédictions en temps réel avec l'API `Predict`. Pour plus d'informations sur l'utilisation de l'`Predict`API, consultez https://docs.aws.amazon.com/machine-learning/latest/APIReference/API_Predict.html le manuel Amazon Machine Learning API Reference.



Localisation du point de terminaison de prédiction en temps réel (console)

Pour utiliser la console Amazon ML afin de trouver l'URL du point de terminaison d'un modèle de ML, accédez à la page de résumé du modèle de ML du modèle.

Pour localiser l'URL d'un point de terminaison en temps réel

1. Connectez-vous à la console Amazon Machine Learning AWS Management Console et ouvrez-la à l'adresse https://console.aws.amazon.com/machinelearning/.
2. Dans la barre de navigation, dans le menu déroulant Amazon Machine Learning, choisissez ML models.
3. Choisissez le modèle pour lequel vous souhaitez générer des prédictions en temps réel.
4. Dans la page ML model summary, faites défiler l'affichage vers le bas jusqu'à ce que vous voyiez la section Predictions.
5. L'URL du point de terminaison du modèle est répertoriée dans Real-time prediction. Utilisez cette URL comme URL Endpoint Url pour vos appels de prédiction en temps réel. Pour plus d'informations sur la façon d'utiliser le point de terminaison pour générer des prédictions, consultez https://docs.aws.amazon.com/machine-learning/latest/APIReference/API_Predict.html le manuel Amazon Machine Learning API Reference.

Localisation du point de terminaison de prédiction en temps réel (API)

Lorsque vous créez un point de terminaison en temps réel à l'aide de l'opération `CreateRealtimeEndpoint`, l'URL et l'état de ce point de terminaison vous sont renvoyés dans la réponse. Si vous avez créé le point de terminaison en temps réel à l'aide de la console ou si vous souhaitez récupérer l'URL et l'état d'un point de terminaison que vous avez créé précédemment, appelez l'opération `GetMLModel` avec l'ID du modèle à interroger pour les prédictions en temps réel. Les informations du point de terminaison figurent dans la section `EndpointInfo` de la réponse. Pour un modèle doté d'un point de terminaison en temps réel associé, la section `EndpointInfo` peut ressembler à ceci :

```
"EndpointInfo":{
    "CreatedAt": 1427864874.227,
    "EndpointStatus": "READY",
    "EndpointUrl": "https://endpointUrl",
    "PeakRequestsPerSecond": 200
}
```

Un modèle sans point de terminaison en temps réel renverrait ce qui suit :

```
EndpointInfo":{
    "EndpointStatus": "NONE",
    "PeakRequestsPerSecond": 0
}
```

Création d'une demande de prédiction en temps réel

Un exemple de charge utile de demande `Predict` peut ressembler à ceci :

```
{
    "MLModelId": "model-id",
    "Record":{
        "key1": "value1",
        "key2": "value2"
    },
    "PredictEndpoint": "https://endpointUrl"
}
```

Le `PredictEndpoint` champ doit correspondre au `EndpointUrl` champ de la `EndpointInfo` structure. Amazon ML utilise ce champ pour acheminer la demande vers les serveurs appropriés du parc de prévisions en temps réel.

`MLModelId` est l'identifiant d'un modèle formé précédemment avec un point de terminaison en temps réel.

`Record` est un mappage de noms de variables et de valeurs de variables. Chaque paire représente une observation. La `Record` carte contient les entrées de votre modèle Amazon ML. Il est analogue à une ligne de données individuelle dans votre jeu de données de formation, sans la variable cible. Quel que soit le type de valeurs figurant dans les données d'entraînement, `Record` contient un string-to-string mappage.

> **Note**
>
> Vous pouvez omettre les variables pour lesquelles vous n'avez pas de valeur, bien que cela puisse réduire la précision de votre prédiction. Plus vous pouvez inclure de variables, plus votre modèle est précis.

Le format de la réponse renvoyée par les demandes `Predict` dépend du type de modèle qui est interrogé pour la prédiction. Dans tous les cas, le champ `details` contient des informations sur la demande de prédiction, notamment le champ `PredictiveModelType` avec le type de modèle.

L'exemple suivant illustre une réponse pour un modèle binaire :

```
{
    "Prediction":{
        "details":{
            "PredictiveModelType": "BINARY"
        },
        "predictedLabel": "0",
        "predictedScores":{
            "0": 0.47380468249320984
        }
    }
}
```

Notez le `predictedLabel` champ qui contient l'étiquette prédite, dans ce cas 0. Amazon ML calcule l'étiquette prédite en comparant le score de prédiction au seuil de classification :

- Vous pouvez obtenir le seuil de classification actuellement associé à un modèle ML en inspectant le `ScoreThreshold` champ dans la réponse à l'`GetMLModel`opération ou en consultant les informations du modèle dans la console Amazon ML. Si vous ne définissez pas de seuil de score, Amazon ML utilise la valeur par défaut de 0,5.
- Vous pouvez obtenir le score de prédiction exact pour un modèle de classification binaire en inspectant le mappage `predictedScores`. Dans ce mappage, l'étiquette prédite est associée au score de prédiction exact.

Pour plus d'informations sur les prédictions binaires, consultez Interprétation des prédictions.

L'exemple suivant illustre une réponse pour un modèle de régression. Remarquez que la valeur numérique prédite se trouve dans le champ `predictedValue` :

```
{
    "Prediction":{
        "details":{
            "PredictiveModelType": "REGRESSION"
        },
        "predictedValue": 15.508452415466309
```

```
    }
}
```

L'exemple suivant illustre une réponse pour un modèle multiclasse :

```
{
    "Prediction":{
        "details":{
            "PredictiveModelType": "MULTICLASS"
        },
        "predictedLabel": "red",
        "predictedScores":{
            "red": 0.12923571467399597,
            "green": 0.08416014909744263,
            "orange": 0.22713537514209747,
            "blue": 0.1438363939523697,
            "pink": 0.184102863073349,
            "violet": 0.12816807627677917,
            "brown": 0.10336143523454666
        }
    }
}
```

Comme pour les modèles de classification binaire, l'étiquette/la classe prédite se trouve dans le champ `predictedLabel`. Vous pouvez mieux comprendre à quel point la prédiction est liée à chaque classe en examinant le mappage `predictedScores`. Plus le score d'une classe au sein de ce mappage est élevé et plus la prédiction est fortement liée à la classe, la valeur la plus haute étant sélectionnée comme `predictedLabel`.

Pour plus d'informations sur les prédictions multiclasses, consultez Analyse du modèle multiclasse.

Suppression d'un point de terminaison en temps réel

Lorsque vous avez terminé vos prédictions en temps réel, supprimez le point de terminaison en temps réel pour ne pas encourir de frais supplémentaires. Les frais cessent de s'accumuler dès que vous supprimez votre point de terminaison.

Pour supprimer un point de terminaison en temps réel

1. Connectez-vous à la console Amazon Machine Learning AWS Management Console et ouvrez-la à l'adresse https://console.aws.amazon.com/machinelearning/.

2. Dans la barre de navigation, dans le menu déroulant Amazon Machine Learning, choisissez ML models.
3. Choisissez le modèle qui n'a plus besoin de prédictions en temps réel.
4. Dans la page de rapport du modèle d'apprentissage-machine, sous Prédictions, choisissez Résumé.
5. Choisissez Delete real-time endpoint (Supprimer le point de terminaison en temps réel).
6. Dans la boîte de dialogue Delete real-time endpoint (Supprimer le point de terminaison en temps réel), choisissez Supprimer.

Gestion des objets Amazon ML

Amazon ML fournit quatre objets que vous pouvez gérer via la console Amazon ML ou l'API Amazon ML :

- Sources de données
- Des modèles d'apprentissage-machine
- Evaluations
- Des prédictions par lots

Chaque objet a un objectif distinct dans le cycle de vie du développement d'une application d'apprentissage-machine, et chaque objet a des attributs et des fonctionnalités spécifiques qui s'appliquent uniquement à cet objet. Malgré ces différences, vous gérez ces objets de manière similaire. Par exemple, vous utilisez des processus presque identiques pour répertorier les objets, récupérer leurs descriptions et les mettre à jour ou les supprimer.

Les sections suivantes décrivent les opérations de gestion communes à ces quatre objets, ainsi que leurs différences.

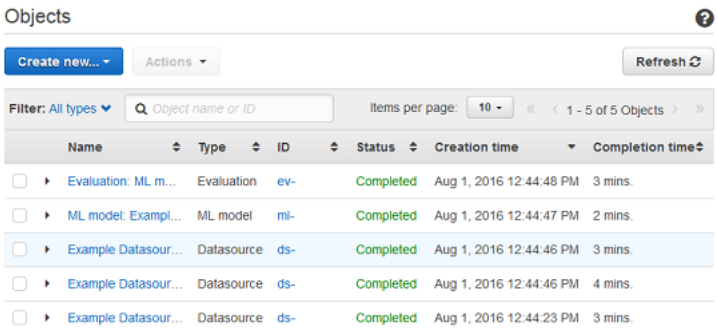
Rubriques

Liste des objets

Pour obtenir des informations détaillées sur vos sources de données Amazon Machine Learning (Amazon ML), vos modèles de machine learning, vos évaluations et vos prédictions par lots, listez-les. Pour chaque objet, vous verrez ses nom, type, ID, code d'état et heure de création. Vous pouvez également voir des détails spécifiques à un type d'objet particulier. Par exemple, vous pouvez voir l'analyse des données d'une source de données.

Liste des objets (console)

Pour consulter la liste des 1 000 derniers objets que vous avez créés, ouvrez le tableau de bord Objects dans la console Amazon ML. Pour afficher le tableau de bord Objects, connectez-vous à la console Amazon ML.



Pour voir plus de détails sur un objet, y compris des détails spécifiques à ce type d'objet, choisissez le nom ou l'ID de l'objet. Par exemple, pour voir l'analyse des données d'une source de données, choisissez le nom de la source de données.

Les colonnes du tableau de bord Objets montre les informations suivantes sur chaque objet.

Nom

Nom de l’objet.

Type

Type de l'objet. Les valeurs valides incluent Datasource, ML model, Evaluation et Batch prediction.

> **Note**
>
> Pour voir si un modèle est configuré pour prendre en charge les prédictions en temps réel, accédez à la page ML model summary en choisissant le nom ou l'ID de modèle.

ID

ID de l'objet.

Statut

Statut de l'objet. Les valeurs incluent En suspens, En cours, Terminé et Echec. Si l'état est Echec, vérifiez vos données et réessayez.

Heure de création

Date et heure auxquelles Amazon ML a terminé de créer cet objet.

Délai d'achèvement

Le temps qu'il a fallu à Amazon ML pour créer cet objet. Vous pouvez utiliser le délai d'achèvement d'un modèle pour estimer la durée de formation d'un nouveau modèle.

ID de source de données

Pour les objets qui ont été créés à l'aide d'une source de données, tels que les modèles et les évaluations, il s'agit de l'identifiant de la source de données. Si vous supprimez la source de données, vous ne pouvez plus utiliser les modèles d'apprentissage-machine créés avec cette source de données pour créer des prédictions.

Triez par n'importe quelle colonne en choisissant l'icône de triangle double en regard de l'en-tête de la colonne.

Liste des objets (API)

Dans l'API Amazon ML, vous pouvez répertorier les objets, par type, en effectuant les opérations suivantes :

- `DescribeDataSources`
- `DescribeMLModels`
- `DescribeEvaluations`
- `DescribeBatchPredictions`

Chaque opération inclut des paramètres de filtrage, de tri et de pagination par le biais d'une longue liste d'objets. Il n'y a pas de limite au nombre d'objets auxquels vous pouvez accéder via l'API. Pour limiter la taille de la liste, utilisez le paramètre `Limit`, qui peut accepter une valeur maximale de 100.

La réponse de l'API à une commande `Describe*` inclut un jeton de pagination (`nextPageToken`), le cas échéant, et une brève description de chaque objet. Les descriptions des objets comprennent

les mêmes informations pour chaque type d'objet qui s'affiche dans la console, y compris des détails spécifiques au type d'objet.

> **Note**
>
> Même si la réponse inclut moins d'objets que la limite spécifiée, elle peut inclure un jeton `nextPageToken` qui indique que d'autres résultats sont disponibles. Même une réponse qui contient 0 élément peut contenir un jeton `nextPageToken`.

Pour plus d'informations, consultez le manuel Amazon ML API Reference.

Récupération des descriptions d'objet

Vous pouvez consulter les descriptions détaillées d'un objet quelconque via la console ou l'API.

Descriptions détaillées dans la console

Pour voir les descriptions sur la console, accédez à la liste d'un type spécifique d'objet (source de données, modèle d'apprentissage-machine, évaluation ou prédiction par lots). Ensuite, localisez la ligne de la table qui correspond à l'objet, en parcourant la liste ou en recherchant son nom ou son ID.

Descriptions détaillées à partir de l'API

Chaque type d'objet a une opération qui récupère tous les détails d'un objet Amazon ML :

- GetDataSource
- Obtenez MLModel
- GetEvaluation
- GetBatchPrediction

Chaque opération accepte exactement deux paramètres : l'ID de l'objet et un indicateur booléen appelé Verbose. Les appels avec l'indicateur Verbose défini sur true incluent des détails supplémentaires sur l'objet, ce qui se traduit par des latences plus élevées et des réponses plus volumineuses. Pour savoir quels champs sont inclus lorsque vous définissez l'indicateur Verbose, consultez la Référence d'API Amazon ML.

Mise à jour d'objets

Chaque type d'objet comporte une opération qui met à jour les détails d'un objet Amazon ML (voir Amazon ML API Reference) :

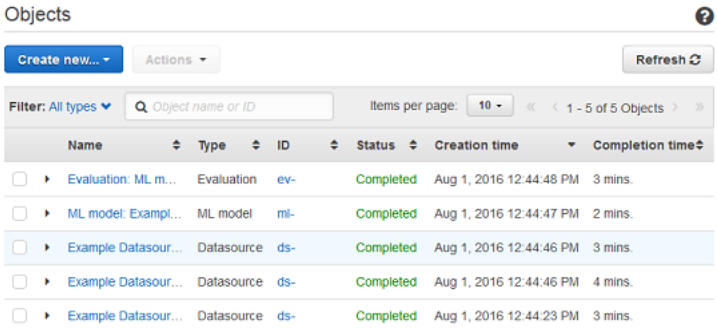
- UpdateDataSource
- Mettre à jour MLModel
- UpdateEvaluation
- UpdateBatchPrediction

Chaque opération nécessite l'ID de l'objet pour spécifier l'objet qui est en cours de mise à jour. Vous pouvez mettre à jour les noms de tous les objets. Vous ne pouvez pas mettre à jour d'autres propriétés des objets pour les sources de données, les évaluations et les prédictions par lots. Pour les modèles ML, vous pouvez mettre à jour le ScoreThreshold champ, à condition qu'aucun point de terminaison de prédiction en temps réel ne soit associé au modèle ML.

Suppression d'objets

Lorsque vous n'avez plus besoin de vos sources de données, modèles d'apprentissage-machine, évaluations ou prédictions par lots, vous pouvez les supprimer. Bien que la conservation des objets Amazon ML soit gratuite, hormis les prédictions par lots une fois que vous en avez fini avec eux, la suppression d'objets permet de simplifier votre espace de travail et de le gérer. Vous pouvez supprimer un ou plusieurs objets à l'aide de la console Amazon Machine Learning (Amazon ML) ou de l'API.

> **Warning**
>
> Lorsque vous supprimez des objets Amazon ML, l'effet est immédiat, permanent et irréversible.



Suppression d'objets (console)

Vous pouvez utiliser la console Amazon ML pour supprimer des objets, y compris des modèles. La procédure que vous utilisez pour supprimer un modèle dépend de si vous utilisez ou non ce modèle pour générer des prédictions en temps réel. Pour supprimer un modèle servant à générer des prédictions en temps réel, commencez par supprimer le point de terminaison en temps réel.

Pour supprimer des objets Amazon ML (console)

1. Connectez-vous à la console Amazon Machine Learning AWS Management Console et ouvrez-la à l'adresse https://console.aws.amazon.com/machinelearning/.
2. Sélectionnez les objets Amazon ML que vous souhaitez supprimer. Pour sélectionner plusieurs objets, utilisez la touche Maj. Pour désélectionner tous les objets sélectionnés, utilisez les boutons

   ou

   
3. Pour Actions, choisissez Supprimer.
4. Dans la boîte de dialogue, choisissez Delete (Supprimer) pour supprimer le modèle.

Pour supprimer un modèle Amazon ML avec un point de terminaison en temps réel (console)

1. Connectez-vous à la console Amazon Machine Learning AWS Management Console et ouvrez-la à l'adresse https://console.aws.amazon.com/machinelearning/.
2. Sélectionnez le modèle à supprimer.
3. Pour Actions, choisissez Delete real-time endpoint (Supprimer le point de terminaison en temps réel).

4. Choisissez Delete (Supprimer) pour supprimer le point de terminaison.
5. Sélectionnez le modèle à nouveau.
6. Pour Actions, choisissez Supprimer.
7. Choisissez Supprimer pour supprimer le modèle.

Suppression d'objets (API)

Vous pouvez supprimer des objets Amazon ML à l'aide des appels d'API suivants :

- `DeleteDataSource` - Accepte le paramètre `DataSourceId`.
- `DeleteMLModel` - Accepte le paramètre `MLModelId`.
- `DeleteEvaluation` - Accepte le paramètre `EvaluationId`.
- `DeleteBatchPrediction` - Accepte le paramètre `BatchPredictionId`.

Pour plus d'informations, consultez la Référence d'API Amazon Machine Learning.

Surveillance d'Amazon ML avec Amazon CloudWatch Metrics

Amazon ML envoie automatiquement des métriques à Amazon CloudWatch afin que vous puissiez recueillir et analyser les statistiques d'utilisation de vos modèles de ML. Par exemple, pour suivre les prévisions par lots et en temps réel, vous pouvez surveiller la PredictCount métrique en fonction de la RequestMode dimension. Les statistiques sont automatiquement collectées et envoyées à Amazon CloudWatch toutes les cinq minutes. Vous pouvez surveiller ces métriques à l'aide de la CloudWatch console Amazon, de l'AWS CLI ou d'AWS SDKs.

Les métriques Amazon ML signalées par le biais de ce service ne sont pas facturées CloudWatch. Si vous définissez des alarmes sur les métriques, vous serez facturé au CloudWatch tarif standard.

Pour plus d'informations, consultez la liste des métriques Amazon ML dans Amazon CloudWatch Namespaces, Dimensions et Metrics Reference du manuel Amazon CloudWatch Developer Guide.

Journalisation des appels d'API Amazon ML avec AWS CloudTrail

Amazon Machine Learning (Amazon ML) est intégré AWS CloudTrailà un service qui fournit un enregistrement des actions effectuées par un utilisateur, un rôle ou AWS un service dans Amazon ML. CloudTrail capture tous les appels d'API pour Amazon ML sous forme d'événements. Les appels capturés incluent des appels provenant de la console Amazon ML et des appels de code vers les opérations de l'API Amazon ML. Si vous créez un suivi, vous pouvez activer la diffusion continue d' CloudTrail événements vers un compartiment Amazon S3, y compris des événements pour Amazon ML. Si vous ne configurez pas de suivi, vous pouvez toujours consulter les événements les plus récents dans la CloudTrail console dans Historique des événements. À l'aide des informations collectées par CloudTrail, vous pouvez déterminer la demande envoyée à Amazon ML, l'adresse IP à partir de laquelle la demande a été effectuée, l'auteur de la demande, la date à laquelle elle a été faite, ainsi que des informations supplémentaires.

Pour en savoir plus CloudTrail, notamment comment le configurer et l'activer, consultez le guide de AWS CloudTrail l'utilisateur.

Informations sur Amazon ML dans CloudTrail

CloudTrail est activé sur votre AWS compte lorsque vous le créez. Lorsqu'une activité événementielle prise en charge se produit dans Amazon ML, cette activité est enregistrée dans un CloudTrail événement avec d'autres événements de AWS service dans l'historique des événements. Vous pouvez consulter, rechercher et télécharger les événements récents dans votre AWS compte. Pour plus d'informations, consultez la section Affichage des événements à l'aide de l'historique des CloudTrail événements.

Pour un enregistrement continu des événements de votre AWS compte, y compris des événements pour Amazon ML, créez un historique. Un suivi permet CloudTrail de fournir des fichiers journaux à un compartiment Amazon S3. Par défaut, lorsque vous créez un journal d'activité dans la console, il s’applique à toutes les régions AWS. Le journal enregistre les événements de toutes les régions de la AWS partition et transmet les fichiers journaux au compartiment Amazon S3 que vous spécifiez. En outre, vous pouvez configurer d'autres AWS services pour analyser plus en détail les données d'événements collectées dans les CloudTrail journaux et agir en conséquence. Pour plus d’informations, consultez les ressources suivantes :

- Vue d’ensemble de la création d’un journal d’activité

- CloudTrail Services et intégrations pris en charge
- Configuration des notifications Amazon SNS pour CloudTrail
- Réception de fichiers CloudTrail journaux de plusieurs régions et réception de fichiers CloudTrail journaux de plusieurs comptes

Amazon ML prend en charge l'enregistrement des actions suivantes sous forme d'événements dans des fichiers CloudTrail journaux :

- AddTags
- CreateBatchPrediction
- CreateDataSourceFromRDS
- CreateDataSourceFromRedshift
- CreateDataSourceFromS3
- CreateEvaluation
- CréerMLModel
- CreateRealtimeEndpoint
- DeleteBatchPrediction
- DeleteDataSource
- DeleteEvaluation
- SuppressionMLModel
- DeleteRealtimeEndpoint
- DeleteTags
- DescribeTags
- UpdateBatchPrediction
- UpdateDataSource
- UpdateEvaluation
- Mettre à jourMLModel

Les opérations Amazon ML suivantes utilisent des paramètres de demande contenant des informations d'identification. Avant que ces demandes ne soient envoyées à CloudTrail, les informations d'identification sont remplacées par trois astérisques (« *** ») :

- CreateDataSourceFromRDS
- CreateDataSourceFromRedshift

Lorsque les opérations Amazon ML suivantes sont effectuées avec la console Amazon ML, l'attribut n'`ComputeStatisticsest` pas inclus dans le `RequestParameters` composant du CloudTrail journal :

- CreateDataSourceFromRedshift
- CreateDataSourceFromS3

Chaque événement ou entrée de journal contient des informations sur la personne ayant initié la demande. Les informations relatives à l'identité permettent de déterminer les éléments suivants :

- Si la demande a été faite avec les informations d'identification de l'utilisateur root ou AWS Identity and Access Management (IAM).
- Si la demande a été effectuée avec les informations d'identification de sécurité temporaires d'un rôle ou d'un utilisateur fédéré.
- Si la demande a été faite par un autre AWS service.

Pour plus d'informations, consultez la section Élément userIdentity CloudTrail .

Exemple : entrées dans le fichier journal Amazon ML

Un suivi est une configuration qui permet de transmettre des événements sous forme de fichiers journaux à un compartiment Amazon S3 que vous spécifiez. CloudTrail les fichiers journaux contiennent une ou plusieurs entrées de journal. Un événement représente une demande unique provenant de n'importe quelle source et inclut des informations sur l'action demandée, la date et l'heure de l'action, les paramètres de la demande, etc. CloudTrail les fichiers journaux ne constituent pas une trace ordonnée des appels d'API publics, ils n'apparaissent donc pas dans un ordre spécifique.

L'exemple suivant montre une entrée de CloudTrail journal illustrant l'action.

```
{
    "Records": [
        {
```

```
            "eventVersion": "1.03",
            "userIdentity": {
                "type": "IAMUser",
                "principalId": "EX_PRINCIPAL_ID",
                "arn": "arn:aws:iam::012345678910:user/Alice",
                "accountId": "012345678910",
                "accessKeyId": "EXAMPLE_KEY_ID",
                "userName": "Alice"
            },
            "eventTime": "2015-11-12T15:04:02Z",
            "eventSource": "machinelearning.amazonaws.com",
            "eventName": "CreateDataSourceFromS3",
            "awsRegion": "us-east-1",
            "sourceIPAddress": "127.0.0.1",
            "userAgent": "console.amazonaws.com",
            "requestParameters": {
                "data": {
                    "dataLocationS3": "s3://aml-sample-data/banking-batch.csv",
                    "dataSchema": "{\"version\":\"1.0\",\"rowId\":null,\"rowWeight
\":null,
                        \"targetAttributeName\":null,\"dataFormat\":\"CSV\",
                        \"dataFileContainsHeader\":false,\"attributes\":[
                          {\"attributeName\":\"age\",\"attributeType\":\"NUMERIC\"},
                          {\"attributeName\":\"job\",\"attributeType\":\"CATEGORICAL
\"},
                          {\"attributeName\":\"marital\",\"attributeType\":
\"CATEGORICAL\"},
                          {\"attributeName\":\"education\",\"attributeType\":
\"CATEGORICAL\"},
                          {\"attributeName\":\"default\",\"attributeType\":
\"CATEGORICAL\"},
                          {\"attributeName\":\"housing\",\"attributeType\":
\"CATEGORICAL\"},
                          {\"attributeName\":\"loan\",\"attributeType\":\"CATEGORICAL
\"},
                          {\"attributeName\":\"contact\",\"attributeType\":
\"CATEGORICAL\"},
                          {\"attributeName\":\"month\",\"attributeType\":\"CATEGORICAL
\"},
                          {\"attributeName\":\"day_of_week\",\"attributeType\":
\"CATEGORICAL\"},
                          {\"attributeName\":\"duration\",\"attributeType\":\"NUMERIC
\"},
```

```
                            {\"attributeName\":\"campaign\",\"attributeType\":\"NUMERIC
\"},
                            {\"attributeName\":\"pdays\",\"attributeType\":\"NUMERIC\"},
                            {\"attributeName\":\"previous\",\"attributeType\":\"NUMERIC
\"},
                            {\"attributeName\":\"poutcome\",\"attributeType\":
\"CATEGORICAL\"},
                            {\"attributeName\":\"emp_var_rate\",\"attributeType\":
\"NUMERIC\"},
                            {\"attributeName\":\"cons_price_idx\",\"attributeType\":
\"NUMERIC\"},
                            {\"attributeName\":\"cons_conf_idx\",\"attributeType\":
\"NUMERIC\"},
                            {\"attributeName\":\"euribor3m\",\"attributeType\":\"NUMERIC
\"},
                            {\"attributeName\":\"nr_employed\",\"attributeType\":
\"NUMERIC\"}
                        ],\"excludedAttributeNames\":[]}"
                },
                "dataSourceId": "exampleDataSourceId",
                "dataSourceName": "Banking sample for batch prediction"
            },
            "responseElements": {
                "dataSourceId": "exampleDataSourceId"
            },
            "requestID": "9b14bc94-894e-11e5-a84d-2d2deb28fdec",
            "eventID": "f1d47f93-c708-495b-bff1-cb935a6064b2",
            "eventType": "AwsApiCall",
            "recipientAccountId": "012345678910"
        },
        {
            "eventVersion": "1.03",
            "userIdentity": {
                "type": "IAMUser",
                "principalId": "EX_PRINCIPAL_ID",
                "arn": "arn:aws:iam::012345678910:user/Alice",
                "accountId": "012345678910",
                "accessKeyId": "EXAMPLE_KEY_ID",
                "userName": "Alice"
            },
            "eventTime": "2015-11-11T15:24:05Z",
            "eventSource": "machinelearning.amazonaws.com",
            "eventName": "CreateBatchPrediction",
            "awsRegion": "us-east-1",
```

```
            "sourceIPAddress": "127.0.0.1",
            "userAgent": "console.amazonaws.com",
            "requestParameters": {
                "batchPredictionName": "Batch prediction: ML model: Banking sample",
                "batchPredictionId": "exampleBatchPredictionId",
                "batchPredictionDataSourceId": "exampleDataSourceId",
                "outputUri": "s3://EXAMPLE_BUCKET/BatchPredictionOutput/",
                "mLModelId": "exampleModelId"
            },
            "responseElements": {
                "batchPredictionId": "exampleBatchPredictionId"
            },
            "requestID": "3e18f252-8888-11e5-b6ca-c9da3c0f3955",
            "eventID": "db27a771-7a2e-4e9d-bfa0-59deee9d936d",
            "eventType": "AwsApiCall",
            "recipientAccountId": "012345678910"
        }
    ]
}
```

Marquage de vos objets Amazon ML

Organisez et gérez vos objets Amazon Machine Learning (Amazon ML) en leur attribuant des métadonnées à l'aide de balises. Une balise est une paire clé-valeur que vous définissez pour un objet.

En plus d'utiliser des balises pour organiser et gérer vos objets Amazon ML, vous pouvez les utiliser pour classer et suivre vos coûts AWS. Lorsque vous appliquez des balises à vos objets AWS, y compris aux modèles d'apprentissage-machine, votre rapport de répartition des coûts AWS comprend l'utilisation et les coûts regroupés par balises. En appliquant des balises qui représentent des catégories métier (telles que les centres de coûts, les noms d'applications ou les propriétaires), vous pouvez organiser vos coûts entre plusieurs services. Pour plus d'informations, consultez Utilisation des identifications de répartition des coûts pour les rapports de facturation personnalisés dans le Guide de l'utilisateur AWS Billing .

Table des matières

Principes de base des identifications

Utilisez des balises pour classer par catégories vos objets afin de faciliter leur gestion. Par exemple, vous pouvez classer les objets par objectif, propriétaire ou environnement. Ensuite, vous pouvez définir un ensemble de balises vous permettant d'effectuer le suivi des modèles par propriétaire et application associée. Voici quelques exemples :

- Projet : nom du projet
- Propriétaire : nom
- Objectif : prédictions marketing
- Application : nom de l'application
- Environnement : production

Vous utilisez la console ou l'API Amazon ML pour effectuer les tâches suivantes :

- Ajouter des balises à un objet
- Afficher les balises de vos objets
- Modifier les balises de vos objets
- Supprimer les balises d'un objet

Par défaut, les balises appliquées à un objet Amazon ML sont copiées dans les objets créés à l'aide de cet objet. Par exemple, si une source de données Amazon Simple Storage Service (Amazon S3) possède la balise « Coût marketing : campagne marketing ciblée », un modèle créé à l'aide de cette source de données comportera également la balise « Coût marketing : campagne marketing ciblée », tout comme l'évaluation du modèle. Cela vous permet d'utiliser des balises pour effectuer le suivi des objets associés, tels que tous les objets utilisés pour une campagne marketing. En cas de conflit entre les sources de balises, telles qu'un modèle avec la balise « Coût marketing : campagne marketing ciblée » et une source de données avec la balise « Coût marketing : clients marketing cibles », Amazon ML applique la balise à partir du modèle.

Restrictions liées aux balises

Les restrictions suivantes s'appliquent aux balises :

Restrictions de base:

- Le nombre maximum de balises par objet est de 50.
- Les clés et valeurs de balise sont sensibles à la casse.
- Vous ne pouvez pas changer ni modifier les balises d'un objet supprimé.

Restrictions relatives aux clés de balise:

- Chaque clé de balise doit être unique. Si vous ajoutez une balise avec une clé qui est déjà en cours d'utilisation, la nouvelle balise remplace la paire clé-valeur existante pour cet objet.
- Une clé de balise ne peut pas commencer par `aws:`, car ce préfixe est réservé à AWS. AWS crée à votre place des balises qui commencent par ce préfixe, mais vous ne pouvez pas les modifier ou supprimer ces balises.
- Les clés de balise doivent comporter entre 1 et 128 caractères Unicode.
- Les clés de balise doivent comporter les caractères suivants : lettres Unicode, chiffres, espaces et les caractères spéciaux suivants : `_ . / = + - @`.

Restrictions relatives à la valeur de balise:

- Les valeurs de balise doivent comporter entre 0 et 255 caractères Unicode.
- Les valeurs de balise peuvent être vides. Si tel n'est pas le cas, elles doivent être composées des caractères suivants : lettres Unicode, chiffres, espaces et les caractères spéciaux suivants : `_ . / = + - @`.

Marquage d'objets Amazon ML (console)

Vous pouvez afficher, ajouter, modifier et supprimer des balises à l'aide de la console Amazon ML.

Pour afficher les balises d'un objet (console)

1. Connectez-vous à la console Amazon Machine Learning AWS Management Console et ouvrez-la à l'adresse https://console.aws.amazon.com/machinelearning/.
2. Dans la barre de navigation, développez le sélecteur de région et choisissez une région.
3. Dans la page Objets, choisissez un objet.
4. Faites défiler la page jusqu'à la section Balises de l'objet choisi. Les balises de cet objet sont répertoriées en bas de la section.

Pour ajouter une balise à un objet (console)

1. Connectez-vous à la console Amazon Machine Learning AWS Management Console et ouvrez-la à l'adresse https://console.aws.amazon.com/machinelearning/.
2. Dans la barre de navigation, développez le sélecteur de région et choisissez une région.
3. Dans la page Objets, choisissez un objet.
4. Faites défiler la page jusqu'à la section Balises de l'objet choisi. Les balises de cet objet sont répertoriées en bas de la section.
5. Sélectionnez Ajouter ou modifier des balises
6. Sous Ajouter une balise, spécifiez la clé de balise dans le champ Clé. Spécifiez éventuellement une valeur de balise dans le champ Valeur, puis choisissez Apply changes.

   Si le bouton Apply changes n'est pas activé, la clé ou la valeur de la balise que vous avez spécifiée ne répond pas aux restrictions de balise. Pour de plus amples informations, veuillez consulter Restrictions liées aux balises.

7. Pour afficher votre nouvelle balise dans la liste, dans la section Balises, actualisez la page.

Pour modifier une balise (console)

1. Connectez-vous à la console Amazon Machine Learning AWS Management Console et ouvrez-la à l'adresse https://console.aws.amazon.com/machinelearning/.
2. Dans la barre de navigation, développez le sélecteur de région et sélectionnez une région.
3. Dans la page Objets, choisissez un objet.
4. Faites défiler la page jusqu'à la section Balises de l'objet choisi. Les balises de cet objet sont répertoriées en bas de la section.
5. Sélectionnez Ajouter ou modifier des balises
6. Sous Applied tags, modifiez une valeur de balise dans le champ Value, puis choisissez Apply changes.

   Si le bouton Apply changes n'est pas activé, la valeur de balise que vous avez spécifiée ne répond pas aux restrictions de balise. Pour de plus amples informations, veuillez consulter Restrictions liées aux balises.
7. Pour afficher votre balise mise à jour dans la liste, dans la section Balises, actualisez la page.

Pour supprimer une balise d'un objet (console)

1. Connectez-vous à la console Amazon Machine Learning AWS Management Console et ouvrez-la à l'adresse https://console.aws.amazon.com/machinelearning/.
2. Dans la barre de navigation, développez le sélecteur de région et choisissez une région.
3. Dans la page Objets, choisissez un objet.
4. Faites défiler la page jusqu'à la section Balises de l'objet choisi. Les balises de cet objet sont répertoriées en bas de la section.
5. Sélectionnez Ajouter ou modifier des balises
6. Sous Applied tags, choisissez la balise que vous souhaitez supprimer, puis choisissez Apply changes.

Marquage d'objets Amazon ML (API)

Vous pouvez ajouter, répertorier et supprimer des balises à l'aide de l'API Amazon ML. Pour obtenir des exemples, consultez la documentation suivante :

AddTags

Ajoute ou modifie des balises pour l'objet spécifié.

DescribeTags

Répertorie les balises de l'objet spécifié.

DeleteTags

Supprime les balises de l'objet spécifié.

Référence Amazon Machine Learning

Rubriques

Octroi à Amazon ML des autorisations nécessaires pour lire vos données depuis Amazon S3

Pour créer un objet source de données à partir de vos données d'entrée dans Amazon S3, vous devez accorder à Amazon ML les autorisations suivantes sur l'emplacement S3 où vos données d'entrée sont stockées :

- GetObjectautorisation sur le compartiment et le préfixe S3.
- ListBucketautorisation sur le compartiment S3. Contrairement aux autres actions, des autorisations ListBucketdoivent être accordées à l'échelle du compartiment (plutôt que sur le préfixe). Cependant, vous pouvez délimiter l'autorisation à un préfixe spécifique à l'aide d'une clause Condition.

Si vous utilisez la console Amazon ML pour créer la source de données, ces autorisations peuvent être ajoutées au compartiment pour vous. Vous serez invité à confirmer si vous souhaitez les ajouter au fur et à mesure que vous terminerez les étapes de l'assistant. L'exemple de politique suivant montre comment autoriser Amazon ML à lire des données depuis l'emplacement d'échantillonnage s3 ://*examplebucket*/*exampleprefix*, tout en limitant l'ListBucketautorisation au seul chemin d'entrée. *exampleprefix*

```
{
    "Version": "2008-10-17",
    "Statement": [
    {
        "Effect": "Allow",
        "Principal": { "Service": "machinelearning.amazonaws.com" },
        "Action": "s3:GetObject",
        "Resource": "arn:aws:s3:::examplebucket/exampleprefix/*"
        "Condition": {
            "StringEquals": { "aws:SourceAccount": "123456789012" }
            "ArnLike": { "aws:SourceArn": "arn:aws:machinelearning:us-
east-1:123456789012:*" }
        }
    },
    {
        "Effect": "Allow",
        "Principal": {"Service": "machinelearning.amazonaws.com"},
        "Action": "s3:ListBucket",
        "Resource": "arn:aws:s3:::examplebucket",
        "Condition": {
            "StringLike": { "s3:prefix": "exampleprefix/*" }
            "StringEquals": { "aws:SourceAccount": "123456789012" }
            "ArnLike": { "aws:SourceArn": "arn:aws:machinelearning:us-
east-1:123456789012:*" }
        }
    }]
}
```

Pour appliquer cette stratégie à vos données, vous devez modifier la déclaration de stratégie associée au compartiment S3 où vos données sont stockées.

Pour modifier la stratégie d'autorisations pour un compartiment S3 (en utilisant l'ancienne console)

1. Connectez-vous à la console Amazon S3 AWS Management Console et ouvrez-la à l'adresse https://console.aws.amazon.com/s3/.
2. Sélectionnez le nom du compartiment dans lequel vos données résident.
3. Choisissez Propriétés.
4. Choisissez Edit bucket policy.
5. Entrez la stratégie illustrée ci-dessus, en la personnalisant en fonction de vos besoins, puis choisissez Save.

6. Choisissez Enregistrer.

Pour modifier la stratégie d'autorisations pour un compartiment S3 (en utilisant la nouvelle console)

1. Connectez-vous à la console Amazon S3 AWS Management Console et ouvrez-la à l'adresse https://console.aws.amazon.com/s3/.
2. Choisissez le nom du compartiment, puis Permissions.
3. Choisissez Stratégie de compartiment.
4. Entrez la stratégie illustrée ci-dessus, en la personnalisant en fonction de vos besoins.
5. Choisissez Enregistrer.

Octroi d'autorisations à Amazon ML pour fournir en sortie des prédictions dans Amazon S3

Pour fournir en sortie les résultats de l'opération de prédiction par lots dans Amazon S3, vous devez accorder à Amazon ML les autorisations suivantes sur l'emplacement de sortie, qui est fourni en entrée à l'opération de création de prédiction par lots :

- GetObjectautorisation sur votre compartiment et votre préfixe S3.
- PutObjectautorisation sur votre compartiment et votre préfixe S3.
- PutObjectAclsur votre compartiment et votre préfixe S3.
  - Amazon ML a besoin de cette autorisation pour pouvoir accorder l' bucket-owner-full-controlautorisation ACL prédéfinie à votre compte AWS, une fois les objets créés.
- ListBucketautorisation sur le compartiment S3. Contrairement aux autres actions, des autorisations ListBucketdoivent être accordées à l'échelle du compartiment (plutôt que sur le préfixe). Toutefois, vous pouvez délimiter l'autorisation à un préfixe spécifique à l'aide d'une clause Condition.

Si vous utilisez la console Amazon ML pour créer la demande de prédiction par lots, ces autorisations peuvent être ajoutées au compartiment pour vous. Vous serez invité à confirmer si vous voulez les ajouter à la fin de la procédure dans l'assistant.

L'exemple de politique suivant montre comment autoriser Amazon ML à écrire des données sur l'emplacement d'échantillon s3://examplebucket/exampleprefix, tout en limitant l'ListBucketautorisation au seul chemin d'entrée du préfixe d'exemple et en autorisant Amazon ML à définir l'objet put ACLs sur le préfixe de sortie :

```
{
    "Version": "2008-10-17",
    "Statement": [
    {
        "Effect": "Allow",
        "Principal": { "Service": "machinelearning.amazonaws.com"},
        "Action": [
            "s3:GetObject",
            "s3:PutObject"
       ],
        "Resource": "arn:aws:s3:::examplebucket/exampleprefix/*"
        "Condition": {
            "StringEquals": { "aws:SourceAccount": "123456789012" }
            "ArnLike": { "aws:SourceArn": "arn:aws:machinelearning:us-
east-1:123456789012:*" }
        }
    },
    {
        "Effect": "Allow",
        "Principal": { "Service": "machinelearning.amazonaws.com"},
        "Action": "s3:PutObjectAcl",
        "Resource": "arn:aws:s3:::examplebucket/exampleprefix/*",
        "Condition": {
            "StringEquals": { "s3:x-amz-acl":"bucket-owner-full-control" }
            "StringEquals": { "aws:SourceAccount": "123456789012" }
           "ArnLike": { "aws:SourceArn": "arn:aws:machinelearning:us-
east-1:123456789012:*" }
       }
    },
    {
       "Effect": "Allow",
        "Principal": {"Service": "machinelearning.amazonaws.com"},
       "Action": "s3:ListBucket",
        "Resource": "arn:aws:s3:::examplebucket",
        "Condition": {
            "StringLike": { "s3:prefix": "exampleprefix/*" }
            "StringEquals": { "aws:SourceAccount": "123456789012" }
            "ArnLike": { "aws:SourceArn": "arn:aws:machinelearning:us-
east-1:123456789012:*" }
        }
    }]
}
```

Pour appliquer cette stratégie à vos données, vous devez modifier la déclaration de stratégie associée au compartiment S3 où vos données sont stockées.

Pour modifier la stratégie d'autorisations pour un compartiment S3 (en utilisant l'ancienne console)

1. Connectez-vous à la console Amazon S3 AWS Management Console et ouvrez-la à l'adresse https://console.aws.amazon.com/s3/.
2. Sélectionnez le nom du compartiment dans lequel vos données résident.
3. Choisissez Propriétés.
4. Choisissez Edit bucket policy.
5. Entrez la stratégie illustrée ci-dessus, en la personnalisant en fonction de vos besoins, puis choisissez Save.
6. Choisissez Enregistrer.

Pour modifier la stratégie d'autorisations pour un compartiment S3 (en utilisant la nouvelle console)

1. Connectez-vous à la console Amazon S3 AWS Management Console et ouvrez-la à l'adresse https://console.aws.amazon.com/s3/.
2. Choisissez le nom du compartiment, puis Permissions.
3. Choisissez Stratégie de compartiment.
4. Entrez la stratégie illustrée ci-dessus, en la personnalisant en fonction de vos besoins.
5. Choisissez Enregistrer.

Contrôle de l'accès aux ressources Amazon ML à l'aide d'IAM

AWS Identity and Access Management (IAM) vous permet de contrôler en toute sécurité l'accès aux services et ressources AWS pour vos utilisateurs. Avec IAM, vous pouvez créer et gérer des utilisateurs, des groupes et des rôles AWS, et utiliser des autorisations pour autoriser ou refuser leur accès aux ressources AWS. En utilisant IAM avec Amazon Machine Learning (Amazon ML), vous pouvez contrôler si les utilisateurs de votre organisation peuvent utiliser des ressources AWS spécifiques et s'ils peuvent effectuer une tâche à l'aide d'actions spécifiques de l'API Amazon ML.

IAM vous permet de :

- Créer des utilisateurs et des groupes sous votre compte AWS.

- Attribuer des informations d'identification de sécurité uniques à chaque utilisateur de votre compte AWS
- Contrôler les autorisations de chaque utilisateur pour exécuter les tâches à l'aide des ressources AWS
- Partager aisément vos ressources AWS avec les utilisateurs de votre compte AWS.
- Créer des rôles pour votre compte AWS et gérer leurs autorisations pour définir les utilisateurs ou services qui peuvent les endosser.
- Vous pouvez créer des rôles dans IAM et gérer des autorisations pour contrôler quelles opérations peuvent être effectuées par l'entité, ou le service AWS, qui endosse le rôle. Vous pouvez également définir l'entité qui est autorisée à endosser le rôle.

Si votre organisation possède déjà des identités IAM, vous pouvez les utiliser pour accorder des autorisations afin d'effectuer des tâches à l'aide des ressources AWS.

Pour de plus amples informations sur IAM, consultez le Guide de l'utilisateur IAM.

Syntaxe de la politique IAM

Une politique IAM est un document JSON qui se compose d'une ou de plusieurs déclarations. Chaque déclaration a la structure suivante :

```
{
    "Statement":[{
        "Effect":"effect",
        "Action":"action",
        "Resource":"arn",
        "Condition":{
            "condition operator":{
                "key":"value"
            }
        }
    }]
}
```

Une déclaration de stratégie inclut les éléments suivants :

- Effect : contrôle l'autorisation requise pour utiliser les ressources et les actions d'API que vous allez spécifier plus tard dans la déclaration. Les valeurs valides sont `Allow` et `Deny`. Comme, par

défaut, les utilisateurs IAM n'ont pas la permission d'utiliser les ressources et les actions d'API, toutes les demandes sont refusées. Une valeur `Allow` explicite remplace la valeur par défaut. Une valeur `Deny` explicite remplace toute valeur `Allows`.
- Action : désigne la ou les actions d'API spécifiques pour lesquelles vous accordez ou refusez l'autorisation.
- Resource : la ressource affectée par l'action. Pour spécifier une ressource dans la déclaration, vous utilisez son nom Amazon Resource Name (ARN).
- Condition (facultatif) : contrôle à quel moment la stratégie sera effective.

Pour simplifier la création et la gestion des politiques IAM, vous pouvez utiliser le générateur de politiques AWS et le simulateur de politiques IAM.

Spécification des actions de politique IAM pour Amazon ML MLAmazon

Dans une déclaration de politique IAM, vous pouvez spécifier une action d'API pour tout service prenant en charge l'IAM. Lorsque vous créez une déclaration de politique pour les actions d'API Amazon ML, ajoutez-la `machinelearning:` au début du nom de l'action d'API, comme indiqué dans les exemples suivants :

- `machinelearning:CreateDataSourceFromS3`
- `machinelearning:DescribeDataSources`
- `machinelearning:DeleteDataSource`
- `machinelearning:GetDataSource`

Pour spécifier plusieurs actions dans une seule déclaration, séparez-les par des virgules :

```
"Action": ["machinelearning:action1", "machinelearning:action2"]
```

Vous pouvez aussi spécifier plusieurs actions à l'aide de caractères génériques. Par exemple, vous pouvez spécifier toutes les actions dont le nom commence par le mot « Get » :

```
"Action": "machinelearning:Get*"
```

Pour spécifier toutes les actions Amazon ML, utilisez le caractère générique* :

```
"Action": "machinelearning:*"
```

Pour obtenir la liste complète des actions de l'API Amazon ML, consultez le manuel Amazon Machine Learning API Reference.

Spécification ARNs des ressources Amazon ML dans les politiques IAM

Les déclarations de politique IAM s'appliquent à une ou plusieurs ressources. Vous spécifiez les ressources pour vos politiques en fonction de leur valeur ARNs.

Pour spécifier les ressources ARNs destinées à Amazon ML, utilisez le format suivant :

"Ressource": `arn:aws:machinelearning:region:account:resource-type/identifier`

Les exemples suivants montrent comment spécifier common ARNs.

ID de source de données : `my-s3-datasource-id`

```
"Resource":
arn:aws:machinelearning:<region>:<your-account-id>:datasource/my-s3-datasource-id
```

ID du modèle d'apprentissage-machine : `my-ml-model-id`

```
"Resource":
arn:aws:machinelearning:<region>:<your-account-id>:mlmodel/my-ml-model-id
```

ID de prédiction par lots : `my-batchprediction-id`

```
"Resource":
arn:aws:machinelearning:<region>:<your-account-id>:batchprediction/my-batchprediction-
id
```

ID d'évaluation : `my-evaluation-id`

```
"Resource": arn:aws:machinelearning:<region>:<your-account-id>:evaluation/my-
evaluation-id
```

Exemples de politiques pour Amazon MLs

Exemple 1 : autoriser des utilisateurs à lire les métadonnées des ressources d'apprentissage-machine

La politique suivante permet à un utilisateur ou à un groupe de lire les métadonnées des sources de données, des modèles de machine learning, des prédictions par lots et des évaluations en effectuant DescribeDataSourcesdes GetEvaluationactions MLModels DescribeBatchPredictionsDescribeEvaluationsGetDataSourceMLModel, GetBatchPredictionDescribe,,,, Get sur les ressources spécifiées. Les autorisations des opérations Describe* ne peuvent pas être restreintes à une ressource en particulier.

```
{
    "Version": "2012-10-17",
    "Statement": [{
        "Effect": "Allow",
        "Action": [
            "machinelearning:Get*"
        ],
        "Resource": [
            "arn:aws:machinelearning:<region>:<your-account-id>:datasource/S3-DS-ID1",
            "arn:aws:machinelearning:<region>:<your-account-id>:datasource/REDSHIFT-DS-
ID1",
            "arn:aws:machinelearning:<region>:<your-account-id>:mlmodel/ML-MODEL-ID1",
            "arn:aws:machinelearning:<region>:<your-account-id>:batchprediction/BP-
ID1",
            "arn:aws:machinelearning:<region>:<your-account-id>:evaluation/EV-ID1"
        ]
    },
    {
        "Effect": "Allow",
        "Action": [
            "machinelearning:Describe*"
        ],
        "Resource": [
            "*"
        ]
    }]
}
```

Exemple 2 : autoriser des utilisateurs à créer des ressources d'apprentissage-machine

La stratégie suivante autorise un utilisateur ou un groupe à créer des sources de données d'apprentissage-machine, des modèles d'apprentissage-machine, des prédictions par lots et des évaluations en effectuant les actions `CreateDataSourceFromS3`, `CreateDataSourceFromRedshift`, `CreateDataSourceFromRDS`, `CreateMLModel`, `CreateBatchPrediction` et `CreateEvaluation`. Vous ne pouvez pas limiter les autorisations pour ces actions à une ressource spécifique.

```
{
    "Version": "2012-10-17",
    "Statement": [{
        "Effect": "Allow",
        "Action": [
            "machinelearning:CreateDataSourceFrom*",
            "machinelearning:CreateMLModel",
            "machinelearning:CreateBatchPrediction",
            "machinelearning:CreateEvaluation"
        ],
        "Resource": [
            "*"
        ]
    }]
}
```

Exemple 3 : autoriser des utilisateurs à créer et supprimer des points de terminaison en temps réel, et à réaliser des prédictions en temps réel sur un modèle d'apprentissage-machine

La stratégie suivante autorise des utilisateurs ou des groupes à créer et supprimer des points de terminaison en temps réel, et à réaliser des prédictions en temps réel pour un modèle d'apprentissage-machine spécifique, en effectuant les actions `CreateRealtimeEndpoint`, `DeleteRealtimeEndpoint` et `Predict` sur ce modèle.

```
{
    "Version": "2012-10-17",
    "Statement": [{
        "Effect": "Allow",
        "Action": [
            "machinelearning:CreateRealtimeEndpoint",
            "machinelearning:DeleteRealtimeEndpoint",
            "machinelearning:Predict"
        ],
        "Resource": [
```

```
            "arn:aws:machinelearning:<region>:<your-account-id>:mlmodel/ML-MODEL"
        ]
    }]
}
```

Exemple 4 : autoriser des utilisateurs à mettre à jour et supprimer des ressources spécifiques

La stratégie suivante autorise un utilisateur ou un groupe à mettre à jour et supprimer des ressources spécifiques dans votre compte AWS en leur donnant l'autorisation d'effectuer les actions `UpdateDataSource`, `UpdateMLModel`, `UpdateBatchPrediction`, `UpdateEvaluation`, `DeleteDataSource`, `DeleteMLModel`, `DeleteBatchPrediction` et `DeleteEvaluation` sur ces ressources dans votre compte.

```
{
    "Version": "2012-10-17",
    "Statement": [{
        "Effect": "Allow",
        "Action": [
            "machinelearning:Update*",
            "machinelearning:DeleteDataSource",
            "machinelearning:DeleteMLModel",
            "machinelearning:DeleteBatchPrediction",
            "machinelearning:DeleteEvaluation"
        ],
        "Resource": [
            "arn:aws:machinelearning:<region>:<your-account-id>:datasource/S3-DS-ID1",
            "arn:aws:machinelearning:<region>:<your-account-id>:datasource/REDSHIFT-DS-
ID1",
            "arn:aws:machinelearning:<region>:<your-account-id>:mlmodel/ML-MODEL-ID1",
            "arn:aws:machinelearning:<region>:<your-account-id>:batchprediction/BP-
ID1",
            "arn:aws:machinelearning:<region>:<your-account-id>:evaluation/EV-ID1"
        ]
    }]
}
```

Exemple 5 : Autoriser n'importe quel Amazon MLaction

La politique suivante permet à un utilisateur ou à un groupe d'utiliser n'importe quelle action Amazon ML. Etant donné que cette stratégie accorde un accès complet à l'ensemble de vos ressources d'apprentissage-machine, limitez-la aux seuls administrateurs.

```
{
    "Version": "2012-10-17",
    "Statement": [{
        "Effect": "Allow",
        "Action": [
            "machinelearning:*"
        ],
        "Resource": [
            "*"
        ]
    }]
}
```

Prévention du cas de figure de l’adjoint désorienté entre services

Le problème de député confus est un problème de sécurité dans lequel une entité qui n’est pas autorisée à effectuer une action peut contraindre une entité plus privilégiée à le faire. En AWS, l'usurpation d'identité interservices peut entraîner un problème de confusion chez les adjoints. L’usurpation d’identité entre services peut se produire lorsqu’un service (le service appelant) appelle un autre service (le service appelé). Le service appelant peut être manipulé et ses autorisations utilisées pour agir sur les ressources d’un autre client auxquelles on ne serait pas autorisé d’accéder autrement. Pour éviter cela, AWS fournit des outils qui vous aident à protéger vos données pour tous les services avec des principaux de service qui ont eu accès aux ressources de votre compte.

Nous recommandons d'utiliser les clés contextuelles de condition `aws:SourceAccount`globale `aws:SourceArn`et les clés contextuelles dans les politiques de ressources afin de limiter les autorisations qu'Amazon Machine Learning accorde à un autre service à la ressource. Si la valeur `aws:SourceArn` ne contient pas l'ID du compte, tel qu'un ARN de compartiment Amazon S3, vous devez utiliser les deux clés de contexte de condition globale pour limiter les autorisations. Si vous utilisez les deux clés de contexte de condition globale et que la valeur `aws:SourceArn` contient l'ID de compte, la valeur `aws:SourceAccount` et le compte dans la valeur `aws:SourceArn` doivent utiliser le même ID de compte lorsqu'ils sont utilisés dans la même instruction de politique. Utilisez `aws:SourceArn` si vous souhaitez qu'une seule ressource soit associée à l'accès entre services. Utilisez `aws:SourceAccount` si vous souhaitez autoriser l’association d’une ressource de ce compte à l’utilisation interservices.

Le moyen le plus efficace de se protéger contre le problème de député confus consiste à utiliser la clé de contexte de condition globale `aws:SourceArn` avec l’ARN complet de la ressource. Si vous

ne connaissez pas l'ARN complet de la ressource ou si vous spécifiez plusieurs ressources, utilisez la clé de contexte de condition globale `aws:SourceArn` avec des caractères génériques (*) pour les parties inconnues de l'ARN. Par exemple, `arn:aws:`*`servicename`*`:*:`*`123456789012`*`:*`.

L'exemple suivant montre comment vous pouvez utiliser les clés de contexte de condition `aws:SourceAccount` globale `aws:SourceArn` et les clés de contexte dans Amazon ML pour éviter le problème de confusion lié aux adjoints lors de la lecture des données d'un compartiment Amazon S3.

```
{
    "Version": "2008-10-17",
    "Statement": [
    {
        "Effect": "Allow",
        "Principal": { "Service": "machinelearning.amazonaws.com" },
        "Action": "s3:GetObject",
        "Resource": "arn:aws:s3:::examplebucket/exampleprefix/*"
        "Condition": {
            "StringEquals": { "aws:SourceAccount": "123456789012" }
            "ArnLike": { "aws:SourceArn": "arn:aws:machinelearning:us-
east-1:123456789012:*" }
        }
    },
    {
        "Effect": "Allow",
        "Principal": {"Service": "machinelearning.amazonaws.com"},
        "Action": "s3:ListBucket",
        "Resource": "arn:aws:s3:::examplebucket",
        "Condition": {
            "StringLike": { "s3:prefix": "exampleprefix/*" }
            "StringEquals": { "aws:SourceAccount": "123456789012" }
            "ArnLike": { "aws:SourceArn": "arn:aws:machinelearning:us-
east-1:123456789012:*" }
        }
    }]
}
```

Gestion des dépendances des opérations asynchrones

Le succès des opérations par lots dans Amazon ML dépend d'autres opérations. Pour gérer ces dépendances, Amazon ML identifie les demandes dotées de dépendances et vérifie que ces

opérations ont été réalisées. Si les opérations n'ont pas été réalisées, Amazon ML met de côté les demandes initiales jusqu'à ce que les opérations dont elles dépendent soient terminées.

Il existe des dépendances entre les opérations par lots. Par exemple, pour pouvoir créer un modèle d'apprentissage-machine, vous devez avoir créé une source de données avec laquelle vous pouvez former le modèle d'apprentissage-machine. Amazon ML ne peut pas former un modèle d'apprentissage-machine si aucune source de données n'est disponible.

Toutefois, Amazon ML prend en charge la gestion des dépendances pour les opérations asynchrones. Par exemple, vous n'avez pas à attendre le calcul de statistiques de données pour envoyer une demande de formation d'un modèle d'apprentissage-machine sur la source de données. Au lieu de cela, dès que la source de données a été créée, vous pouvez envoyer une demande de formation d'un modèle d'apprentissage-machine à l'aide de la source de données. En fait, Amazon ML ne commence pas véritablement l'opération de formation tant que les statistiques de la source de données n'ont pas été calculées. La MLModel demande de création est placée dans une file d'attente jusqu'à ce que les statistiques soient calculées ; une fois cela fait, Amazon ML tente immédiatement d'exécuter l'MLModel opération de création. De même, vous pouvez envoyer des demandes d'évaluation et de prédiction par lots pour les modèles d'apprentissage-machine qui n'ont pas terminé la formation.

Le tableau suivant indique les conditions requises pour effectuer différentes actions Amazon ML :

Afin de …	Vous devez avoir …
Création d'un modèle ML (créerMLModel)	Une source de données avec des statistiq ues de données calculées
Création d'une prédiction par lots (createBa tchPrediction)	Une source de données Modèle ML
Création d'une évaluation par lots (createBa tchEvaluation)	Une source de données Modèle ML

Vérification de l'état d'une demande

Lorsque vous soumettez une demande, vous pouvez vérifier son statut à l'aide de l'API Amazon Machine Learning (Amazon ML). Par exemple, si vous soumettez une `createMLModel` demande,

vous pouvez vérifier son statut en utilisant l'`describeMLModel`appel. Amazon ML répond avec l'un des statuts suivants.

Statut	Définition
PENDING	La demande est en cours de validation par Amazon ML. OU Amazon ML attend que des ressources de calcul deviennent disponibles avant d'exécuter la requête. Cela peut se produire lorsque votre compte a dépassé le nombre maximal de demandes d'opérations par lots pouvant s'exécuter simultanément. Si tel est le cas, le statut passe au InProgressmoment où les autres demandes en cours sont terminées ou annulées. OU Amazon ML attend la fin d'une opération par lots dont dépend votre demande.
EN COURS	Votre demande est toujours en cours d'exécution.
TERMINÉ	La demande s'est terminée et l'objet est prêt à être utilisé (modèles d'apprentissage-machine et sources de données) ou affiché (prédictions par lots et évaluations).
ÉCHEC	Il y a un problème avec les données que vous avez fournies ou vous avez annulé l'opération. Par exemple, si vous essayez de calculer des données statistiques sur une source de données qui a échoué avant de se terminer, vous pouvez recevoir un message d'état Non valide ou Erreur. Le message d'erreur explique pourquoi l'opération n'a pas abouti.
SUPPRIMÉ	L'objet a déjà été supprimé.

Amazon ML fournit également des informations sur un objet, par exemple la date à laquelle Amazon ML a fini de créer cet objet. Pour de plus amples informations, veuillez consulter Liste des objets.

Limites du système

Afin de fournir un service fiable et robuste, Amazon ML impose certaines limites sur les demandes que vous envoyez au système. La plupart des problèmes d'apprentissage-machine s'inscrivent dans ces contraintes. Toutefois, si vous trouvez que votre utilisation d'Amazon ML est restreinte par ces limites, vous pouvez contacter le service client AWS et demander à augmenter ces limites. Par exemple, vous pouvez être limité à cinq tâches exécutables simultanément. Si vous constatez que vous avez souvent des tâches en file d'attente, en attente de ressources en raison de cette limite, il est probablement judicieux d'augmenter cette limite pour votre compte.

Le tableau ci-dessous indique les limites par défaut par compte dans Amazon ML. Toutes ces limites ne peuvent pas être augmentées par le service client AWS.

Type de limite	Limite du système
Taille de chaque observation	100 Ko
Taille des données de formation *	100 Go
Taille des données d'entrée de prédiction par lots	1 To
Taille des données d'entrée de prédiction par lots (nombre d'enregis trements)	100 millions
Nombre de variables dans un fichier de données (schéma)	1 000
Complexité de la recette (nombre de variables de sortie traitées)	10 000
TPS pour chaque point de terminaison de prédiction en temps réel	200
TPS totale pour tous les points de terminaison de prévision en temps réel	10 000
RAM totale pour tous les points de terminaison de prévision en temps réel	10 Go
Nombre de tâches simultanées	25
Durée d'exécution maximale de toute tâche	7 jours

Type de limite	Limite du système
Nombre de classes pour les modèles ML multiclasses	100
Taille du modèle ML	Minimum de 1 Mo, maximum de 2 Go
Nombre de balises par objet	50

- La taille de vos fichiers de données est limitée afin de garantir que les tâches se terminent dans un délai raisonnable. Les tâches qui s'exécutent depuis plus de sept jours sont automatiquement arrêtées, entraînant l'état ECHEC.

Noms et IDs pour tous les objets

Chaque objet dans Amazon ML doit avoir un identifiant, ou ID. La console Amazon ML génère des valeurs d'ID pour vous, mais si vous utilisez l'API, vous devez générer vos propres ID. Chaque ID doit être unique parmi tous les objets Amazon ML du même type dans votre compte AWS. En d'autres termes, vous ne pouvez pas avoir deux évaluations portant le même ID. Il est possible d'avoir une évaluation et une source de données avec le même ID, bien que cela ne soit pas recommandé.

Nous vous recommandons d'utiliser des identifiants générés de façon aléatoire pour vos objets, préfixés avec une courte chaîne pour identifier leur type. Par exemple, lorsque la console Amazon ML génère une source de données, elle lui attribue un identifiant unique et aléatoire tel que « WIu WiOx DS-Zsc F ». Cet ID est suffisamment aléatoire pour éviter tout conflit avec un utilisateur individuel, et il est également compact et lisible. Le préfixe « ds- » est utilisé pour des raisons de commodité et de clarté, mais il n'est pas obligatoire. Si vous n'êtes pas certain de ce qu'il convient d'utiliser pour vos chaînes d'ID, nous vous recommandons d'utiliser des valeurs UUID hexadécimales (comme 28b1e915-57e5-4e6c-a7bd-6fb4e729cb23), qui sont disponibles instantanément dans tout environnement de programmation moderne.

Les chaînes d'ID peuvent contenir des lettres ASCII, des chiffres, des tirets et des traits de soulignement, et peuvent comporter jusqu'à 64 caractères. Il est possible et peut-être pratique d'encoder des métadonnées dans une chaîne d'ID. Mais cela n'est pas recommandé, car une fois qu'un objet a été créé, son ID ne peut pas être modifié.

Les noms des objets vous permettent d'associer facilement des métadonnées conviviales à chaque objet. Vous pouvez mettre à jour les noms une fois qu'un objet a été créé. Vous pouvez ainsi faire en sorte que le nom de l'objet reflète un aspect particulier de votre flux de travail d'apprentissage-machine. Par exemple, vous pouvez nommer initialement un modèle d'apprentissage-machine « essai 3 », puis ultérieurement renommer le modèle « modèle de production final ». Un nom peut correspondre à n'importe quelle chaîne de votre choix, avec 1 024 caractères au maximum.

Durées de vie des objets

Tout objet source de données, modèle d'apprentissage-machine, évaluation ou prédiction par lots que vous créez avec Amazon ML sera disponible pour que vous l'utilisiez pendant au moins deux ans après sa création. Amazon ML peut automatiquement supprimer des objets qui n'ont pas été consultés ni utilisés pendant plus de deux ans.

Ressources

Les ressources connexes suivantes peuvent s'avérer utiles lors de l'utilisation de ce service.

- Informations sur les produits Amazon ML : capture toutes les informations pertinentes sur les produits Amazon ML dans un emplacement central.
- Amazon ML FAQs : couvre les principales questions posées par les développeurs à propos de ce produit.
- Exemple de code Amazon ML : exemples d'applications utilisant Amazon ML. Vous pouvez utiliser les exemples de code comme point de départ pour créer vos propres applications d'apprentissage-machine.
- Référence d'API Amazon ML — Décrit en détail toutes les opérations d'API pour Amazon ML. Elle fournit également des exemples de demandes et de réponses pour les protocoles de services web pris en charge.
- Centre de ressources pour développeurs AWS : fournit un point de départ central pour trouver de la documentation, des exemples de code, des notes de publication et d'autres informations qui vous aideront à créer des applications innovantes avec AWS.
- Formations et cours AWS : liens vers des cours spécialisés et basés sur les rôles, ainsi que des ateliers d'autoformation pour vous aider à perfectionner vos compétences AWS et à acquérir une expérience pratique.
- Outils de développement AWS – Liens vers des outils de développement et les ressources qui fournissent une documentation, des exemples de code, des notes de mise à jour et d'autres informations pour vous aider à développer des applications innovantes avec AWS.
- Centre de support AWS — Le centre de création et de gestion de vos dossiers de support AWS. Comprend également des liens vers d'autres ressources utiles, telles que des forums, des informations techniques FAQs, l'état de santé des services et AWS Trusted Advisor.
- AWS Support : page Web principale contenant des informations sur AWS Support one-on-one, un canal d'assistance rapide destiné à vous aider à créer et à exécuter des applications dans le cloud.
- Contactez-nous : point de contact central pour les demandes concernant la facturation AWS, votre compte, les événements, les abus et d'autres problèmes.
- Conditions d'utilisation du site AWS – Informations détaillées sur nos droits d'auteur et notre marque, sur vos compte, licence et accès au site, et sur d'autres sujets.

Historique du document

Le tableau suivant décrit les modifications importantes apportées à la documentation dans cette version d'Amazon Machine Learning (Amazon ML).

- Version API : 2015-04-09
- Dernière mise à jour de la documentation : 02/08/2016

Modification	Description	Date de modification
Ajout de métriques	Cette version d'Amazon ML ajoute de nouvelles métriques pour les objets Amazon ML. Pour de plus amples informations, veuillez consulter Liste des objets.	2 août 2016
Suppression de plusieurs objets	Cette version d'Amazon ML ajoute la possibilité de supprimer plusieurs objets Amazon ML. Pour de plus amples informations, veuillez consulter Suppression d'objets.	20 juillet 2016
Ajout du balisage	Cette version d'Amazon ML ajoute la possibilité d'appliquer des balises aux objets Amazon ML. Pour de plus amples informations, veuillez consulter Marquage de vos objets Amazon ML.	23 juin 2016
Copier des sources de données Amazon Redshift	Cette version d'Amazon ML ajoute la possibilité de copier les paramètres de source de données Amazon Redshift vers une nouvelle source de données Amazon Redshift. Pour plus d'informations sur la copie des paramètres de source de données Amazon Redshift, consultez. Copie d'une source de données (console)	11 avril 2016

Modification	Description	Date de modification
Ajout de la réorganisation	Cette version d'Amazon ML ajoute la possibilité de mélanger vos données d'entrée. Pour plus d'informations sur l'utilisation du paramètre Shuffle type (Type de réorganisation), consultez Type de réorganisation des données de formation.	5 avril 2016
Création de sources de données améliorée avec Amazon Redshift	Cette version d'Amazon ML ajoute la possibilité de tester vos paramètres Amazon Redshift lorsque vous créez une source de données Amazon ML dans la console afin de vérifier que la connexion fonctionne. Pour de plus amples informations, veuillez consulter Création d'une source de données avec Amazon Redshift Data (console).	21 mars 2016
Conversion améliorée du schéma de données Amazon Redshift	Cette version d'Amazon ML améliore la conversion des schémas de données Amazon Redshift (Amazon Redshift) en schémas de données Amazon ML. Pour plus d'informations sur l'utilisation d'Amazon Redshift avec Amazon ML, consultez. Création d'une source de données Amazon ML à partir des données d'Amazon Redshift	9 février 2016
CloudTrail journalisation ajoutée	Cette version d'Amazon ML ajoute la possibilité de consigner les demandes à l'aide de AWS CloudTrail (CloudTrail). Pour plus d'informations sur l'utilisation de la CloudTrail journalisation, consultezJournalisation des appels d'API Amazon ML avec AWS CloudTrail.	10 décembre 2015

Modification	Description	Date de modification
DataRearrangement Options supplémentaires ajoutées	Cette version d'Amazon ML ajoute la possibilité de diviser vos données d'entrée de manière aléatoire et de créer des sources de données complémentaires. Pour plus d'informations sur l'utilisation du `DataRearrangement` paramètre, consultezRéorganisation des données. Pour obtenir des informations sur la façon d'utiliser les nouvelles options pour la validation croisée, consultez Validation croisée.	3 décembre 2015
Essai d'utilisation des prédictions en temps réel	Cette version d'Amazon ML permet d'essayer des prédictions en temps réel dans la console de service. Pour plus d'informations sur la façon d'essayer des prédictions en temps réel, consultez Demande de prédiction en temps réel le manuel Amazon Machine Learning Developer Guide.	19 novembre 2015
Nouvelle région	Cette version d'Amazon ML ajoute le support pour la région UE (Irlande). Pour plus d'informations sur Amazon ML dans la région UE (Irlande), Régions et points de terminaison consultez le manuel Amazon Machine Learning Developer Guide.	20 août 2015
Première version	Il s'agit de la première version du manuel Amazon ML Developer Guide.	9 avril 2015