Previsão da demanda por capacidade de frete usando a Amazon AI SageMaker

AWS Recomendações

AWS Recomendações: Previsão da demanda por capacidade de frete usando a Amazon AI SageMaker

Table of Contents

Previsão da demanda por capacidade de frete usando a Amazon AI SageMaker

Tianxia Jia e Hengzhi Chen, Amazon Web Services ()AWS

Maio de 2024 (histórico do documento)

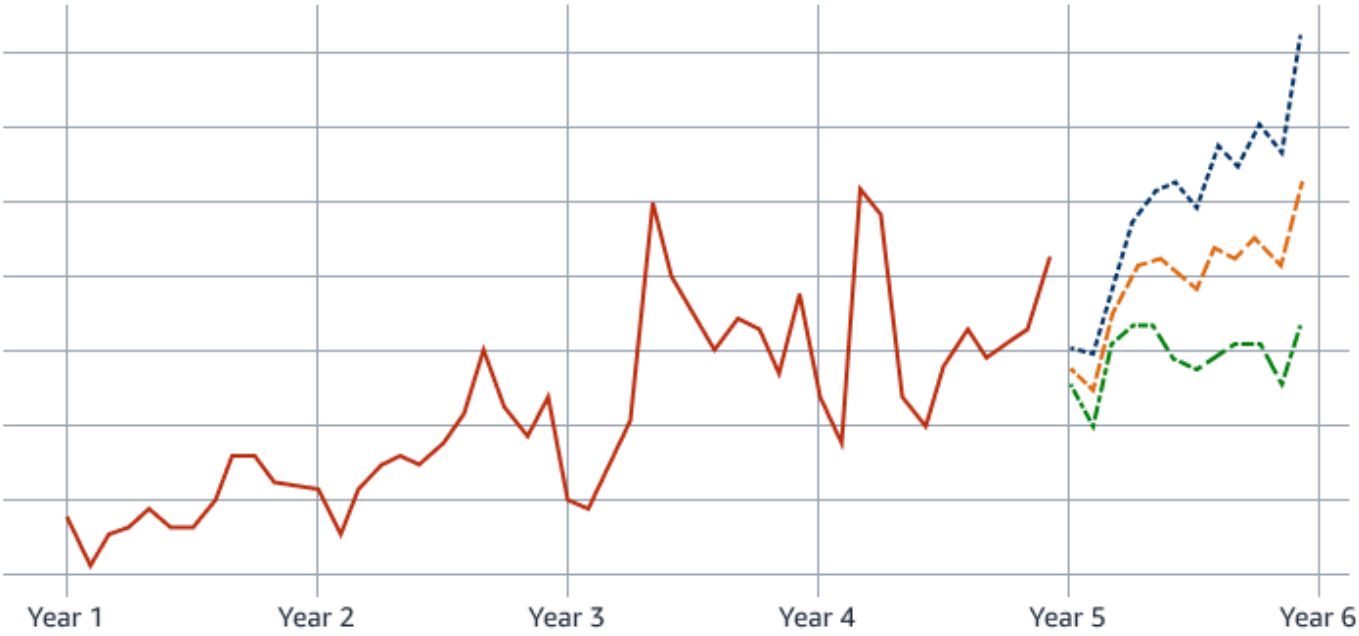
A previsão de demanda é fundamental no setor de transporte e logística, especialmente durante períodos de restrições na cadeia de suprimentos. Estimativas precisas da demanda de frete beneficiam empresas envolvidas na logística e na cadeia de suprimentos, como aquelas que transportam contêineres e cargas aéreas através das fronteiras. Isso ajuda as empresas a gerenciar com eficácia o custo de proteger sua rede de transporte, o que as ajuda a gerenciar os custos de envio e maximizar a receita e o lucro.

Um modelo de aprendizado de máquina (ML) capaz de fazer uma previsão precisa depende de dados de treinamento de alta qualidade. Para a previsão da demanda, os dados de treinamento podem incluir o volume histórico da demanda e outros dados gerados internamente que podem estar relacionados ao volume, como preço, estoque e número de funcionários da equipe de vendas. Além disso, dados externos, como concorrentes, ambiente de mercado, feriados, clima e macroeconomia, também podem afetar o volume da demanda. Esses fatores de dados internos e externos podem ser usados como recursos em um modelo de ML.

Depois que todos os recursos forem identificados, a empresa também pode querer fornecer informações sobre alguns desses recursos que estão sob seu controle. Por exemplo, a empresa pode definir o preço do frete com antecedência ou decidir quando fazer promoções e descontos. Esses tipos de entradas do usuário podem ser incorporados ao modelo ao fazer a previsão.

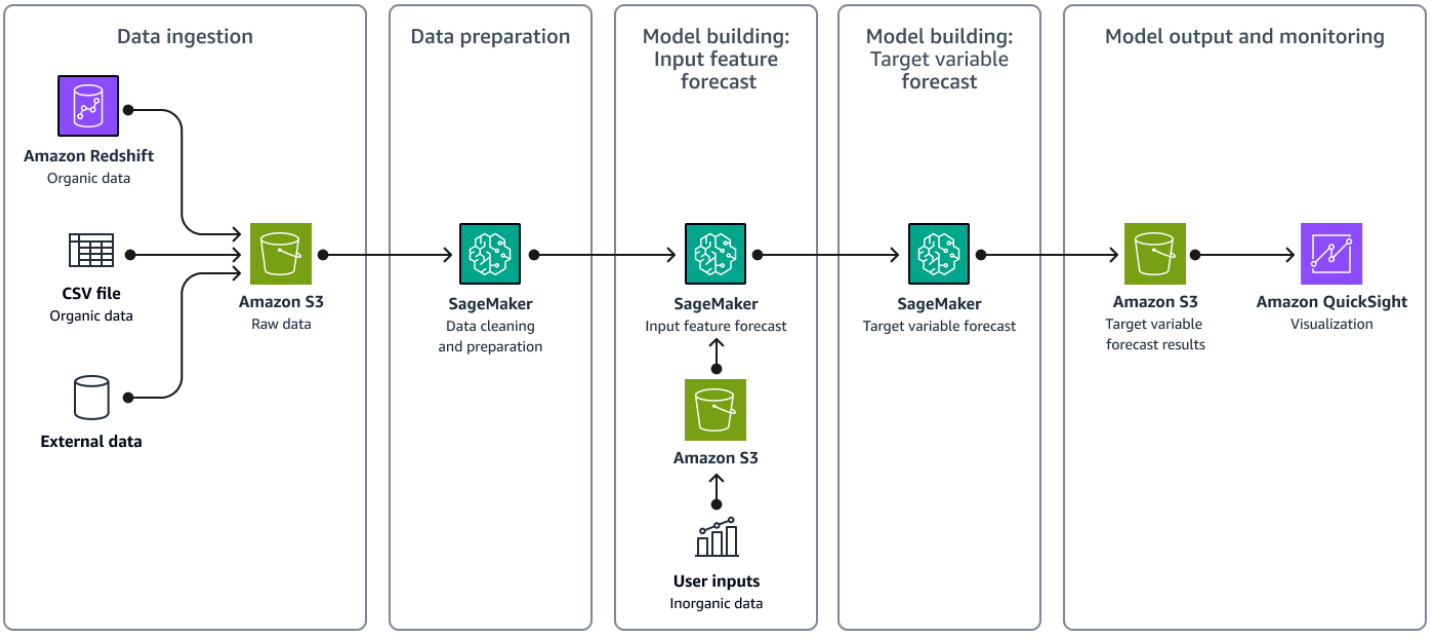
Este guia descreve uma estratégia para criar uma solução AWS que faça uma previsão precisa da demanda logística usando um modelo de ML. Você treina o modelo em um conjunto de dados histórico que contém volume de demanda e características relacionadas à demanda. Esses recursos e métricas incluem dados orgânicos internos e dados externos. A solução também oferece flexibilidade para que o usuário e o analista de negócios forneçam informações, que podem ser incorporadas ao modelo de previsão.

A imagem a seguir mostra um exemplo de uma série temporal histórica e um intervalo previsto de 12 meses. Você pode usar as recomendações deste guia para criar um modelo de ML que produza esse tipo de previsão.


Year 1
Year 2
Year 3
Year 4
Year 5
Year 6

Arquitetura para prever a demanda de frete

A imagem a seguir mostra o fluxo de trabalho da solução, incluindo ingestão de dados, preparação de dados, criação de modelos e saída e monitoramento finais.



A arquitetura da solução inclui os seguintes componentes principais:

1. Ingestão de dados — Você armazena dados orgânicos e dados externos no Amazon Simple Storage Service (Amazon S3).
2. Preparação de dados — A Amazon SageMaker AI limpa os dados e os prepara para o treinamento do modelo de ML. Para obter mais informações, consulte Preparar dados na documentação de SageMaker IA.
3. Construção do modelo: previsão do recurso de entrada — a SageMaker IA usa Prophetpara gerar uma previsão de série temporal para cada recurso de entrada. Você examina os resultados da previsão. Se necessário, você fornece informações do usuário para substituir a previsão do recurso.
4. Construção do modelo: previsão da variável alvo — A SageMaker IA cria um modelo de regressão para inferência usando os recursos de entrada modificados.
5. Saída e monitoramento do modelo — O modelo de regressão envia os resultados da previsão para o Amazon S3. Você pode visualizar a previsão no Amazon Quick. Os analistas podem

monitorar os resultados da previsão e avaliar a precisão comparando a previsão com o volume real da demanda.

Todo o pipeline de processamento, desde a ingestão de dados até a saída final do modelo, pode ser orquestrado para ser executado automaticamente. Por exemplo, você pode configurá-lo para ser executado automaticamente mensalmente para uma previsão de demanda mensal. Se precisar de previsões para mais de um produto, você pode executar o pipeline em paralelo para vários produtos. Para obter mais informações, consulte Implementar MLOPs na documentação de SageMaker IA.

Dados para previsão da demanda de frete

Dados de alta qualidade são essenciais para que qualquer modelo de ML faça previsões e previsões significativas. Para previsões de demanda, o conjunto de dados consiste em qualquer dado relevante que possa afetar a demanda final. Esses dados podem vir de várias fontes. Você pode classificar esses dados em duas categorias, dados internos e externos.

Dados internos

Os dados internos são dados orgânicos gerados pela empresa. Esses dados geralmente são armazenados em um data warehouse, como o Amazon Redshift.

Você pode gerar ou extrair diretamente os valores de saída de destino das tabelas no data warehouse que contêm volumes históricos de produtos de interesse. Para companhias marítimas, as saídas ou os valores-alvo podem estar em unidades de carga total de contêineres para transporte marítimo ou peso total para carga aérea.

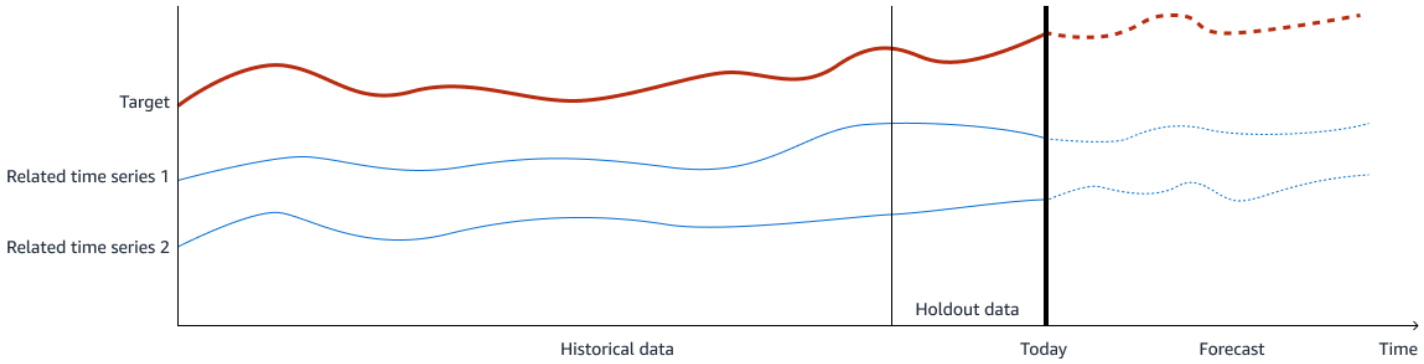
Você também pode gerar várias métricas comerciais históricas. Eles podem ser usados como recursos no modelo de aprendizado de máquina ao prever a demanda. Exemplos de recursos incluem preço histórico, custo, capacidade e estoque.

Dados externos

Fontes de dados externas podem ser usadas como recursos adicionais para melhorar a precisão da previsão. Exemplos de fontes de dados externas incluem dados meteorológicos, dados macroeconômicos, dados do setor e dados de mercado. Esses fatores podem ter impacto direto ou indireto no setor de logística e transporte, afetando, portanto, a demanda. Por exemplo, a taxa de frete do mercado fornece uma referência do mercado global de frete, o que acaba afetando a demanda específica da empresa. Dados macroeconômicos, como dados de importação e exportação das principais economias, também podem ser usados como uma medida da atividade do mercado. Para incorporar essas fontes de dados externas, você pode usar várias APIs para ingerir dados. Por exemplo, St. Louis Fed fornece o acesso Federal Reserve Economic Data (FRED) APIa dados macroeconômicos e National Oceanic and Atmospheric Administration (NOAA) fornece o acesso Climate Data Online (CDO) APIa dados meteorológicos mundiais.

Modelos de aprendizado de máquina para prever a demanda de frete

A imagem a seguir mostra um exemplo dos dados de treinamento. O alvo é o que você deseja prever, e as séries temporais 1 e 2 relacionadas são recursos de entrada relevantes para prever o alvo. Os dados históricos são usados para treinamento e validação, e você retém um período dos dados históricos para validação do modelo.



Na previsão de demanda, a produção (ou meta) é o volume de demanda que você deseja prever. Os recursos de entrada são dados de séries temporais relacionados à saída. Para treinar um modelo de ML para fazer uma previsão precisa do volume de demanda, dois modelos de aprendizado de máquina são necessários na solução. O primeiro modelo faz uma previsão de séries temporais para os recursos de entrada, incluindo dados internos e externos. O segundo modelo faz a previsão final da demanda usando todos os recursos. Ao usar esses dois modelos juntos, você pode capturar com eficácia a tendência da série temporal e a relação entre o alvo e as entradas.

Modelo de ML para a previsão do recurso de entrada

Os recursos de entrada incluem dados históricos de séries temporais internos e externos. Para fazer previsões para cada recurso, você pode usar um modelo de série temporal unidimensional (1D). Existem vários algoritmos disponíveis. Por exemplo, Prophetfunciona melhor com séries temporais com fortes efeitos sazonais e várias temporadas de dados históricos. Uma previsão é gerada para cada recurso individual.

Modelo de ML para a previsão da variável alvo

O modelo ML para a saída, ou o volume de demanda, foi criado para capturar a relação entre todos os recursos e a saída. Você pode usar vários modelos de regressão supervisionada, comolasso, ridge regressionrandom forest, e. XGBoost Ao criar o modelo e encontrar os melhores parâmetros e hiperparâmetros, você pode usar dados de retenção. Os dados de retenção são uma parte dos dados históricos rotulados que são retidos do conjunto de dados usado para treinar um modelo de aprendizado de máquina. Você pode usar dados de retenção para avaliar o desempenho do modelo comparando as previsões com os dados de retenção.

Entradas do usuário para prever a demanda de frete

Embora o futuro da maioria dos recursos seja desconhecido, pode haver alguns recursos sobre os quais a empresa tenha controle. Por exemplo, o preço é uma característica que geralmente tem uma forte relação com o volume de demanda. Como a empresa define os preços, você sabe quando os preços aumentarão ou quando haverá um desconto ou promoção. Além disso, o tamanho da equipe de vendas também pode afetar o volume de demanda, e as empresas controlam o tamanho de suas equipes de vendas. Para esses pontos de dados que a empresa gerencia, você pode fornecer informações do usuário. Na etapa de modelagem de ML, os modelos de séries temporais 1D fornecem uma previsão para cada recurso e, em seguida, os usuários podem examinar esses valores previstos e substituí-los pelas entradas do usuário. Essas entradas sobrescritas são então usadas no modelo ao fazer a previsão final.

Essa etapa de entrada do usuário pode ser crítica em situações em que as previsões do modelo de série temporal 1D não correspondem às expectativas da empresa em relação ao comportamento futuro do recurso. Você pode sobrescrever esses valores previstos, o que pode melhorar a previsão geral da produção.

Melhores práticas para um modelo de ML que prevê a demanda de frete

Seguindo essas melhores práticas, você pode aprimorar a precisão, a confiabilidade e a interpretabilidade do seu modelo de aprendizado de máquina para prever a demanda de frete, levando a uma melhor tomada de decisão e eficiência operacional:

- Qualidade e pré-processamento de dados — Certifique-se de que os dados usados para treinar o modelo sejam de alta qualidade e livres de erros, valores faltantes e inconsistências. As etapas de pré-processamento de dados, como tratamento de valores ausentes, detecção de valores discrepantes e engenharia de recursos, desempenham um papel crucial na melhoria da precisão do modelo.
- Dados históricos suficientes — Ter dados históricos suficientes é essencial para capturar padrões, tendências e sazonalidade. No entanto, também é importante considerar a relevância e a atualidade dos dados históricos. Se houve mudanças significativas no mercado, nas operações comerciais ou em fatores externos, os dados mais antigos podem não ser representativos do cenário atual. Nessa situação, dê maior peso aos dados mais recentes.
- Seleção e engenharia de recursos — Identificar recursos relevantes e criar novos recursos a partir de dados existentes pode melhorar significativamente o desempenho do modelo. Colabore estreitamente com especialistas do domínio para usar seus conhecimentos e insights ao selecionar os recursos apropriados. Além disso, considere realizar uma análise da importância dos recursos para identificar os recursos mais influentes e, potencialmente, remover recursos redundantes ou irrelevantes.
- Modelos de conjunto — Em vez de confiar em um único modelo, considere usar técnicas de conjunto que combinem as previsões de vários modelos. Os modelos de conjunto podem superar os modelos individuais e fornecer previsões mais robustas e precisas.
- Avaliação e validação do modelo — Avalie e valide regularmente o desempenho do modelo usando métricas apropriadas, como erro quadrático médio (MSE), erro percentual absoluto médio (MAPE) ou qualquer outra métrica específica do domínio. Use validação cruzada ou validação de retenção para avaliar os recursos de generalização do modelo.
- Monitoramento e reciclagem contínuos — Os padrões de demanda de frete podem mudar com o tempo devido a vários fatores, como condições econômicas, dinâmica do mercado ou mudanças nas operações comerciais. Monitore continuamente o desempenho do modelo e treine-o periodicamente com os dados mais recentes para melhorar sua precisão e relevância.

- IA explicável — Os modelos de previsão de demanda devem ser interpretáveis e explicáveis, especialmente nos casos em que as partes interessadas precisam entender o raciocínio por trás das previsões. Técnicas como análise de importância de características, gráficos de dependência parcial e explicações aditivas de Shapley (SHAP) podem ajudar a explicar as decisões do modelo.
- Incorpore o conhecimento do domínio — Colabore estreitamente com especialistas do domínio e partes interessadas da empresa para incorporar seus conhecimentos e insights ao processo de modelagem. Sua experiência no domínio pode ajudar a identificar possíveis vieses, interpretar resultados e tomar decisões informadas com base nas previsões.
- Análise de cenários e simulações hipotéticas — incorpore a capacidade de realizar análises de cenários e simulações hipotéticas à solução de previsão. Isso permite que as partes interessadas explorem o efeito de diferentes decisões de negócios ou fatores externos na previsão da demanda, permitindo uma tomada de decisão mais informada.
- Pipeline automatizado e escalável — Crie um pipeline automatizado e escalável para ingestão de dados, pré-processamento, treinamento de modelos e implantação. Isso executa o processo de previsão de forma consistente e eficiente, especialmente ao lidar com vários produtos ou regiões.

Próximas etapas

Antes de implementar essa solução de previsão de demanda AWS, é recomendável avaliar o problema que está tentando resolver. É uma boa ideia reunir os proprietários da empresa e os cientistas de dados para debater se o problema pode ser resolvido por meio de um modelo de ML. É fundamental entender quais conjuntos de dados você tem e a extensão dos dados históricos disponíveis. Também é importante que os proprietários da empresa colaborem com os cientistas de dados para fornecer conhecimento do domínio, identificar recursos úteis e ajudar a criar esses recursos. A confiabilidade do modelo aumenta com o número de recursos relevantes que você pode criar, o que fornece uma previsão mais precisa.

Para desenvolver essa arquitetura AWS, comece configurando Conta da AWS e provisionando os serviços necessários, como o Amazon Simple Storage Service (Amazon S3) para armazenamento de dados e o SageMaker Amazon AI para treinamento de modelos de aprendizado de máquina. Em seguida, identifique e colete as fontes de dados internas e externas que serão usadas como recursos de entrada para o modelo de previsão. Armazene esses dados no Amazon S3 e use os recursos de processamento de dados na SageMaker IA para pré-processar e preparar os dados para o treinamento do modelo. Na SageMaker IA, use o ajuste automático do modelo e os recursos de treinamento distribuído para treinar e otimizar os modelos de previsão. Você também pode usar Serviços da AWS como AWS Step Functions ou AWS Lambda para configurar um pipeline que retreine periodicamente os modelos de previsão com os dados mais recentes. Depois de treinar novamente, inicie um trabalho de transformação em lote na SageMaker IA para gerar os resultados da previsão, que você armazena no Amazon S3. Use o Amazon Quick para visualizar e monitorar os resultados da previsão gerados a partir do trabalho de transformação em lote.

Recursos

Documentação do Amazon Quick

- O que é o Amazon Quick?
- Trabalhando com fontes de dados

Documentação da Amazon SageMaker AI

- O que é Amazon SageMaker AI?
- Prepare os dados
- Execute o ajuste automático do modelo
- Use a transformação em lote

AWS exemplos de código

- Repositório de exemplos de SageMaker IA da Amazon () GitHub

Histórico do documento

A tabela a seguir descreve alterações significativas feitas neste guia. Se quiser ser notificado sobre futuras atualizações, você pode assinar um RSSfeed.

Alteração	Descrição	Data
Publicação inicial	—	14 de maio de 2024

As traduções são geradas por tradução automática. Em caso de conflito entre o conteúdo da tradução e da versão original em inglês, a versão em inglês prevalecerá.