02/09/2024

Relatório

1 - Objetivo do projeto

Construir uma solução de análise moderna para a Adventure Works, uma indústria de bicicletas, utilizando dados de vendas e produtos. A missão é fornecer os meios para transformar a empresa em uma organização orientada por dados, capacitando a tomada de decisões estratégicas e operacionais por meio de uma plataforma, baseada nos princípios do Modern Data Stack. O projeto será dividido em três partes principais:

1. Modelagem dimensional com dbt
2. Visualização de dados
3. Análise de dados e previsão de demanda

Principais interessados e benefícios esperados

- João Muller – Diretor de Inovação: Como patrocinador do projeto, João Muller busca modernizar a infraestrutura de dados, posicionando a empresa como líder em inovação e tecnologia no setor.
- Carlos Silveira – CEO: Carlos Silveira vê o projeto como um diferencial estratégico, que fortalecerá a vantagem competitiva da empresa e a sua posição no mercado.
- Silvana Teixeira – Diretora Comercial: Silvana está interessada em observar resultados tangíveis que comprovem a eficácia da transformação para uma cultura orientada por dados e seu impacto nas ações comerciais.
- Nilson Ramos – Diretor de TI: Nilson espera que o projeto facilite o acesso e a gestão dos dados, otimizando a infraestrutura de TI e simplificando o processo de obtenção de informações.
- Gabriel Santos – Analista de TI: Gabriel busca melhorias na administração dos bancos de dados, com ênfase na manutenção da integridade e na eficiência da gestão das informações.
- Luís Soares – Gestor de Planejamento de Demanda: Luís espera que o projeto forneça análises preditivas robustas para otimizar o planejamento de produção, ajudando na tomada de decisões e na contratação de soluções completas para implementação.

Divisão do projeto

1. Transformação dos dados:
    - Modelagem e transformação de dados para o data warehouse.
    - Criação de tabelas de fatos e dimensões para responder às perguntas de negócios.
    - Implementação de boas práticas de modelagem dimensional.

2. Visualização dos dados:
    - Desenvolvimento de dashboards interativos para visualização dos dados.
    - Aplicação de boas práticas de visualização para facilitar a análise e interpretação dos dados.
    - Criação de relatórios que atendam às necessidades das partes interessadas.

3. Análise de previsão:
    - Desenvolvimento e validação de modelos preditivos para prever a demanda de produtos.
    - Análise de sazonalidade e comparação de crescimento entre diferentes regiões.
    - Utilização de técnicas estatísticas e machine learning para melhorar a precisão das previsões.

2 - Ferramentas utilizadas

1. Coleta e armazenamento de dados
    - Fonte dos dados: Infraestrutura SAP, CRM Salesforce, web analytics e outros.
    - Plataforma de armazenamento: Google BigQuery
2. Transformação de dados
    - Ferramenta utilizada: dbt core
3. Visualização e dashboard
    - Ferramenta utilizada: Power BI
4. Análise de dados e previsão
    - Ferramenta utilizada: Python

3 - Modelagem dimensional

Para construir a infraestrutura moderna de analytics para a Adventure Works, optou-se pelo Google BigQuery como a plataforma para configurar o data warehouse. A modelagem dos dados foi realizada com base no dicionário de dados disponível publicamente, o Adventure Works dataedo, que forneceu as informações necessárias para estruturar o modelo de dados de forma eficiente.

Modelo Dimensional

Fontes de Dados

As tabelas de dimensões são construídas a partir das tabelas staging do banco de dados da Adventure Works, garantindo que os dados sejam preparados e limpos antes de serem incorporados ao modelo dimensional. As chaves substitutas (SK) são geradas para garantir a unicidade dos registros e são utilizadas para criar relacionamentos entre as tabelas de dimensões e fatos.

Tabelas de Dimensões

- **dim_dates**: Representa as datas de forma detalhada, seguindo as boas práticas descritas no Date dimension Medium da Indicium . Inclui informações como ano, mês, dia, e atributos relacionados a datas que facilitam análises temporais.
- **dim_customer**: Contém informações sobre os clientes.
- **dim_locales**: Armazena dados sobre as localizações geográficas, que podem incluir região, cidade e país.
- **dim_personcreditcard**: Inclui detalhes sobre os cartões de crédito utilizados nas compras.
- **dim_products**: Detalha os produtos vendidos, com informações como ID do produto, nome, categoria e preço.
- **dim_salesreason**: Explica o motivo das vendas.

Tabelas de fato e agregadas

- **fct_ordersales**: Combina informações das tabelas salesorderdetail e salesorderheader. Cada registro na tabela de fatos representa um pedido, com detalhes como quantidade e preço, possibilitando o cálculo de receita e análise de desempenho de vendas.
- **agg_sales_locales**: Tabela agregada que sumariza as vendas por localizações, permitindo análises de desempenho por região e vendedor.
- **agg_sales_seller**: Tabela agregada que resume as vendas por vendedor, oferecendo uma visão detalhada sobre o desempenho de vendas.

Chaves e Relacionamentos

- Chaves Substitutas (SK): Criadas para assegurar a unicidade nas tabelas de dimensões e usadas para mapear as chaves estrangeiras (FK) nas tabelas de fatos.
- Relacionamentos: Os relacionamentos entre as tabelas de fatos e dimensões são preservados através do mapeamento das chaves substitutas e estrangeiras, garantindo a integridade referencial do modelo dimensional.

Materialização

Optou-se pela materialização do tipo `tabela` para armazenar os dados de forma persistente no data warehouse. Esta abordagem permite que os dados sejam rapidamente acessados e consultados, eliminando a necessidade de recalcular ou recriar os dados a cada consulta.

Documentação e testes

- Foi criado um arquivo YAML para cada dimensão, contendo uma descrição detalhada do modelo e dos atributos de cada coluna.
- Cada arquivo YAML também inclui a definição de testes para validar a integridade e a precisão dos dados nas tabelas dimensionais. Esses testes verificam aspectos como a unicidade dos registros, a conformidade com os formatos esperados e a consistência dos dados.

3 - Visualização de dados

Foram desenvolvidos dois dashboards, um de nível operacional e outro de nível executivo, ambos voltados para apoiar a tomada de decisões estratégicas.

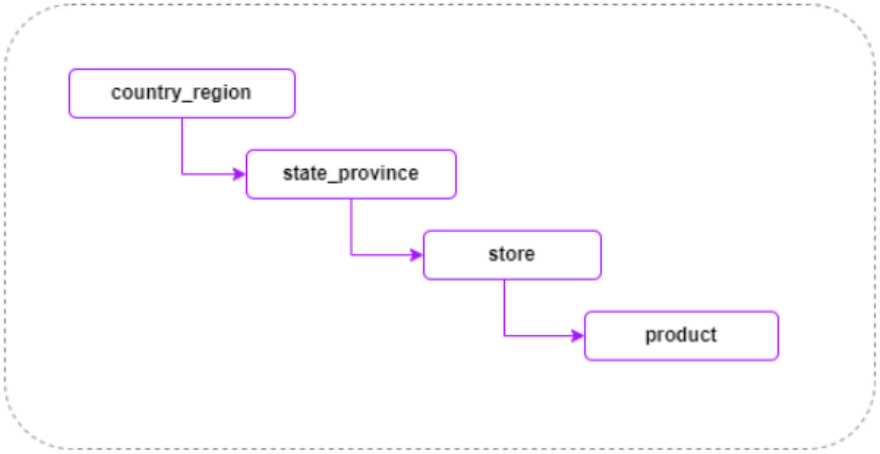
A explicação completa dos dashboards pode ser conferida neste link para o vídeo de apresentação do dashboard.

4 - Análise de dados e previsão de demanda

Previsão de Demanda e Análise de Sazonalidade

Para a previsão da demanda dos próximos 3 meses e a identificação da sazonalidade, foi utilizado um modelo de séries temporais hierárquicas. Foi constatado que alguns registros estavam com a feature de nome da loja ausente. Em tais casos, os pedidos foram identificados como realizados online e rotulados adequadamente.

Uma análise temporal das vendas revelou uma tendência geral de crescimento. Para a modelagem, definimos um nível hierárquico que inclui região, estado, loja e produto. A estrutura hierárquica pode ser visualizada no diagrama abaixo:

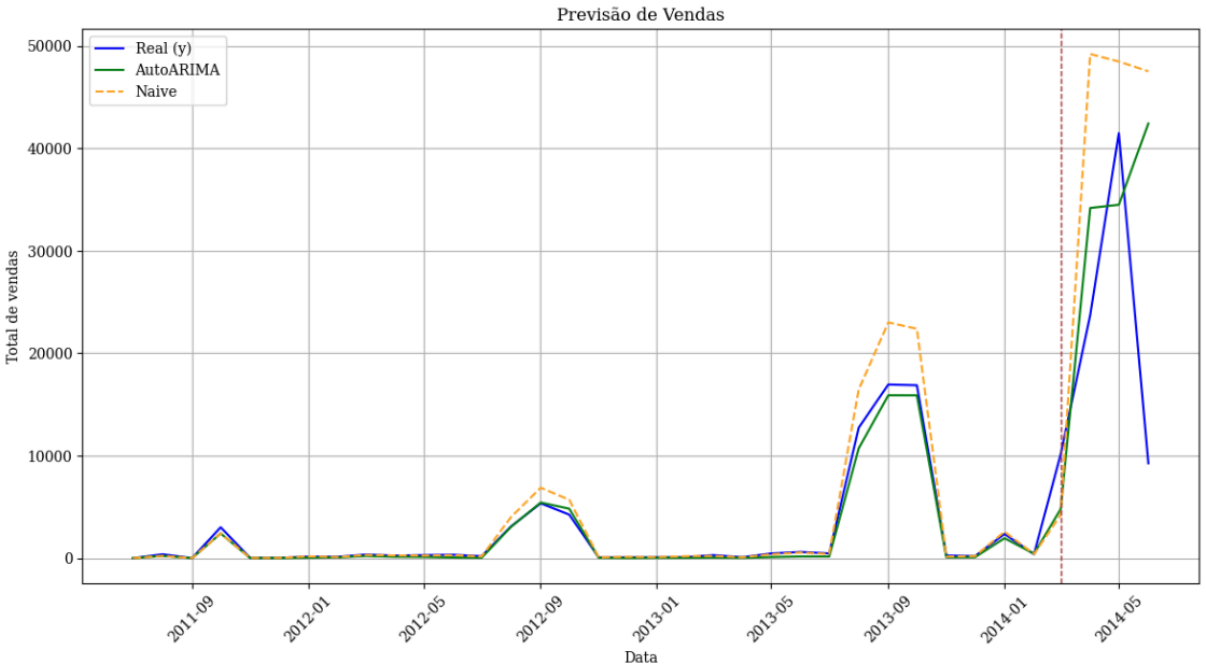


Modelagem e Resultados

O modelos de séries temporais hierárquicas foi aplicado utilizando o pacote statsforecast, adotando o modelo Naive como baseline e o AutoARIMA para as previsões. Seguimos as etapas recomendadas no curso de Previsão de Demanda, apresentado durante uma Tech Friday. A seleção desses modelos foi fundamentada em suas características específicas:

- Naive: Utilizado como baseline, é eficaz mesmo com dados esparsos, oferecendo um ponto de referência simples para comparar com modelos mais avançados.
- AutoARIMA: Eficaz na modelagem de padrões e sazonalidades, mesmo com variações no número de observações.

Visualização temporal da previsão



Os resultados das métricas de avaliação para ambos os modelos foram:

Modelo	MAE	MSE	RMSE
AutoARIMA	9.83	15226.29	123.39
Naive	11.52	27072.95	164.54

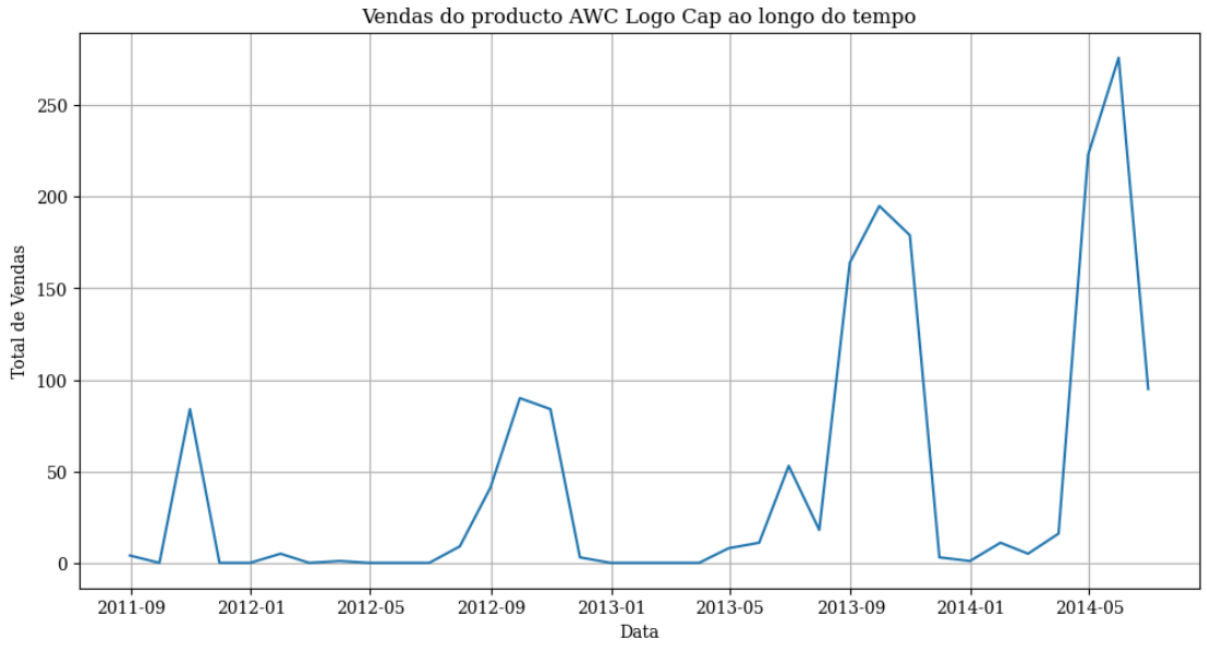
- MAE (Erro Absoluto Médio): Mede a média das diferenças absolutas entre as previsões e os valores reais.
- MSE (Erro Quadrático Médio): Mede a média dos erros quadráticos. Penaliza mais fortemente grandes erros.
- RMSE (Raiz do Erro Quadrático Médio): É a raiz quadrada do MSE. Fornece uma medida da magnitude média dos erros em unidades originais, facilitando a interpretação.

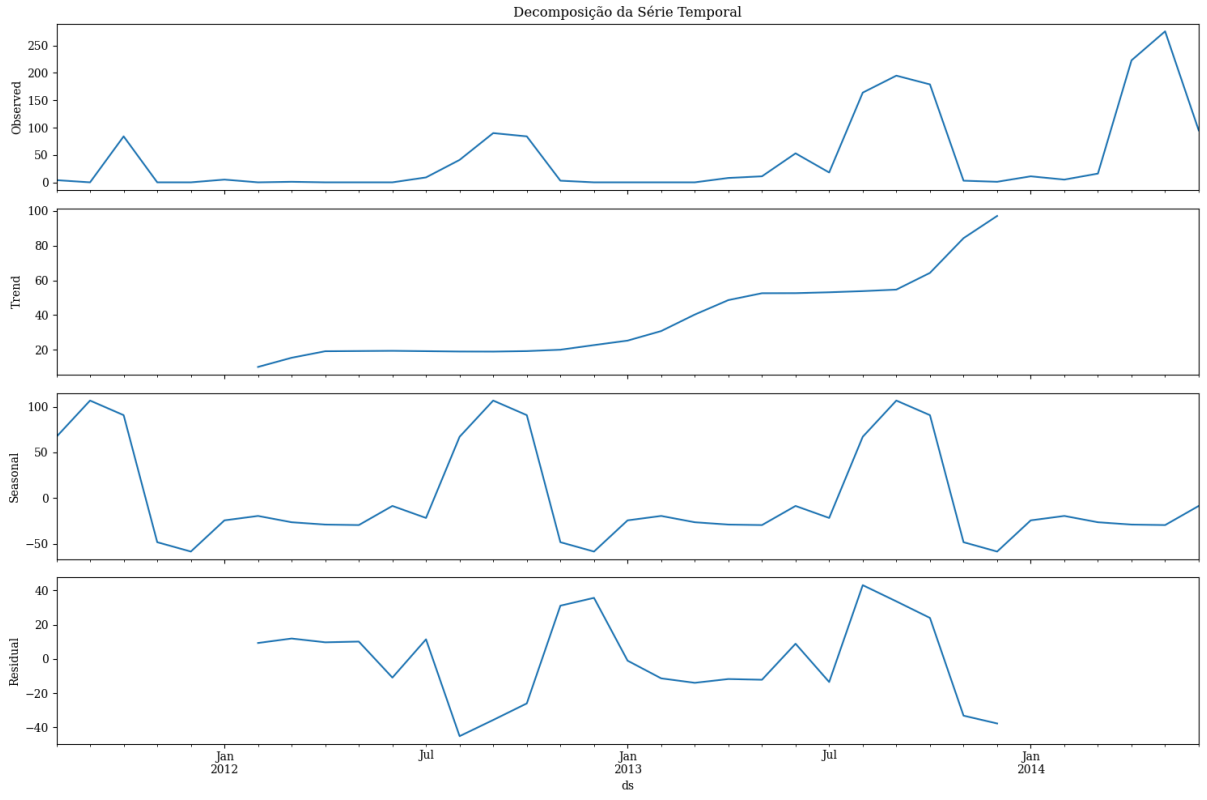
Comparando os gráficos das previsões com os dados reais, o modelo AutoARIMA se aproximou mais dos dados reais na tendência de subida, mas não conseguiu capturar bem a queda subsequente.

Análise de Sazonalidade

Focamos no produto mais vendido, o "**AWC Logo Cap**". Que pode trazer conhecimento estratégico para a empresa. O gráfico de sazonalidade revelou um padrão com amplitude de aproximadamente 100 unidades.

As vendas máximas atingem 250 unidades, e a amplitude sazonal representa cerca de 40% do pico máximo de vendas. A sazonalidade foi identificada principalmente nos meses de outubro e novembro, coincidindo com uma proximidade de períodos de alta demanda como Black Friday e Natal.





Modelo de Regressão

Com os dados hierarquizados pelo método ‘aggregate’ da biblioteca ‘scikit-hts’, optamos por utilizar o XGBoost Regressor devido à sua habilidade superior em modelar relações complexas e não lineares. Para garantir que a divisão dos dados por nível hierárquico seja adequada, estabelecemos que cada grupo deve conter pelo menos 4

registros. Essa abordagem assegura que haja um número suficiente de dados para o treinamento e validação do modelo em cada nível hierárquico, o que é fundamental para obter previsões confiáveis.

Para o XGBoost Regressor, foram criadas features adicionais de defasagens (lags) para capturar a dinâmica temporal das séries temporais. As previsões foram realizadas para um horizonte de 3 meses, e utilizamos as mesmas métricas de avaliação aplicadas ao modelo de séries temporais hierárquicas, como MAE, MSE e RMSE. Podemos comparar na tabela abaixo:

Modelo	MAE	MSE	RMSE
AutoARIMA	9.83	15226.29	123.39
Naive	11.52	27072.95	164.54
XGBoostRegressor	8.21	10785.31	103.85

Com base nas métricas de avaliação, o XGBoost Regressor apresentou os menores valores de MAE, MSE e RMSE, indicando que, para este conjunto de dados específico, ele conseguiu fornecer previsões mais precisas e com menor erro do que o AutoARIMA.

Crescimento de Demanda por Região

Para analisar o crescimento da demanda entre os grupos de províncias dos EUA e outros países, comparamos a demanda prevista para os últimos três meses com a demanda real dos três meses anteriores. Os resultados são apresentados na tabela abaixo:

Grupo	Demanda Anterior	Demanda Prevista	Crescimento
Outros Países	13595	62855.23	362%
EUA Províncias	16617	48248.82	190%

O grupo "Outros Países" apresentou um crescimento percentual maior em demanda, com 362% em comparação aos 190% das províncias dos EUA.

Estimativa de Zíperes Necessários

Para estimar o número de zíperes necessários para os próximos 3 meses, filtramos os dados da tabela de previsões para produtos que contêm a palavra "luva". Considerando que cada par de luvas necessita de 2 zíperes, chegamos à seguinte conclusão:

- **Demanda de Luvas:** 261 pares
- **Zíperes Necessários:** 261 x 2 = 522

Portanto, o fabricante deve solicitar aproximadamente 522 zíperes para atender à demanda projetada.