Nome: Guilherme Montes de Luca

Curso/Turma: Ciência da Computação

Data de Entrega: 25/06/2025

1 Introdução

No cenário atual de alta competitividade no mercado, a capacidade de prever a demanda futura de produtos tornou-se um diferencial estratégico fundamental para empresas que buscam otimizar sua cadeia de suprimentos, reduzir custos e aumentar a eficiência operacional. Com o avanço da ciência de dados e a disponibilidade de dados ricos e variados, tornou-se possível construir modelos preditivos que auxiliam na tomada de decisões de forma mais precisa e fundamentada.

Este trabalho tem como tema a **previsão de demanda futura de produtos** com base em dados históricos de vendas e fatores contextuais diversos, como promoções, sazonalidade, índice econômico, variações de preço, clima e tipo de loja. O dataset analisado contém registros de vendas de produtos ao longo do tempo, incorporando variáveis que influenciam diretamente o comportamento do consumidor e o desempenho comercial.

2 Objetivo

O **objetivo principal** deste projeto é aplicar técnicas de ciência de dados para analisar os padrões de vendas e desenvolver um modelo preditivo capaz de estimar a demanda futura (*future_demand*) dos produtos. Para isso, serão realizadas etapas de limpeza e tratamento dos dados, análise exploratória, aplicação de modelos estatísticos e de aprendizado de máquina, e discussão dos resultados obtidos. A proposta visa demonstrar a importância e aplicabilidade da ciência de dados no apoio à gestão comercial e à previsão de mercado.

3 Descrição do Conjunto de Dados

O conjunto de dados utilizado neste trabalho é composto por 5.000 registros, cada um representando informações relacionadas à venda de um determinado produto em uma data específica. As variáveis presentes no dataset abrangem aspectos temporais, comerciais, econômicos, climáticos e categóricos que influenciam o comportamento de compra dos consumidores.

Variáveis de entrada:

- **data**: data da venda.
- **product_id**: identificador único do produto.
- **sales_units**: número de unidades vendidas no dia.

- **holiday_season**: indicador binário se o dia foi durante uma temporada de feriados.
- **promotion_applied**: indicador binário se havia promoção aplicada.
- **competitor_price_index**: índice de preços da concorrência.
- **economic_index**: indicador da situação econômica.
- **weather_impact**: impacto das condições climáticas nas vendas.
- **price**: preço unitário do produto.
- **discount_percentage**: percentual de desconto aplicado.
- **sales_revenue**: receita gerada pelas vendas.
- **region_Europe**, **region_North America**: variáveis booleanas indicando a região da venda.
- **store_type_Retail**, **store_type_Wholesale**: tipo de loja (varejo ou atacado).
- **category_Cabinets**, **category_Chairs**, **category_Sofas**, **category_Tables**: categorias de produtos.

Variável-alvo:

- **future_demand**: quantidade estimada de demanda futura para o produto em questão.

Este conjunto de dados oferece uma base sólida para explorar relações entre variáveis e construir modelos de previsão de demanda, com potencial para apoiar estratégias comerciais baseadas em dados.

4 Pré-Processamento dos Dados

4.1 Etapas de Limpeza e Tratamento dos Dados

O processo de preparação dos dados envolveu as seguintes etapas:

- Conversão da coluna `date` para o formato `datetime`, possibilitando a extração de atributos temporais como `mês`, `dia da semana` e `fim de semana`.
- Conversão de variáveis booleanas para inteiros (`True/False` para `1/0`).
- Remoção de colunas não preditivas como `product_id` e `sales_revenue` no modelo preditivo.
- Normalização de variáveis contínuas (`price`, `sales_units`, `discount_percentage`, etc.) com *StandardScaler* para padronizar as escalas.
- Separação do conjunto de dados em treino e teste, utilizando a divisão de 80/20.

5 Análise Exploratória dos Dados

5.1 Estatísticas Descritivas

As variáveis quantitativas principais apresentaram os seguintes valores médios e desvios padrão:

- **Média:**
  - `sales_units`: valor médio das unidades vendidas.
  - `price`: média dos preços dos produtos.
  - `discount_percentage`: média dos percentuais de desconto aplicados.
  - `future_demand`: média da demanda futura observada.
- **Desvio padrão:** semelhante às variáveis acima, indicando a dispersão dos dados.

5.2 Distribuições Univariadas

Foram utilizados histogramas e curvas de densidade para analisar as distribuições:

- **sales_units:** distribuição assimétrica com cauda à direita, indicando produtos com alta venda esporádica.
- **price:** preços concentrados em uma faixa média, com poucos valores extremos.
- **discount_percentage:** a maioria das observações com 0%, com alguns picos em promoções mais altas.
- **future_demand:** distribuição semelhante à de `sales_units`, sugerindo similaridade na estrutura.

5.3 Análises Bivariadas e Gráficas

- **Gráfico de dispersão:** foi analisada a relação entre:
  - `price` e `sales_revenue`, indicando tendência de crescimento da receita com aumento do preço (até certo ponto).
  - `discount_percentage` e `future_demand`, sem correlação clara visível.
- **Boxplots:**
  - `sales_units` por região (Europa e América do Norte).
  - Comparação entre `store_type` (Atacado vs Varejo).
  - `sales_units` por categoria de produto (`Cabinets`, `Chairs`, `Sofas`, `Tables`).
- **Violinplot:** distribuição das vendas com e sem promoção aplicada (`promotion_applied`), evidenciando possíveis diferenças de dispersão.
- **Mapa de Correlação (Heatmap):** análise de correlação entre variáveis numéricas. As variáveis com maior correlação com `future_demand` foram destacadas visualmente, embora nenhuma correlação extremamente forte tenha sido identificada.

Essas análises forneceram insights fundamentais para a seleção de variáveis e entendimento da estrutura do dataset, sendo base essencial para a modelagem preditiva subsequente.

6 Aplicação da Técnica Estatística ou Preditiva

6.1 Modelagem

Nesta etapa, foi aplicada uma técnica de regressão para prever a variável `future_demand`, utilizando o algoritmo **Random Forest Regressor**, uma técnica baseada em árvores de decisão com múltiplos estimadores (floresta de árvores).

O processo de modelagem seguiu os seguintes passos:

- Exclusão das colunas não preditivas: `date`, `product_id`, `sales_revenue`, e a própria variável alvo `future_demand`.
- Conversão de variáveis booleanas em variáveis inteiras (`True/False` para `1/0`).
- Normalização das variáveis contínuas com `StandardScaler`, garantindo média 0 e desvio padrão 1, aplicando especialmente às colunas:
  - `sales_units`, `competitor_price_index`, `economic_index`, `weather_impact`, `price`, `discount_percentage`.
- Divisão do conjunto de dados em treino (80%) e teste (20%).
- Treinamento do modelo com `RandomForestRegressor`, utilizando 100 estimadores e `random_state = 42`.

6.2 Avaliação do Modelo

O modelo foi avaliado com as métricas de regressão clássicas:

- **RMSE (Root Mean Squared Error)**: 56,47
  Indica que o erro médio quadrático das previsões é de aproximadamente 56 unidades de demanda.
- **MAE (Mean Absolute Error)**: 48,64
  O erro médio absoluto da previsão é de cerca de 48 unidades.
- **R^2 (Coeficiente de Determinação)**: $-0{,}0424$
  O valor negativo de R^2 demonstra que o modelo performa pior do que uma simples média constante, evidenciando um caso de **underfitting**.

6.3 Importância das Variáveis

Foi também realizada uma análise de importância das variáveis com base nos critérios internos do modelo *Random Forest*, os quais avaliam a redução de impureza (Gini ou MSE) associada a cada variável.

As variáveis mais relevantes identificadas foram:

- **category_Tables, category_Sofas, category_Chairs** — indicando forte impacto da categoria do produto na demanda.

- **region_North America** — sugerindo variação geográfica significativa no comportamento de compra.
- Variáveis temporais como `month` e `day_of_week` e econômicas como `economic_index` também apresentaram importância intermediária.

As variáveis com menor impacto incluíram:

- `promotion_applied`, `holiday_season`, `weather_impact` e `discount_percentage`.
- `sales_units` — surpreendentemente, teve baixa importância na previsão da demanda futura, possivelmente por problemas de defasagem temporal.

A Figura 1 ilustra a importância relativa de cada variável segundo o modelo.

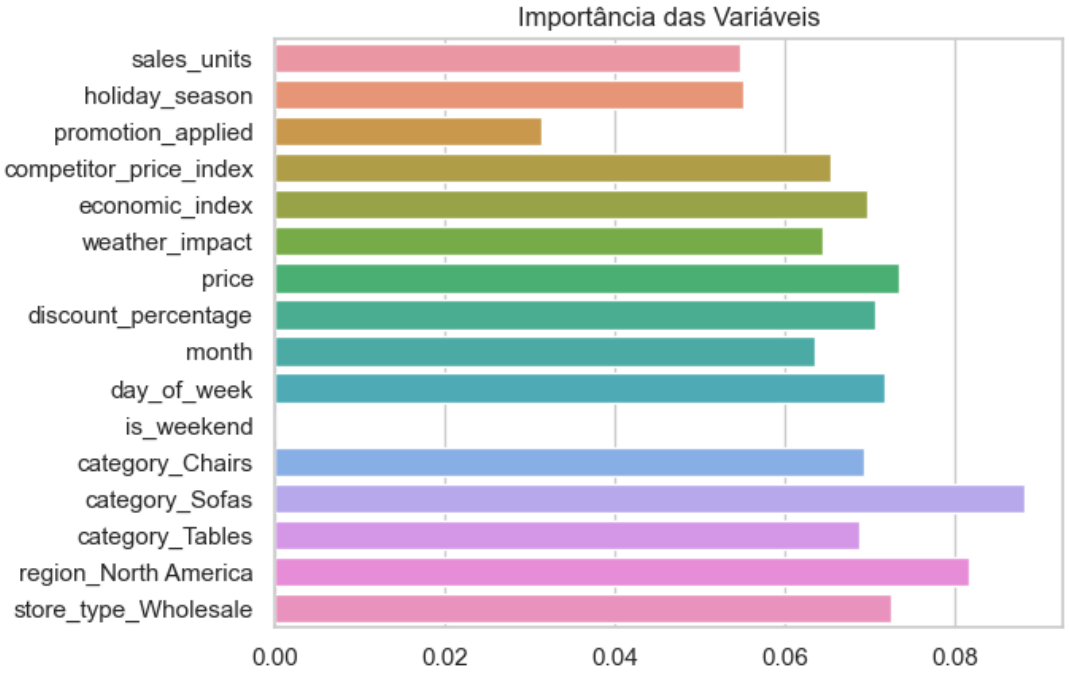

Figura 1: Importância das Variáveis segundo o Random Forest

7 Discussão dos Resultados

7.1 Desempenho do Modelo

Os resultados obtidos com o modelo *RandomForestRegressor* indicam um desempenho insatisfatório na previsão da variável *future_demand*. O valor de $R^2 = -0.0424$ aponta que o modelo teve um desempenho inferior ao de um modelo base que simplesmente prevê a média da demanda para todos os casos. Além disso, os erros de previsão, com $RMSE = 56{,}47$ e $MAE = 48{,}64$, mostram que o modelo está cometendo erros significativos em termos absolutos.

Esse cenário caracteriza um caso de **underfitting**, no qual o modelo não consegue capturar padrões relevantes nos dados, possivelmente por simplicidade excessiva do modelo ou ausência de variáveis preditoras expressivas.

7.2 Impacto de Promoções e Condições Climáticas

A análise de importância das variáveis mostrou que atributos como *promotion_applied*, *holiday_season*, *weather_impact* e *discount_percentage* tiveram contribuição limitada para a previsão. Isso pode indicar que:

- Esses fatores não têm impacto significativo na demanda futura neste conjunto de dados;
- Ou, mais provavelmente, que essas variáveis não foram bem representadas (por exemplo, o índice climático pode ser muito genérico, e feriados não foram contextualizados localmente).

7.3 Variáveis Mais Relevantes

A análise de importância das variáveis indicou que as mais relevantes foram:

- **category_Tables**, **category_Sofas**, **category_Chairs**: a categoria do produto foi o fator mais determinante para prever a demanda futura.
- **region_North America**: a região de venda teve impacto importante, possivelmente por volume ou padrões sazonais.

Variáveis com importância média (entre 0,06 e 0,08):

- *store_type_Wholesale*, *day_of_week*, *month*, *price*, *economic_index*, *competitor_price_index*.

Variáveis com baixa importância (abaixo de 0,05):

- *promotion_applied*, *holiday_season*, *weather_impact*, *discount_percentage*, *sales_units*.

7.4 Sugestões para Melhorias

1. **Engenharia de Atributos**:
   - Criar variáveis derivadas de *date*, como: mês, dia da semana, fim de semana e feriado.
   - Gerar interações como: *promotion_applied* × *discount_percentage*, ou *price_relative* = *price* / *competitor_price_index*.
2. **Modelos Avançados**:
   - Utilizar algoritmos como *XGBoost*, *LightGBM* ou *CatBoost*, mais eficazes para dados tabulares com ruído e relações não-lineares.
3. **Transformação da Variável Alvo**:
   - Aplicar transformações como $\log(1 + \textit{future\_demand})$ para reduzir o impacto de outliers.
4. **Análise de Dados e Balanceamento**:
   - Avaliar a distribuição da variável alvo e remover ou tratar valores extremos.
   - Agrupar produtos semelhantes por subcategoria ou faixa de preço para melhor generalização.

5. **Incorporação de Dados Externos**:
    - Incluir dados de feriados oficiais, indicadores econômicos do setor, clima local e ações promocionais da concorrência.

8 Conclusão

O modelo inicial apresenta limitações relevantes, com baixo poder preditivo e fraca representação de variáveis influentes como promoções e clima. Em contrapartida, características como **categoria do produto** e **região de venda** foram as mais determinantes na previsão da demanda futura.

Com a adoção de técnicas avançadas de modelagem, enriquecimento de dados e melhorias na engenharia de atributos, há potencial significativo para aprimorar a acurácia das previsões e produzir insights mais úteis para o planejamento de vendas.