Previsão de Demanda de Produtos em Varejo Utilizando Inteligência Artificial

Pedro Catarino[1], Daniel Kabadayan[1], Rafael Yun[1]

[1]Faculdade de Computação e Informática – Universidade Presbiteriana Mackenzie (UPM)
São Paulo – SP – Brazil

10395215@mackenzista.com.br, 10332633@mackenzista.com.br, 10313890@mackenzista.com.br

Abstract. *This project aims to develop a predictive model using Artificial Intelligence to forecast product demand in the retail sector. The goal is to help e-commerce companies optimize inventory management and improve logistical efficiency, using historical sales data and machine learning algorithms. The study explores the AI technique, decision tree. The implementation was done using Python and the Scikit-learn library, as well as others to help with data manipulation, such as Pandas and NumPy, based on a public sales dataset.*

Resumo. *Este projeto visa desenvolver um modelo preditivo utilizando Inteligência Artificial para previsão da demanda de produtos no setor de varejo. O objetivo é ajudar empresas de e-commerce a otimizar o gerenciamento de estoques e melhorar a eficiência logística, utilizando dados históricos de vendas e algoritmos de aprendizado de máquina. O estudo explora a técnica de IA, árvore de decisão. A implementação foi feita usando Python e a biblioteca Scikit-learn, bem como outras para o auxílio da manipulação dos dados, como Pandas e NumPy, com base em um dataset público de vendas.*

1. Introdução

a. Contextualização

O setor de varejo enfrenta desafios significativos para gerenciar de forma eficiente seus estoques e prever a demanda futura de produtos. A falta de precisão nas previsões pode levar a problemas como excesso de estoque, rupturas de produtos, desperdícios e custos adicionais. Com o aumento da competitividade no mercado e a necessidade de otimização de processos, a previsão de demanda com o uso de Inteligência Artificial surge como uma solução estratégica para melhorar a tomada de decisões.

b. Justificativa

O uso de IA para prever a demanda de produtos tem o potencial de transformar o varejo, melhorando a precisão das previsões e permitindo um melhor planejamento de compras, logística e marketing. A aplicação de técnicas de aprendizado de máquina ajuda a identificar padrões complexos nos dados de vendas, o que não seria possível com métodos tradicionais. Assim, as empresas podem melhorar a alocação de recursos e reduzir custos operacionais.

c. Objetivo

Este projeto tem como objetivo desenvolver um modelo preditivo que permita prever para a demanda de produtos em uma loja de varejo utilizando dados históricos de vendas obtidos do dataset público Walmart Sales Forecast (AHMEDOV, 2024).

d. Opção do Projeto

Este projeto se enquadra na "Opção Framework", empregando ferramentas de Machine Learning como Scikit-learn para resolver um problema de regressão.

2. Descrição do Problema

O problema abordado é a falta de precisão na previsão da demanda de produtos no varejo, o que pode gerar falhas no gerenciamento de estoque e impactar diretamente os custos e as receitas das empresas. Ao prever corretamente a demanda de produtos em períodos futuros, as empresas podem otimizar seus processos de compra e evitar problemas como falta de produtos em prateleira ou excesso de estoque.

3. Dataset

O dataset escolhido para este projeto é um conjunto de dados de vendas no varejo disponível publicamente no Kaggle, denominado *Walmart Sales Forecast* (AHMEDOV, 2024).

Os três datasets fornecem informações complementares para o modelo de previsão de demanda. O stores.csv contém dados sobre as lojas, como o ID, o tipo e o tamanho de cada loja, permitindo analisar como essas características influenciam as vendas.

O train.csv traz o histórico de vendas semanais, com o ID da loja, o departamento, a data, as vendas semanais (que serão a variável alvo do modelo) e indicações sobre feriados, que são essenciais para o treinamento do modelo.

Já o features.csv oferece variáveis externas, como temperatura, preço do combustível, índice de preços ao consumidor, taxa de desemprego e feriados, que ajudam a enriquecer o modelo capturando influências sazonais e econômicas que afetam o comportamento do consumidor.

4. Metodologia

A abordagem escolhida para este projeto é baseada em técnicas de aprendizado de máquina para realizar a previsão de demanda de produtos em uma loja de varejo. Também será utilizado o Erro Médio Quadrático Raiz (RMSE) para servir de métrica para avaliar o desempenho dos modelos de previsão, pois mede a diferença entre os valores reais (dataset) e os valores previstos pelo modelo. Ele pode ser calculado da seguinte forma:

$$RMSE = \sqrt{\frac{SSE_w}{W}} = \sqrt{\frac{1}{W}\sum_{i=1}^{n} w_i u_i^2}$$

Em que SSE_w se refere ao total ponderado dos quadrados; W ao peso total da população; N ao número de observações; w_i é o peso da i-ésima observação; e u_i diz respeito ao erro associado à i-ésima observação.

O RMSE é amplamente utilizado em problemas de predição por ser uma métrica que penaliza maiores diferenças entre os valores reais e previstos, conforme discutido por Martelotte et al. (2019) e pela documentação da SAP (2023). O processo metodológico será dividido em várias etapas, descritas a seguir:

Coleta e preparação dos dados

Os dados foram coletados a partir de datasets públicos que incluem vendas semanais de produtos em uma rede de varejo. Esses datasets incluem atributos como data, quantidade vendida, categoria de produtos, localização da loja e promoções. Primeiramente, os dados passarão por um processo de limpeza e tratamento. Isso inclui o tratamento de valores ausentes, padronização das variáveis, e a aplicação de técnicas de transformação para variáveis categóricas. A análise exploratória (distribuição de vendas, sazonalidade, correlação entre variáveis, etc.) será feita para garantir que os dados estejam prontos para a modelagem.

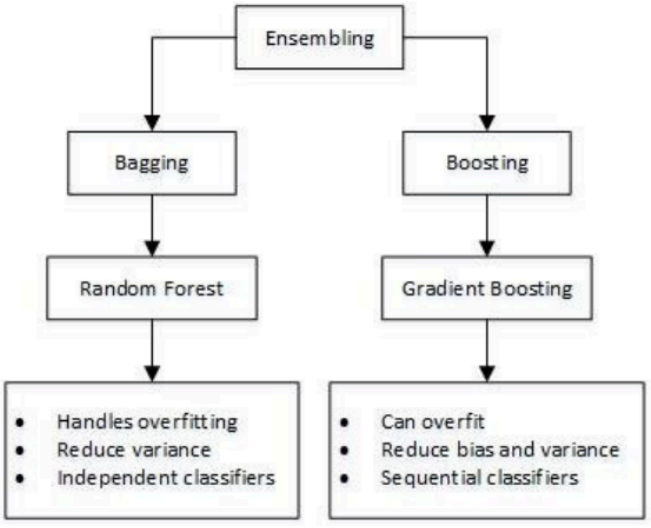
Divisão dos Dados

O dataset será dividido em dois subconjuntos: 30% para o treinamento do modelo e 70% para o teste, garantindo uma validação adequada do desempenho do modelo. Técnicas de validação cruzada podem ser utilizadas para aumentar a robustez da avaliação dos modelos.

Seleção de Algoritmos

Para este projeto, o algoritmo principal de aprendizado de máquina utilizado será a árvore de decisão. Usaremos dois modelos: o Random Forest e o Gradient Boosting, que são modelos simples e eficientes para previsões contínuas. O Random Forest combina várias árvores de decisão para reduzir o risco de overfitting e melhorar a precisão geral, enquanto o Gradient Boosting constrói sequencialmente modelos, corrigindo os erros de predições anteriores, como descrito por Idrees (2023).

Imagem 1 - Diagrama de comparação entre os algoritmos



Fonte: ResearchGate - Hari Suparwito

Ferramentas

O desenvolvimento será feito utilizando Python para desenvolvimento e as bibliotecas Scikit-learn para implementar os modelos propostos. A análise exploratória, visualização e preparação dos dados serão realizados principalmente com Pandas, Matplotlib e Numpy.

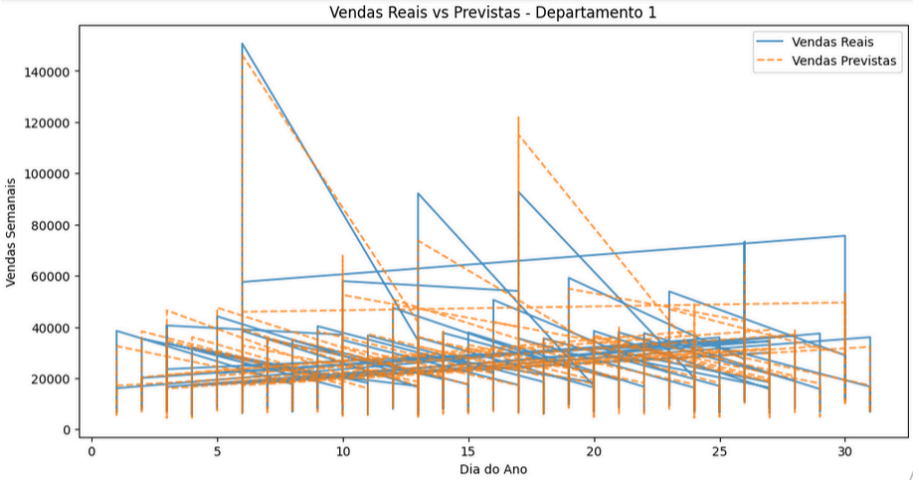
O ambiente a executar o modelo será no Jupyter Notebook.

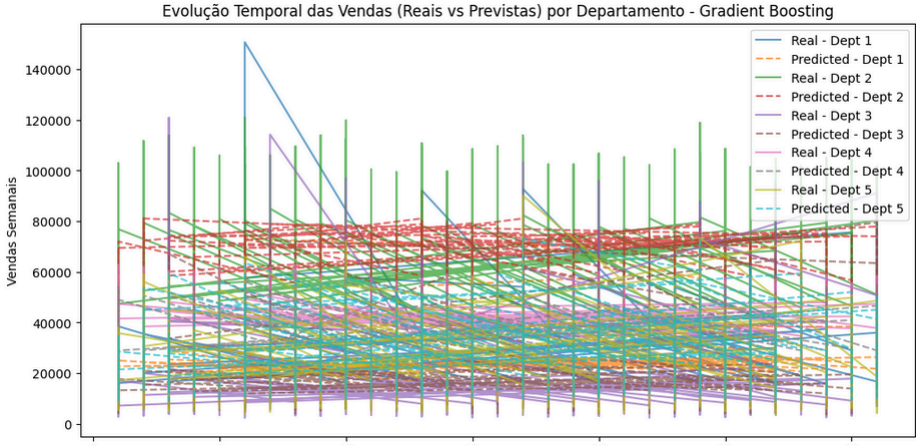
5. Resultados

Os modelos desenvolvidos, Random Forest Regressor e Gradient Boosting Regressor, foram aplicados com o objetivo de prever vendas semanais em lojas de departamentos. Utilizando um conjunto de dados que inclui informações históricas de vendas, feriados e datas, foi possível gerar previsões que acompanham, em boa parte, os padrões reais de vendas. O modelo Random Forest apresentou um erro médio quadrático raiz (RMSE) de 4041.70, indicando uma precisão sólida para o contexto do varejo. Já o Gradient Boosting alcançou um RMSE de 7185.02, demonstrando menor capacidade de prever com exatidão os valores reais.

Os resultados foram visualizados por meio de gráficos comparativos, que serão mostrados abaixo:

Imagens 2 e 3 - Comparação entre os Modelos Random Forest e Gradient Boosting

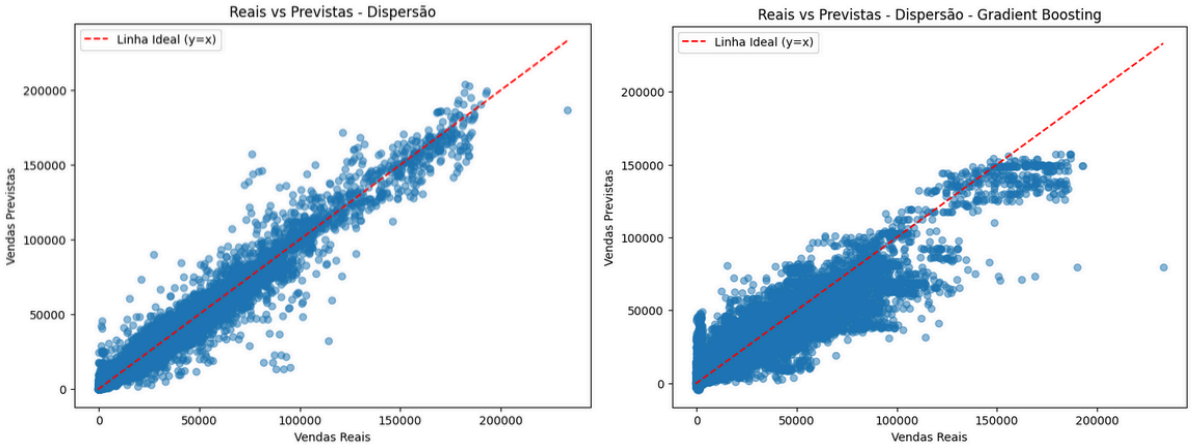




Fonte: Elaboração própria

Os gráficos acima mostram que, para o modelo Random Forest, as previsões seguiram de perto as vendas reais na maior parte dos períodos. Contudo, em momentos de picos acentuados, como feriados, o modelo apresentou algumas discrepâncias. Já o modelo Gradient Boosting mostrou uma maior discrepância em geral, especialmente em departamentos com alta variabilidade de vendas.

Imagens 4 e 5 - Gráficos de Dispersão Random Forest e Gradient Boosting



Fonte: Elaboração própria

Nesse gráfico de dispersão entre vendas reais e previstas, observou-se uma boa concentração de pontos ao longo da linha ideal ($y = x$) para o Random Forest, enquanto o Gradient Boosting apresentou maior dispersão, indicando previsões menos consistentes. Esses gráficos evidenciam que o Random Forest foi mais eficiente em capturar os padrões gerais, enquanto que o Gradient Boosting enfrentou dificuldades em prever eventos atípicos.

Os outros gráficos também mostram informações relevantes que ajudam a corroborar com as informações mencionadas acima. No geral, a maioria das previsões se manteve próximo do real, mas houve erros significativos em picos como feriados. Além disso, o modelo conseguiu prever tendências de altas e baixas, o que mostra sua aplicabilidade prática para o planejamento e organização de estoque, como foi proposto.

6. Conclusão

Os resultados deste projeto demonstram a relevância de modelos baseados em aprendizado de máquina, como Random Forest e Gradient Boosting, para prever a demanda de produtos no varejo. Utilizando dados históricos enriquecidos por variáveis externas, foi possível desenvolver um modelo que captura padrões complexos de vendas, otimizando a tomada de decisão em processos logísticos e de gerenciamento de estoques.

Embora ambos os modelos tenham apresentado desempenhos sólidos, o Random Forest destacou-se como a abordagem mais consistente, com um RMSE consideravelmente menor e maior aderência aos dados reais, especialmente em

cenários de vendas regulares. As limitações observadas nos picos de vendas, como feriados, ressaltam a necessidade de ajustes futuros, seja pela inclusão de mais variáveis ou pelo emprego de técnicas avançadas de otimização.

Este estudo reafirma o papel estratégico da Inteligência Artificial no varejo, especialmente em um cenário competitivo e altamente dinâmico. A utilização de modelos preditivos permite que as empresas não apenas reduzam custos operacionais, mas também aprimorem a experiência do cliente por meio de um planejamento mais eficiente de estoques e logística.

7. Referências

AHMEDOV, Aslan. **Walmart Sales Forecast**. Kaggle Datasets. Disponível em: https://www.kaggle.com/datasets/aslanahmedov/walmart-sales-forecast/data. Acesso em: 21 nov. 2024.

IDREES, H. *Gradient Boosting vs. Random Forest: Which Ensemble Method Should You Use?* Medium, 2023. Disponível em: https://medium.com/@hassaanidrees7/gradient-boosting-vs-random-forest-which-ensemble-method-should-you-use-9f2ee294d9c6. Acesso em: 22 nov. 2024.

MARTELOTTE, A. et al. *Modelos e métodos de previsão de séries temporais*. 2019. Disponível em: https://edisciplinas.usp.br/pluginfile.php/7462505/mod_resource/content/0/Apostila_Cap3.pdf. Acesso em: 23 nov. 2024.

SAP SE. *Root mean square error (RMSE)*. 2023. Disponível em: https://help.sap.com/docs/SAP_PREDICTIVE_ANALYTICS/41d1a6d4e7574e32b815f1cc87c00f42/5e5198fd4afe4ae5b48fefe0d3161810.html. Acesso em: 24 nov. 2024.

SUPARWITO, H. The difference between random forest and gradient boosting machine. ResearchGate, 2020. Disponível em: https://www.researchgate.net/figure/The-difference-between-random-forest-and-gradient-boosting-machine_fig1_342548825. Acesso em: 21 nov. 2024.

8. Bibliografia

AGGARWAL, Charu C. Artificial Intelligence: A Textbook. New York: Springer: 2021.

CHOLLET, François. Deep Learning with Python, 2ed. Shelter Island: Manning, 2021.

GÉRON, Aurélien. Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 2 ed. Sebastopol: O'Reilly, 2019.

9. Link do GitHub

https://github.com/RafaelMackCC/Projeto-1-de-IA

10. Link do Vídeo

https://youtu.be/Hu6UE223EMs