RELATÓRIO DE DESENVOLVIMENTO DE UM SISTEMA DE PREVISÃO DE LUCRATIVIDADE DE UMA LOJA DE VAREJO

Paulo Afonso/BA
maio/2025

KAÍQUE PEREIRA DOS SANTOS

RELATÓRIO DE DESENVOLVIMENTO DE UM SISTEMA DE PREVISÃO DE LUCRATIVIDADE DE UMA LOJA DE VAREJO

Relatório apresentado ao Corpo Docente do Centro Universitário do Rio São Francisco - Unirios, como pré-requisito de avaliação para a Disciplina de Inteligência Artificial, do curso de Bacharelado em Sistema de Informação, sob orientação do Professor Mestre Douglas Costa Braga

Paulo Afonso/BA
maio/2025

SUMÁRIO

1. INTRODUÇÃO

Este relatório detalha o processo de desenvolvimento de um sistema de Machine Learning para prever a lucratividade de vendas no setor de varejo. O objetivo central foi criar um sistema completo, desde a análise de dados até a implementação de um modelo preditivo com uma interface web interativa.

2. DESCRIÇÃO DA METODOLOGIA ADOTADA

A metodologia foi estruturada em etapas sequenciais para garantir um desenvolvimento robusto e organizado, partindo da análise dos dados até a disponibilização do modelo final em uma aplicação web, assim como proposto nas instruções para a realização deste projeto.

1. **Análise e Preparação dos Dados**: O projeto utilizou um conjunto de dados histórico de vendas de uma loja de varejo (SampleSuperstore.csv), obtido na plataforma Kaggle (disponível aqui). Foi escolhido outro dataset ao invés do disponibilizado no AVA somente com o objetivo de realizar um desenvolvimento sem a interferência direta dos exemplos disponibilizados. A primeira etapa consistiu na preparação dos dados, que incluiu a remoção de quaisquer registros com valores ausentes para garantir a qualidade e a integridade dos dados de entrada do modelo.

2. **Engenharia de Features e Definição do Problema**: Para transformar a análise em um problema de classificação, uma variável-alvo binária chamada High_Profit foi criada. Uma venda foi classificada como de "Alto Lucro" (High_Profit = 1) se seu lucro (Profit) estivesse acima do 75º percentil de todos os lucros do dataset. Caso contrário, foi classificada como "Baixo Lucro" (High_Profit = 0). Essa abordagem permite ao modelo focar em prever não o valor exato do lucro, mas sim a probabilidade de uma venda ser mais lucrativa.

3. **Codificação de Variáveis Categóricas**: As variáveis categóricas, como Region, Category, Sub-Category, e State, foram convertidas para um formato numérico utilizando a técnica LabelEncoder. Cada categoria única em uma coluna recebeu um número inteiro correspondente, tornando-as compreensíveis para o algoritmo de Machine Learning. Os encoders foram salvos para serem reutilizados na aplicação final.

4. **Treinamento e Validação do Modelo**: O conjunto de dados foi dividido em 80% para treinamento e 20% para teste. As features selecionadas para o modelo incluíram Sales, Quantity, Discount, e as versões codificadas das variáveis categóricas. A divisão dos dados garante que o modelo seja avaliado em dados não vistos, fornecendo uma estimativa mais realista de seu desempenho no mundo real.

5. **Desenvolvimento da Interface Web**: Para a implantação, uma aplicação web foi desenvolvida utilizando o framework Flask. Uma interface de usuário (index.html) permite que o usuário

insira os dados de uma venda hipotética (categoria, região, vendas, desconto, etc.). Esses dados são enviados para o servidor, processados, e o modelo retorna uma previsão de "Alto Lucro" ou "Baixo Lucro" com um percentual de confiança.

3. JUSTIFICATIVA PARA A ESCOLHA DO ALGORITMO

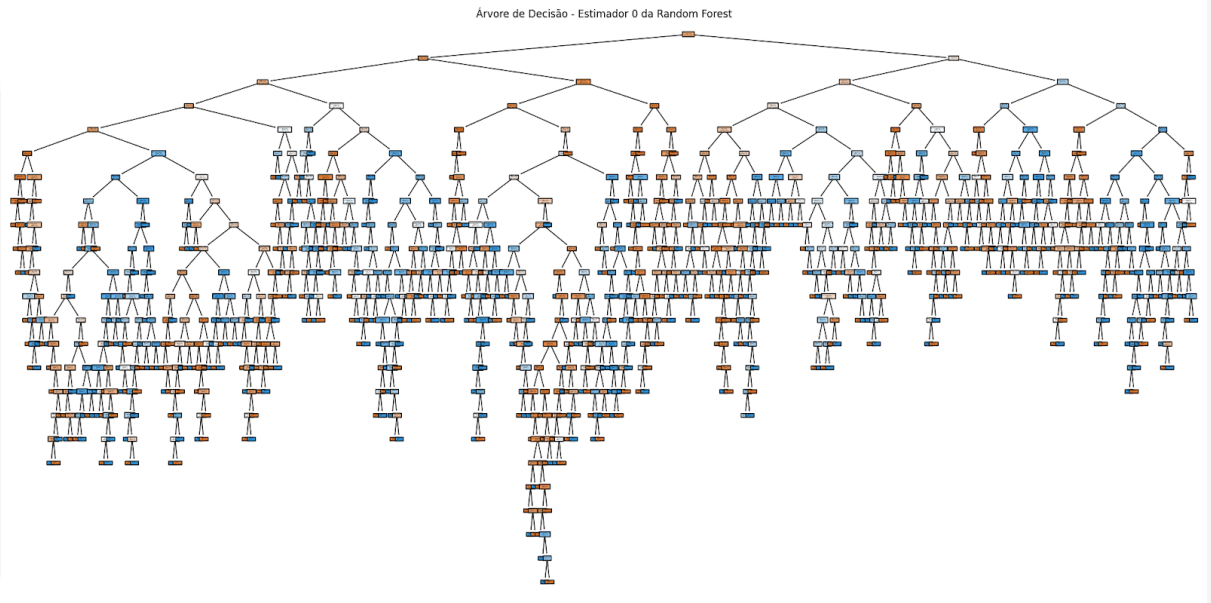
A abordagem inicial do projeto foi tentar prever o valor exato do lucro (Profit) utilizando modelos de **regressão**. No entanto, os resultados obtidos foram consistentemente negativos, indicando que os modelos não conseguiam aprender os padrões nos dados de forma eficaz. Foram testados diversos algoritmos de regressão, que apresentaram as seguintes métricas de R-quadrado (R2):

- **Decision Tree Regressor**: R2 entre -0.54 e -0.70
- **Random Forest Regressor**: R2 de -0.12
- **Gradient Boosting Regressor**: R2 de -0.004

Um valor de R2 negativo significa que o modelo de regressão performou **pior do que um modelo básico** que simplesmente prevê o valor médio do lucro para todas as vendas. Essa performance insatisfatória demonstrou que prever o valor numérico exato do lucro era uma tarefa inviável com as features disponíveis.

Diante disso, a estratégia foi alterada para uma **tarefa de classificação**. Para esta tarefa, o algoritmo escolhido foi o **Random Forest Classifier**.

Este algoritmo funciona criando um "comitê" de múltiplas árvores de decisão e usando a "votação" da maioria para determinar o resultado final. A imagem abaixo ilustra uma **única árvore de decisão** das centenas que compõem a floresta final do modelo.



A escolha do Random Forest foi fundamentada nas seguintes razões:

Robustez e Prevenção de Overfitting: Uma única árvore de decisão, como a mostrada acima, pode se tornar excessivamente complexa e se ajustar perfeitamente aos dados de treino, incluindo seus ruídos (overfitting). Ao utilizar uma floresta de árvores, o Random Forest combina as previsões de muitas árvores ligeiramente diferentes, corrigindo os erros individuais e resultando em um modelo muito mais robusto e generalizável. O alto desempenho resultante dessa escolha é detalhado na seção a seguir.

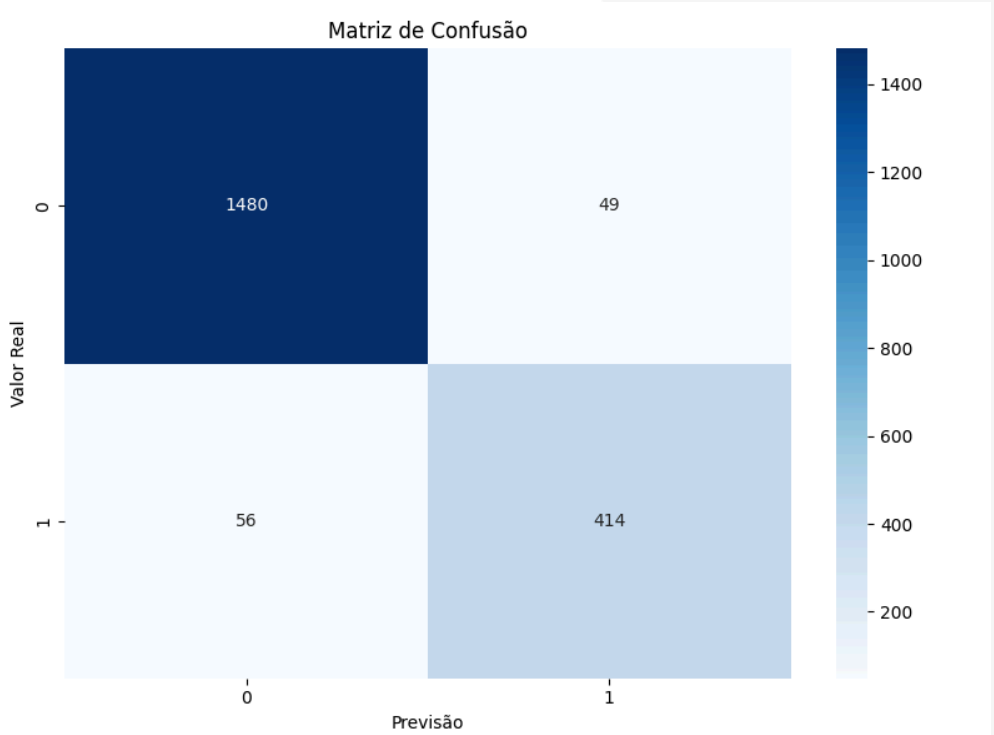
Alto Desempenho e Precisão: Geralmente, o Random Forest oferece uma alta acurácia de classificação, o que foi comprovado nos resultados.

Análise de Importância das Features: Uma vantagem intrínseca do Random Forest é sua capacidade de medir a importância de cada feature, como foi detalhado na seção de insights.

4. RESULTADOS OBTIDOS COM INTERPRETAÇÕES

O desempenho do modelo final de Random Forest foi avaliado no conjunto de teste, que representa 20% dos dados totais e não foi utilizado durante o treinamento. Os resultados quantitativos demonstram a alta eficácia do modelo na tarefa de classificação.

A acurácia geral do modelo foi de **94,75%**, indicando que ele classifica corretamente quase 95 em cada 100 casos. Um detalhamento mais profundo é fornecido pela Matriz de Confusão abaixo e pela tabela de métricas de classificação a seguir:A matriz visualiza os seguintes resultados:



- **Verdadeiros Negativos (Previsto 0, Real 0)**: 1480 acertos.
- **Verdadeiros Positivos (Previsto 1, Real 1)**: 414 acertos.
- **Erros**: O modelo cometeu 49 erros de Falso Positivo (previu lucro alto, mas era baixo) e 56 erros de Falso Negativo (previu lucro baixo, mas era alto).

Classe	Precision	Recall	F1-Score	Support
0 (Baixo Lucro)	0.96	0.97	0.97	1529
1 (Alto Lucro)	0.89	0.88	0.89	470
Weighted Avg	0.95	0.95	0.95	1999

4.1. Interpretação dos Resultados:

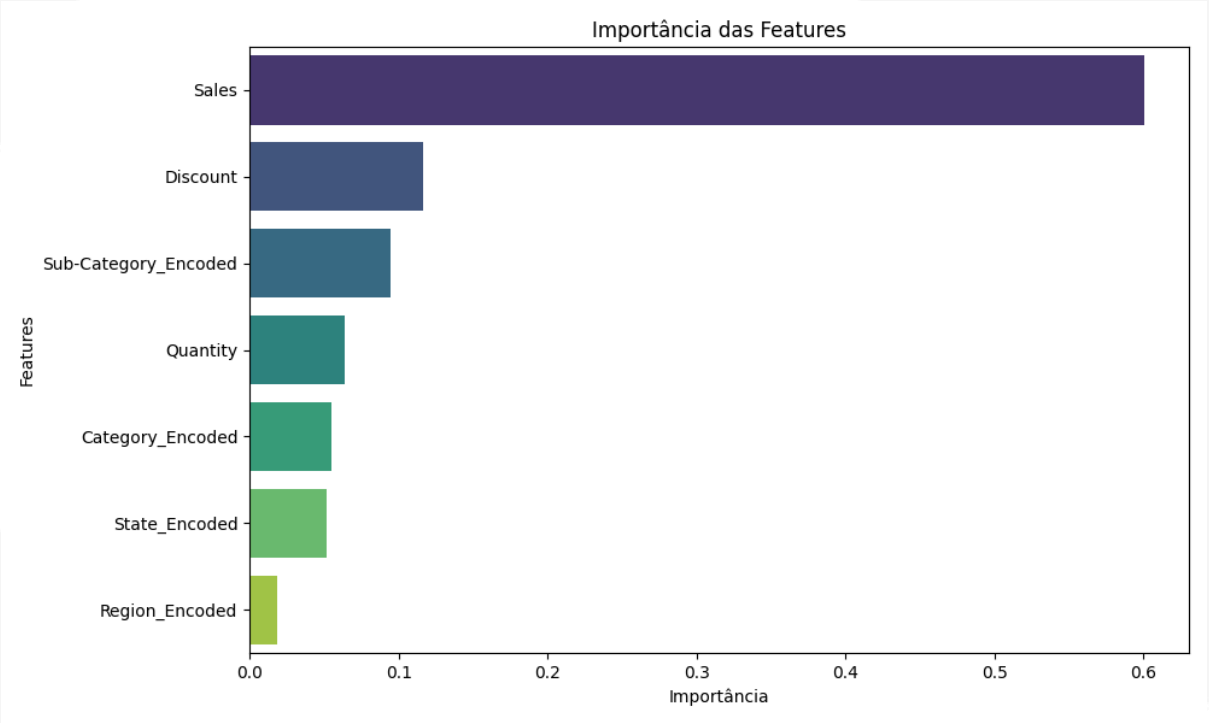
Classe 1 (Alto Lucro):

- **Precisão (0.89)**: Quando o modelo prevê "Alto Lucro", essa previsão está correta em 89% das vezes. Este é um indicador de alta confiança.
- **Recall (0.88)**: O modelo identificou corretamente 88% de todas as vendas que realmente eram de "Alto Lucro", perdendo poucas oportunidades.

Classe 0 (Baixo Lucro): O modelo é extremamente preciso e sensível para identificar vendas de baixo lucro, o que é crucial para evitar estratégias não rentáveis. Além de sua alta performance, o modelo também nos permite extrair inteligência de negócio a partir de sua estrutura interna, como veremos a seguir

5. INSIGHTS DE NEGÓCIO EXTRAÍDOS

A implementação do modelo e a análise da importância das features permitem extrair insights valiosos para a tomada de decisão estratégica no varejo. O gráfico abaixo, gerado pelo modelo, mostra o peso de cada variável na decisão.



Foco em Vendas e Descontos: O gráfico confirma que Sales (Vendas) é o fator de maior impacto, seguido por Discount (Desconto). O insight principal é que não basta ter um volume alto de vendas; a lucratividade é extremamente sensível aos descontos aplicados. O modelo pode ser usado para simular o ponto de equilíbrio ideal entre atrair vendas com descontos e manter uma alta lucratividade.

Estratégia de Produto e Geográfica: Sub-Category, Category e State são os próximos fatores mais importantes. Isso indica que a lucratividade varia significativamente dependendo do **tipo de produto** vendido e da **localização**. A empresa pode usar o sistema para focar os esforços de marketing e gestão de estoque nas combinações de produto-estado mais rentáveis.

Ferramenta de Suporte à Decisão: A interface web funciona como uma ferramenta de suporte para as equipes de vendas. Antes de lançar uma promoção, um gestor pode inserir os parâmetros (produto,

estado, valor e um novo desconto) para obter uma previsão instantânea sobre o potencial de lucratividade, permitindo decisões mais informadas e baseadas em dados.

6. LIMITAÇÕES E TRABALHOS FUTUROS

Apesar dos resultados positivos, o projeto possui limitações e oportunidades para melhorias futuras.

Definição Relativa de Lucro: A classificação de "Alto Lucro" baseia-se em um percentil (75%) dos dados existentes, o que é uma medida relativa. No futuro, o limiar poderia ser definido com base em metas de negócio absolutas ou KPIs financeiros (ex: uma margem de lucro mínima específica).

Análise de Sazonalidade: O modelo atual prevê a lucratividade de transações individuais e não captura padrões sazonais ou tendências ao longo do tempo. Um trabalho futuro poderia envolver a criação de um modelo de séries temporais para prever as vendas e a lucratividade em diferentes períodos do ano.

Engenharia de Features Avançada: Futuramente, o modelo poderia ser aprimorado com features mais complexas, como informações demográficas dos clientes, dados sobre a concorrência, ou o tempo de vida do cliente (Customer Lifetime Value - CLV).