Relatório Técnico

Relatório apresentado como exigência do
processo seletivo de estágio da VExpensens.

Yuri Fernandes Pereira
2025

1. Introdução

Este relatório apresenta o processo de análise, modelagem e otimização de modelos preditivos para um conjunto de dados fornecido. O objetivo é comparar diferentes técnicas de aprendizado de máquina que preveem se um usuário de um site imobiliário irá ou não comprar uma casa. Isto foi feito a partir de um dataset que conta com diferentes atributos relacionados aos usuários do site, tais como renda anual, tempo em que navegou pelo site, entre outros.

2. Análise exploratória e pré-processamento dos dados

O conjunto de dados apresentava:

- 200 entradas
- 6 colunas (Idade, Renda Anual, Gênero, Tempo no Site, Anúncio Clicado e Compra)
- Valores ausentes em diversas colunas
- Outliers na coluna “Tempo no Site”

O tratamento de cada variável explicativa foi feito de maneira diferente:

Idade: Valores nulos foram substituídos pela mediana das idades.

Renda Anual: Valores nulos foram substituídos pela média das rendas.

Gênero: A partir da análise da distribuição da variável “Comprou” em cada gênero, optou-se pela remoção da coluna Gênero, pois chegou-se à conclusão de que o sexo de um usuário não influencia na sua decisão de comprar ou não uma casa.

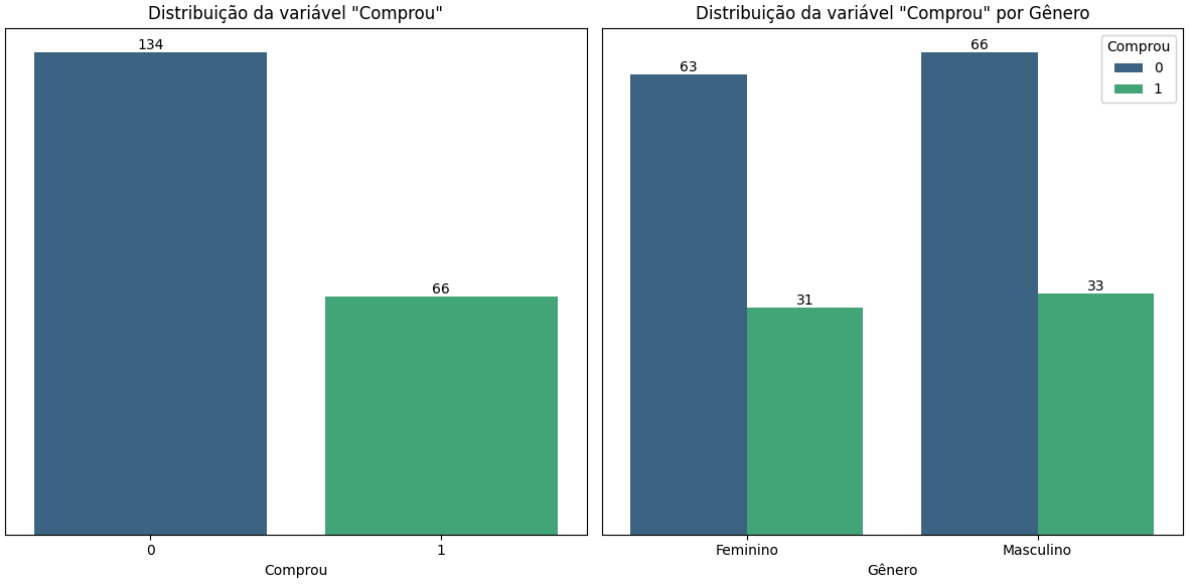

Figura 1: Distribuição da variável “Comprou” é a mesma independente do gênero

Tempo no Site: Valores discrepantes (-1) foram substituídos pela média dos tempos no site.

Anúncio Clicado: Usuários com valores nulos foram removidos do dataset. Mapeamento de Sim/Não para 1/0.

O dataset foi dividido em 70% das amostras para o conjunto de treino e 30% das amostras para o conjunto de teste, mantendo a distribuição da variável-alvo em ambos os conjuntos.

3. Treinamento e avaliação dos modelos

Foram escolhidos três modelos de machine learning para comparação de desempenho: Regressão Logística, Árvore de Decisão e Random Forest. A métrica de avaliação do desempenho foi o **Recall da Classe 1**, pois foi priorizado um modelo capaz de identificar o maior número de clientes que são potenciais compradores. Isso significa que em uma campanha de marketing, por exemplo, o modelo vai ser capaz de garantir que a maioria dos usuários-alvo serão alcançados.

Tabela 1: Desempenho de cada modelo antes da otimização

Modelo	**Recall Classe 1**
Regressão Logística	6%
Árvore de Decisão	61%
Random Forest	22%

4. Otimização dos hiperparâmetros

Utilizou-se o Grid Search e o Cross Validation como técnicas de otimização. Por ser um dataset pequeno, o conjunto de treino foi dobrado em apenas 3 folds no Cross Validation.

Tabela 2: Desempenho de cada modelo após otimização

Modelo	**Recall Classe 1**
Regressão Logística	78%
Árvore de Decisão	89%
Random Forest	67%

5. Resultados e Conclusão

A aplicação de técnicas de otimização de hiperparâmetros, combinada com o Cross Validation, resultou em um aumento de performance para os três modelos, considerando que a métrica mais importante nesse problema é o recall da classe 1.

Para os três modelos, identificou-se que o **tempo** que um usuário passa navegando pelo site é o fator mais importante para o modelo decidir se ele vai ou não comprar uma casa.

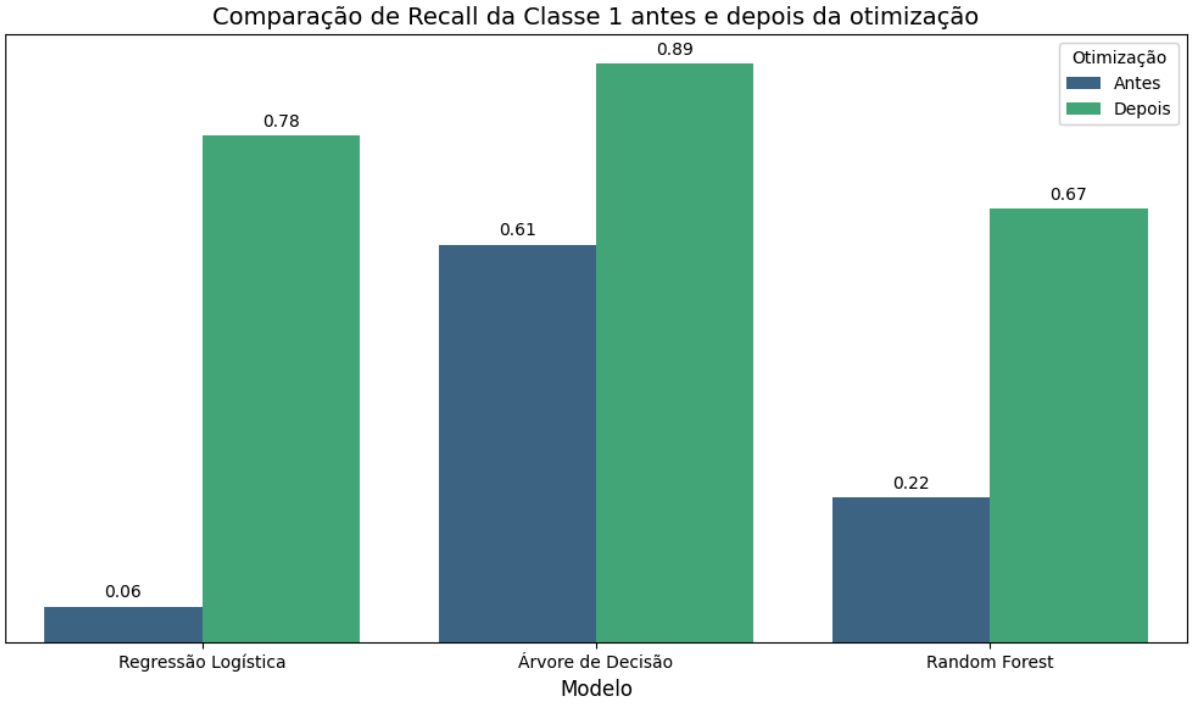

Figura 2: Recall da classe 1 para cada modelo, antes e depois da otimização

O gráfico de comparação de modelos antes/depois da otimização mostra que a Árvore de Decisão Otimizada é o modelo com melhor performance para este dataset, com um recall de 89% para a classe 1. Isso significa que de cada 100 usuários que vão comprar uma casa, o modelo é capaz de identificar 89 deles, favorecendo, por exemplo, campanhas de marketing a alcançar a maioria dos usuários-alvo.

Embora seja esperado que o Random Forest tenha um desempenho superior a uma única Árvore de Decisão, existem cenários em que o Random Forest tem um desempenho inferior. O Random Forest combina várias árvores de decisão (geralmente dezenas ou centenas), o que exige mais dados para gerar estimativas confiáveis. Com um conjunto de dados pequeno, o Random Forest pode não ser capaz de capturar informações significativas e pode "subutilizar" os dados disponíveis, enquanto uma única árvore pode ajustar-se melhor.

Referências

DIDÁTICA TECH. **O que é e como funciona o algoritmo RandomForest.** Disponível em: https://didatica.tech/o-que-e-e-como-funciona-o-algoritmo-randomforest/. Acesso em: 06 jan. 2025.

GUIMARAES, N. **Regressão Logística: Como usá-la em análise de dados.** Disponível em: medium.com/@nara.guimaraes/regress%C3%A3o-log%C3%ADstica-como-usu%C3%A1-la-em-an%C3%A1lise-de-dados-3fdb6be3a255. Acesso em: 06 jan. 2025.

IBM. **Decision Trees.** Disponível em: https://www.ibm.com/docs/en/db2/12.1?topic=building-decision-trees. Acesso em: 06 jan. 2025.