Relatório - Dataset de Vendas

Arthur Antonio Rezende Pereira

January 8, 2025

1 Introdução

Neste relatório, é apresentado a análise e modelagem de dados realizada com o objetivo de prever a variável *Compra* (0 ou 1) utilizando um conjunto de dados contendo informações como idade, gênero, renda anual, tempo no site, entre outros. A modelagem foi conduzida utilizando técnicas de aprendizado supervisionado, com destaque para *Random Forest* e *Logistic Regression*.

2 Exploração de Dados

Antes da modelagem, realizei uma análise exploratória dos dados para compreender suas principais características e identificar padrões relevantes. O conjunto de dados original continha 200 registros, mas, após o pré-processamento, esse número foi reduzido para 183 devido à presença de dados ausentes ou inconsistências.

As estatísticas básicas (média, variança e quartis) do dados númericos do conjunto estão presentes na tabela 1. Com essas informações, é possível observar que pssivelmente não há outliers no conjunto.

Table 1: Estatísticas de Idade, Renda Anual, Tempo no Site e Compra

Idade	**Renda Anual (em $)**	**Tempo no Site (min)**	**Compra (0 ou 1)**
172	172	172	172
39.24	58,779.07	17.99	0.31
12.56	26,095.15	7.13	0.47
18	30,000	5.05	0
29	30,000	12.61	0
39	50,000	18.29	0
51	70,000	24.00	1
59	100,000	29.85	1

Gráficos foram gerados para ilustrar as distribuições e relações entre as variáveis:

- Distribuição da variável *Idade* por *Compra* (Figura 1 e 2)

  É possível perceber que não há uma diferença clara na distribuição dos dados em relação à idade. No histograma, não se observa uma tendência associada a ela, e nos box-plots, as caixas são similares.

Figure 1: Histograma

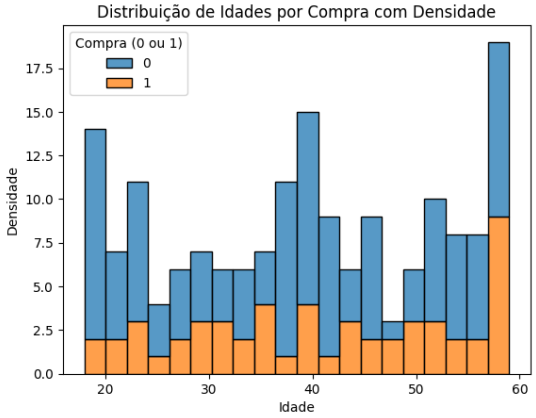
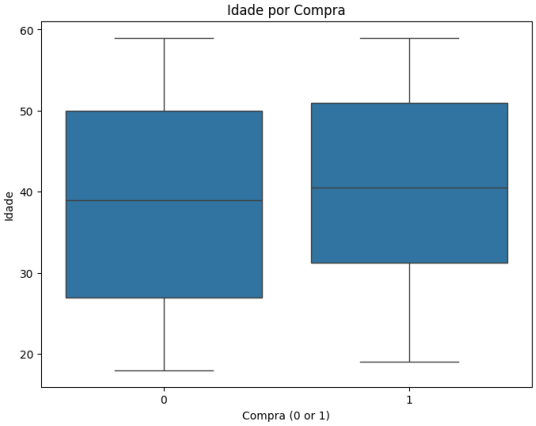


Figure 2: BoxPlot



- Relação entre variáveis categóricas e *Compra* (Figuras 3, 4 e 5)

  As três variáveis não apresentam uma tendência clara, e as proporções entre compra e não-compra parecem ser similares. A variável mais discrepante é a *Anúncio Clicado* (Figura 4), onde a proporção de vendas aumenta quando o anúncio é clicado. A que menos apresenta influência, visualemente falando, é a variável gênero.

Figure 3: Renda

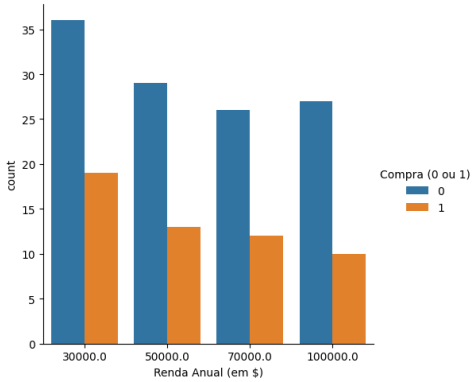


Figure 4: Anúncio Clicado

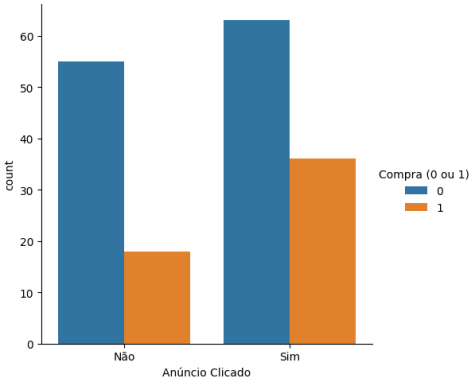
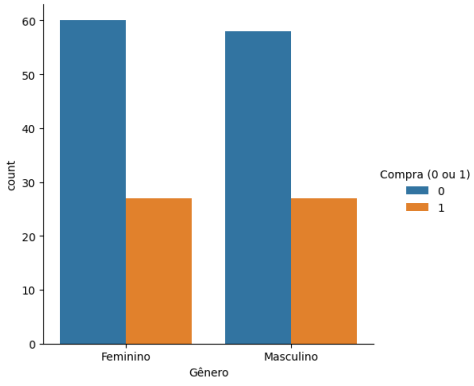


Figure 5: Gênero



- Comparação de *Tempo no Site* por *Compra* (Figura 6 e 7)

Neste gráfico, observa-se um leve aumento na distribuição de compradores com maior tempo de permanência no site. O padrão identificado no gráfico anterior também se repete aqui: a mediana e os quantis de tempo dos compradores são um pouco maiores. Dessa forma, a variável tempo no site se destaca como a mais promissora, sugerindo uma possível relação positiva com as compras.

Figure 6: Histograma

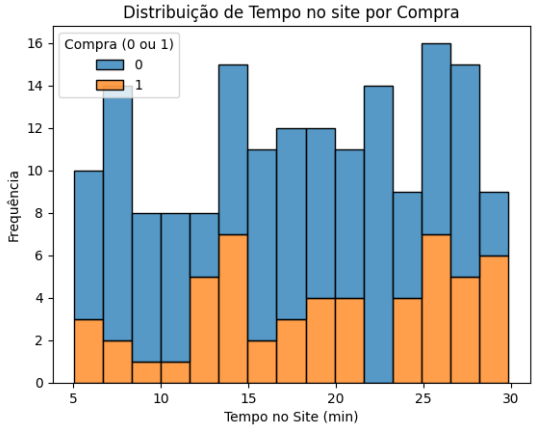
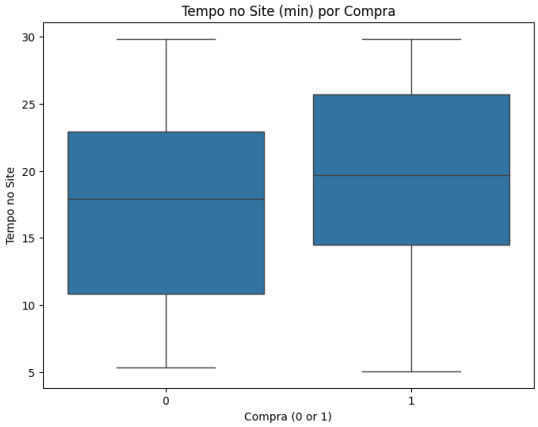


Figure 7: BoxPlot



3 Metodologia

3.1 Pré-processamento

Os dados passaram por um processo de limpeza e transformação que incluiu:

- Substituição de valores nulos na variável *Idade* pela mediana.
- Exclusão de valores nulos nas variáveis categóricas.
- Remoção de valores negativos na variável *Tempo no Site (min)*.
- Aplicação do *Label Encoder* na variável *Renda*.
- Normalização das variáveis independentes com *StandardScaler*.
- Aplicação do *SMOTE* para balancear classe da variável resposta.

3.2 Divisão dos Dados

O conjunto de dados foi dividido em dois subconjuntos:

- **Treinamento:** 80% dos dados.
- **Teste:** 20% dos dados.

A divisão foi realizada utilizando *train_test_split* do `scikit-learn`.

3.3 Modelos Utilizados

Foram treinados dois modelos principais:

1. **Random Forest Classifier**: Com ajuste de pesos para lidar com o desbalanceamento de classes.
2. **Regressão Logística**: Utilizada como baseline para comparação de desempenho.

4 Resultados

4.1 Random Forest

Após a otimização de hiperparâmetros utilizando *GridSearchCV*, os seguintes valores foram escolhidos:

- Número de estimadores: 50.
- Profundidade máxima: 10.
- Tamanho mínimo de divisão: 2.

Os resultados no conjunto de teste foram:

- **Acurácia:** 48.57%.

O modelo não teve um bom desempenho nos dados de teste, devido ao overfitting ocorrido durante o treinamento. Dessa forma, tentei solucionar o problema com outro modelo.

4.2 Importância das Features

Embora o desempenho baixo do modelo, foi possível tirar conclusões das hipóteses levantadas a respeito da importancia das features:

1. *Tempo no Site (min)* (0.41).
2. *Idade* (0.32).
3. *Renda Categórica* (0.15).
4. *Homem* (0.06).
5. *Anúncio Clicado* (0.05).

4.3 Regressão Logística

A regressão logística apresentou desempenho similar, com uma acurácia de 57% no conjunto de teste. Com certa dificuldade de classificar a classe 1.

4.4 PCA e Random Forest

Com os problemas nos métodos anteriores, optei por tentar aplicar o PCA em conjunto com o Random Forest, utilizando validação cruzada para avaliar e treinar o modelo. A escolha pela validação cruzada foi motivada pela limitação do método tradicional de divisão em conjunto de treino e teste, que pode ser influenciado pela maneira como os dados são divididos. A validação cruzada oferece uma avaliação mais robusta e confiável, pois utiliza todo o conjunto de dados, tanto para treino quanto para teste, e avalia o modelo em várias iterações, o que diminui a variabilidade nos resultados e reduz o risco de overfitting. Isso é especialmente relevante para garantir que o modelo tenha um bom desempenho em dados não vistos, tornando-o mais generalizável.

Dessa forma, os resultados foram positivos, após o obtendo os seguintes valores:

- *AUC médio* (0.77).
- *Acurácia média* (0.70).
- *Média Recall* (0.70).
- *Média Precision* (0.71).
- *Média F1-Score* (0.69).

5 Conclusão

Após a implementação e análise dos modelos de aprendizado de máquina, observou-se que o desempenho do Random Forest foi insatisfatório, com uma acurácia de 0.48 nos dados de teste, devido a um possível overfitting. A análise da importância das features revelou que "Tempo no Site (min)" e "Idade" foram as características mais impactantes para o modelo. A Regressão Logística também não apresentou resultados significativamente melhores, com acurácia de 0.57, especialmente com dificuldades na classificação da classe 1.

A tentativa de melhorar os resultados com a combinação de PCA e Random Forest mostrou-se mais eficaz. A utilização de validação cruzada foi crucial para mitigar o risco de overfitting e fornecer uma avaliação mais robusta. Com essa abordagem, o modelo alcançou uma acurácia média de 0.7 e boas métricas de desempenho, como AUC médio de 0.77, Recall médio de 0.70, Precision média de 0.71 e F1-Score médio de 0.69, o que indica um desempenho consideravelmente melhor e mais generalizável para novos dados.

Se houvesse mais tempo para dedicação, eu gostaria de explorar com mais profundidade as interações entre as variáveis, criando novas features, além de estudar a exclusão de variáveis com pouca importância. Também seria interessante testar o modelo com um conjunto de dados maior para avaliar sua real potência.