Desafio Prático | Estágio em Ciência de Dados

Introdução

O presente relatório tem como objetivo analisar um conjunto de dados com foco em prever o comportamento de compra dos clientes.

Metodologias

1. *Pandas (import pandas as pd)*

Por que escolhi? O Pandas é ótimo para lidar com tabelas de dados. Com ele, posso abrir o arquivo, olhar as informações e mexer nos dados, como limpar, filtrar ou organizar. É o ponto de partida para a análise.

2. Scikit-learn

É uma biblioteca que tem várias coisas prontas para trabalhar com modelos de machine learning. Usei:

train_test_split: Para dividir os dados em treino e teste. Assim, eu treino o modelo com uma parte e vejo se ele funciona com a outra.

LogisticRegression: Escolhi porque é um modelo simples e fácil de entender. Perfeito para começar e ver se os dados têm algum padrão.

accuracy_score, classification_report, confusion_matrix: Essas ferramentas me ajudam a avaliar se o modelo está bom, mostrando números e relatórios sobre como ele está se saindo.

SimpleImputer: Serve para preencher dados que estão faltando. Não dá para treinar um modelo com dados incompletos, então isso é essencial.

StandardScaler: Ele deixa os dados todos numa mesma escala, o que ajuda o modelo a não se confundir.

3. SMOTE

Por que usei? Às vezes, tem muito mais exemplos de uma classe do que da outra, como 90% de um e só 10% de outro. O SMOTE cria mais exemplos da classe menor para balancear os dados e ajudar o modelo a funcionar melhor.

4. Matplotlib e Seaborn

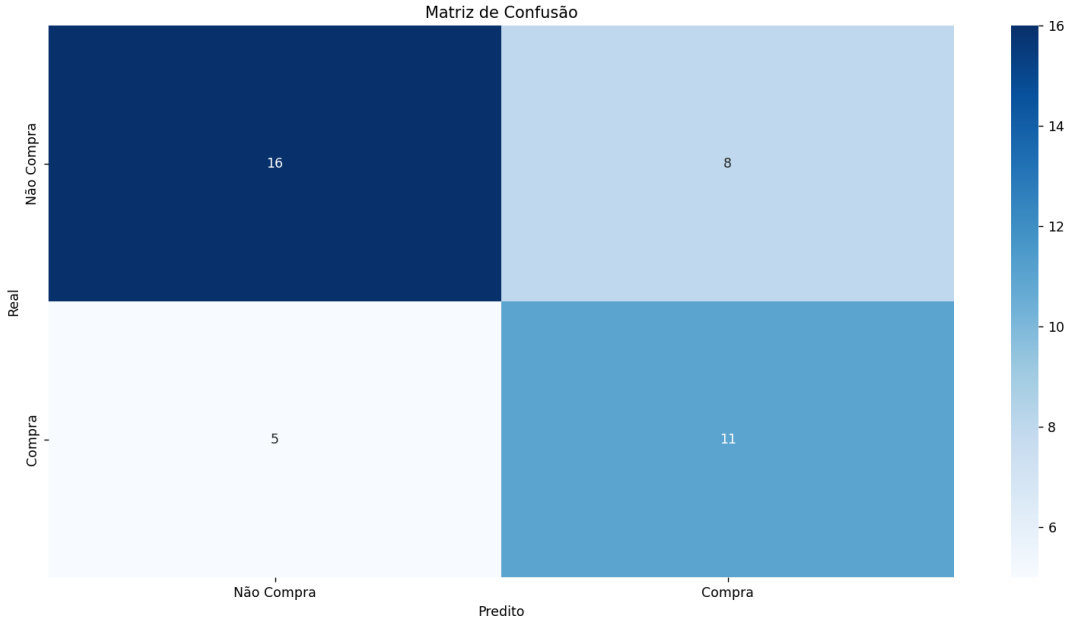
Por que usei? Essas bibliotecas são para fazer gráficos. Com elas, consigo visualizar os dados, como as relações entre variáveis ou os resultados do modelo. Ajuda muito a entender e explicar o que está acontecendo.

5. NumPy (import numpy as np)

Por que usei? O NumPy é como uma calculadora muito rápida e eficiente. Ele é usado para fazer contas e mexer em números grandes dentro do código.

Resultados

A matriz de confusão apresentada mostra como o modelo de classificação está se saindo ao prever se um cliente irá **comprar** ou **não comprar**. Ela compara os valores reais com os valores previstos pelo modelo, dividindo os resultados em 4 categorias:



1. **Verdadeiro Negativo (16)**: O modelo previu corretamente que 16 clientes não iriam comprar.
2. **Falso Positivo (8)**: O modelo previu que 8 clientes iriam comprar, mas, na verdade, eles não compraram.
3. **Falso Negativo (5)**: O modelo previu que 5 clientes não iriam comprar, mas, na verdade, eles compraram.
4. **Verdadeiro Positivo (11)**: O modelo previu corretamente que 11 clientes iriam comprar.

Interpretação

- O modelo acertou **27 vezes** no total (16 + 11) e errou **13 vezes** (8 + 5).
- A maioria dos erros vem de prever que alguém compraria (8 erros de falso positivo).
- O modelo conseguiu identificar uma boa quantidade de compradores (11 acertos).

Análise dos dados

Visualizando apenas os dados iniciais, foi possível seguir com a tarefa de criação de um modelo, entretanto, quando realizamos uma análise de base de conversão, outras informações seriam úteis, dentre eles:

- Valor e localização do imóvel, analisar região x valor de imóvel é importante;
- Quantidade de cliques por anúncio, com isso podemos analisar se determinados anúncios geram mais conversões e
- Mais dados de compradores, para um treinamento mais assertivo.

Cliques em anúncios parecem ser um fator positivo para uma conversão, assim, com divulgação ativa, podemos ter mais conversões.

Exemplo: Se um usuário X preenche um formulário de interesse em um site e clica em um anúncio, podemos nos basear em seus gostos e realizar novas sugestões personalizadas, além de sugestões no próximo site, podemos tentar engajar de maneira externa.

Pode ser estudado a quantidade média de cliques que o usuário faz antes de fazer uma conversão, com base nesse volume, podemos tentar medir a quantidade de mensagens ativas que podemos enviar para levar a uma conversão, caso o engajamento aumente com os disparos de mensagens, dependendo do custo da campanha, disparo ativo pode levar a mais conversões.

Conclusão

Com base nos resultados obtidos, podemos concluir que o modelo apresentou um desempenho satisfatório, com acertos relevantes em suas previsões. No entanto, ainda existem oportunidades de melhoria. Embora o modelo tenha

conseguido identificar corretamente muitos dos casos, a ocorrência de erros, como previsões incorretas em algumas situações, aponta para a necessidade de ajustes no processo de treinamento ou a exploração de abordagens alternativas para alcançar uma performance mais robusta.