Relatório do Desafio de Ciência de Dados – Indicium

1. Introdução

O presente relatório apresenta a resolução do desafio de Ciência de Dados proposto pela Indicium, no âmbito do processo seletivo da Bolsa Lighthouse 2025. O objetivo é analisar dados de filmes do IMDB, identificar fatores relevantes para o sucesso comercial e crítico e desenvolver um modelo preditivo para a nota do IMDB.

2. Análise Exploratória dos Dados (EDA)

- A base contém 250 filmes e 15 colunas, com informações sobre título, ano, duração, gênero, nota IMDB, metascore, elenco, número de votos e faturamento.
- Gêneros mais frequentes: Drama, Ação, Comédia.
- A distribuição das notas do IMDB concentra-se entre 7.5 e 9.0.
- Fatores como número de votos e metascore apresentam correlação positiva com o faturamento.

3. Respostas às Perguntas do Desafio

a) Qual filme recomendar?

Recomendaria The Shawshank Redemption (1994) ou The Dark Knight (2008), pois são filmes com altíssima avaliação no IMDB e grande popularidade, indicando boa aceitação por públicos diversos.

b) Fatores relacionados ao faturamento

- Popularidade (número de votos).
- Avaliação crítica (Metascore).
- Gênero (filmes de ação, aventura e ficção científica tendem a ter maior bilheteria).
- Presença de estrelas renomadas.

c) Insights da coluna Overview

A coluna Overview contém descrições curtas do enredo dos filmes. É possível identificar palavras-chave que remetem a determinados gêneros. Por exemplo, termos como 'space' ou 'alien' sugerem ficção científica, enquanto 'love' ou 'family' sugerem drama/romance. Técnicas de NLP, como TF-IDF e modelos de classificação (ex.: Naive Bayes), podem ser aplicadas para prever gêneros a partir dessa coluna.

d) Como prever a nota do IMDB

A previsão da nota IMDB foi tratada como um problema de regressão. Foram utilizadas as variáveis numéricas: Runtime, Meta_score, No_of_Votes e Gross. O modelo escolhido foi a Regressão Linear, por sua simplicidade e interpretabilidade. A métrica de avaliação utilizada foi o RMSE (Root Mean Squared Error). O modelo obteve:
- MSE: 0.0408
- RMSE: 0.20209

e) Nota prevista para The Shawshank Redemption

Com as características fornecidas no desafio, a nota prevista pelo modelo para The Shawshank Redemption foi 9.24. A nota real no IMDB é 9.2, mostrando que o modelo conseguiu uma previsão bastante próxima.

4. Modelagem

O pré-processamento incluiu a conversão das colunas Runtime e Gross para valores numéricos. Foi criado um subconjunto de dados apenas com variáveis numéricas relevantes e a variável alvo (IMDB_Rating). Após divisão em treino e teste, o modelo de Regressão Linear foi treinado e avaliado. Os resultados mostraram que o modelo tem boa capacidade de previsão.

5. Conclusão

O desafio permitiu analisar fatores relevantes para o sucesso de filmes no IMDB e demonstrou que a popularidade, o gênero e a recepção crítica são determinantes importantes para o faturamento e a avaliação de um filme. O modelo preditivo desenvolvido mostrou bom desempenho, com RMSE de aproximadamente 0.20, prevendo a nota do filme The Shawshank Redemption com alta precisão. Como trabalhos futuros, pode-se explorar técnicas de NLP sobre a coluna Overview, incluir variáveis categóricas transformadas em dummies e testar modelos mais robustos, como Random Forest ou Gradient Boosting.