Relatório Interpretativo – Projeto IMDb
Desafio Lighthouse Indicium + PProductions

1. Interpretação e Hipóteses Testáveis

1.1 Principais padrões observados

A partir da análise exploratória e visual dos dados, foram identificados os seguintes padrões:

- Filmes com orçamento elevado tendem a ter maior visibilidade e potencial de faturamento.
- Atores e diretores recorrentes aparecem com frequência em filmes bem avaliados.
- Gêneros como drama, ação e aventura são predominantes na base e apresentam boa aceitação.
- Filmes com notas altas no IMDb nem sempre correspondem aos de maior bilheteria.
- A classificação indicativa influencia o alcance do público, sendo mais ampla em filmes com classificação livre ou PG.

1.2 Hipóteses formuladas

Com base nos padrões observados, foram formuladas as seguintes hipóteses:

- H1: Filmes com maior orçamento apresentam maior bilheteria.
- H2: Filmes com atores recorrentes têm maior nota média no IMDb.
- H3: Filmes de gêneros populares (ação, aventura, drama) recebem mais votos do público.
- H4: Filmes com classificação indicativa mais acessível têm maior número de visualizações.
- H5: Filmes premiados no Oscar têm maior nota no IMDb, mas não necessariamente maior bilheteria.

Essas hipóteses podem ser testadas com modelos estatísticos ou algoritmos de machine learning em etapas futuras do projeto.

1.3 Limitações da análise

Alguns fatores limitam a profundidade da análise:

- A bilheteria não está disponível para todos os filmes da base.
- A classificação indicativa pode variar conforme o país de origem.
- A coluna “Overview” não foi explorada com técnicas de NLP, mas possui potencial para inferência de gênero e análise temática.
- Filmes lançados diretamente em plataformas de streaming não possuem dados de bilheteria, o que pode distorcer comparações.

1.4 Sugestões para aprofundamento

- Aplicar modelos preditivos para estimar bilheteria ou nota IMDb com base nas variáveis disponíveis.
- Utilizar técnicas de NLP para classificar o gênero dos filmes a partir da sinopse.
- Investigar correlações entre prêmios recebidos e sucesso comercial.
- Avaliar o impacto da classificação indicativa na popularidade dos filmes.

2. Respostas às Perguntas do Desafio

2.1 Qual filme você recomendaria para uma pessoa que você não conhece?
Com base nos dados analisados, o filme recomendado é **Forrest Gump,** sustentado por múltiplos fatores:

- Gênero: Drama, o mais frequente na base e com ampla aceitação.
- Ator: Tom Hanks, presente entre os atores mais recorrentes e bem avaliados.
- Diretor: Robert Zemeckis, cineasta premiado e reconhecido.
- Nota IMDb: 8.8, entre as maiores da base.
- Classificação indicativa: PG-13, acessível para públicos diversos.

Essa combinação torna Forrest Gump uma escolha segura e universal, ideal para recomendação sem conhecer o perfil da pessoa.

2.2 Quais são os principais fatores relacionados à expectativa de faturamento de um filme?

- Orçamento alto: Filmes com mais investimento tendem a faturar mais.
- Popularidade: Muitos votos e nota alta no IMDb indicam interesse do público.
- Elenco e direção famosos: Nomes conhecidos atraem mais espectadores.
- Gênero comercial: Ação, aventura e ficção científica costumam gerar mais receita.
- Classificação indicativa acessível: Filmes para todos os públicos têm maior alcance. Esses fatores aumentam a chance de sucesso financeiro, mas não garantem bilheteria alta.

2.3 Quais insights podem ser tirados da coluna Overview? É possível inferir o gênero do filme a partir dela?
A coluna Overview traz a sinopse dos filmes, com palavras que indicam temas e estilo. É possível identificar padrões de linguagem que sugerem o gênero (ex: “crime”, “amor”, “guerra”, “fuga”). Com técnicas de processamento de linguagem natural (NLP), é viável classificar o gênero com boa precisão.
Apesar de não ter sido aplicada essa análise no projeto, a coluna tem potencial para gerar insights valiosos.

2.4 Explique como você faria a previsão da nota do imdb a partir dos dados. Quais variáveis e/ou suas transformações você utilizou e por quê? Qual tipo

de problema estamos resolvendo (regressão, classificação)? Qual modelo melhor se aproxima dos dados e quais seus prós e contras? Qual medida de performance do modelo foi escolhida e por quê?

A previsão da nota do IMDb foi tratada como um problema de regressão, pois o objetivo é estimar uma variável contínua (nota entre 0 e 10).
Variáveis utilizadas e transformações aplicadas:

- Released_Year: mantida como variável numérica.
- Runtime: convertido de string para número inteiro (minutos).
- Meta_score: mantido como numérico.
- No_of_Votes e Gross: transformados com logaritmo natural (usando np.log1p) para reduzir a dispersão e a influência de outliers.
- Genre: tratado com One-Hot Encoding, separando os gêneros por vírgula e criando uma coluna binária para cada um. Essas variáveis foram selecionadas por representarem características relevantes que influenciam a avaliação de um filme, como popularidade, desempenho comercial, crítica especializada e gênero.

Modelo escolhido: Random Forest Regressor
Prós:

- Capta relações não lineares entre variáveis
- Robusto a outliers e ruído
- Não exige normalização
- Funciona bem com grande número de variáveis

Contras:

- Menor interpretabilidade
- Pode demandar mais tempo de processamento
- Tende a superestimar em casos com poucos dados

Métricas de avaliação:

- MSE (Erro Quadrático Médio): 0.0351
- R^2 (Coeficiente de Determinação): 0.5810

Esses resultados indicam desempenho razoável na previsão da nota do IMDb com os dados disponíveis.

2.5 Qual seria a nota do IMDb para um filme com as seguintes características?
Filme: The Shawshank Redemption Características fornecidas:

- Released_Year: 1994
- Runtime: 142 min
- Genre: Drama
- Meta_score: 80.0
- No_of_Votes: 2.343.110
- Gross: 28.341.469

Resultado da previsão:
A nota prevista pelo modelo foi **8.79.**

Nota real no IMDb: **9.3/10**

Interpretação:
O modelo reconheceu o filme como altamente avaliado, embora não tenha atingido exatamente o valor real. A diferença sugere que há espaço para melhorias, como o uso de variáveis adicionais (sinopse vetorizada, elenco codificado, dados de crítica especializada).

Conclusão:
Apesar da diferença, o modelo demonstrou capacidade de generalização e entregou uma previsão consistente. O refinamento será considerado em versões futuras, mas o desempenho atual já é satisfatório para uma primeira entrega.

3. Recomendação Estratégica para a PProductions

Com base na análise exploratória, visualizações, interpretações e respostas às perguntas do desafio, foi possível identificar os seguintes pontos-chave:

- Gêneros com maior aceitação: Drama, Ação e Aventura se destacam tanto em popularidade quanto em presença na base.
- Perfis de sucesso: Filmes com orçamento elevado, atores e diretores renomados, e classificação indicativa acessível tendem a ter maior alcance e faturamento.
- Popularidade e engajamento: Métricas como número de votos e nota IMDb são bons indicadores de interesse do público.
- Exemplo de referência: Forrest Gump representa um equilíbrio entre qualidade, apelo emocional, elenco forte e sucesso comercial — sendo uma inspiração para futuras produções. Dessa forma, recomenda-se que o próximo filme da PProductions seja:
- De gênero dramático com elementos universais (como romance ou superação).
- Com elenco reconhecido e direção experiente.
- Com classificação indicativa ampla, para atingir diferentes faixas etárias.
- Com investimento estratégico em produção e divulgação, visando alto engajamento. Essa abordagem aumenta as chances de sucesso crítico e comercial, alinhando dados históricos com tendências de mercado.

4. Referência Técnica

Para detalhes técnicos, gráficos, código-fonte e execução completa do projeto, consulte o notebook completo disponível no repositório:
GitHub:https://github.com/Cantalixto/desafio-lighthouse-cd
Esse repositório contém:

- O notebook de análise e preparação dos dados
- O notebook de construção e avaliação do modelo preditivo
- Scripts e funções utilizadas no projeto
- Gráficos gerados durante a análise exploratória
- Testes práticos com previsão de nota IMDb