Esse relatório visa responder às questões do desafio técnico proposto.

1. Análise Exploratória de Dados (EDA)

A análise iniciou-se com a preparação e limpeza dos dados para explorar possíveis hipóteses do DataFrame.

Primeiramente, as colunas com dados faltantes (Certificate, Meta-score e Gross) foram tratadas. A coluna Certificante preenchida com a moda, já que a coluna é categórica, enquanto Meta-score e Gross foram preenchidas com a mediana para lidar melhor com os outliers. Além disso, houve conversões nos tipos, para possibilitar a melhor exibição nos gráficos, análise estatística e cálculos.

A análise inicial revelou que o sucesso de um filme, que é medida pelas avaliações das críticas, não está diretamente ligado com seu sucesso financeiro. O principal fator que afeta o faturamento de um filme é sua popularidade (quantidade de votos), com uma relação de 0,60 na matriz de correlação. No entanto, a relação entre a avaliação do IMDB e o faturamento é fraca. Portanto, filmes que geram mais engajamento e discussão (resultando em mais votos) tendem a faturar mais. A popularidade de um filme são indicadores de bilheterias mais fortes do que sua avaliação crítica ou do público. O mesmo padrão segue para a análise de diretores e atores, onde os mais aclamados pela crítica não são, necessariamente, os que geram maior receita.

O gênero Drama é o que se destaca, sendo ele o mais comum no conjunto de dados. Aparecendo em franquias bem consolidadas como “The Dark Night” e “The Godfather”, indicando uma preferência do público por esse gênero. No entanto, o gênero Drama domina apenas em quantidade, enquanto gêneros como Aventura, Ação e SCI-FI possuem maior média de faturamento.

Além disso, foram exploradas as diversas classificações presentes no DataFrame. Após agrupar as diferentes classificações (EUA, Índia, etc.) em categorias globais (G, PG, R), notou-se que filmes com classificações que sugerem orientação dos pais, foram os que alcançaram a maior média de faturamento, possivelmente por atingirem um público mais amplo.

2.

a) Com base nas análises, o melhor filme a se recomendar é Dark Knight, devido a seu sucesso em bilheteria, nas críticas do público e por ser dirigido por Christopher Nolan, o diretor mais aclamado, além de juntar os gêneros mais populares como Ação, Crime, e Drama.

b) A principal coluna que está relacionada ao faturamento de um filme, conforme visto anteriormente, é o número de votos. Além disso, atores e diretores populares como Robert Downey Jr, Chris Evans e Anthony Russo, presentes em franquias de grande sucesso (Vingadores) possuem uma forte influência no engajamento e consequentemente no número de votos.

c) Baseado nas palavras utilizadas podemos extrair emoções, temas e narrativas, e com base nisso, poderíamos fazer uma relação para verificar o que interessa mais ao público. É totalmente possível inferir o gênero do filme com base na sinopse, é um problema de classificação, onde é possível associar palavras a gêneros específicos e daí tirar uma conclusão.

3.

A nota do IMDB, por ser um valor numérico e contínuo, se encaixaria num problema de regressão. Nesse caso, eu utilizaria variáveis que intuitivamente influenciaram, além de utilizar meus insights da análise anterior. Assim, as variáveis de entrada seriam: Genre, Released_Year, No_of_Votes, Runtime e Gross. Para isso, seria necessário pré processar os dados, como Genre é uma variável categórica, eu utilizaria a técnica One-Hot-Encoding, que visa transformar os gêneros em colunas numéricas, isso se deve ao fato de que o modelo de regressão apenas aceita variáveis numéricas. Além disso, as demais colunas utilizadas foram transformadas em numéricas. O modelo escolhido foi o RandomForest, pois ele é robusto e consegue extrair relações não lineares entre as colunas, além de possuir um alto desempenho. O maior problema de se usar esse modelo é seu alto custo computacional. As medidas de desempenho utilizadas foram: RMSE(Raiz do Erro Quadrático Médio), R^2(Coeficiente de Determinação) e MAE(Erro Médio Absoluto). Elas foram utilizadas pois, RMSE é a métrica mais comum em problemas de regressão e é muito útil para identificar a presença de grandes erros, pois eles são amplificados pelo processo de elevação ao quadrado. R^2 é uma forma de explicar o quão bem as variáveis de entrada conseguem prever a variável de saída. Já o MAE, pela sua facilidade de interpretação, pois ele estão na mesma unidade da variável preditora.

4.

A nota foi de 8.81