Desafio Cientista de Dados — Análise Exploratória (EDA) e Modelagem Preditiva de Ratings IMDB

Resumo executivo

Este projeto analisou uma base cinematográfica, enriquecida com informações do TMDB (the movie database), e desenvolveu um modelo para prever as notas do IMDB, métrica crucial para o estúdio PProductions, que precisa selecionar qual o tipo de filme deve ser produzido levando em consideração insights de rentabilidade e popularidade.

Para o projeto usamos os seguintes dados:

Informações Básicas

- Title: Nome completo do filme
- Year: Ano de lançamento (importante para análise temporal)
- Certificate: Classificação etária (ex: PG, PG-13, R)

Metadados Técnicos

- Runtime: Duração em minutos
- Genre: Gênero cinematográfico

Avaliações e Engajamento

- Rating: Nota do IMDB
- Overview: Sinopse e descrição
- scoreAvg: Média ponderada das críticas especializadas
- Votes: Volume total de votos

Métricas Financeiras

- Revenue: Receita doméstica do filme
- tmdb_revenue: Receita global total
- tmdb_budget: Orçamento total de produção

Dados TMDB (Enriquecimento)

- tmdb_popularity: Índice de popularidade na plataforma
- tmdb_vote_count: Número de avaliações no TMDB
- tmdb_vote_average: Nota média no TMDB

Com esses dados foi realizada uma Análise Exploratória (EDA) para entender variáveis como receita, orçamento, gêneros e engajamento do público. Em seguida, técnicas de engenharia de features foram aplicadas, incluindo TF-IDF para extrair os insights da coluna sinopse (Overview).

Três modelos principais foram usados:

- DecisionTreeRegressor (baseline)
- TF-IDF + Regressão Linear / OneVsRest (para Overview)
- LightGBM Regressor (modelo final)

O LightGBM se destacou, com desempenho robusto no conjunto de teste:

- RMSE: 0.1143
- MAE: 0.0768
- R^2: 0.8009

O modelo é capaz de explicar 80% da variabilidade das notas IMDB, com erro médio de apenas 0,08 pontos, sendo uma boa ferramenta para apoiar decisões de investimento e priorização de projetos do estúdio.

Principais observações da EDA

- **Distribuição das notas (Rating):** a maior parte dos filmes concentra-se entre 6.0 e 8.0. Valores extremos são pouco frequentes.
- **Receita (Revenue):** altamente enviesada, poucos filmes dominam quase toda a bilheteria

- **Correlação entre variáveis:**

  Votes e Revenue = 0.59: quanto mais votos é provável que o faturamento seja maior.

  Votes e Rating = 0.48: maior engajamento está associado a notas mais altas.

  tmdb_budget e tmdb_revenue = 0.78: orçamento fortemente ligado à receita. Quanto mais o estúdio investe no filme, maior é a bilheteria.

  tmdb_popularity e tmdb_revenue = 0.66 : popularidade tem impacto direto na receita.

- **Gêneros:** o dataset apresenta predominância de Dramas; contudo, Ação / Aventura/Ficção Científica tendem a gerar maior receita média, principalmente quando combinados a grandes orçamentos e franquias.

- **Overview (texto)** :As sinopses apresentam sinal semântico coerente com os gêneros (ex.: space, alien estão relacionados à Ficção Científica; murder, detective estão associados a Crime/Thriller).

  O modelo TF-IDF combinado com OneVsRest alcançou micro F1 = 0.409, valor limitado para certos gêneros, principalmente os menos frequentes.

Correlação e Insights sobre os novos dados

Há uma relação entre receita e número de votos (0.603), assim como entre orçamento e receita (0.783).

Filmes com maior orçamento e maior popularidade prévia costumam ter bilheterias maiores.

Popularidade pré-lançamento apresenta correlação significativa com a arrecadação (0.657), evidenciando a importância do marketing para a receita.

Existe forte conexão entre popularidade e votos (0.770), reforçando novamente a importância do marketing do filme.

O orçamento também está ligado ao número de votos (0.523), sugerindo que grandes produções chamam mais atenção.

Esses dados e insights indicam que filmes com maior orçamento, mais divulgação e maior popularidade têm maiores chances de sucesso comercial.

Orçamento influencia diretamente a receita

Hipótese: Filmes com maior orçamento (tmdb_budget) costumam gerar mais receitas (Revenue e tmdb_revenue).

Análise:

- **Qualidade da produção:** orçamentos maiores possuem diretores renomados, elenco famoso, tecnologia de ponta para efeitos visuais, bom marketing.

- **Marketing e divulgação:** campanhas mais extensas aumentam visibilidade e alcance, estimulando maior engajamento do público.

- **Distribuição:** recursos elevados permitem lançamentos em mais salas de cinema ao redor do mundo.

- **Relação entre dados:** a correlação de 0.783 entre orçamento e receita confirma o impacto positivo dos investimentos.

Conclusão:

O orçamento é um dos fatores centrais para o sucesso financeiro de um filme, mas não pode ser considerado o único (gênero, popularidade e críticas também influenciam).

Gênero influencia diretamente a receita

Hipótese: Filmes de Ação, Aventura e Ficção Científica tendem a gerar maior receita que outros gêneros.

Análise:

- **Apelo comercial:** esses gêneros atraem públicos diversos em escala global, com maior potencial de bilheteria.

- **Franquias e produtos derivados:** costumam se expandir em sequências, brinquedos, jogos e outros itens licenciados, ampliando receita

- **Experiência coletiva:** filmes de Ação e Aventura se tornam grandes eventos culturais, incentivando a ida ao cinema em grupo.

Conclusão:

Embora o dataset tenha predominância de Dramas, Ação, Aventura e Ficção Científica se destacam como gêneros mais lucrativos, ainda mais quando são combinados a orçamentos elevados e a criação de franquias. (Ex: Marvel, DC)

Respostas às perguntas do desafio:

Qual filme você recomendaria para uma pessoa que você não conhece?

Recomendação: The Goodfather.

Este longa tem a maior nota do IMDB(9.2) , além disso possui um scoreAvg (100.0) elevado, atores e direção renomados, e uma quantidade expressiva de votos. Essas características sugerem grande popularidade entre o público, tornando-o um filme de destaque quando se trata de recomendações.

Quais são os principais fatores relacionados com alta expectativa de faturamento?

Com base na EDA e nos dados adicionados do TMDB, os fatores mais relacionados a alto faturamento são:

Orçamento (tmdb_budget): possui alta correlação com receita, ou seja, mais investimento possibilita melhores produções e divulgação.

Gênero: Ação, Aventura e Sci-Fi tendem a ser mais famosos, e tem publico abrangente.

Popularidade pré-lançamento (tmdb_popularity): filmes mais populares antes de estrear arrecadam mais.

Elenco e diretor: nomes reconhecidos aumentam alcance comercial.

O que insights podem ser tirados com a coluna Overview? Dá para inferir gênero a partir dela?

A sinopse revela tema, tom e vocabulário característico. Para ilustrar, em filmes de suspense e horror, é comum vermos termos como "assassino" aparecendo mais vezes do que outros, como "namoro". Ademais, o tamanho do texto e a forma como ele é escrito,se é mais sério, tocante, ou com jargões dão indicações importantes sobre quem o filme busca atingir e a qual tipo específico ele pertence.

Como você faria a previsão da nota do IMDB a partir dos dados?

Para prever a nota do IMDB, considerando que se trata de um problema de regressão devido à natureza contínua da variável target (Rating), cria-se um pipeline que combina engenharia de features e modelagem. O pipeline aplicaria Target Encoding para diretor e gênero, capturando o desempenho histórico médio, Frequency Encoding para o elenco, refletindo sua popularidade, processamos a sinopse (Overview) com TF-IDF que consegue extrair padrões semânticos relevantes. Essas features são adicionadas com as variáveis numéricas. E por fim tem se o modelo final, LightGBM, ele foi escolhido devido à sua eficiência em dados tabulares, capacidade de lidar com valores ausentes. O resultado é um modelo robusto, com R^2 de 0,80 e RMSE de 0,11, capaz de prever ratings com alta precisão, oferecendo suporte confiável para decisões estratégicas de produção cinematográfica.

Predição solicitada — The Shawshank Redemption

{'Series_Title': 'The Shawshank Redemption',
'Released_Year': '1994',
'Certificate': 'A',
'Runtime': '142 min',
'Genre': 'Drama',
'Overview': 'Two imprisoned men bond over a number of years, finding solace and eventual redemption through acts of common decency.',
'Meta_score': 80.0,
'Director': 'Frank Darabont',
'Star1': 'Tim Robbins',
'Star2': 'Morgan Freeman',
'Star3': 'Bob Gunton',
'Star4': 'William Sadler',
'No_of_Votes': 2343110,
'Gross': '28,341,469'}

Após examinar o padrão estabelecido e levando em conta o cenário de 1994(período em que as avaliações cinematográficas variam entre 7,6 e 8,9), estima-se que Um Sonho de Liberdade alcance uma nota superior a 8,5. Essa projeção alinha-se com a percepção de que filmes do gênero Drama geralmente obtêm classificações mais elevadas. Adicionalmente, a direção de Frank Darabont, a presença marcante de Tim Robbins e Morgan Freeman no elenco, e a riqueza da narrativa fortalecem a crença de que o filme será muito bem avaliado pelo IMDB.