Relatório Desafio

Indicium para PProductions

Título: Relatório de Análise Cinematográfica
Da Plataforma IMDb

Autor: Andriel José da Silva

10/07/2024

andrieldata@proton.me

1. Introdução

Este relatório tem como finalidade fornecer uma análise detalhada para o desenvolvimento de um novo filme pela PProductions. O contexto do projeto envolve entender as tendências passadas e atuais do mercado cinematográfico, bem como avaliar o desempenho de filmes anteriores para tomar decisões informadas. O estudo visa identificar gêneros promissores, elencos e diretores eficazes, além de entender as preferências do público e a concorrência.

2. Metodologia

Para realizar esta análise, utilizamos diversas ferramentas e fontes de dados. Os dados internos foram coletados do IMDB,abrangendo varias métricas. Ferramentas de visualização de dados como Matplotlib e Seaborn foram utilizadas para criar gráficos e tabelas. A seguir, descrevemos os passos e as ferramentas utilizadas no processamento e análise dos dados:

1. ***Importação das Bibliotecas**: Utilizamos pandas para manipulação de dados, Matplotlib e Seaborn para visualização, e Scikit-Learn para modelagem preditiva.*
2. ***Carregamento dos Dados**: Os dados foram carregados a partir de um arquivo CSV.*
3. ***Limpeza e Pré-processamento**: As colunas foram convertidas para tipos de dados apropriados, e valores faltantes foram tratados.*

3. Respostas as Perguntas

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

	Series_Title	Released_Year	Certificate	Runtime	Genre	IMDB_Rating
0	The Godfather	1972.0	A	175.0	Crime, Drama	9.2
1	The Dark Knight	2008.0	UA	152.0	Action, Crime, Drama	9.0
2	The Godfather: Part II	1974.0	A	202.0	Crime, Drama	9.0
3	12 Angry Men	1957.0	U	96.0	Crime, Drama	9.0
4	The Lord of the Rings: The Return of the King	2003.0	U	201.0	Action, Adventure, Drama	8.9

De acordo com a minha análise acima eu recomendaria "The Godfather" (1972), pois tem a maior classificação IMDB de 9.2 e é amplamente considerado um clássico.

3. Respostas as Perguntas

2. **Quais são os principais fatores que estão relacionados com alta expectativa de faturamento de um filme?**

Para maximizar a expectativa de faturamento de um filme, vou considerar os seguintes fatores criados baseando-nos nos gráficos e dados analisados no projeto:

1. Escolha de gêneros populares como Adventure, Animation, e Thriller.
2. Otimize a duração do filme para entre 100 e 150 minutos.
3. Almeje classificações etárias amplas, como U/A e U, ou PG-13 para atrair um público diversificado.
4. Engaje o público desde cedo para gerar altos números de votos e discussões e, foque em qualidade para receber boas avaliações no IMDb.
5. Escolha diretores renomados ou bem avaliados para atrair mais público.

Esses fatores, combinados, podem aumentar significativamente as chances de um filme alcançar um alto faturamento.

3. Quais insights podem ser tirados com a coluna Overview? É possível inferir o gênero do filme a partir dessa coluna?

Atualmente, a coluna "Overview" não está fornecendo informações suficientes para inferir corretamente o gênero do filme, devido às limitações na vetorização, modelo simples de regressão logística.

Acredito que existem sim possíveis melhorias para melhorar a inferência do gênero do filme a partir da coluna "Overview", mas não cheguei nesse nível ainda de implementar modelos complexos de machine learning.

3. 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?

Pré-processamento e Transformações

SimpleImputer: Usado para lidar com valores faltantes, preenchendo-os com a média das respectivas variáveis.

Train-Test Split: Dividimos o conjunto de dados em treino e teste (80% e 20%) para avaliar a performance do modelo de maneira justa.

Tipo de Problema

Regressão: Estamos prevendo uma variável contínua (nota do IMDb), o que caracteriza o problema como um problema de regressão.

Modelo Utilizado

Linear Regression: Foi utilizado para treinar o modelo de previsão.

Prós: Simplicidade, interpretabilidade, fácil de treinar e testar.

Contras: Pode não capturar relações não lineares entre as variáveis, sensível a outliers.

Medida de Performance

Mean Squared Error (MSE): Mede a média dos erros quadrados entre os valores preditos e reais, fornecendo uma indicação de quão longe as previsões estão dos valores reais.

R-squared (R^2): Indica a proporção da variabilidade dos dados explicada pelo modelo, ajudando a entender o poder explicativo do modelo.

Avaliação e Resultados

MSE: 0.0487, indicando a magnitude dos erros de previsão.
R^2: 0.2578, mostrando que aproximadamente 25.78% da variabilidade na avaliação do IMDb é explicada pelas variáveis utilizadas.

Resumo

Para prever a nota do IMDb, usamos variáveis como ano de lançamento, duração, meta score, número de votos e receita bruta. Utilizamos uma regressão linear para resolver o problema de regressão. A performance foi avaliada usando MSE e R^2, com MSE de 0.0487 e R^2 de 0.2578, indicando que o modelo explica cerca de 25.78% da variabilidade na nota. Modelos mais complexos podem melhorar a precisão. O modelo de regressão linear foi capaz de fornecer previsões razoáveis, mas com limitações devido ao baixo R^2.