RESPOSTAS

quinta-feira, 4 de julho de 2024 20:37

2- Responder as seguintes perguntas:

A- Qual filme seria recomendado para uma pessoa que não conheço

The Lord of the Rings: The Fellowship of the Ring.
Esse filme foi escolhido selecionando a melhor relação entre popularidade(número de votos) e a avaliação do IMDB mais alta. Também foi levado em conta que já que não conheço a pessoa devo escolher um filme com classificação indicativa Livre.

B- Quais são os principais fatores que estão relacionados com alta expectativa de faturamento de um filme ?

Baseado no dataset do IMDB os principais fatores são: Número de votos, Ano de lançamento, se o filme foi dirigido pelo diretor Christopher Nolan, Se o ator Elijah Wood atuou, Tempo de duração do filme e se o ator Mark Hamill atuou no filme.
Basedo na combinação com os data sets do oscar temos outros fatores como:
Quantidade de indicações do filme ao oscar, quantidade de oscars ganhado pelo filme, se o ator/atriz ou diretor/diretoras já foram premiados pelo oscar anteriormente

C- Quais insights podem ser tirados com a coluna Overview? é possível inferir o gênero do filme a partir dela?

Sim nós temos algumas correlações por exemplo, filmes do Gênero "Crime" possuem palavras como young,murder,crime,family com frequência. Já filmes do gênero "Action": must, man, world, former. Filmes de Biografia: story, life, war. Comédia tem friends com frequência já drama tem love e woman.

Acesse a nuvem de palavras completa:
https://github.com/IanPerigoVianna/IMDB_PREDICTOR/tree/main/WordCloud

3- Fazer a previsão da nota do IMDB a partir

dos dados

A- Quais variáveis e transformações foram utilizadas

Foram utilizadas as variáveis numéricas (Faturamento, número de votos, duração do filme, ano de lançamento, média ponderada das críticas) Foi combinado com dois data sets de filmes, atores/atrizes, diretores/diretoras ganhadores do ocars e foram utilizado as variáveis (Quantidade de oscars ganho pelo filme, quantidade de indicações, se alguém do elenco já havia ganhado algum oscar)

B- Qual tipo de problema está sendo resolvido

Um problema de regressão onde queremos prever um valor contínuo (Nota do IMDB)

C- Qual modelo melhor se aproximou dos dados, prós e contras?

O modelo que mais se aproximou dos dados foi o regressor da biblioteca XGBoost. Esse modelo foi treinado com 9 features no eixo X e nosso alvo (IMDB_Rating) no eixo Y.
As features foram: Oscars_Ganhos, Indicacoes_Oscar, movie_oscar_winner, previous_oscar_winner, Released_Year, Runtime, Meta_score, No_of_Votes, Gross.
Prós:
Maior acurácia: Leva em conta se o elenco e o filme já ganharam o Oscar, o que pode ser um forte indicador de qualidade.
Capacidade de lidar com dados complexos: XGBoost pode capturar relações não lineares entre as features e a variável alvo, resultando em melhores previsões.
Eficiência computacional: XGBoost é altamente otimizado para velocidade e desempenho, utilizando técnicas de paralelismo e processamento distribuído.
Contras:
Necessidade de atualizações constantes: O data set com os ganhadores do Oscar precisa ser atualizado regularmente. Filmes recentes que ainda não foram atualizados na lista do Oscar podem afetar o desempenho do modelo.
Complexidade do modelo: XGBoost pode ser mais difícil de interpretar em comparação com modelos mais simples, o que pode dificultar a compreensão dos fatores que influenciam as previsões.
Dependência de tuning de hiper parâmetros: A eficácia

do XGBoost pode depender significativamente da escolha de hiper parâmetros, o que requer tempo e recursos para otimização adequada.

D- Qual medida de performance foi escolhida

A medida de performance escolhida foi a Acurácia, calculada pela subtração da porcentagem da Média Absoluta do Erro (MAPE) da perfeição (100%) e o R^2 Score.

Acurácia: Indicador que mostra o quão próximo o modelo está das previsões corretas. Uma acurácia alta significa que o modelo tem um bom desempenho geral.

MAPE (Mean Absolute Percentage Error): Mede a precisão do modelo, expressando o erro absoluto médio como uma porcentagem das observações reais. É útil por ser uma métrica intuitiva e fácil de interpretar.

R^2 Score: Avalia a proporção da variabilidade total dos dados explicada pelo modelo. Um R^2 próximo de 1 indica um modelo que explica bem a variabilidade dos dados, enquanto um valor próximo de 0 indica o contrário.