Documentação do desafio Indicium de Ciência de Dados
Concorrente: João Gilberto Pelisson Casagrande

Início - Respostas:

1). Faça uma análise exploratória dos dados (EDA), demonstrando as principais características entre as variáveis e apresentando algumas hipóteses relacionadas. Seja criativo!

R:
Suponho que as principais características sejam a nota do IMDB e o Meta Score, com certeza estes estão relacionados entre si, além disso suponho também que o diretor e um ou outro ator de peso influencie bastante na nota, sendo ao meu ver talvez a peça mais importante com relação a nota, por isso criei uma variável chamada “Director_Competence”, que representa a média da avaliação crítica e pública de todos os filmes que cada diretor realizou. Por último, o certificado também acho que vale a pena comentar pois mostra a quantas pessoas tal filme é acessível, mas não acho que importe para nota do filme e nem para faturamento, demonstrarei com uma correlação entre a nota do IMDB e o quanto vendeu o filme.
Além dessas hipóteses iniciais, penso que seria válido observar o “No_of_Votes”, pois a distribuição das notas deve variar ao longo dos anos, verificar a correlação entre número de votos e nota. Acredito que filmes mais recentes podem ter um comportamento diferente de filmes mais antigos, tanto em popularidade quanto em faturamento, algo que poderia ser avaliado cruzando ano de lançamento com “No_of_Votes” e “Gross”, devido a inflação e etc.

2. Responda também às seguintes perguntas: a. Qual filme você recomendaria para uma pessoa que você não conhece? b. Quais são os principais fatores que estão relacionados com alta expectativa de faturamento de um filme? c. Quais insights podem ser tirados com a coluna Overview? É possível inferir o gênero do filme a partir dessa coluna?

R:
a.Se eu não conheço a pessoa e seus gostos, e quero recomendar um filme que ela vá gostar, o mais lógico seria eu recomendar um filme com uma nota relativamente boa em relação a quantas pessoas assistiram o filme. Seria como se, neste caso, quanto maior o número de pessoas que viu o filme melhor, e se ele manteve mesmo assim uma nota boa, melhor ainda. Neste caso, eu fiz um sistema simples de pontuação contínua utilizando a nota do IMDB e o No_of_Votes, expressa da seguinte forma:

Recommended = (IMDB_Rating+Pontuation)/2

Pontuation = min(10,max(1,No_of_Votes/k))

Neste caso K é um fator de escala, no caso foi utilizado 200000. Então 200k votos daria +1 ponto até um máximo de 10. Exemplo: 50k votos da aproximadamente 1.25 → vira 1. 1.2M votos ≈ 6 → vira 6. 2M votos ≈ 10 → satura no máximo. Utilizei esta formula pois assim

evita-se que os blockbusters dominem esse critério só por venderem muito, apesar de serem sim os melhores a se recomendar quando além de venderem muito, tem uma nota boa. Neste caso, o programa que fiz deixou The Dark Knight em primeiro lugar para se recomendar.

b. Com relação ao faturamento do filme, temos o fator acessibilidade ao público, sendo essa uma correlação entre a faixa etária dele (certificate), o assunto/tema que ele trata, super heróis por exemplo tende a ter uma aderência muito grande do público, o gênero do filme, ação e aventura tendem a ter também uma aderência grande, e por fim a duração dele. Também diria que o investimento do filme e no marketing são extremamente importantes para esse resultado, mas a tabela não os traz, então por motivos de estratégia de recursos, trabalharemos somente com o que temos em mãos. Outro ponto a considerar seria o ano de lançamento. O mercado atual de cinema movimenta muito mais dinheiro do que décadas atrás, então filmes mais recentes têm naturalmente maiores chances de faturar alto, mesmo que em termos proporcionais isso varie.

c. O que podemos tirar de insights do overview é por exemplo se o filme em questão é uma sequência ou não, se ele vai tratar de um assunto mais acessível e interessante ao público e por fim, sim, podemos inferir o gênero do filme através do que é dito na coluna de overview, filmes de herói, por exemplo, são via de regra de ação, etc, porém não é uma certeza. A análise de texto dessa coluna poderia ser feita com técnicas de NLP (processamento de linguagem natural), como TF-IDF ou embeddings de palavras, para identificar padrões de gênero. Isso não garante 100%, mas aumentaria a precisão da inferência.

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?

R:

Primeiramente faria uma nota média para cada diretor representando a média dos filmes que ele dirigiu com a nota do IMDB e do Meta Score (“Director_Competence”). Depois dessa média da nota entre os envolvidos, eu faria uma correlação com a Meta_score e o No_of_Votes do filme, e o restante das informações pode ser levado em conta mas com um peso muito menor, ou até desconsiderados por serem de menor importância neste aspecto da nota, por motivos de prática aqui vou desconsiderá-los.

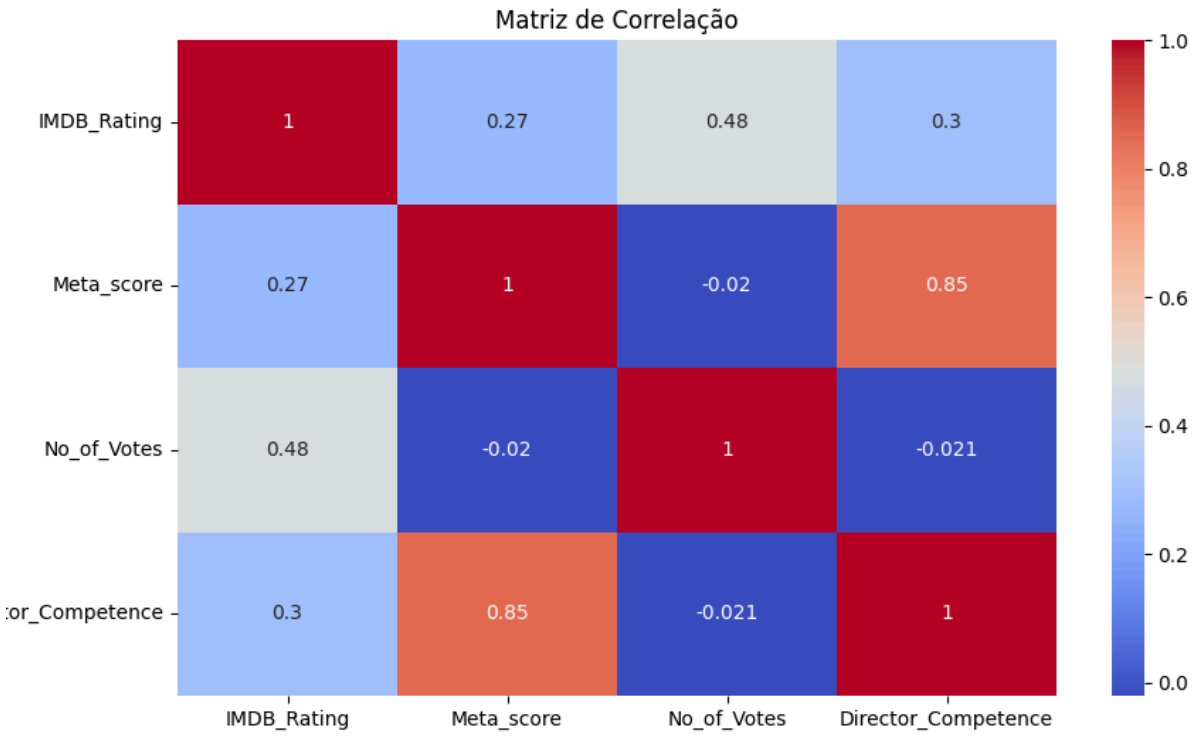
Esse é claramente um problema de regressão, já que queremos prever um valor contínuo, no caso, a nota. Além da média de notas por diretor e atores, outras variáveis que poderiam entrar no modelo são o número de votos, o gênero do filme e até o ano de lançamento. Para os modelos, e para este teste, eu começaria com uma regressão linear simples para ter uma base de comparação, mas também, com tempo, testaria algoritmos mais robustos como Random Forest ou Gradient Boosting, que conseguem capturar relações não lineares entre variáveis. Como métrica de performance, escolheria RMSE (Root Mean Squared Error) ou MAE (Mean Absolute Error), que mostram o quão distante em média as previsões ficam das notas reais.

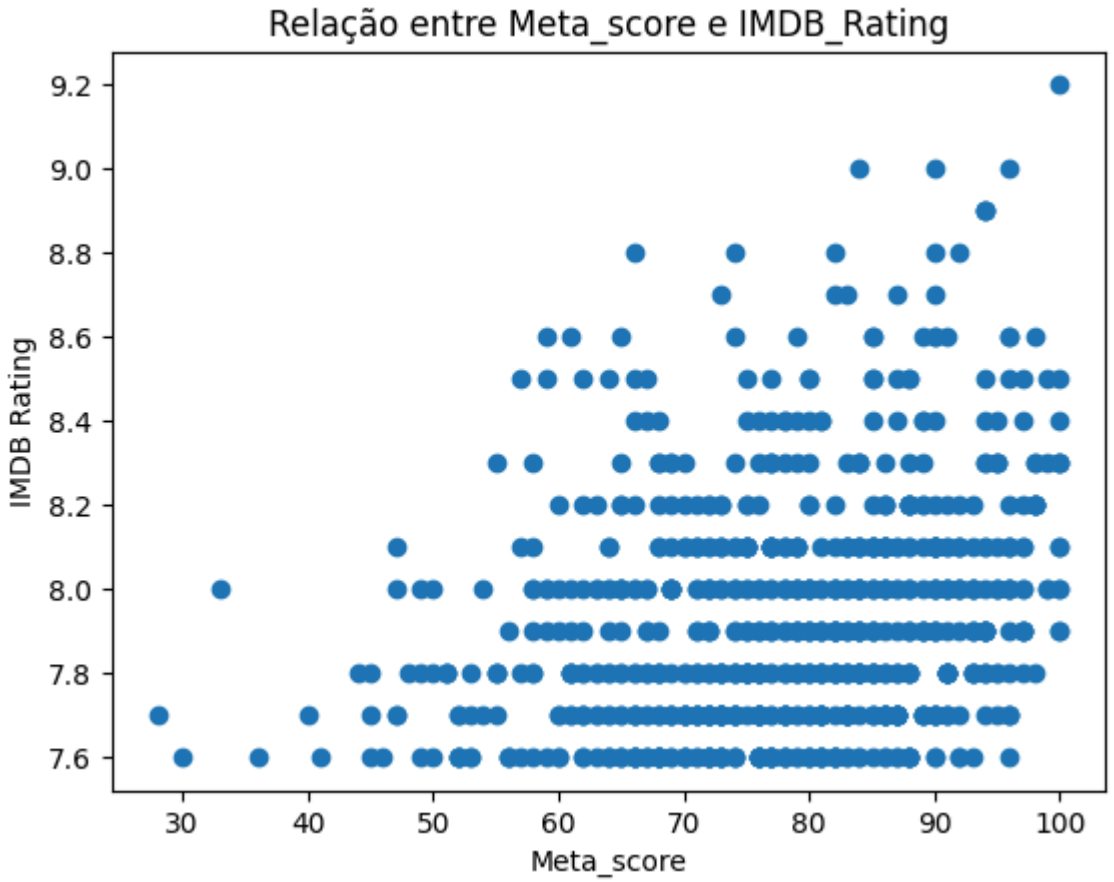
4. Supondo um filme com as seguintes características, qual seria sua nota do IMDB?

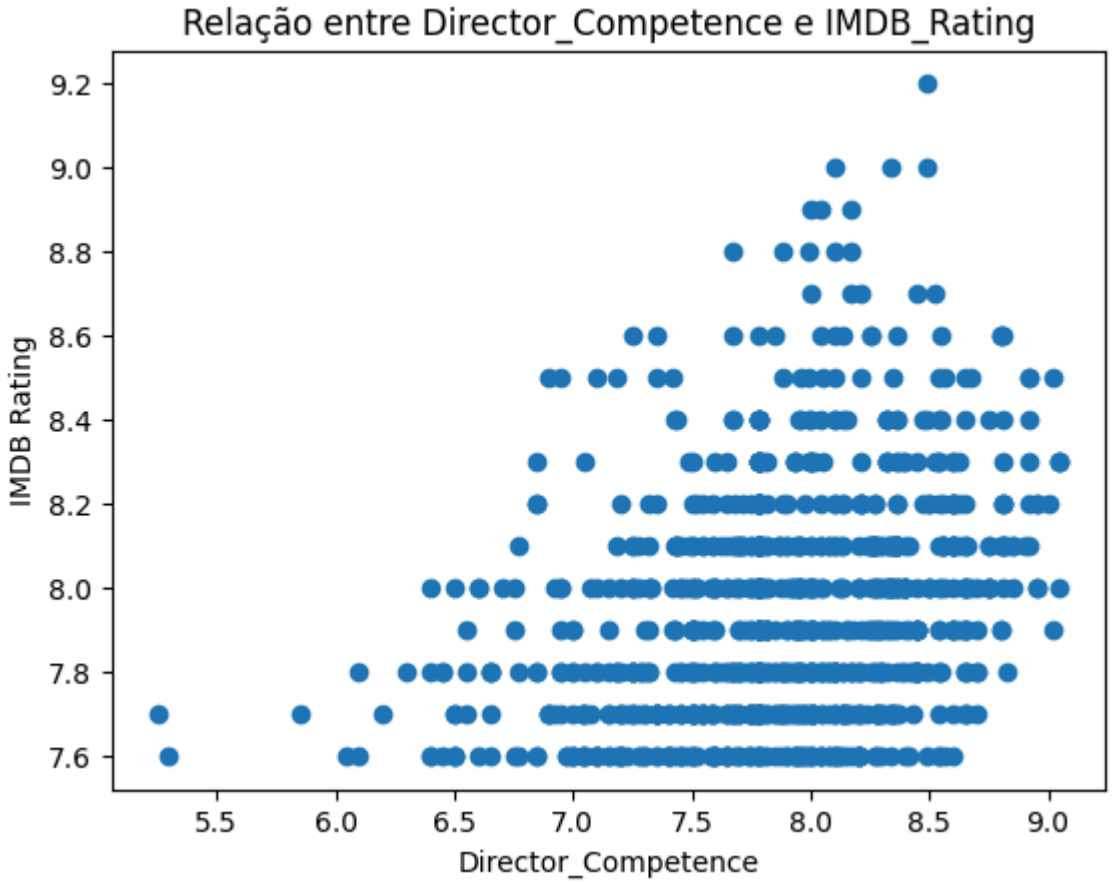
{'Series_Title': 'The Shawshank Redemption',
'Released_Year': '1994',
'Certificate': 'A',
'Runtime': '142 min',
'Genre': 'Drama',
'Overview': 'Two imprisoned men bond over a number of years, finding solace and eventual redemption through acts of common decency.',
'Meta_score': 80.0,
'Director': 'Frank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

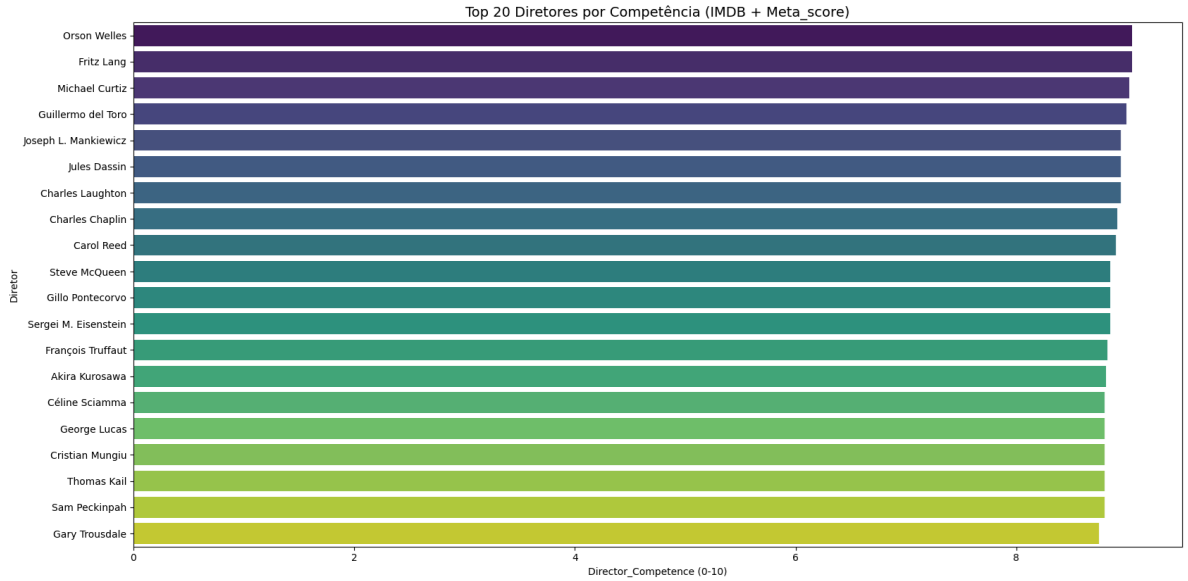
R:
Utilizando o método que criei na resposta 3, aproximadamente 8.83.

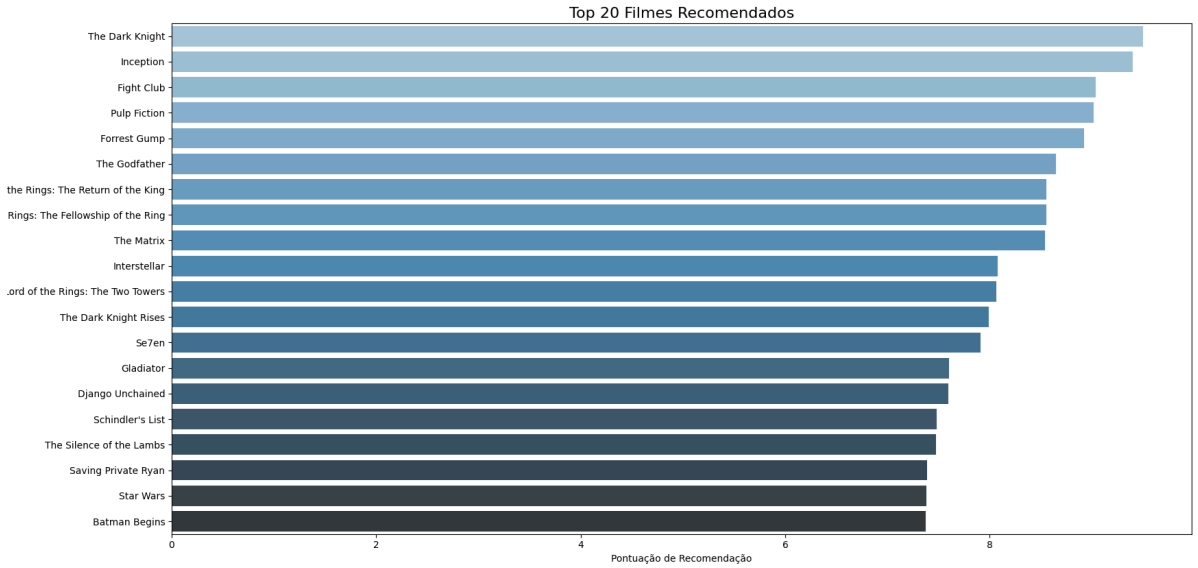
Resultados Gráficos e suas Interpretações:




Relação entre Meta_score e IMDB_Rating
IMDB Rating
Meta_score
9.2
9.0
8.8
8.6
8.4
8.2
8.0
7.8
7.6
30
40
50
60
70
80
90
100


Relação entre Director_Competence e IMDB_Rating
IMDB Rating
Director_Competence
9.2
9.0
8.8
8.6
8.4
8.2
8.0
7.8
7.6
5.5
6.0
6.5
7.0
7.5
8.0
8.5
9.0






```
Certificate A - Correlação IMDB_Rating x Gross: 0.09
Certificate UA - Correlação IMDB_Rating x Gross: 0.13
Certificate U - Correlação IMDB_Rating x Gross: 0.07
Certificate PG-13 - Correlação IMDB_Rating x Gross: -0.21
Certificate R - Correlação IMDB_Rating x Gross: 0.15
Certificate PG - Correlação IMDB_Rating x Gross: 0.07
Certificate G - Correlação IMDB_Rating x Gross: -0.34
Certificate Passed - Correlação IMDB_Rating x Gross: -0.63
Certificate GP - Correlação IMDB_Rating x Gross: -1.00
Certificate Approved - Correlação IMDB_Rating x Gross: -0.13
```

```
MAE: 0.1703484102179093
RMSE: 0.20846083157145134
R²: 0.5049323569214028
```

```
Nota prevista para 'The Shawshank Redemption': 8.83
```

Com a variável criada "Director_Competence" utilizada para avaliar e pontuar os diretores, foi possível enriquecer a análise de dados e aumentar a precisão do programa, podendo observar como a qualidade média dos diretores tem correlação com as notas do IMDB. A variável de competência variou entre 5.25 e 9.05, com média próxima de 7.88. Isso significa que, em geral, os diretores presentes no conjunto mantêm uma boa consistência na produção de filmes bem avaliados. Podemos ver, por exemplo, que o ranking dos diretores por competência trouxe de fato nomes clássicos e renomados, como Orson Welles, Fritz Lang, Akira Kurosawa e Charles Chaplin, todos reconhecidos pela consistência em manter alta qualidade em seus trabalhos. O gráfico mostra que diretores com uma longa carreira e relativamente estáveis em seus trabalhos naturalmente ficam melhores posicionados.

Podemos observar alguns padrões interessantes a partir da análise de correlação geral, a relação entre IMDB_Rating e Meta_score foi positiva, mas fraca (0.27), indicando que críticos e público tendem a concordar em parte, mas cada um avalia aspectos distintos de um filme. Já a correlação entre IMDB_Rating e No_of_Votes (0.48) mostrou que filmes com mais votos tendem a ter avaliações levemente melhores, o que sugere que a popularidade está relacionada a uma percepção de maior qualidade. Além disso podemos ver que a competência do diretor está fortemente diretamente relacionada ao Meta_score (0.85), já que também é parcialmente calculada por ele, mas enquanto a sua relação com o IMDB_Rating a correlação foi mais moderada (0.30), o que reforça que diretores mais competentes tendem a entregar filmes melhores, mas ainda há espaço para variações.

Nos gráficos de dispersão, essas tendências ficaram evidentes. Tanto na relação entre Meta_score e IMDB_Rating quanto na de Director_Competence e IMDB_Rating, observou-se um padrão ascendente, embora com bastante dispersão. Isso significa que embora haja tendência de que notas mais altas em uma métrica acompanhem a outra, ainda existem diversos filmes que fogem dessa regra, mostrando que qualidade pode ser percebida de formas diferentes e que informações como o quanto foi gasto em marketing de um filme entre outras variáveis também são importantes. Também quando analisamos a relação entre IMDB_Rating e Gross (bilheteria) por classificação indicativa (Certificate), os resultados mostraram correlações baixas ou até negativas. Por exemplo, em filmes UA e R, as correlações foram positivas, ainda que fracas (0.13 e 0.15, respectivamente). Já em classificações mais antigas como Passed e GP, os valores foram negativos fortes (-0.63 e -1.00), o que sugere problemas de representatividade, já que eram categorias pouco usadas e com poucos registros. Esses resultados nos mostram que, novamente, resultado por sí só não é um bom preditor de bilheteria, pois fatores como marketing, época de lançamento, apelo comercial, contexto social, etc, pesam muito mais.

Quanto à modelagem, a regressão linear, apesar de simples, conseguiu resultados satisfatórios. O modelo alcançou um MAE ≈ 0.17 e RMSE ≈ 0.20, o que demonstra erros pequenos da previsão de notas. O $R^2 \approx 0.50$ mostrou que cerca de metade da variabilidade das notas do IMDB pode ser explicada pelas variáveis consideradas (Meta_score, número de votos e competência do diretor), o que é razoável dado o caráter simples do programa e do modelo utilizado, e a falta de informações mais complexas como marketing, época de lançamento, contexto social do filme, etc. A exemplo temos o filme que foi dado para predição, The Shawshank Redemption, onde a nota prevista foi de 8.83, bastante próxima da avaliação real (9.3), o que mostra a coerência do modelo e uma boa predição.

Por fim, a criação de um sistema de recomendação baseado em pontuação contínua (IMDB + número de votos) trouxe uma lista muito consistente de grandes filmes, como The Dark Knight, Inception, Fight Club, Pulp Fiction, Forrest Gump e The Godfather. O que seria

de fato ideal como recomendação a uma pessoa desconhecida. Essa abordagem acabou por equilibrar popularidade e qualidade crítica, o que acaba sendo uma visão prática do que se recomendar para o público num geral.

De forma geral, os resultados indicam que o modelo e as análises são capazes de capturar relações relevantes dentro do dataset. Embora não seja possível prever com alta precisão o sucesso comercial de um filme a partir de somente de suas notas e das informações presentes, porém as métricas de avaliação, especialmente quando combinadas à competência do diretor, fornecem bons indicadores de qualidade percebida tanto por público quanto por crítica.