Desafio Cientista de Dados

Desafio

Você foi alocado em um time da Indicium contratado por um estúdio de Hollywood chamado *PProductions,* e agora deve fazer uma análise em cima de um banco de dados cinematográfico para orientar qual tipo de filme deve ser o próximo a ser desenvolvido. Lembre-se que há muito dinheiro envolvido, então a análise deve ser muito detalhada e levar em consideração o máximo de fatores possíveis (a introdução de dados externos é permitida - e encorajada).

Entregas

1. **Faça uma análise exploratória dos dados (EDA), demonstrando as principais características entre as variáveis e apresentando algumas hipóteses relacionadas. Seja criativo!**
   A base de dados é o arquivo 'desafio_indicium_imdb.csv', possuindo 15 colunas relacionadas aos dados de filmes. São 998 linhas x 15 colunas, cujas informações são de caráter numérico (float) e cadeias de caracteres (strings). Segue abaixo exemplo de elemento qualquer:

   'Series_Title': 'The Godfather',
   'Released_Year': '1972',
   'Certificate': 'A',
   'Runtime': '175 min',
   'Genre': 'Crime, Drama',
   'IMDB_Rating': 9.2,

'Overview': 'An organized crime dynasty's aging patriarch transfers control of his clandestine empire to his reluctant son.',

'Meta_score': 100.0,

'Director': 'Francis Ford Coppola',

'Star1': 'Marlon Brando',

'Star2': 'Al Pacino',

'Star3': 'James Caan',

'Star4': 'Diane Keaton',

'No_of_Votes': 1620367,

'Gross': '134,966,411'

Com base nesses dados, como poderíamos inferir a nota do IMDB de um filme qualquer? A primeira constatação são os números que apontam para o sucesso. Um filme aclamado possui uma elevada nota no imdb (IMDB_Rating) - sendo esta a variável de interesse, boa avaliação das críticas (Meta_score), recebeu muitos votos (No_of_Votes) - tornando-se popular, por consequência angariando um alto faturamento (Gross), entretando esta última encontra-se no formato de string. Então, ater-se as variáveis numéricas indicam um caminho para encontrar uma solução numérica (IMDB_Rating). As variáveis strings são pouco determinantes para encontrar uma solução, pois o nome do título ou a presença de uma estrela não determina a certeza do sucesso. Também, não desejamos classificar um filme, mas sim determinar a sua nota. Portanto, tratar o problema como Regressão é uma excelente alternativa. Devemos desenvolver um algoritmo que faça um tratamento prévio dos dados e nos aponte algumas estatísticas para análise, aplicando o modelo de Regressão Linear para treinamento e teste, fazendo as previsões adequadamente. O modelo dever ser avaliado usando a métrica do Erro Quadrático Médio (MSE) e possibilitar a inserção de novos dados (edições no código), já que o objetivo é prever a nota de qualquer filme.

Durante a execução do código, algumas estatísticas são exibidas. Segue exemplo abaixo:

	IMDB_Rating	Meta_score	No_of_Votes
count	999.000000	842.000000	9.990000e+02
mean	7.947948	77.969121	2.716214e+05
std	0.272290	12.383257	3.209126e+05
min	7.600000	28.000000	2.508800e+04
25%	7.700000	70.000000	5.547150e+04
50%	7.900000	79.000000	1.383560e+05
75%	8.100000	87.000000	3.731675e+05
max	9.200000	100.000000	2.303232e+06

Na imagem acima, podemos visualizar algumas estatísticas descritivas, como média, desvio padrão, mínimo e máximo para as variáveis numéricas de interesse.

Também, visualizamos alguns histogramas. No código, a partir das imagens geradas, é possível inferir visualmente para onde os valores numéricos convergem, facilitando a interpretação da base de dados.

Durante o processo da Análise Exploratória dos Dados, o algoritmo foi elaborado de modo a verificar valores ausente e preenchê-los, se houver necessidade. Segue abaixo os resultados:

```
Series_Title       0
Released_Year      0
Certificate      101
Runtime            0
Genre              0
IMDB_Rating        0
Overview           0
Meta_score       157
Director           0
Star1              0
Star2              0
Star3              0
Star4              0
No_of_Votes        0
Gross            169
dtype: int64
```

Na execução do código, precisamos transformar as variáveis categóricas (strings) em variáveis dummy (indicadoras, de caráter numérico binário), para então seguir com o ajuste e transformação das features do conjunto de dados. Os demais passos estão documentados no código, como também os resultados obtidos.

2. **Responda também às seguintes perguntas:**
    a. **Qual filme você recomendaria para uma pessoa que você não conhece?**

    Recomendaria um filme que possui alta avaliação no IMDB.

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

A nota no IMDB (IMDB_Rating), a média ponderada de todas as críticas (Meta_score) e número de votos (No_of_Votes).

c. **Quais insights podem ser tirados com a coluna *Overview?* É possível inferir o gênero do filme a partir dessa coluna?**
   A coluna overview fornece um resumo da trama relacionada ao filme, escrita de modo objetivo para fixar a atenção do espectador. Sim, é possível inferir o gênero a partir da coluna.

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?**
   Para fazer a previsão da nota do IMDB de um filme, é interessante analisar as características da variável de interesse (IMDB_Rating). Trata-se de um valor numérico, logo é possível determinar a solução com a aplicação de um modelo de Regressão Linear, sendo o modelo que melhor se aproxima dos dados (relação das variáveis independentes e dependente). Prós e contras da Regressão Linear: fácil implementação e formulação matemática simples; as relações lineares entre as variáveis independentes e a dependente podem não ser capturadas adequadamente, também sendo limitada a relacionamentos lineares.
   As variáveis utilizadas da base de dados são de valor numérico (float), como 'IMDB_Rating', 'Meta_score' e 'No_of_Votes', e as variáveis categóricas (strings), a exemplo de 'Series_Title' são transformadas em variáveis dummy de modo a serem empregadas no treinamento do modelo. A transformação é necessária por que a Regressão Linear não trabalha com valores não numéricos. A medida de performance do modelo escolhida foi o Erro Quadrático Médio (MSE), que avalia a precisão de um modelo de regressão: quanto menor o valor obtido, melhor a qualidade do ajuste, possuindo uma interpretação muito simples.

4. **Supondo um filme com as seguintes características:**

```
{'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'}
```

Qual seria a nota do IMDB?

Previsão de IMDB_Rating: 8.855721879186872.

Dicionário dos dados

A base de dados de treinamento contém 15 colunas. Seus nomes são auto-explicativos, mas, caso haja alguma dúvida, a descrição das colunas é:

Series_Title – Nome do filme

Released_Year - Ano de lançamento

Certificate - Classificação etária

Runtime – Tempo de duração

Genre - Gênero

IMDB_Rating - Nota do IMDB

Overview - *Overview* do filme

Meta_score - Média ponderada de todas as críticas

Director – Diretor

Star1 - Ator/atriz #1

Star2 - Ator/atriz #2

Star3 - Ator/atriz #3

Star4 - Ator/atriz #4

No_of_Votes - Número de votos

Gross - Faturamento