Relatório de Análise do Projeto

Akyla de Aquino Pinto

Github: Lighthouse

Email: akylaaquino@hotmail.com

LinkedIn: Akyla

Análise exploratória dos dados, características das variáveis e hipóteses relacionadas.

Relatório

O relatório detalhado das etapas de desenvolvimento e tratamento dos dados foi feita no script principal main.ipynb.

Questionamentos da Entrega

1. Faça uma análise exploratória dos dados (EDA), demonstrando as principais características entre as variáveis e apresentando algumas hipóteses relacionadas. Seja criativo!
    1. R: Análise e relatório feito no arquivo jupyter.

2. Responda também às seguintes perguntas:
    a. Qual filme você recomendaria para uma pessoa que você não conhece?
        1. The Lords of the Rings: The Return of the King, segundo a seção 1.4.3
    b. Quais são os principais fatores que estão relacionados com alta expectativa de faturamento de um filme?
        1. Segundo a seção 1.5.2, o número de votos, filmes do gênero ação e aventura e o filme ter faixa etária para maiores de 12 anos, isso influência positivamente nas vendas do filme.
        2. Quanto ao que influencia negativamente as vendas, temos o gênero de filmes do tipo Drama
    c. Quais insights podem ser tirados com a coluna *Overview*? É possível inferir o gênero do filme a partir dessa coluna?
        1. A coluna Overview apresenta um breve texto sobre o filme, realizando tratamento de dados com processamento de linguagem natural é possivel encontrar palavras chaves para identificar o gênero do filme, por exemplo.

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?
    1. Conforme a seção 2, Quanto ao modelo, iremos trabalhar com resultados que são gerados por meio de regressão. Porque? por

que nesse caso não faz sentido usarmos um modelo de classificação, pois nosso objetivo não é classificar algo, mas sim inferir um resultado com base nos dados existentes.

2. Nesse caso, iremos pegar as variáveis que foram transformadas para inteiro. No caso de modelos preditivos, como Regressão Linear ou Polinomial, precisamos analisar variáveis que não estejam correlacionadas linearmente através do Coeficiente de Pearson, ou seja, na matriz de correlação elas devem estar próximas de 0. Para evitar variáveis altamente correlacionadas pode ajudar a reduzir multicolinearidade, que pode afetar negativamente a precisão dos coeficientes estimados.
3. Foi utilizado medidores de desempenho, como o erro médio quadrático (MSE) e R2, para os modelos utilizados. Onde para MSE, quanto mais próximo de zero, melhor, pois indica menor erro nas previsões. Um MSE de 0 seria ideal, o que indicaria que o modelo está fazendo previsões perfeitas. Para o R2, quanto mais próximo de 1, melhor, pois indica que o modelo está explicando uma maior parte da variabilidade nos dados. Um R2 de 0 significa que o modelo não é capaz de explicar nenhuma variabilidade nos dados.
4. Dos nossos resultados, o modelo Random Forest, foi o que se saiu melhor.

4. Segundo a seção 3, o resultado para imdb com os dados do exemplo seria de aproximadamente 8,811.

.