Relatório Previsão da Nota do IMDb

Para prever a nota do IMDb usando as variáveis relevantes, foi necessário seguir uma abordagem estruturada que envolve a seleção de variáveis, construção de modelos, avaliação e seleção do modelo mais adequado. Abaixo estão os passos detalhados:

Seleção de Variáveis Relevantes

Primeiramente, analisamos as variáveis disponíveis no conjunto de dados para identificar quais poderiam ter um impacto significativo na nota do IMDb. As variáveis selecionadas foram:

- **'Series_Title'**: Não usaremos no modelo, pois o nosso primeiro modelo será regressão linear e usará apenas números, e esta coluna seria convertida para um identificador numérico apenas.
- **'Released_Year'**: Deixaremos este dado no modelo para análise futura. Talvez algum ano tenha produzido filmes melhores que outros.
- **'Certificate'**: Deixaremos a classificação etária no modelo. Possivelmente filmes com classificação mais ampla tenham mais expectadores, o que pode gerar mais avaliações.
- **'Runtime'**: Deixaremos este dado no modelo, pois a duração dos filmes pode influenciar nas indicações e avaliações.
- **'Genre'**: Deixaremos este dado no modelo porque o gênero define o tipo de assunto do filme, que está diretamente ligado ao gosto das pessoas, influenciando na avaliação.
- **'IMDB_Rating'**: Deixaremos no nosso modelo, será o nosso dado alvo (Nota do IMDb).
- **'Overview'**: O resumo do filme será retirado do modelo, pois é um conjunto de palavras que no nosso modelo de regressão linear não se aplica, já que iremos codificar dados não numéricos.
- **'Meta_score'**: Usaremos este dado no modelo, pois é a média ponderada de todas as críticas dos filmes.
- **'Director'**: Usaremos este dado no modelo, pois o diretor do filme influencia na produção do filme.
- **'Star1'**: Usaremos este dado no modelo, pois o ator/atriz principal influencia na produção do filme.
- **'Star2'**: Usaremos este dado no modelo, pois o ator/atriz secundário influencia na produção do filme.
- **'Star3'**: Usaremos este dado no modelo, pois o ator/atriz terciário influencia na produção do filme.
- **'Star4'**: Usaremos este dado no modelo, pois o ator/atriz quaternário influencia na produção do filme.
- **'No_of_Votes'**: Usaremos este dado no modelo, pois o número de votos indica a quantidade de pessoas que assistiram e decidiram classificar o filme.

- **'Gross'**: Usaremos este dado no modelo, pois o faturamento do filme indica que muitas pessoas assistiram ao filme.

As variáveis foram escolhidas com base em sua potencial influência na percepção da qualidade do filme, refletida na nota do IMDb.

Construção e Avaliação dos Modelos

- **Regressão Linear**:

A regressão linear é um modelo simples que assume uma relação linear entre as variáveis independentes e a variável dependente. Apesar de sua simplicidade, serve como uma boa linha de base para comparações.

- **Árvores de Decisão**:

As árvores de decisão são modelos não lineares que particionam os dados em subconjuntos baseados em valores de variáveis explicativas, criando uma árvore de decisões. Elas são interpretáveis e podem capturar relações complexas.

- **Random Forest**:

O Random Forest é um ensemble de árvores de decisão, que melhora a robustez e precisão ao reduzir o overfitting. Ele combina as previsões de várias árvores de decisão para obter uma previsão final.

- **Gradient Boosting**:

O Gradient Boosting é um método de ensemble que cria modelos de forma sequencial, onde cada novo modelo corrige os erros do modelo anterior. É poderoso para capturar padrões complexos nos dados.

Seleção da Métrica de Desempenho

Para avaliar os modelos, utilizamos duas métricas principais:

- **Root Mean Squared Error (RMSE)**:

A RMSE mede a raiz quadrada da média dos erros quadráticos, fornecendo uma medida da magnitude do erro. É uma métrica comum em problemas de regressão e foi escolhida porque penaliza fortemente grandes erros, ajudando a identificar modelos que preveem bem a maioria dos pontos.

Relatório Previsão da Nota do IMDb

- **R^2 (Coeficiente de Determinação)**:

O R^2 mede a proporção da variabilidade da variável dependente que é explicada pelas variáveis independentes no modelo. É uma métrica útil para entender o quão bem o modelo está explicando a variabilidade dos dados.

Ponto de Partida

- **Objetivo**: Criar um modelo capaz de prever a nota do IMDb de um filme com base em suas características principais, tais como ano de lançamento, duração, gênero, diretor, atores principais, número de votos e faturamento.
- **Dados Utilizados**: O dataset continha 999 filmes com diversas colunas, incluindo características categóricas e numéricas.

Processo para Regressão Linear

Link Notebook:
https://github.com/Raul-Lemelle/lighthouse_desafio_ciencia_dados/blob/main/notebooks/analysis_regressao_linear_imdb.ipynb

1. Preparação dos Dados:
    - As colunas foram selecionadas e pré-processadas adequadamente para garantir que os dados fossem aptos para o modelo de regressão linear.
    - Foi utilizado o SimpleImputer para lidar com valores nulos e o OneHotEncoder para transformar variáveis categóricas.
2. Treinamento do Modelo:
    - O modelo de regressão linear foi treinado utilizando um pipeline que integra o pré-processamento e o treinamento de forma eficiente.
    - A divisão dos dados em conjuntos de treinamento e teste permitiu avaliar a performance do modelo.
3. Avaliação:
    - O modelo foi avaliado usando o coeficiente de determinação (R^2), Root Mean Squared Error **(**RMSE), que indicou a qualidade da previsão.
4. Previsão para Novo Filme:
    - A previsão para um novo filme, "The Shawshank Redemption", foi realizada com sucesso, demonstrando a aplicabilidade prática do modelo.

Relatório Previsão da Nota do IMDb

Previsão Realizada: A nota prevista do IMDb para o filme "The Shawshank Redemption" foi coerente com a expectativa, considerando suas características de alta qualidade e histórico de sucesso.

Ao comparar a nota do IMDb no site oficial (9.3) com a previsão feita pelo nosso modelo de regressão linear (9.1), observamos uma diferença mínima de apenas 0.2 pontos. Esta pequena variação indica que nosso modelo está bastante próximo da avaliação oficial, o que sugere que ele é capaz de capturar bem as características e fatores que influenciam as notas dos filmes no IMDb.

Link IMDb do Filme:

The Shawshank Redemption](https://www.imdb.com/title/tt0111161/?ref_=chttp_t_1

Processo para Random Forest Regressor

Link Notebook:
https://github.com/Raul-Lemelle/lighthouse_desafio_ciencia_dados/blob/main/notebooks/analysis_random_forest_regressor_imdb.ipynb

1. Preparação dos Dados:
    - Limpeza e conversão das colunas `Released_Year`, `Runtime` e `Gross` para tipos numéricos apropriados.
    - Codificação de variáveis categóricas (`Certificate`, `Genre`, `Director`, `Star1`, `Star2`, `Star3`, `Star4`) usando `LabelEncoder`.
    - Imputação de valores faltantes com `IterativeImputer`.
    - Remoção de colunas desnecessárias como `Unnamed: 0`, `Series_Title` e `Overview`.

2. Treinamento do Modelo:
    - Divisão dos dados em conjuntos de treino e teste.
    - Treinamento do modelo usando as features preparadas e o target `IMDB_Rating`.
    - Avaliação do modelo usando o coeficiente de determinação (R^2), Root Mean Squared Error (RMSE), que indicou a qualidade da previsão.

Relatório Previsão da Nota do IMDb

3. Previsão para Novo Filme:
    - Carregamento do modelo treinado a partir de um arquivo `.pkl`.
    - Processamento de novos dados de filmes usando a função:

      Tivemos problemas na predição do novo filme devido a um erro: após o processamento pela função do nosso pipeline, as colunas do novo filme não eram compatíveis com as do modelo treinado. Esse desajuste causava erros na etapa de predição, pois as features do novo filme não correspondiam às esperadas pelo modelo.