Eduardo Darrazão - 1906399

Marcelo Guimarães da Costa - 1937570

Leandro Batista de Almeida - Professor

Big Data e Aplicações

12 de dezembro de 2023

Projeto 2: Desenvolvimento de um modelo de machine learning

Este projeto é a parte 2 de um projeto final de duas partes, sendo o objetivo final uma apresentação para a classe que envolveria todos os processos desde a seleção de um *dataset* até a utilização de modelos de *AutoML* sobre o *dataset* escolhido.

Nesta parte do projeto, os processos incluem:

1 - Geração de dataset para ML

- Feature engineering (criação de possíveis campos de interesse)
- Adaptação dos campos à necessidade dos potenciais algoritmos (se necessário)

2 - Seleção e execução de algoritmos

- Análise de algoritmos que possam resolver o problema estipulado
- Testar ao menos dois algoritmos disponíveis
- Treinamento dos modelos

3 - Comparação de modelos

- Avaliar a acurácia e outras métricas de cada modelo testado
- Comparar e justificar as conclusões obtidas

1 – GERAÇÃO DE DATASET PARA ML

Primeiramente, definimos que nosso objetivo é utilizar o *dataset* para a previsão de *averageRating* de um filme. Em segundo lugar, conseguimos inferir que campos numéricos são mais apropriados como entrada para a maioria dos modelos de previsão que iremos utilizar posteriormente, portanto, utilizamos alguns processos para criar *features* numéricas adicionais.

A primeira *feature* criada é bastante simples, apenas uma booleana que é o resultado da comparação da igualdade entre os campos *primaryTitle* e *originalTitle*, que tem por objetivo indicar se o título mais popular é o original. Esta feature foi chamada de *popularIsOriginal*.

Depois trabalhamos com o processo de dumificação da coluna multivariada genres. Por se tratar de uma coluna multivariada, a transformação desta coluna para uma coluna numérica indicando o número da categoria não seria apropriado, então utilizamos múltiplas colunas booleanas que indicam se o número pertence ao gênero da coluna. Segue um pedaço do código utilizado, e exemplos de saída.

```
# Dividir a coluna 'genres' por vírgulas e expandir em colunas
genres_split = title_basics_filtered.withColumn('genres', split('genres', ','))

# Usar a função explode() para criar múltiplas linhas para cada gênero
genres_exploded = genres_split.withColumn('genre', explode('genres'))

# Criar dummies para cada gênero usando pivot()
dummies = genres_exploded.groupBy('tconst').pivot('genre').agg(lit(1)).fillna(0)
```

```
df[df['tconst']=='tt0001184'][['genres','Action', 'Adult',
       'Adventure', 'Animation', 'Biography', 'Comedy', 'Crime', 'Documentary',
       'Drama', 'Family', 'Fantasy', 'Film-Noir']]
```

✓ 0.8s Python

	genres	Action	Adult	Adventure	Animation	Biography	Comedy	Crime	Documentary	Drama	Family	Fantasy	Film-Noir
0	Adventure,Drama	0	0	1	0	0	0	0	0	1	0	0	0

Por fim, determinamos que informações quantitativas como *averageRating* e *averageNumberOfVotes* dos atores participantes de um filme seriam relevantes para a popularidade do mesmo. Para calcular estes valores, selecionamos a média de *averageRating* e *numberOfVotes* de todos os filmes que cada ator participava, e fizemos uma média destes valores para cada ator presente em cada filme. De forma análoga, fizemos o mesmo para produtores (*producers)* e equipe (*crew*).

IMAGENS

2 – SELEÇÃO E EXECUÇÃO DE ALGORITMOS

3 – COMPARAÇÃO DE MODELOS