ICMC USP SÃO CARLOS
Instituto de Ciências Matemáticas e de Computação
| Universidade de São Paulo |

Predizendo notas IMDB através da Mineração de Dados e de Algoritmos Regressores

Eduardo Rodrigues Amaral, Gabriela Rodrigues do Prado, Laís Saloum Deghaide, Otto Cruz Fernandes

Trabalho da Disciplina de Inteligência Artificial – SCC0230, Profa. Solange Oliveira Rezende

INTRODUÇÃO

CONTEXTUALIZAÇÃO

Um filme de sucesso não apenas entretém o público, mas também permite que as empresas cinematográficas obtenham lucros tremendos. Muitos fatores como bons diretores, atores experientes são consideráveis para criar bons filmes. No entanto, por mais que diretores e atores famosos possam trazer uma boa receita de bilheteria e/ou uma obra aclamada pela crítica, nem sempre esse é o caso. Questiona-se então, quais os fatores resultam em uma obra bem avaliada pela crítica.

OBJETIVOS

Com base nas informações dos filmes, deseja-se entender quais são os fatores importantes que tornam um filme mais bem-sucedido do que outros. Então, analisa-se que tipo de filme faz mais sucesso, ou seja, obtém maior pontuação no IMDB. Também tem-se como objetivo mostrar os resultados dessa análise de maneira intuitiva, de modo a visualizar o resultado. Neste projeto, as pontuações do IMDB são tomadas como variável de resposta e tem-se como foco as previsões operacionais obtidas a partir da análise do restante das variáveis da base de dados. Também é desejado analisar os resultados de diferentes modelos regressores de modo a identificar qual obtém os melhores resultados para a base em questão.

MATERIAIS E MÉTODOS

Para o desenvolvimento deste trabalho, foi utilizado um notebook Python executado na plataforma Google Colab. Utilizou-se a técnica de Mineração de Dados para descoberta de padrões e obtenção de conhecimento a partir dos dados.

A partir da base de dados com as notas de IMDB, foi realizada uma análise exploratória, identificando as colunas da base, analisando, por exemplo, a média de notas por gênero dos filmes, os países que produzem mais filmes, a quantidade de filmes em preto e branco versus colorido.

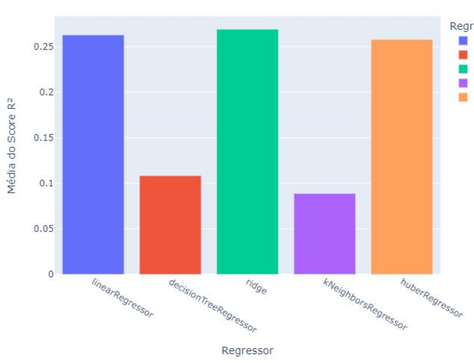
Após esta análise, foi realizado o pré-processamento dos dados. Esta etapa foi responsável pela limpeza e transformação dos dados. Removeu-se duplicatas e valores nulos, em outros casos substituiu-se valores nulos pela média da coluna, reduziu-se o número de valores possíveis de algumas colunas, e, por fim, removeu-se colunas com pouca influência na predição de notas e criou-se novas colunas a partir de outras para auxiliar a análise.

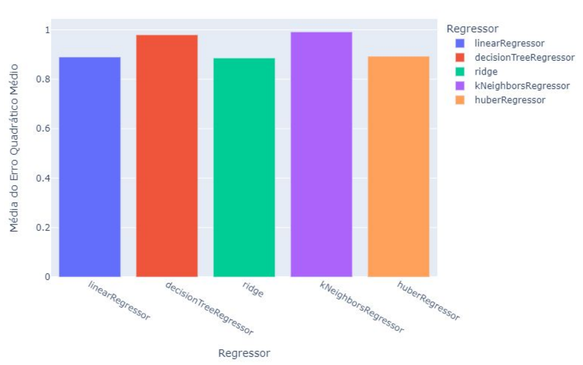
Com essas informações, foi feita a predição do *score* IMDB através de algoritmos de regressão. Os algoritmos foram testados por meio da Validação Cruzada com 5 folds, de forma a assegurar a confiabilidade dos resultados. Antes de todos os treinos, os dados foram normalizados. Reduziu-se a dimensionalidade dos dados para 250 com o uso do PCA. Por fim testou-se 5 diferentes algoritmos de regressão.

O resultado dos algoritmos foram comparados com base no Erro Quadrático Médio e Score R^2 de cada um.

RESULTADOS E DISCUSSÃO

- Os algoritmos testados foram: Regressão Linear, Regressão por Árvore de Decisão, Regressão Ridge, Regressão KNN, Regressão Huber





CONCLUSÃO

- Observa-se que o melhor regressor em termos do Score R^2 é o Ridge. O erro quadrático médio neste caso é 0,89 e o Score R^2 0,27.
- Também é possível notar que existe espaço para melhora do modelo, de modo a obter resultados mais satisfatórios.
- É importante ressaltar que o principal desafio deste conjunto de dados é seu alto desvio padrão nos valores de algumas colunas, resultando em um grande número de Outliers Verdadeiros. Atacar tal problema é o caminho para melhorar os resultados das predições.
- Ademais, é possível notar que para filmes com scores medianos o modelo obtido performa melhor, para filmes mais bem avaliados a performance abaixa um pouco e para filmes mal avaliados o erro da previsão é o maior dentre todas situações.
- Por fim, tal modelo pode ser utilizado para prever o sucesso crítico de um filme antes de seu lançamento. Como o sucesso depende de um grande número de fatores, muitas vezes não explícitos e transparentes, tal modelo é de grande valia para suportar, de antemão, decisões como investir ou não em certo filme, quanto investir, quais avaliações da crítica esperar, entre outras.