Etapa 3 - Processo Seletivo, Cientista de Dados

Fabrício Rodrigues de Souza

Prevendo intervalos para valores de licenciamento

Este relatório descreve os todos os passos tomados, desde a preparação do ambiente, leitura dos dados, tratamento, treinamento, avaliação até a predição, para o desafio enviado.

Preparação do ambiente

O primeiro passo tomado foi a criação de um ambiente virtual para execução da tarefa. Neste ambiente instalei Python 3.9.13 e as versões mais recentes de todas as bibliotecas utilizadas, o arquivo requirements.txt contem todas a lista de todas as bibliotecas utilizadas e é possível reproduzir o ambiente o utilizando como ponto de inicial de instalação.

Leitura dos dados

A leitura dos dados foi feita utilizando a biblioteca Pandas, através da função read_excel.

Tratamento dos dados

Ao analisar os dados de treino e os dados para a predição foi identificado que as colunas “Total Video Views In Milions” e “Total Hours Watched In Milions” não estavam disponíveis nos dados de predição. Isso nos dá duas opções, utilizar as colunas no treinamento, e imputar os dados de predição, ou seja, completar os dados de treinamento, ou não utilizar os dados no treinamento. Como nenhuma das entradas dos dados de predição possuem estes dados optei por não utilizar os dados no treinamento, visto que o processo de imputação não traria benefícios.

O segundo passo do tratamento dos dados foi a transformação das variáveis categóricas dos dados. O modelo utilizado, que será explicado mais a frente, necessita que as variáveis categóricas estejam no formato ordinal, ou seja, se possuímos três categorias, como, por exemplo, “Gato”, “Cachorro” e “Pato”, devemos transformar, respectivamente, em 0, 1 e 2. Neste desafio utilizei a biblioteca SKLearn para realizar a transformação através de um OrdinalEncoder. Foi criado um encoder para cada uma das variáveis “Release Year”, “Country” e “Genre”.

Após o encoding das variáveis os dados foram separados entre dados de entrada, chamados de “X”, e dados de saída, chamados de y. Os dados de entrada contem as colunas “Genre”, “Release Year”, “Duration Minutes”, “Country” e “Production Cost in Milions”. Os dados de saída, ou seja, os dados que queremos prever, são a coluna “Licensing Cost Milions”.

Em seguida, os dados foram separados entre dados de treino e teste, para isso foi utilizada a função “train_test_split” da biblioteca SKLearn, sendo que o tamanho dos dados de teste é de 10%.

Treinamento do modelo

O modelo escolhido foi um regressor da biblioteca LightGBM. O regressor deste biblioteca utiliza um algoritmo de *boosting* para realizar a regressão. *Boosting* significa utilizar diversos preditores “burros”, não complexos, e utilizar a média deles para tomar decisões. Outra opção muito comum para este tipo de tarefa é a utilização da biblioteca XGBoost. Decidi utilizar o LightGBM porque o XGBoost apresenta um consumo de memória muito grande quando utilizado em conjuntos de dados extensos. Pensando no futuro onde este modelo, mesmo que hipotético, poderia ser executado em uma quantidade massiva de dados, é interessante que utilizemos o LightGBM porque seu consumo de recursos é extremamente menor que o do XGBoost e apresenta resultados tão bons quanto seu competidor.

Uma etapa comum em tarefas de predição utilizando modelos, especialmente modelos baseados em redes neurais, é a otimização de seus hyperparâmetros. No entanto, é comum na comunidade de ciência de dados o fato de modelos de boosting não necessitarem de passar por essa etapa, pois frequentemente os parâmetros padrão da biblioteca apresentam resultados estatisticamente iguais aos de modelos otimizados e economizamos tempo e recursos computacionais.

O modelo final criado consiste de três modelos separados. Um para prever o valor do licenciamento em si e outros dois modelos para prever os extremos do intervalo. Para a predição dos extremos utilizei mais uma vez o lightgbm mas a função utilizada como objetivo é a função de quantil, disponível neste link. A função objetivo necessita de um hyperparâmetro chamado de alpha que determina a confiança do intervalo. Neste desafio a confiança utilizada foi de 95%. Isso significa que o modelo consegue responder à pergunta:

"Qual o intervalo em que o valor real de licenciamento estará dentro dele 95% das vezes?"

Avaliação do modelo

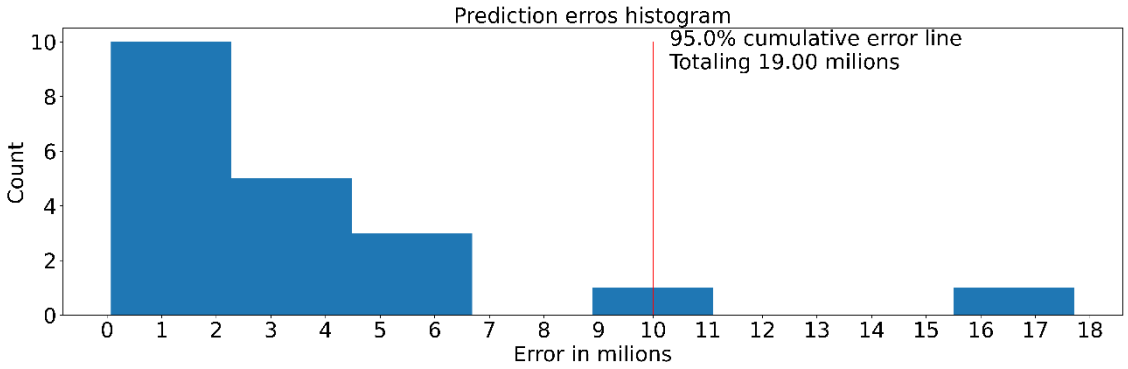
É necessário que nosso modelo seja avaliado, para que possamos entender seus pontos fortes, fracos e onde podemos atuar para melhorá-lo. O primeiro passo para a avaliação é a observação de algumas métricas tradicionais do aprendizado de máquina. Estas duas métricas são o r^2 e o erro absoluto médio.

O r^2 representa, numa escala de 0% a 100% o quanto conseguimos explicar dos dados de saída com os dados de entrada. O modelo implementado apresenta um valor de, aproximadamente, 82% para o r^2 o que significa que nosso modelo consegue explicar 82% dos dados. Este valor é aceitável para um modelo de regressão, pensando que a quantidade de variáveis utilizadas no treinamento não é alta e a quantidade de dados de entrada utilizados também não é massiva.

A segunda métrica, o erro absoluto médio, representa o quanto o modelo erra em média quando comparado aos valores reais, ou

seja, em milhões, o quanto devemos esperar que a predição sugerida esteja errada. O modelo apresenta um erro médio absoluto de aproximadamente três milhões e seiscentos mil (3.600.000,00). Pensando que a maioria dos custos de licenciamento estão na casa de dezenas de milhões, o erro obtido está em uma ordem de magnitude, unidades de milhões, menor do que os custos de licenciamento, o que é considerado aceitável pela comunidade. No entanto seria necessária uma conversa com o time de negócio para confirmar essa afirmação.

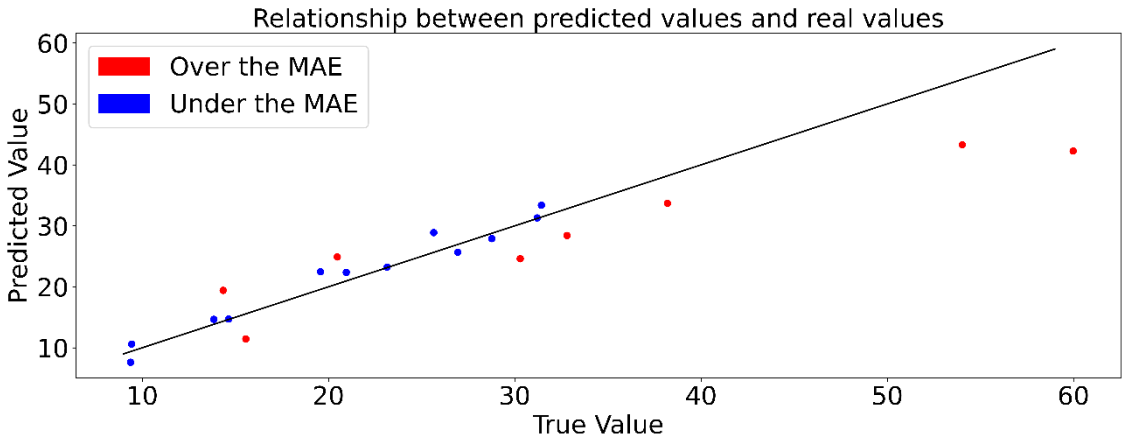
A Figura abaixo demonstra a distribuição dos erros para os dados de teste.



Podemos ver que 95% dos erros alcançam, individualmente, no máximo 10 milhões e que estes mesmos 95% acumulados alcançam aproximadamente 19 milhões.

A próxima Figura apresenta a relação entre os valores previstos e os valores reais. Num cenário ideal todos os pontos estariam sobre a linha preta, que representa o cenário onde o valor predito é igual ao real. Os pontos coloridos de vermelho são pontos em que o valor do erro é maior que o valor do erro médio absoluto e os pontos coloridos de azul são pontos em que o valor do erro é

menor que o valor do erro médio absoluto. Podemos ver que os valores não se distanciam exageradamente da linha preta, exceto por dois pontos.

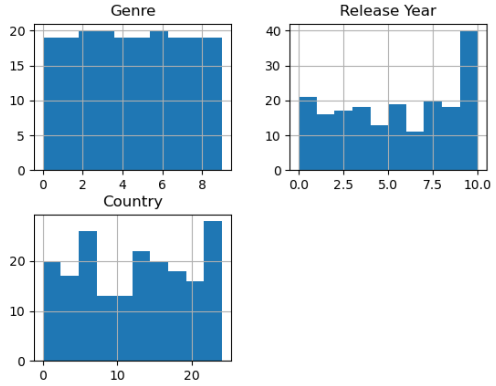


Estes dois pontos representam os títulos “Surprise With Honor” e “Foreigner Of Light”.

Title	Genre	Release Year	Country
Surprise With Honor	3	1	17
Foreigner Of Light	6	0	24

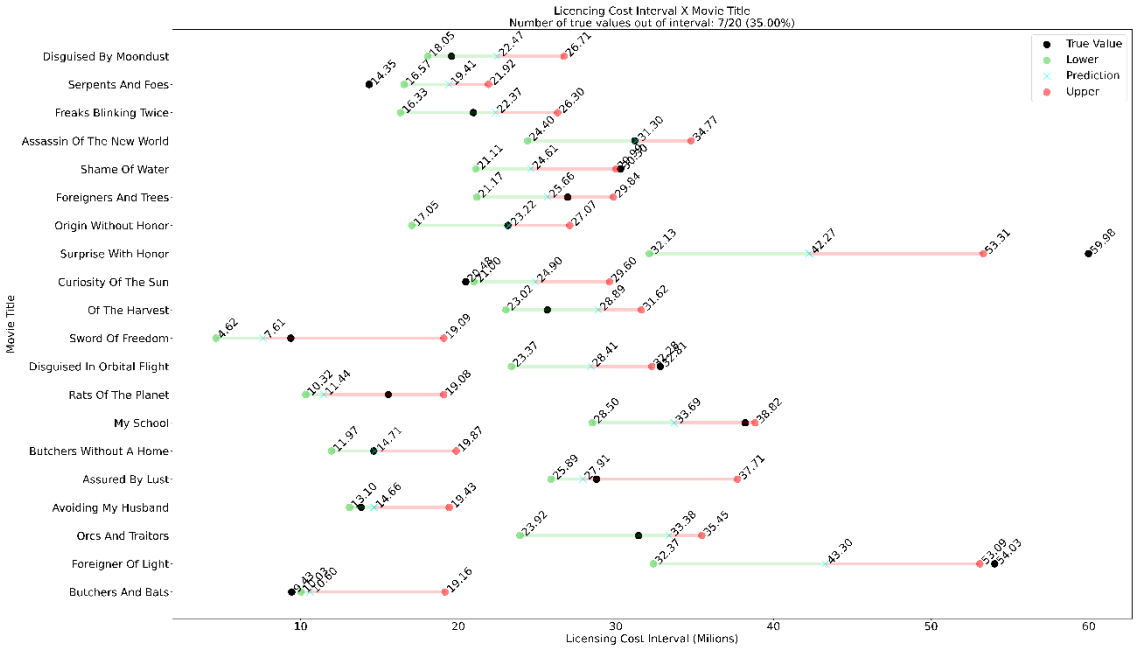
Os dados destes pontos estão disponíveis na tabela acima. Abaixo temos a distribuição dos valores dos dados de treino para cada

uma das variáveis listadas acima.



Ao criamos histogramas destas três variáveis podemos ver que os dados de gênero são extremamente bem distribuídos, o que significa que provavelmente não são o que está causando o erro na predição. Quando analisamos o ano de criação podemos ver que existe um desbalanceamento destes anos. No entanto, a quantidade de exemplos com os anos que os dois outliers pertencem é significativa, então não podemos afirmar que o problema na predição vem da falta de exemplos com os anos afirmados. O mesmo acontece com o país de origem, não existe um problema claro de falta de exemplos. Analisando estas variáveis podemos pensar que o problema venha de variáveis não mensuradas e não utilizadas no modelo, como, por exemplo, presença de atores famosos etc.

O gráfico abaixo apresenta os intervalos previstos para os dados de teste.



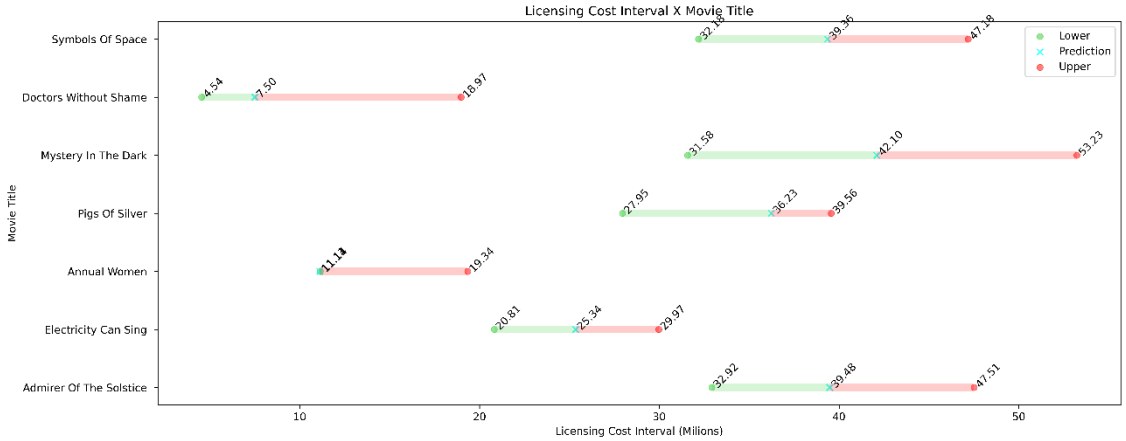
O ponto verde representa o valor mínimo do intervalo, a faixa verde representa todos os valores abaixo do valor sugerido pelo modelo como valor real do custo de licenciamento. O ponto vermelho representa o valor máximo do intervalo, a faixa vermelha representa todos os valores acima do valor sugerido pelo modelo como valor real do custo de licenciamento. A cruz azul representa o valor sugerido pelo modelo e o ponto preto representa o valor real.

Podemos ver que para 65% dos casos de teste o modelo conseguiu criar um intervalo onde o valor real está dentro deste intervalo. Nestes exemplos não é plotado o valor real. Para os outros 35% dos casos podemos ver que o intervalo não inclui o valor real, no entanto para cinco dos sete casos onde o intervalo não inclui o valor real a diferença é pequena.

Com estas análises feitas e os resultados apresentados para os casos de teste, acreditamos que o modelo está validado e pronto para a utilização na predição dos dados do desafio.

Predição dos intervalos

É necessário tratar os dados da mesma maneira como foram tratados os dados de treinamento, isso significa fazer um encoding das variáveis. Após este encoding podemos aplicar o modelo treinado e obter as predições. A figura abaixo mostra os intervalos preditos para os dados.



Com este gráfico a equipe responsável pela decisão de compra dos títulos pode escolher o valor ofertado.

Limitações do modelo

A quantidade de dados utilizada pelo modelo é baixa, mas num cenário real a quantidade de dados disponíveis seria maior, o que significa que a ideia do modelo pode ser reaproveitada.

A quantidade de variáveis disponibilizadas, bem como suas naturezas talvez não sejam as mais indicadas para a predição. Frequentemente vemos notícias sobre filmes que não

recuperaram seu investimento inicial de produção, por isso outras informações como, elenco, dados que nos informam se a obra é uma adaptação de outro formato, se a obra é uma sequência ou não etc., podem ser mais valiosas.