UNIVERSIDADE FEDERAL DO CEARÁ

TECHNICAL REPORT

Aluno: Luciana Sousa Martins

1. Introdução

O conjunto de dados Clean Dataset usado contém informações sobre as opções de reserva de voo no site Easemytrip para viagens entre as 6 principais cidades metropolitanas da Índia. Existem 300261 observações e 11 variáveis no conjunto de dados, são eles:

- *airline (Companhia aérea): O nome da companhia aérea*
- *source_city (Cidade de origem): Cidade de onde parte o voo*
- *departure_time (Horário de partida): Horário de partida*
- *stops (Paradas): Números de paradas entre a cidade de origem e destino*
- *arrival_time (Horário de Chegada): Horário de Chegada*
- *destination_city (Cidade de destino): Cidade onde o voo irá pousar*
- *class (Classe): Classes de assentos*
- *duration (Duração): Tempo total necessário para viajar entre as cidades em horas*
- *price (Preço): Preço do bilhete*
- *flight : Código de voo do avião*
- *days_left (Dias restantes): Subtração da data de viagem pela data de reserva*

2. Observações

Na primeira questão de início houve conversões em 8 variáveis (‘airline’, ‘source_city’, ‘departure_time’, ‘stops’, ‘arrival_time’, ‘destination_city’, ‘class’, ‘duration’) do tipo ‘string’ para do tipo ‘int’ e a exclusão de 2 variáveis ('flight', 'days_left') não muito importantes para a análise. Logo após a análise feita para gerar um gráfico com os resultados dos tributos mais relevantes mostrou que a classe do dataset era a variável “class” com dois valores distintos. Na segunda questão houve problemas no gráfico por conta da variável classe usada, disso houve a mudança de uso de atributos nos códigos em diante, sendo eles: a variável ‘price’ e a variável ‘duration’.

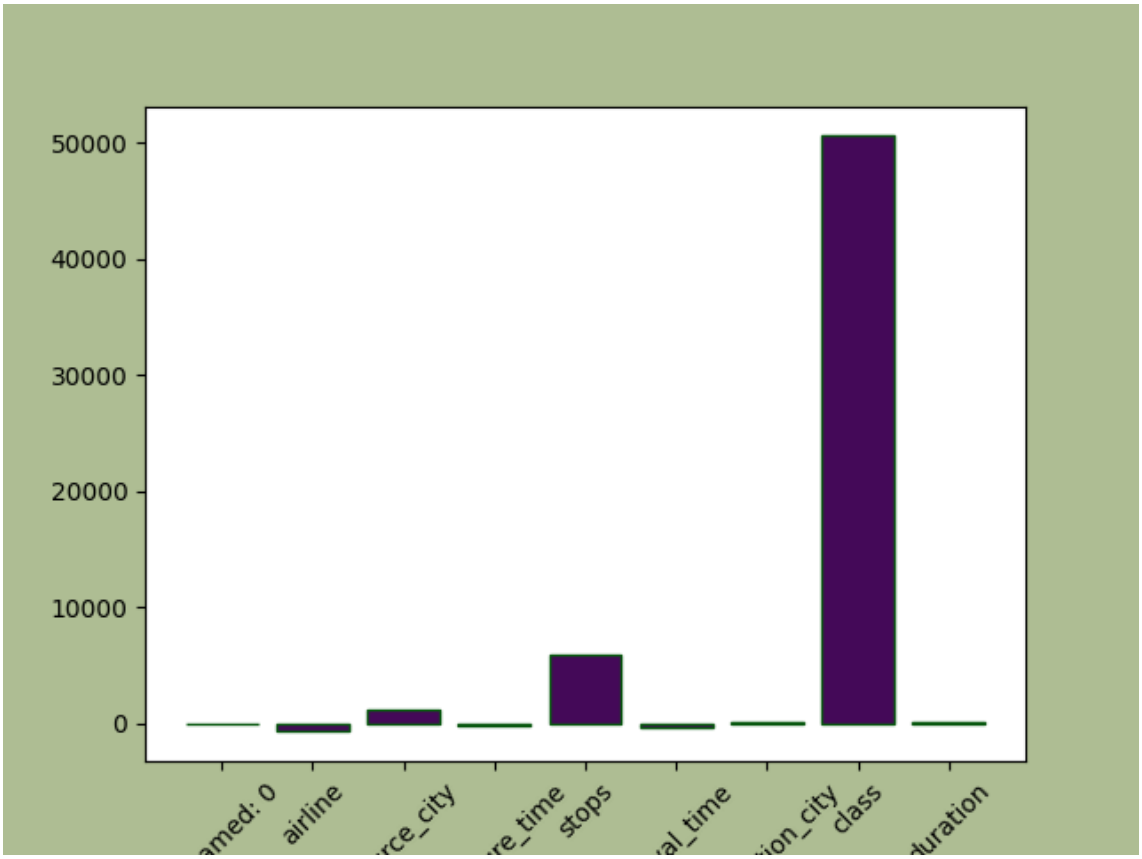
3. Resultados e discussão

1° Questão:

Na primeira questão foi preciso importar bibliotecas necessárias para geração dos códigos, em seguida criar uma variável e atribui-la o caminho do dataset. Como a

maioria dos atributos da tabela estava com tipo 'String', houve 8 conversões de atributos para tipo 'int', exclusão de dois atributos não muito relevantes para a análise.

Logo após, a questão pediu para mostrar o atributo mais importante do dataset, o código utilizou o método Lasso e gerou um gráfico de barras para visualizar os coeficientes de importância dos atributos em relação ao atributo alvo "price" no dataset para realizar uma análise de regressão.

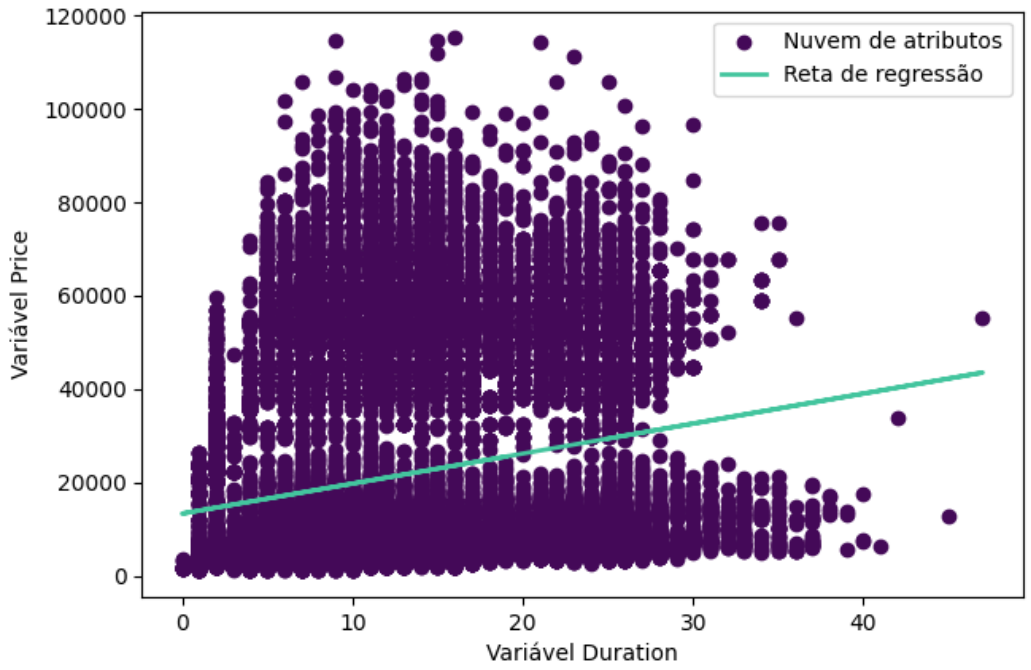


No gráfico mostrado acima revela o atributo "class" como mais importante para uma análise.

2° Questão:

Na segunda questão o código realiza uma regressão linear simples utilizando o atributo "duration" como variável independente e o tributo "price" como variável dependente, o código dividirá o conjunto de dados em treino e teste para avaliação de desempenho

do modelo. Faz a previsão dos valores de "price" para os dados de teste e calcula o coeficiente de determinação para avaliar o desempenho do modelo.



No gráfico acima mostras suas principais características, de direção da linha com uma inclinação positiva indicando uma relação positiva entre as variáveis, a medida que a variável independente aumenta, a variável dependente aumenta também. Tem seus ajustes de pontos pouco distribuídos em relação a linha de regressão, mostrando que o modelo está moderadamente ajustado pois os pontos estão próximos à linha. Portanto, com base nesses padrões e características, podemos concluir que o gráfico de regressão linear mostra uma relação positiva bem ajustada entre as variáveis independentes e dependentes.

Em seguida, a questão pediu para determinar também os valores: RSS, MSE, RMSE e R_squared para a análise de regressão.

As métricas estatística acima representam a avaliação do desempenho de um modelo de regressão:

RSS=29588271327790.26, indica que o modelo está fornecendo uma boa representação dos dados e que há uma quantidade considerável de variação não capturada pelas variáveis independentes.

MSE=492883199.1436135, indica que o modelo não está bem ajustado aos dados e que há uma quantidade substancial de erro entre os valores reais e os valores previstos pelo modelo.

RMSE=22200.97293236523, indica que, em média, os erros entre os valores reais e os valores previstos são relativamente grandes.. Isso sugere que o modelo não está se ajustando bem aos dados e está cometendo erros consideráveis na sua capacidade de prever os valores corretos.

R_squared=0.0438406431879752, indica que o modelo explica aproximadamente 4.38% da variabilidade da variável dependente. Isso sugere que o modelo possui uma capacidade limitada de capturar os padrões ou relacionamentos entre as variáveis.

*RSS (*Soma dos Quadrados dos Erros *)*	*29588271327790.26*
*MSE (*Erro Quadrático Médio*)*	*492883199.1436135*
*RMSE (*Raiz do Erro Quadrático*)*	*22200.97293236523*
*R_squared (*Coeficiente de Determinação*)*	*0.0438406431879752*

Esses resultados sugerem que o modelo não está fornecendo uma boa representação dos dados e não está capturando adequadamente os padrões ou relacionamentos entre as variáveis.

3 Questão:

Na terceira questão, o código realiza o treinamento e a busca em gridsearch cross-validation para verificar qual a melhor parametrização para os regressores de Lasso e Ridge, utilizando validação cruzada.

Após a execução deste código, os objetos 'lasso_grid' e 'ridge_grid' contém os modelos Lasso e Ridge treinados e ajustados aos de treinamento, respectivamente, com os melhores valores de 'alpha' encontrados durante a busca em grid.

A tabela a seguir mostra as melhores configurações para Lasso e Ridge e os melhores scores para cada:

Melhores configurações para Lasso	*alpha: 0.01*
Melhor score para Lasso	*0.04159178529279062*
Melhores configurações para Ridge	*alpha: 10*
Melhor score para Ridge	*0.04159178587252514*

alpha:0.01, indica que o modelo Lasso está aplicando uma regularização relativamente leve aos coeficientes das variáveis independentes.

Score para Lasso: 0.04159178529279062, significa que o modelo Lasso explica aproximadamente 4.16% da variabilidade da variável dependente com base nas variáveis independentes selecionadas.

alpha: 10, indica que o modelo Ridge está aplicando uma regularização relativamente alta aos coeficientes das variáveis independentes.

Score para Ridge: 0.04159178587252514, isso significa que o modelo Ridge explica aproximadamente 4.16% da variabilidade da variável dependente com base nas variáveis independentes selecionadas.

4 Questão:

A 4° questão utilizará KFold e Cross-validation para gerar uma regressão linear utilizando os atributos definidos da questão anterior. O código realiza a validação cruzada para os modelos Lasso e Ridge e imprime o Score médio para cada um deles, isso permite avaliar o desempenho dos modelos com base na métrica de score utilizada.

Score médio para Lasso	*0.042050862336992845*
Score médio para Ridge	*0.042050862336511195*

Os Scores médios do Lasso e do Ridge são próximos de 0, isso indica um fraco desempenho do modelo. Nesse caso, tanto o modelo Lasso e Ridge tiveram scores médios próximos de 0.042, o que sugere que ambos os modelos tiveram um desempenho semelhante durante a validação cruzada.

4. Conclusões

Os resultados não foram satisfatórios. Pois diante da análise feita, podemos concluir que o gráfico de regressão linear mostrou uma relação positiva discreta ajustada entre as variáveis independentes e dependentes, não estava capturando adequadamente os padrões ou relacionamentos entre as variáveis, os modelos: Ridge e Lasso explicaram aproximadamente 4.16% da variabilidade da variável dependente com base nas variáveis independentes selecionadas.

5. Próximos passos

Para uma melhor análise e precisão de resultados para interpretação:

Avaliação de outros modelos: Além dos modelos Lasso e Ridge, explorar outros algoritmos de regressão, como regressão linear simples, regressão polinomial, regressão de árvore de decisão, regressão de floresta aleatória, regressão de redes neurais, entre outros. Comparar o desempenho desses modelos em relação às métricas de avaliação relevantes.

Exploração dos dados: Realizar uma exploração mais aprofundada dos dados, analisando as distribuições das variáveis, identificando possíveis outliers, investigando relações entre as variáveis independentes e o alvo, entre outros. Isso pode fornecer insights adicionais sobre os dados e auxiliar na seleção de recursos relevantes.

Validação externa: Além da validação cruzada realizada até o momento, reservar um conjunto de teste externo para avaliar o desempenho final do modelo selecionado. Isso ajudará a ter uma estimativa mais confiável de como o modelo se comporta em dados não vistos.

link para visualização de códigos:
https://github.com/alysonbz/IA/tree/Luciana_Martins/AV2