UNIVERSIDADE FEDERAL DO CEARÁ

TECHNICAL REPORT

Aluno: Larissa Vitória Vasconcelos Sousa

1. Introdução

Nesse relatório, serão apresentadas a implementação e a análise dos modelos de Regressão Linear, GridSearch Cross-Validation, Lasso e Ridge, KFold, Cross-Validation e métricas como RSS, MSE, RMSE e R_squared utilizando a base de dados *Car Price Prediction.* Essa base de dados é um conjunto abrangente disponibilizado no Kaggle, projetado especificamente para análise de regressão.
O conjunto de dados Car Price Prediction possui diversas colunas informativas, que incluem informações relevantes sobre os veículos.
Algumas das colunas são:

- Marca;
- Preço;
- Número de portas;
- Tipo de combustível;
- Estilo de carroceria;
- Tipo de tração;
- Potência;
- Sistema de combustível;
- Consumo.

Tabela 1 – Conceitos de cada técnica utilizada nesse projeto.

Modelos	Conceito
Regressão Linear	Modelo estatístico que busca encontrar a melhor linha reta para modelar a relação entre uma variável dependente e uma ou mais variáveis independentes.
GridSearch Cross-Validation	Técnica para encontrar os melhores hiperparâmetros de um modelo por meio da avaliação de várias combinações utilizando validação cruzada.
Lasso e Ridge	Técnicas de regularização utilizadas na regressão linear para lidar com multicolinearidade e overfitting, reduzindo a magnitude dos coeficientes.
KFold	Técnica de validação cruzada que divide o conjunto de dados em k partes, usando cada parte como conjunto de teste uma vez para avaliar o modelo.
Cross-Validation	Técnica que envolve a divisão do conjunto de dados em conjuntos de treinamento e teste para estimar o desempenho do modelo e evitar overfitting (fenômeno que ocorre quando um

	modelo de machine learning se ajusta excessivamente aos dados de treinamento, resultando em um desempenho ruim).

Tabela 2 – Conceitos de cada métrica utilizada nesse projeto.

Métricas	Conceito
RSS (Residual Sum of Squares)	Calcula a soma dos quadrados dos resíduos, ou seja, das diferenças entre os valores observados e os valores previstos pelo modelo de regressão. Quanto menor o valor do RSS, melhor o ajuste do modelo aos dados.
MSE (Mean Squared Error)	Calcula o erro médio quadrático, sendo a média dos quadrados dos resíduos. Mede a média das diferenças ao quadrado entre os valores observados e os valores previstos pelo modelo. Quanto menor o valor do MSE, melhor o ajuste do modelo aos dados.
RMSE (Root Mean Squared Error)	É a raiz quadrada do MSE, proporcionando uma medida do erro médio quadrático na mesma escala das variáveis de interesse. O RMSE é uma medida mais intuitiva. Quanto menor o valor do RMSE, melhor o ajuste do modelo aos dados.
R_squared (R^2)	Indica a proporção da variabilidade dos dados que é explicada pelo modelo de regressão. Varia de 0 a 1, sendo que valores mais próximos de 1 indicam um bom ajuste do modelo aos dados. Um valor de R^2 igual a 1 indica que o modelo explica totalmente a variabilidade dos dados, enquanto valores próximos de 0 indicam que o modelo não consegue explicar a variabilidade dos dados. O R^2 também pode ser interpretado como a porcentagem da variabilidade dos dados capturada pelo modelo.

2. Observações

Para encontrar as colunas mais relevantes, foi preciso transformar labels de colunas categóricas em numéricas. Tentei usar LabelEncoder(), sem sucesso. Então, fiz do jeito mais lento (usando map da biblioteca “pandas”), de uma por uma, cada label recebendo um número.

E na geração dos gráficos, na parte da análise exploratória, não consegui consertar o xlabel nos gráficos: normalized-losses e price.

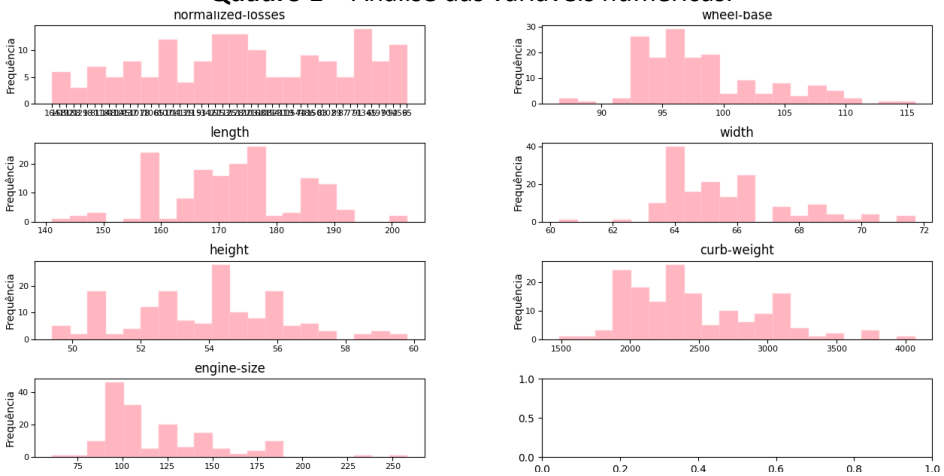
Na questão 2, fiquei com dúvidas em qual usar variável mais relevante usar, o sistema de combustivel (com 6 valores) ou a marca que tem muitos valores. Os dois casos foram analisados.

3. Resultados e discussão

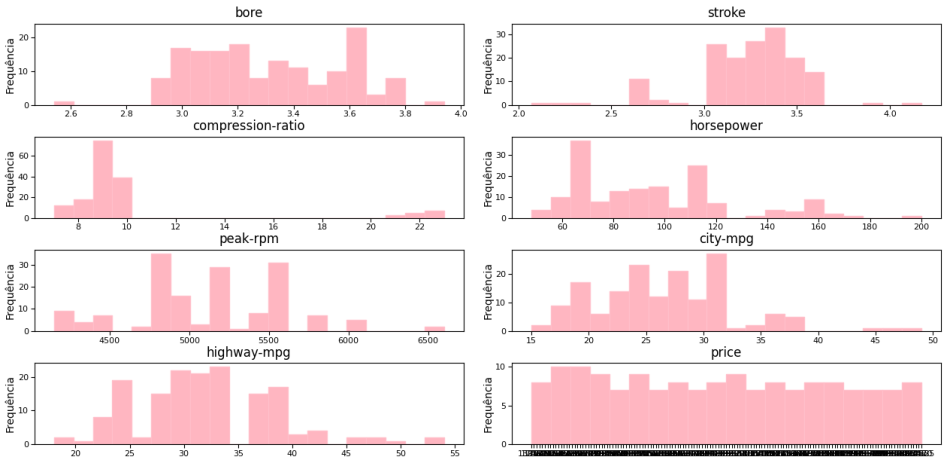
3.1 Questão 1

Antes de verificar qual atributo seria o alvo para regressão do dataset e fazer uma análise de qual atributo seria mais relevante para realizar a regressão do alvo escolhido, foi feita uma breve análise exploratória:

Quadro 1 – Análise das variáveis numéricas.

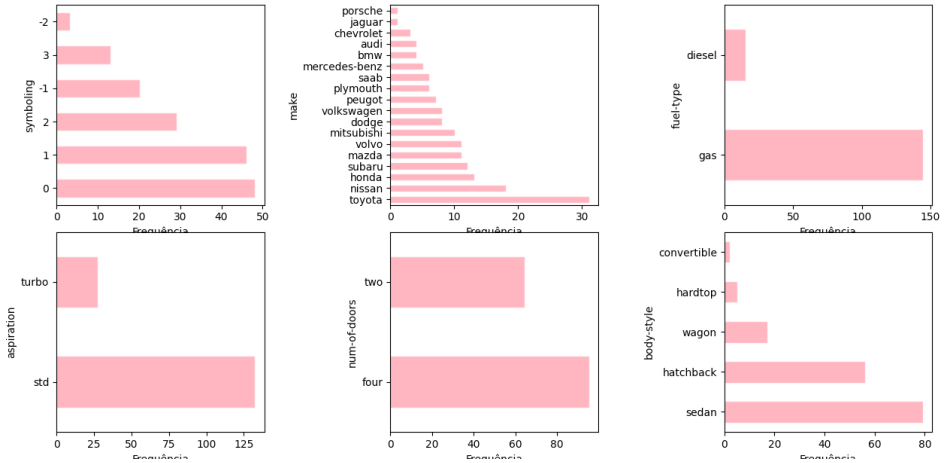


Quadro 2 – Análise das variáveis numéricas.

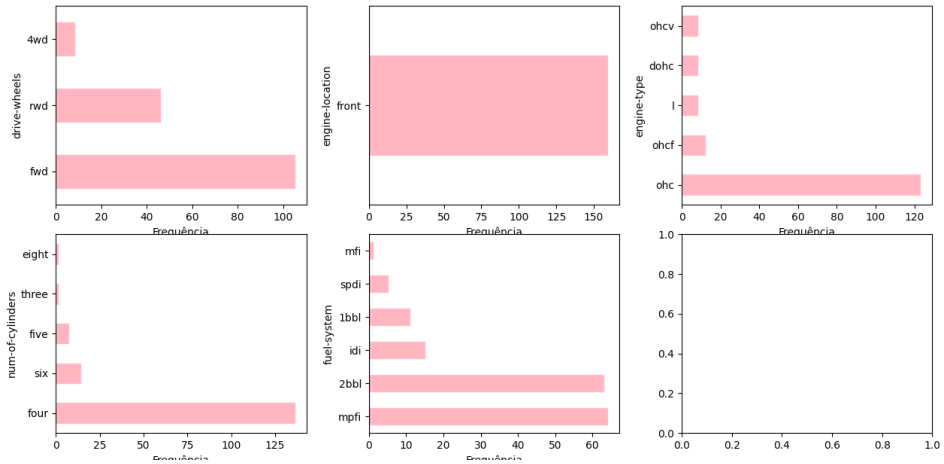


Através de histogramas, vemos as distribuições das variáveis numéricas, como elas se comportam de forma individual.

Quadro 3 – Análise das variáveis categóricas.

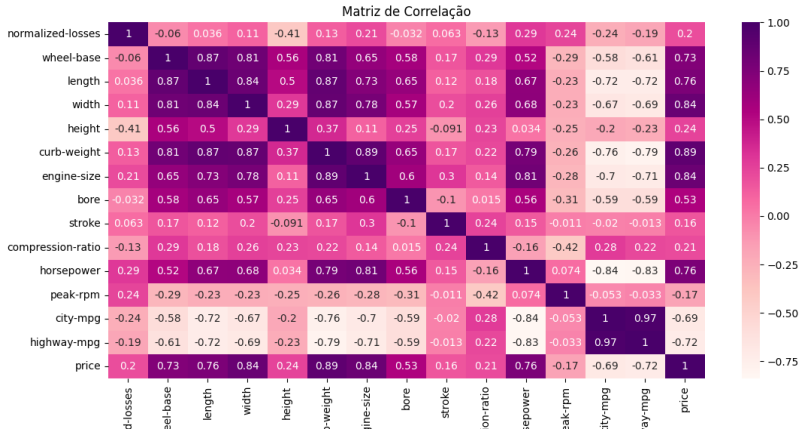


Quadro 4 – Análise das variáveis categóricas.



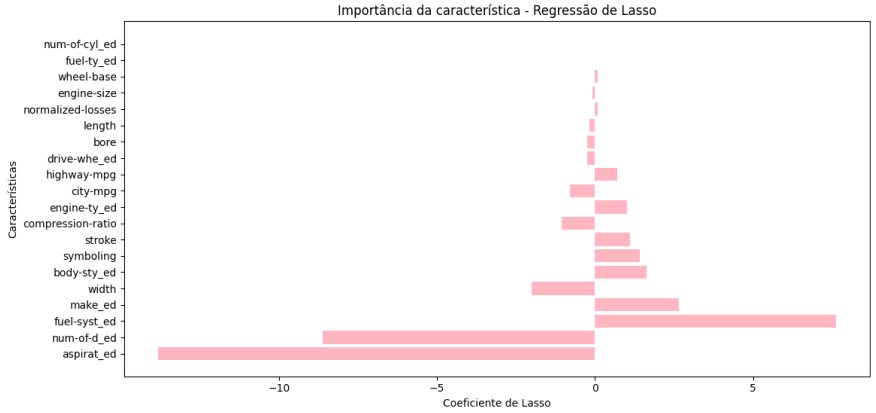
Por meio de gráficos de barras, observamos o comportamento das variáveis categóricas.

Quadro 5 – Matriz de correlação.



Na matriz de correlação, cada célula representa a correlação entre duas variáveis, a intensidade da cor indica o valor da correlação, onde cores mais claras (tons de rosa) indicam uma correlação mais forte (próxima de 1), e cores mais escuras (tons de roxo) indicam uma correlação mais fraca (próxima de 0). Aqui vemos a variável que está mais correlacionada com “price” = a “curb_weight” (0.89), porém, ela é o peso, e acredita-se que a marca seja mais relevante.

Quadro 6 – Colunas relevantes.



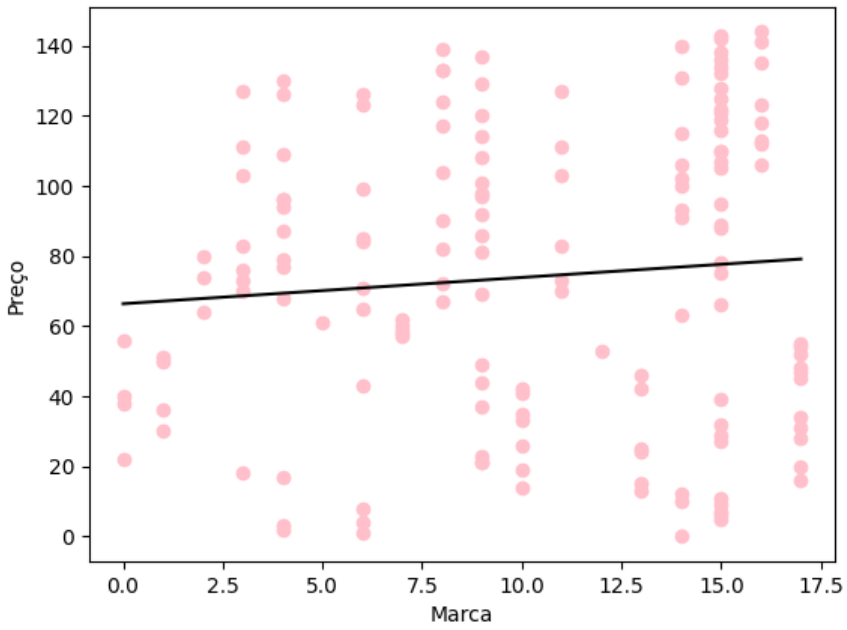
Vê-se que a coluna mais relevante é a “aspirat_ed” (binária) que traduzindo é a variável “aspiração”, em segundo lugar, “num-of-d_ed” (binária), em terceiro, “fuel-syst_ed” (possui seis valores), em quarto, “make_ed” (possui muitos valores), na questão 2 utilizei a marca para a Regressão Linear.

3.2 Questão 2

Previsões da Regressão Linear são as previsões do atributo alvo (preço) feitas pelo modelo de regressão linear com base no atributo mais relevante (“make_ed” = marca) utilizado como entrada. Nesse caso, temos uma lista de cinco valores de previsão:

[66.39672237 66.39672237 66.39672237 66.39672237 67.14729522]

Quadro 7 – Regressão Linear para a marca.



Para plotar o gráfico, não foi possível usar a coluna mais relevante “aspirat_ed” poque ela é binária, da mesma forma, a segunda. Então usei a quarta mais relevante: “marca”.

Tabela 3 – Resultados obtidos através das métricas para a marca.

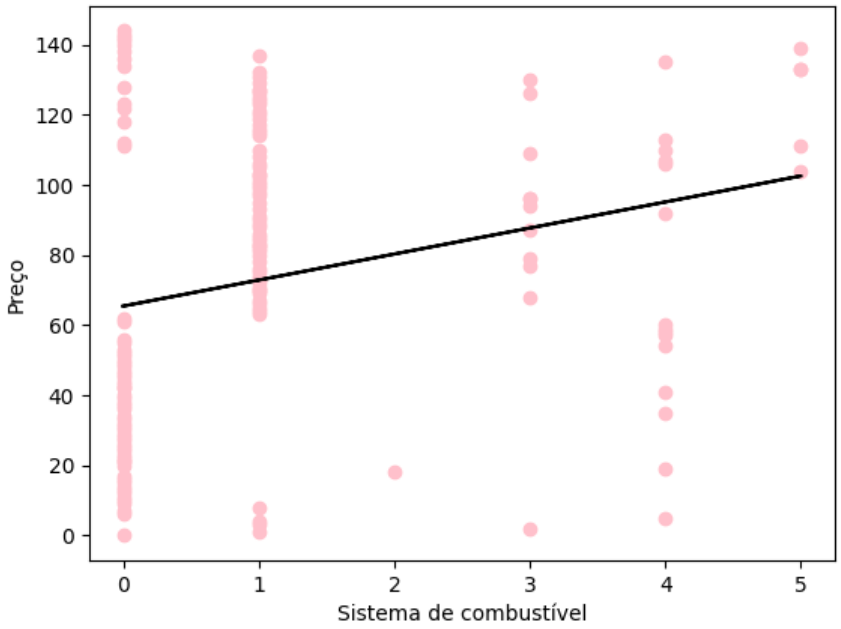
Métricas	Resultados
RSS	272742.61270209553
MSE	1715.3623440383367
RMSE	41.41693305929758
R_squared	0.008064308785086877

Essas são métricas utilizadas para avaliar o desempenho de um modelo de regressão. Essas previsões representam os valores estimados pelo modelo para o preço dos carros

com base nas características do atributo make_ed. Porém, é importante notar que, com base nos resultados da regressão linear e nos valores baixos de R^2, as previsões podem não ser muito precisas ou explicativas.

Na indecisão, foi executado o código e o gráfico com a variável "sistema de combustível", mas a previsões ficaram ainda mais baixas:

Quadro 8 – Regressão Linear para sistema de combustível.



Previsões da Regressão Linear:

[65.43987044 65.43987044 65.43987044 65.43987044 65.43987044]

Tabela 4 – Resultados obtidos através das métricas para o sistema de combustível.

Métricas	Resultados
RSS	257289.94505300355
MSE	1618.1757550503369
RMSE	40.22655534656599
R_squared	0.06426400715183428

3.3 Questão 3

Tabela 5 – Resultados obtidos através dos modelos Lasso e Ridge.

Modelos	Resultados	
	Ajustado	**Afinado**
Lasso	{'alpha': 1e-05}	0.20058929372531328
Ridge	{'alpha': 1e-05, 'solver': 'sag'}	0.20538462067788774

Vemos que, para o modelo Lasso, o melhor valor para o parâmetro "alpha" foi 1e-05 e a pontuação obtida foi de 0.20058929372531328. Já para o Ridge, o melhor valor para o parâmetro "alpha" foi 1e-05 e o melhor valor para o parâmetro "solver" foi 'sag'. A pontuação obtida foi de 0.20538462067788774. Esses resultados indicam que, dentre as configurações testadas, o modelo Ridge obteve uma pontuação melhor do que o modelo Lasso. No entanto, é importante ressaltar que as pontuações são baixas.

3.4 Questão 4

Tabela 6 – Resultados obtidos para CrossValidation, Média e DP.

Cv_score	[0.33008073 0.48982592 0.22985243 -0.16242236 0.08649263 0.34524471]
Média	0.21984567724729787
Desvio padrão	0.21007063755782243

Ao analisar os resultados, a média das pontuações indica o desempenho médio do modelo de Regressão Linear. Nesse caso, a média é de aproximadamente 0.22, o que sugere que o modelo possui um desempenho moderado na previsão do preço dos carros. O desvio padrão das pontuações indica a variabilidade do desempenho do modelo entre os diferentes folds do KFold. Um desvio padrão de 0.21 indica que há uma certa variação nas pontuações, o que pode indicar que o modelo pode se comportar de maneira inconsistente em diferentes conjuntos de dados.

4. Conclusões

As previsões feitas pelo modelo de Regressão Linear indicam que o modelo não está capturando bem a variabilidade dos dados e as previsões podem não ser precisas ou explicativas. Ao ajustar os modelos de Lasso e Ridge utilizando técnicas de regularização, encontramos os melhores valores de hiperparâmetros, embora as pontuações sejam baixas, o modelo de Ridge apresentou um desempenho melhor que o modelo de Lasso. O KFold indicou que o modelo apresenta um desempenho moderado na previsão do preço dos carros, porém, com uma certa variação nas pontuações entre os diferentes

folds. Comparando os desempenhos dos modelos, podemos observar que os modelos Lasso e Ridge apresentaram pontuações um pouco melhores em relação à Regressão Linear simples. No entanto, todas as pontuações estão relativamente baixas, o que sugere que os modelos não estão ajustando bem aos dados. Ou seja, os resultados esperados infelizmente não foram satisfeitos.

5. Próximos passos

Alguns desses pontos seriam interessantes para obter melhores resultados nesse projeto:

- Análise mais detalhada dos dados, explorando a relação entre as variáveis independentes e a variável alvo;
- Além da regressão linear, explorar outras técnicas de regressão mais avançadas, como árvores de decisão, entre outras. Elas podem capturar relações não-lineares, o que pode levar a um melhor desempenho de previsão;
- Além do KFold utilizado, experimentar outras estratégias de validação cruzada, como StratifiedKFold, para avaliar melhor o desempenho dos modelos e verificar sua consistência em diferentes conjuntos de dados.

REFERÊNCIAS

FACELI, Katti; LORENA, Ana C.; GAMA, João; AL, et. Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina. Rio de Janeiro: Grupo GEN, 2021. E-book. ISBN 9788521637509. Disponível em: https://app.minhabiblioteca.com.br/#/books/9788521637509/. Acesso em: 28 mai. 2023.