UNIVERSIDADE FEDERAL DO ABC - UFABC

Bacharelado em Ciência e Tecnologia

Gustavo de Souza Gonçalves - 11201721371

ANÁLISE E MODELAGEM DE PRECIFICAÇÃO CARROS USADOS NO REINO UNIDO

Santo André – SP

2021

1. DADOS

O projeto busca entender como funciona a precificação de carros usados, baseando-se em uma base disponível no kaggle com 100 mil dados coletados no Reino Unido, com uma amostra na figura 1, com as seguintes variáveis:

- **Model** – Modelo do Carro (155 Categorias, não será utilizado pela alta cardinalidade)
- **Year** – Ano de Fabricação
- **Preço** – Preço de Venda
- **Transmission** – Transmissão (4 Categorias)
- **Mileage** – Milhagem do Carro
- **fuelType** – Tipo de combustível (5 Categorias)
- **tax** – Road Tax, taxa paga para rodar com o carro nas estradas
- **mpg** – Milhas por galão
- **Engine Size** – Volume de combustível e ar suportado no motor
- **Car** – Marca/Montadora (7 Categorias)

model	year	price	transmission	mileage	fuelType	tax	mpg	engineSize	car
Passat	2017	12100	Semi-Auto	62776	Diesel	30	60.1	2.0	vw
Kuga	2014	9990	Manual	61590	Diesel	160	47.9	2.0	ford
T-Cross	2019	17666	Manual	3905	Petrol	145	47.9	1.0	vw
Astra	2017	9998	Manual	17579	Diesel	150	74.3	1.6	vauxhall
Fiesta	2013	7000	Manual	49138	Petrol	0	65.7	1.0	ford

Figura 1 – Amostra da base de dados de carros usados vendidos no reino unido

A base não conta com valores nulos, pela figura 2, foram considerados outliers: year, valores fora da faixa [2009,2021], engineSize iguais a 0. Ao utilizar a correlação entre as variáveis numéricas é possível verificar que em relação ao nosso target (price) o engineSize se destaca com um coeficiente de 0.64 e que em geral em módulo nenhuma fica perto de 0, o que pode significar que todas tem alguma influência mesmo que média sobre o target, algo que poderia ser

esperado e que é confirmado pela correção é quão mais novo o carro menor a milhagem, sendo o coeficiente mais forte encontrado entre as variáveis, mas se definirmos um limite de 0.9 para eliminar variáveis semelhantes, nenhuma seria descartada.

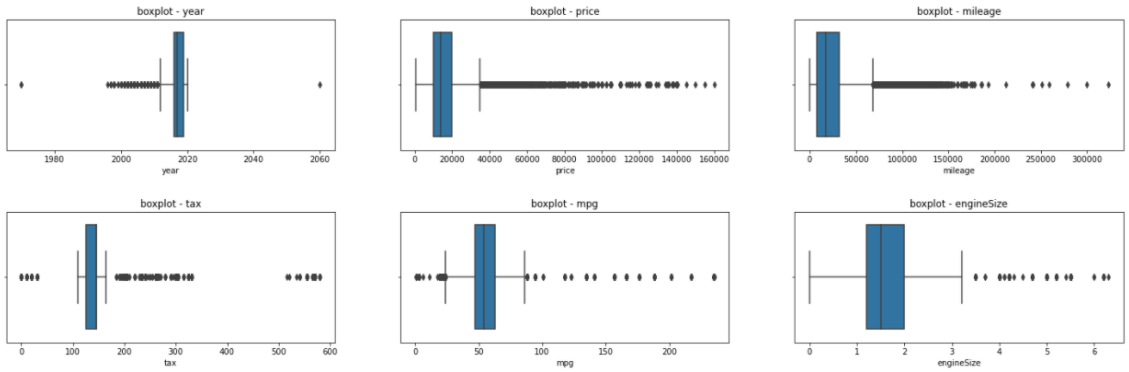

Figura 2 – Gráficos de boxplot para cada variável numérica

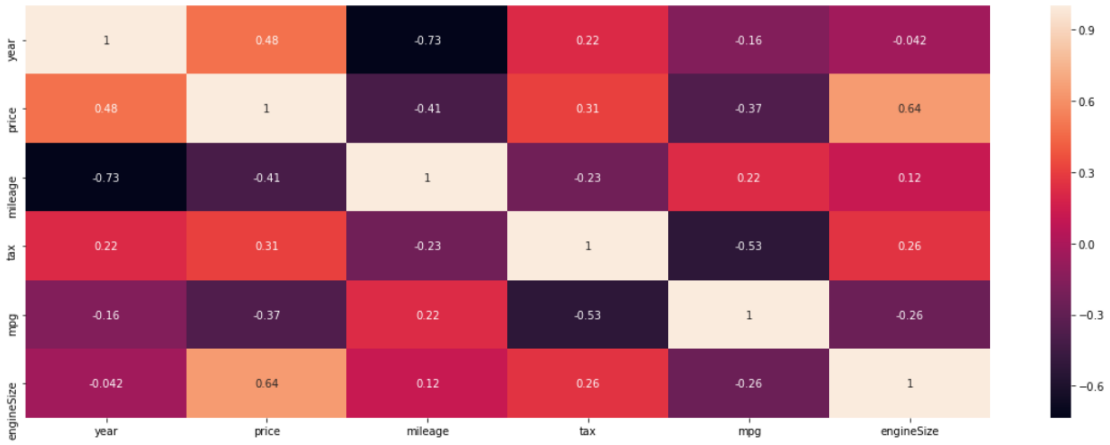

Figura 3 – Mapa de calor da correlação entre as variáveis numéricas

Para as variáveis categorias foi escolhido um pré-tratamento com OneHotEnconding e ao verificar a figura 4 é notável que carros com a transmissão automática ou semi-automática, veículos que funcionam a base de Diesel ou modelo híbrido e que são das marcas Audi ou Merc, tem preço mais elevado.

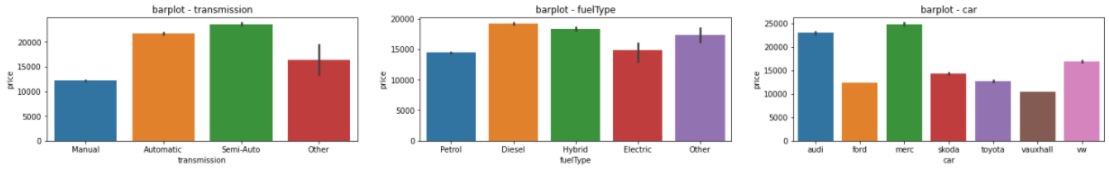

Figura 4 – Gráfico de barras entre o preço e cada variável categórica

2. MODELAGEM

Na tentativa de encontrar os padrões das variáveis combinadas para prever o preço, foi adotado uma estratégia de usar modelos de machine learning, sendo escolhidos 2 modelos com abordagem linear, 1 k-vizinhos e 3 baseados em árvore de decisão, com a métrica de R^2 para regressão.

Como estratégia de validação de dados, foram escolhidos 80% dos dados para treino e validação, e 20% para teste, dentro dos dados de treino para avaliação foi utilizado um kfold com 3 separações que contemplam a busca dos melhores hiper parâmetros com o GridSearchCV e a validação cruzada, a figura 5 e na tabela 1 demostram que nos resultados de treino e validação modelos de árvore de decisão em conjunto tiveram melhores resultados em comparação a outras abordagens.

Figura 5 – Resultados de treino e validação após a melhoria de hiper parâmetros e validação cruzada em 3 partes randomizadas

	Treino		Validação		Teste	
Modelo	**Score Médio**	**Desvio Padrão**	**Score Médio**	**Desvio Padrão**	**Score Médio**	**Desvio Padrão**
GradientBoostingRegressor	0,96	0,001	0,93	0,002	0,892	0,009
LinearSVR	0,765	0,001	0,765	0,001	0,73	0,016
RandomForestRegressor	0,989	0	0,933	0,002	0,9	0,015
ElasticNet	0,788	0,001	0,788	0,002	0,79	0,019
DecisionTreeRegressor	0,934	0,004	0,91	0,005	0,873	0,021
KNeighborsRegressor	0,935	0,001	0,909	0,003	0,889	0,015

Tabela 1 – Resultados de treino e validação após a melhoria de hiper parâmetros e validação cruzada em 3 partes randomizadas

O RandomForest se sagrou como melhor modelo pelo maior R^2 em todas as etapas, então para compreender melhor como funcionam as relações entre o preço e as outras variáveis em termos de importância, é possível utilizar o feature importance, e também para ver se o efeito é positivo ou negativo utilizar os coeficientes do melhor modelo linear, aqui no caso, o ElasticNet.

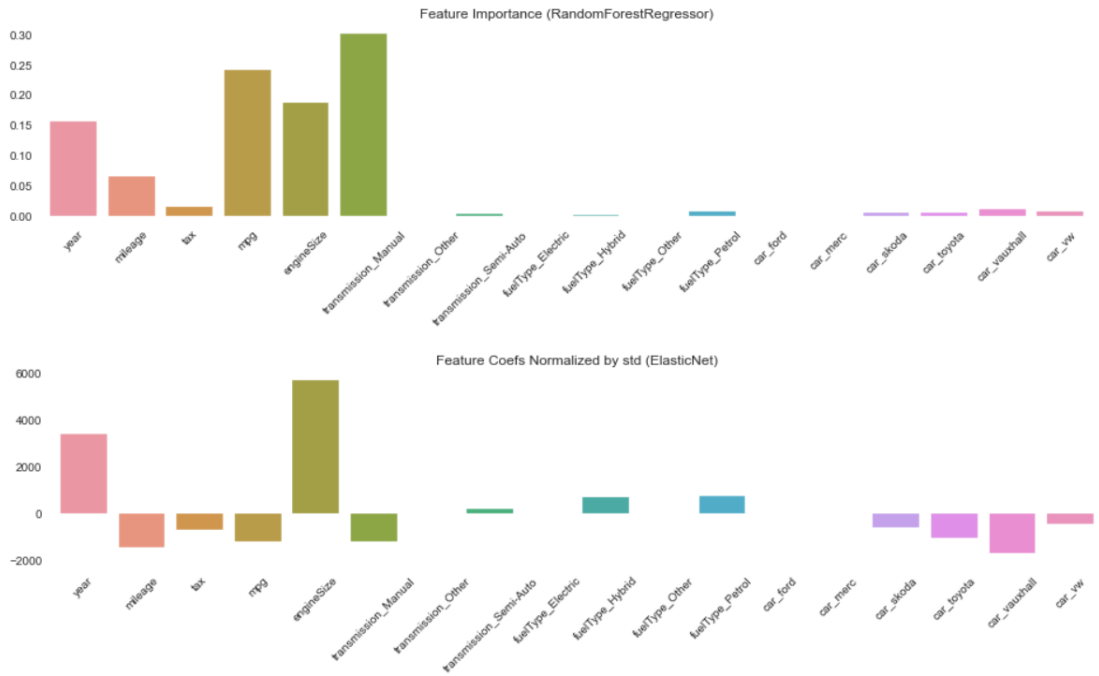

Figura 6 – Importância e coeficientes das variáveis utilizando RandomForest e Elasticnet

As Váriaveis mais importantes são a transmissão manual, mpg, engineSize, year e mileage, portanto para a precificação carros que não tenham transmissão manual, que tenham menos milhas por galão, que tenham o maior volume de combustível, sejam mais recentes e com menor milhagem são os carros mais caros, salta aos olhos que para os modelos o tipo de combustível e a marca não influenciam de maneira significativa.

Diante do exposto, foi possível alcançar um R^2 de 0.9 em todas as etapas com o RandomForest que expõe que é possível fazer uma boa previsão com os dados obtidos para precificação de carros usados sem ocorrer overfitting nem underfitting e que é possível melhorar os resultados nos dados de treino se reduzirmos o número de varíavel buscando escolher as mais importantes além de necessitar de menos features.