Projeto Automobile

Jefferson Kauan Cavalcante Chaves

Itaguaí - Rio de Janeiro
2024

Objetivo

Este projeto foi desenvolvido com o dataset *Automobile* da Kaggle, abrangendo etapas de limpeza, visualização e construção de modelos de machine learning. O principal objetivo foi testar e comparar as métricas de desempenho de dois modelos de aprendizado de máquina distintos. Para obter os melhores resultados, foi realizada uma análise cuidadosa dos dados, incluindo a limpeza e visualização, com o intuito de gerar insights mais aprofundados.

Tratamento dos dados

Esse documento mostrará as principais etapas do processo de limpeza dos dados, pois o passo a passo da resolução está presente no notebook no github. A metodologia do processo foi pensado em identificar valores default e substituí-los pela média e em alguns casos a mediana(A mediana foi escolhida como métrica de substituição por motivos estatísticos, visto que a mediana não sofre alterações provenientes de outliers). Dessa forma, segue abaixo os principais passos.

1. Exclusão da coluna **normalized-losses**
   A coluna normalized-losses não apresentou grande importância para o nosso objetivo final, e sua inconsistência causaria problemas no treinamento do modelo. Nesse caso não foi escolhida alguma técnica de substituição porque a coluna não era relevante ao nosso estudo

```
automobile.head(20)
```

	symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	...	engine-size	fuel-system	bore
0	3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47
1	3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47
2	1	?	alfa-romero	gas	std	two	hatchback	rwd	front	94.5	...	152	mpfi	2.68
3	2	164	audi	gas	std	four	sedan	fwd	front	99.8	...	109	mpfi	3.19
4	2	164	audi	gas	std	four	sedan	4wd	front	99.4	...	136	mpfi	3.19
5	2	?	audi	gas	std	two	sedan	fwd	front	99.8	...	136	mpfi	3.19
6	1	158	audi	gas	std	four	sedan	fwd	front	105.8	...	136	mpfi	3.19
7	1	?	audi	gas	std	four	wagon	fwd	front	105.8	...	136	mpfi	3.19
8	1	158	audi	gas	turbo	four	sedan	fwd	front	105.8	...	131	mpfi	3.13
9	0	?	audi	gas	turbo	two	hatchback	4wd	front	99.5	...	131	mpfi	3.13
10	2	192	bmw	gas	std	two	sedan	rwd	front	101.2	...	108	mpfi	3.5
11	0	192	bmw	gas	std	four	sedan	rwd	front	101.2	...	108	mpfi	3.5
12	0	188	bmw	gas	std	two	sedan	rwd	front	101.2	...	164	mpfi	3.31
13	0	188	bmw	gas	std	four	sedan	rwd	front	101.2	...	164	mpfi	3.31
14	1	?	bmw	gas	std	four	sedan	rwd	front	103.5	...	164	mpfi	3.31
15	0	?	bmw	gas	std	four	sedan	rwd	front	103.5	...	209	mpfi	3.62

2. Transformando os ‘?’ em NaN(not a number)
Após fazer um estudo geral do dataset, percebeu-se que os valores default estavam sendo representados por ‘?’. Dessa forma, todos os ? foram trocados por NaN e depois somados para termos uma noção de quantos espaços NaN tínhamos em cada coluna. Algumas colunas utilizou-se a média para substituição e outras a mediana. Após essa última etapa, o tratamento de dados foi finalizado.

```
#Transformando todos os campos com "?" por nan
automobile_new = automobile_new.replace('?', np.nan)
```

```
#Somando todos os nan de todas as colunas
#Esse metódo facilita o processo de tratamento de dados, visto que sabemos onde é necessário realizar alterações
automobile_new.isna().sum()
```

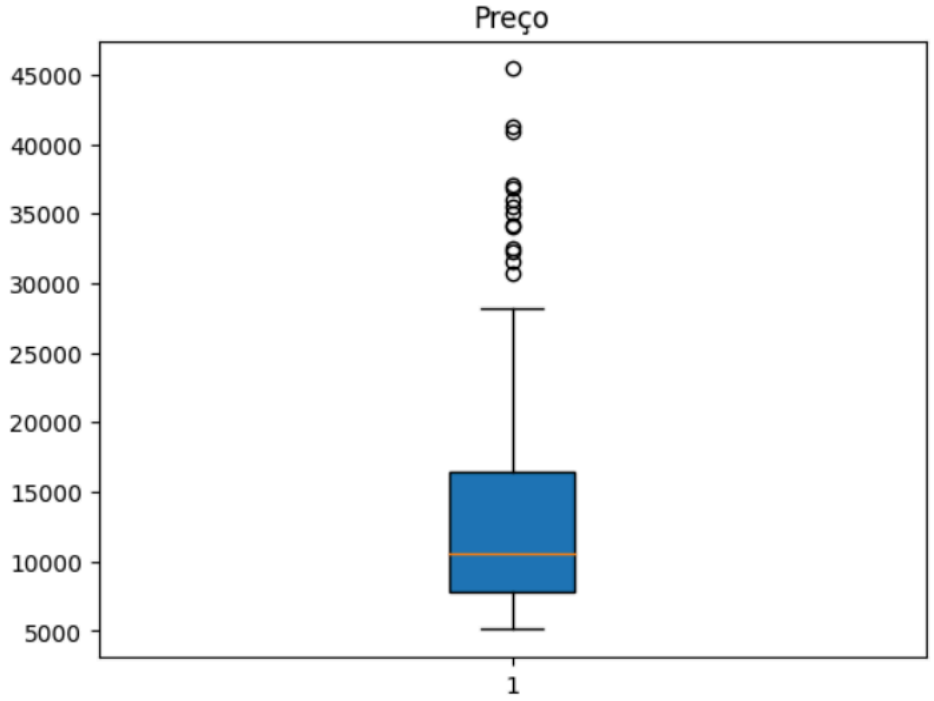
	0
symboling	0
make	0
fuel-type	0
aspiration	0
num-of-doors	2
body-style	0
drive-wheels	0
engine-location	0
wheel-base	0
length	0
width	0
height	0
curb-weight	0
engine-type	0
num-of-cylinders	0
engine-size	0
fuel-system	0
bore	4
stroke	4
compression-ratio	0
horsepower	2
peak-rpm	2
city-mpg	0
highway-mpg	0
price	4

Visualização de dados

Nessa etapa realizamos uma análise exploratória dos dados, visando entender melhor as métricas e gerar insights.

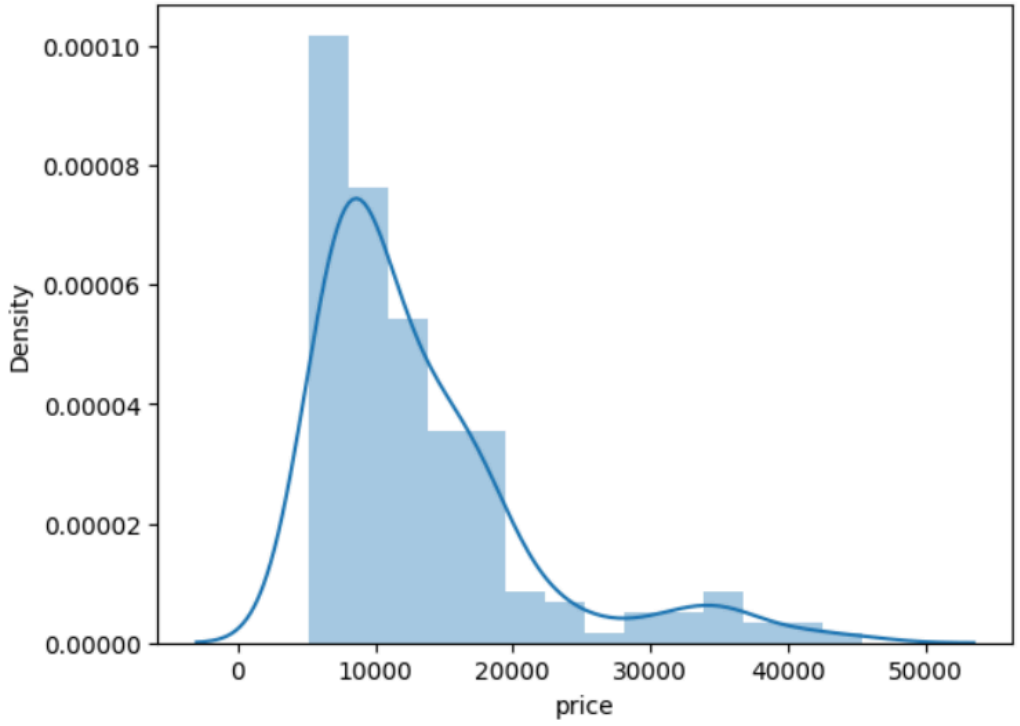
Box plot

Nesse gráfico de box plot podemos descobrir se há outliers em alguma coluna. A variável escolhida para análise foi preço, pois é a que iremos trabalhar mais adiante. Com esse gráfico, percebemos que há valores discrepantes na variável preço, com seu valor máximo girando em torno de $45.000, enquanto a mediana se encontra em aproximadamente $10.000. Isso nos mostra que há valores de automóveis muito discrepantes em relação ao resto.



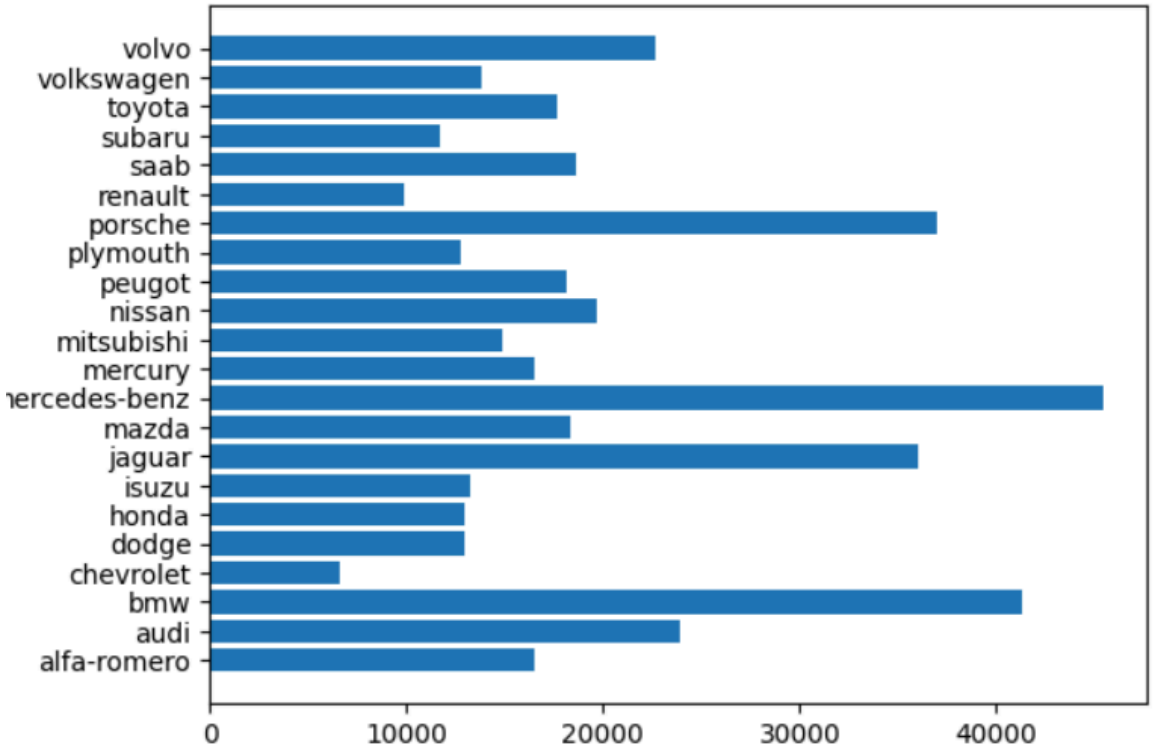
Densidade de preço

A partir desse gráfico, medimos a concentração de faixas de valores dos carros, e percebemos que de fato, a maior concentração se encontra na faixa dos $9.000 aos $20.000. E uma baixíssima proporção de carros com valores acima de $40.000. Isso nos indica que há automóveis com preços muito maiores. Esses valores podem estar associados a questões como: marca, motor, estilo etc. Mais adiante iremos analisar essas relações.



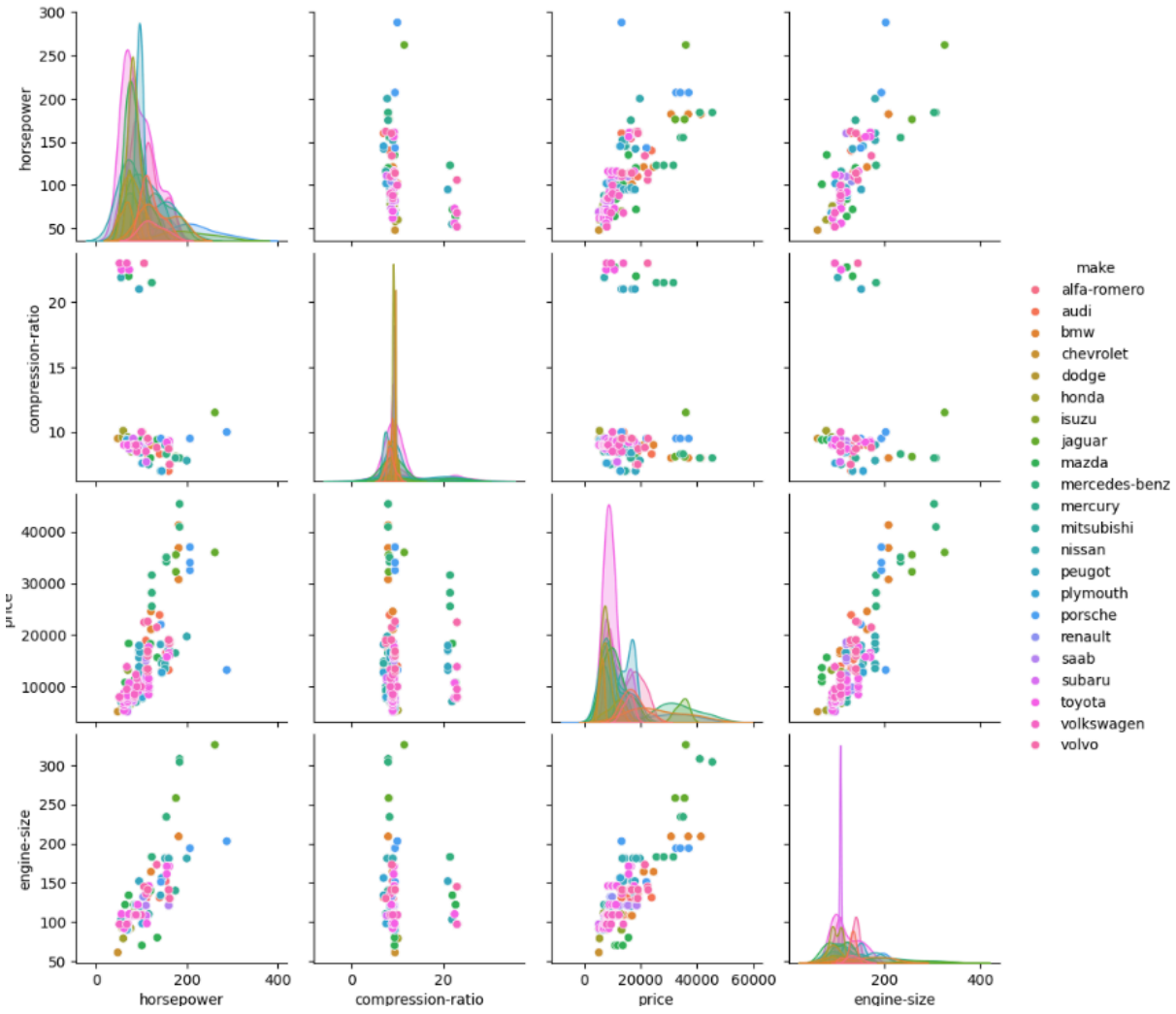
Histograma de marca e preço

A partir desse histograma, percebemos que as marcas mais caras são: Mercedes-benz, bmw e porsche(com muita aproximação ao jaguar). As marcas mais baratas são: Chevrolet, Renault e Subaru. Vale ressaltar que as marcas mais caras são conhecidas por seus estilos de carro esportivos e luxuosos, o que pode ser um dos motivos por figurarem entre as mais caras, enquanto as mais baratas são marcas comuns para carros populares.



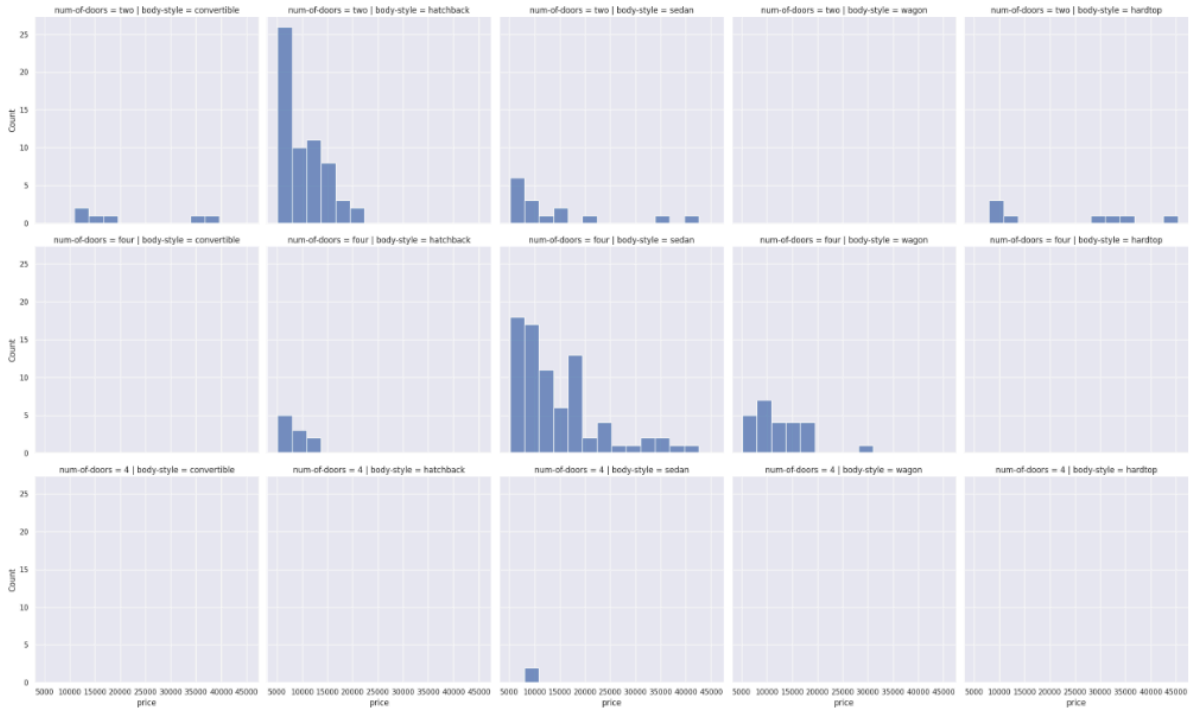
Marcas e variáveis quantitativas

A partir desse gráfico, podemos observar que as marcas mais caras possuem os melhores indicativos com relação a: Horse Power, compression-ratio e engine-size. Isso indica que as marcas mais caras apresentam, em todos os aspectos, indicativos mais altos dessas variáveis. Percebemos também, que as retas mais lineares dizem respeito a correlações entre horsepower, engine-size e preço. Ou seja, as variáveis horsepower e engine-size podem estar diretamente ligadas ao preço.



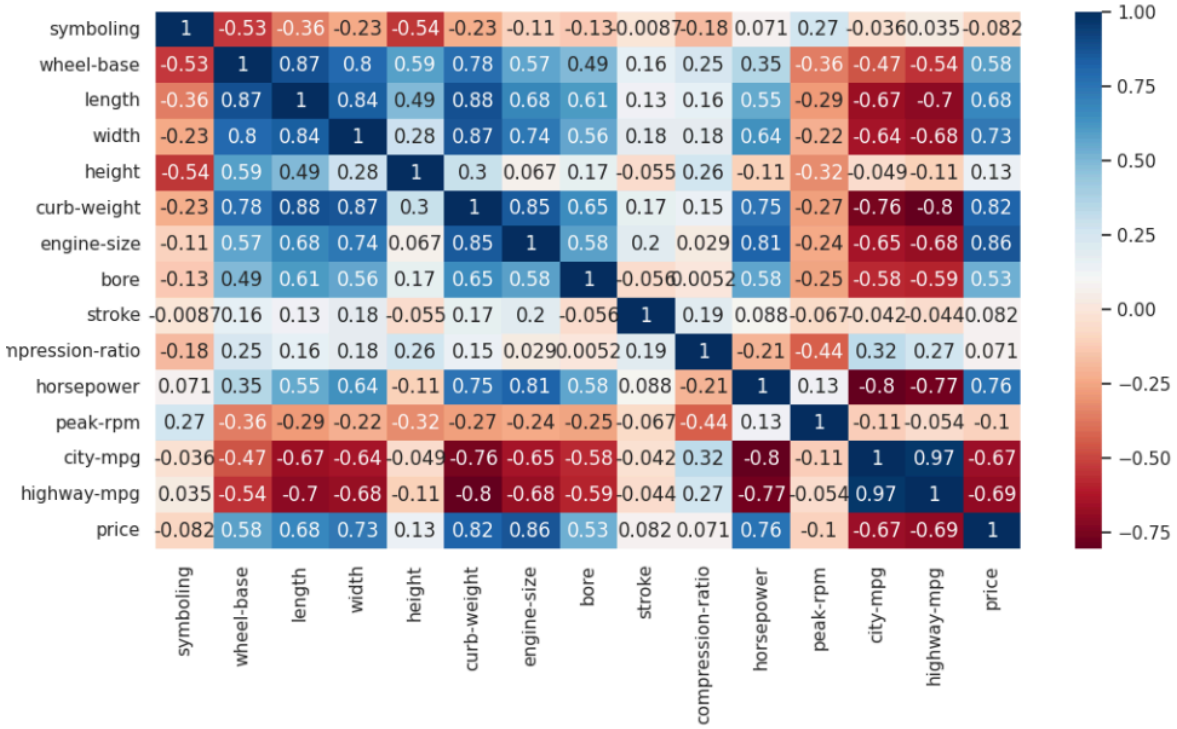
Variáveis qualitativas e preço

Com esse gráfico, analisamos que os carros mais baratos são de dois lugares de estilo hatch, e quatro lugares de estilo sedan. O que corresponde a carros tidos como populares. Já os mais caros possuem mais concentração no estilo duas portas e hardtop, o que pode indicar carros esportivos. Perceba que também há carros caros no estilo sedan e duas portas, e o preço pode ter sido influenciado por tamanho do motor e potência, como também pela marca.



Correlação

O mapa de correlação nos mostra quanto às variáveis quantitativas estão associadas a outras. Dessa forma, percebemos que as variáveis que mais estão associadas ao preço são: engine-size, horsepower, curb-weight, width e length. Essas variáveis dizem respeito à potência e tamanho do motor. Ou seja, no gráfico anterior vimos que carros sedans também se encontram em faixas de preço mais altas. E essa faixa maior de preço pode estar relacionada a essas variáveis citadas.

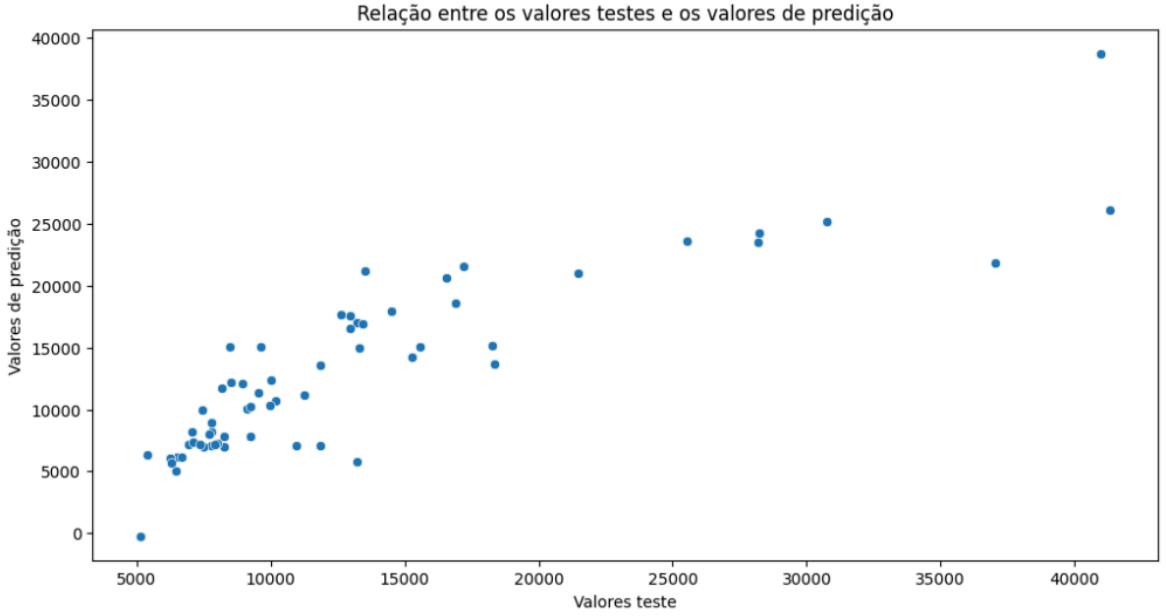


Treinamento do modelo

Esse projeto foi pensado para treinar dois modelos de machine learning e compará-los, visando buscar o melhor resultado possível. Os modelos foram treinados a partir de regressão linear e random forest regression.

Regressão Linear

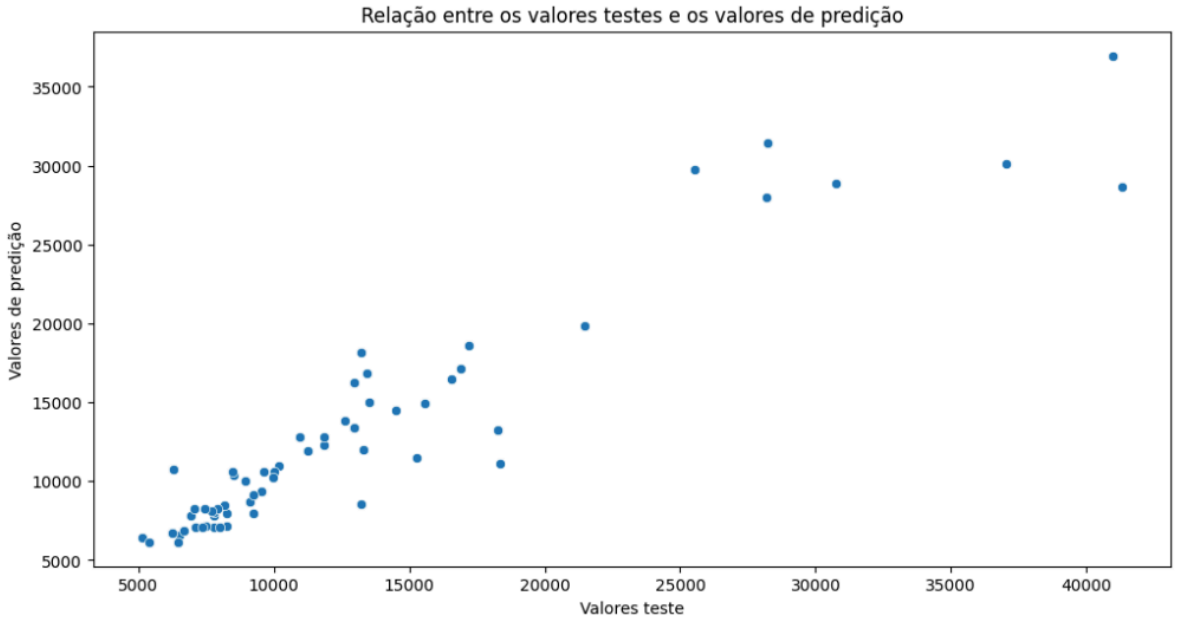
A partir do conceito de regressão linear, dividimos o modelo entre target e feature para treinamento com a intenção de treinar um modelo com maior probabilidade de acerto para prever os preços dos automóveis. Perceba que o modelo observado teve um R2 de 75%, que é um bom indicativo, e uma linearidade interessante apresentada no gráfico abaixo.




```
R2:  0.7564796887691362
RMSE:  4088.296675014038 6
```

Random Forest

Da mesma forma que foi feito o treinamento para o modelo de regressão, dividimos as mesmas categorias para calcular o modelo com random. Perceba que o modelo de random forest se saiu melhor, com um R2 de 88%. Essa questão acontece pelo método de treinamento do random ser diferente da regressão linear.




```
R2:  0.8880575109070539
RSME:  7683243.13569969
```

Comparação

Após, percebemos que o modelo do random performa melhor, fizemos a comparação entre o preço original e o preço predito para compararmos. Perceba que em praticamente todos os casos, o modelo chegou bem próximo do valor real, o que indica um bom desempenho do modelo escolhido.

	Preço Original	Preço calculado
0	30760.000000	28857.362587
1	13207.129353	18157.800000
2	9549.000000	9367.740000
3	11850.000000	12313.180000
4	28248.000000	31441.560000
5	7799.000000	7107.320000
6	7788.000000	7848.520000
7	9258.000000	7975.200000
8	10198.000000	10949.260000
9	7775.000000	8077.940000
10	13295.000000	12038.360000