

Automobile Price Prediction

Dataset consist of various characteristic of a car

Data overview

Range	eIndex: 205 entries	, 0 to 204	
Data	columns (total 26	columns):	
#	Column	Non-Null Count	Dtype
0	symboling	205 non-null	int64
1	normalized-losses	164 non-null	float64
2	make	205 non-null	object
3	fuel-type	205 non-null	object
4	aspiration	205 non-null	object
5	num-of-doors	203 non-null	object
6	body-style	205 non-null	object
7	drive-wheels	205 non-null	object
8	engine-location	205 non-null	object
9	wheel-base	205 non-null	float64
10	length	205 non-null	float64
11	width	205 non-null	float64
12	height	205 non-null	float64
13	curb-weight	205 non-null	int64
14	engine-type	205 non-null	object
15	num-of-cylinders	205 non-null	object
16	engine-size	205 non-null	int64
17	fuel-system	205 non-null	object
18	bore	201 non-null	float64
19	stroke	201 non-null	float64
20	compression-ratio	205 non-null	float64
21	horsepower	203 non-null	float64
22	peak-rpm	203 non-null	float64
23	city-mpg	205 non-null	int64
24	highway-mpg	205 non-null	int64
25	price	201 non-null	float64
dtype	es: float64(11), in	t64(5), object(1	0)

This dataset consist of data From 1985 Ward's Automotive Yearbook from bellow sources:

- 1985 Model Import Car and Truck Specifications,
 1985 Ward's Automotive Yearbook.
- Personal Auto Manuals, Insurance Services Office, 160
 Water Street, New York, NY 10038
- Insurance Collision Report, Insurance Institute for Highway Safety, Watergate 600, Washington, DC 20037

There are 25 columns and 205 rows, with some missing values.

Target: Price (continous)
Categorical features: 10

Numerical features: 14

Problem and Objective

Our target variable (price) is <u>continuous</u>. Therefore, the machine learning problem we face is <u>regression</u>.

Our objective in this analysis focuses on <u>prediction</u> to obtain the <u>highest</u> <u>accuracy</u> from the model

Missing value

	null (sum)	null (%)
normalized-losses	41	20.00
price	4	1.95
stroke	4	1.95
bore	4	1.95
peak-rpm	2	0.98
num-of-doors	2	0.98
horsepower	2	0.98
engine-type	0	0.00
highway-mpg	0	0.00
city-mpg	0	0.00
compression-ratio	0	0.00
fuel-system	0	0.00
engine-size	0	0.00
num-of-cylinders	0	0.00
symboling	0	0.00

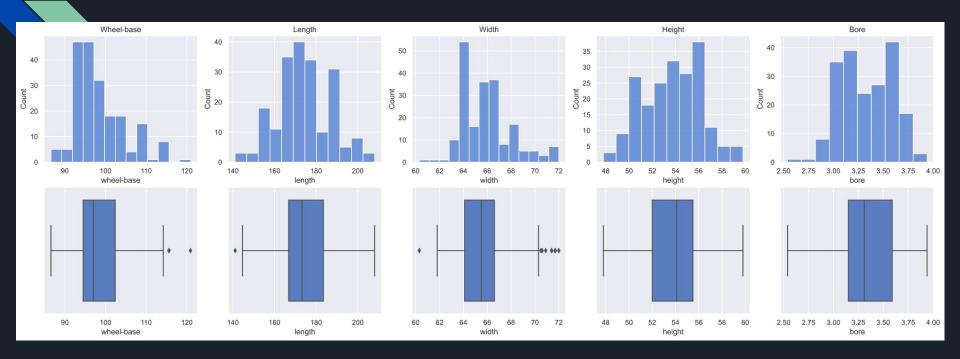
In our data set, there are several missing values with the highest number on normalized-losses feature.

Treatment for missing value:

- Remove rows with missing value in the price variable.
- Discard the normalized-losses column because the amount of missing value is too high.
- Performs imputation on stroke, bore, peak-rpm, num-of-doors, and horsepower features with their median, mean, or mode values (depending on the type and distribution of data).

Exploratory Data Analysis

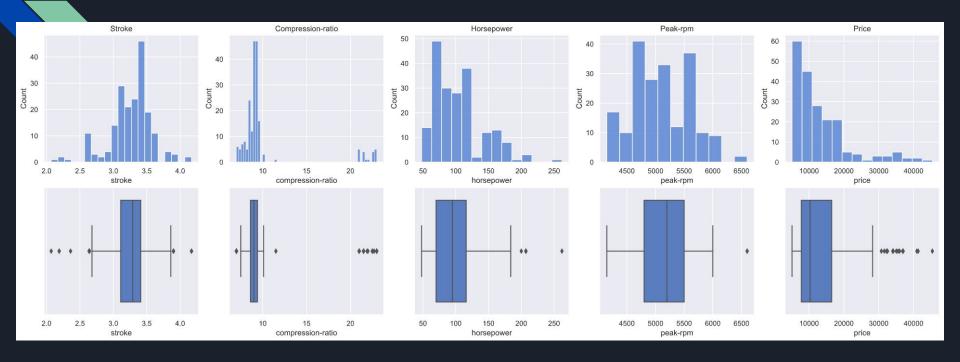
Univariate Analysis (continuous variable)



This is our first 5 continuous variables.

From our analysis, there are 2 variables which have high skew (greater than 0.85). Those variables are <u>wheel-base</u> and <u>width.</u> We will do log transformation to these variables.

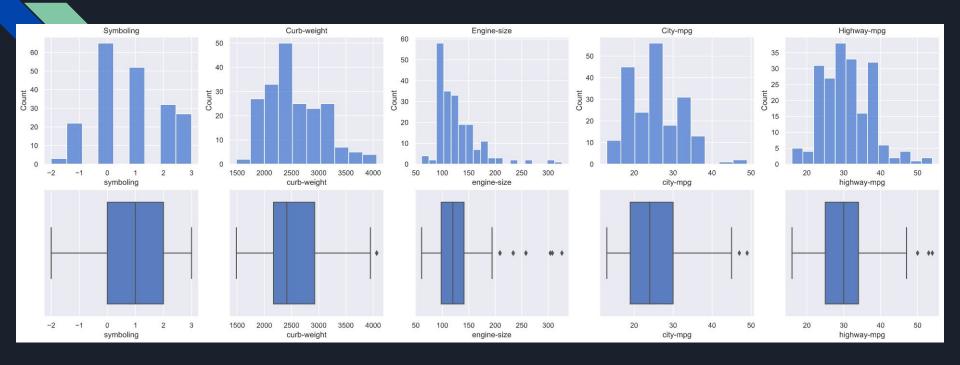
Univariate Analysis (continuous variable)



This is our next 5 continuous variables.

From our analysis, there are 3 variables which have high skew (greater than 0.85). Those variables are <u>compression-ratio</u> and <u>horsepower</u>. We will do log transformation to these variables.

Univariate Analysis (discrete variable)



This is our 5 discrete variables.

From our analysis, there are 1 variables which have high skew (greater than 0.85) which is <u>engine-size</u>. We will do log transformation to these variables.

Correlation Heatmap

symboling -	1.00	-0.13	-0.20	-0.05	-0.53	-0.04	-0.09	0.21	-0.54	-0.37	-0.24	-0.55	-0.23	0.03	-0.03	-0.11	-0.02	-0.14	-0.01	-0.18	80.0	0.28	-0.04	0.04	-0.08
make -	-0.13	1.00	0.17	0.11	0.09		0.60		0.56	0.62	0.67	0.22	0.72	0.29	0.66	0.75			0.01		0.62	-0.15	-0.54	-0.56	0.89
fuel-type -	-0.20	0.17	1.00		0.07	0.12	0.12	-0.04	0.31	0.21	0.24	0.28	0.22	-0.08	0.01	0.07	0.18	0.05	0.24	0.99	-0.17	-0.48			0.11
aspiration -	-0.05	0.11	0.41	1.00	-0.00	-0.05	0.12	-0.06	0.26			0.09		-0.04	-0.03	0.11	0.21	0.23	0.22	0.31	0.25	-0.19	-0.19	-0.24	0.18
num-of-doors -	-0.53	0.09	0.07	-0.00	1.00	0.10	-0.01	-0.11			0.21		0.20	-0.02	0.03	0.05	0.03	0.13	-0.03	0.06	-0.05	-0.19	-0.07	-0.09	0.06
body-style -	-0.04		0.12	-0.05	0.10	1.00	0.28		0.15	0.26	0.19	0.12	0.26	0.11	0.26	0.34	0.32	0.30	-0.04		0.24	-0.10	-0.20	-0.20	0.40
drive-wheels -	-0.09	0.60	0.12	0.12	-0.01	0.28	1.00	0.16				0.05	0.67						-0.03	0.10		-0.09	-0.55	-0.58	0.64
engine-location -	0.21	0.31	-0.04	-0.06	-0.11		0.16	1.00	-0.19	-0.05	-0.05	-0.11	0.05	0.02	0.22	0.20	0.11	0.19	-0.14	-0.02		0.20	-0.16	-0.10	0.33
wheel-base -	-0.54					0.15		-0.19	1.00	0.88	0.81	0.59	0.78	0.23	0.40				0.16		0.37	-0.36	-0.47	-0.54	
length -	-0.37	0.62			0.36	0.26			88.0	1.00	0.86	0.49	0.88	0.33	0.47	0.69	0.64	0.61	0.12			-0.29	-0.67	-0.70	0.69
width -	-0.24	0.67		0.31	0.21	0.19	0.51		0.81	0.86	1.00	0.31	0.87	0.39	0.58	0.73	0.60	0.54	0.19		0.62	-0.25	-0.63	-0.68	0.75
height -	-0.55	0.22		0.09	0.46	0.12	0.05	-0.11	0.59	0.49	0.31	1.00	0.31	-0.07	0.05	0.07	0.19	0.19	-0.06		-0.09	-0.31	-0.05	-0.10	0.14
curb-weight -	-0.23	0.72	0.22	0.32	0.20	0.26	0.67		0.78	0.88	0.87	0.31	1.00	0.51	0.62	0.85	0.66	0.64	0.17		0.76	-0.28	-0.75	-0.79	0.83
engine-type -	0.03	0.29	-0.08		-0.02	0.11			0.23			-0.07		1.00	0.63	0.60			-0.14			-0.03	-0.43	-0.47	0.45
num-of-cylinders -		0.66	0.01		0.03	0.26						0.05	0.62	0.63	1.00	0.75		0.20	0.11		0.63	-0.08	-0.52	-0.56	0.76
engine-size -	10.000	0.75	0.07		0.05	0.34				0.69	0.73	0.07	0.85	0.60	0.75	1.00	0.53		0.21		0.82	-0.26	-0.65	-0.68	0.87
fuel-system -	11.10		0.18		0.03	0.32				0.64	0.60	0.19	0.66	0.35	0.41		1.00	0.48	0.06		0.62	-0.06	-0.59	-0.58	0.59
bore -	-0.14	0.48	0.05		0.13	0.30	0.58	0.19	0.49	0.61	0.54	0.19	0.64	0.23	0.20	0.58	0.48	1.00	-0.06		0.57	-0.28	-0.59	-0.60	0.54
stroke -		0.01	0.24		-0.03	-0.04	-0.03	-0.14	0.16			-0.06		-0.14	0.11	0.21	0.06	-0.06	1.00	0.19	0.10	-0.07	-0.03	-0.04	0.08
compression-ratio -	-0.18	0.14	0.99		0.06	0.13	0.10	-0.02	0.25	0.16	0.19	0.26	0.16	-0.09	-0.01	0.03	0.16	0.00	0.19	1.00	-0.21	-0.44	0.33	0.27	0.07
horsepower -		0.62	-0.17	0.25	-0.05	0.24	0.59		0.37	0.58	0.62	-0.09	0.76	0.56	0.63	0.82	0.62	0.57	0.10	-0.21	1.00	0.11	-0.82	-0.80	0.81
peak-rpm -		-0.15	-0.48	-0.19	-0.19	-0.10	-0.09	0.20	-0.36	-0.29	-0.25	-0.31	-0.28	-0.03	-0.08	-0.26	-0.06	-0.28	-0.07	-0.44	0.11	1.00	-0.12	-0.06	-0.10
city-mpg -		-0.54	0.27	-0.19	-0.07	-0.20	-0.55	-0.16	-0.47	-0.67	-0.63	-0.05	-0.75	-0.43	-0.52	-0.65	-0.59 -0.58	-0.59	-0.03	0.33	-0.82	-0.12	1.00	0.97	-0.69
highway-mpg -	0.04	-0.56			-0.09		-0.58	10000	-0.54 0.58	-0.70	12/232	-0.10	-0.79	0.45	-0.56	-0.68		0.54	-0.04		-0.80		0.97	1.00	100000
price -		0.89	0.11	0.18	0.06	0.40	0.64	0.33		0.69	0.75	0.14	0.83		0.76	0.87	0.59			0.07	0.81	-0.10	-0.69	-0.70	1.00
	symboling	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	length	width	height	curb-weight	engine-type	num-of-cylinders	engine-size	fuel-system	bore	stroke	compression-ratio	horsepower	peak-rpm	city-mpg	highway-mpg	price

Our target variable has a strong correlation with several variables such as:

- make
- engine-size
- curb-weight
- horsepower
- num-of-cylinders
- etc

Problem:

- 0.00

-0.25

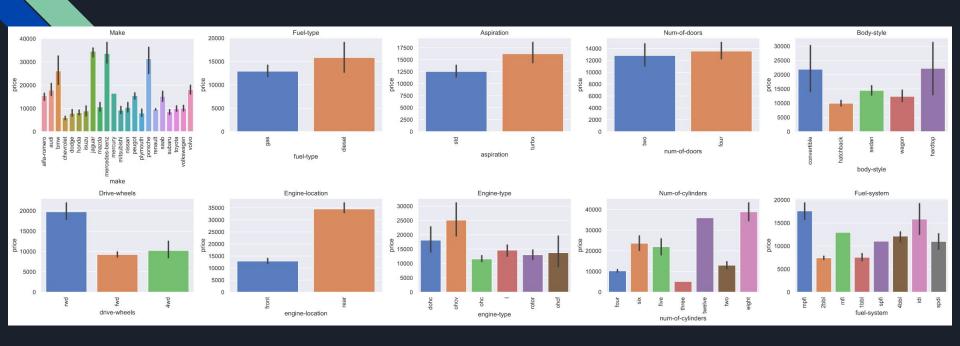
- -0.50

- -0.75

There are 2 independent variables whose correlation is close to perfect, namely <u>fuel-type with</u> <u>compression-ratio</u> and <u>city-mpg with</u> <u>highway-mpg</u>.

We need to remove one of these variables to reduce the multicollinearity problem in the Linear Model.

Multivariate Analysis (categorical variables)



- Diesel and turbo car have higher average car price compared to the other category.
- Car with four doors have slightly higher price compared to car with two doors.
- Car with rwd type have more higher average car price compared to the other type.
- For make variable, there are to much category. We will group the low count car to "Other" category.

Data Preprocessing

Preprocessing

For preprocessing data, we do the following treatment:

- Missing value imputation
 Perform imputation using median, mean, or mode values (depending on the type and distribution of data).
- Outlier / skewness handling Using log transformation.
- Categorical encoding
 Using labels and one-hot encoding.
- 4. Scaling

 Do scaling with minmaxscaler
- Feature engineering Added polynomial features.

Modeling

Base Model (LinearRegression)

	Train R2	Test R2	Train RMSE	Test RMSE
Model				
Linear	0.9707	0.9071	1162.1067	3166.4252

There is no additional polynomial features on the base model. Because the number of features is large, and if you add the polynomial features it will overfit.

Ridge and Lasso Regression

	Train R2	Test R2	Train RMSE	Test RMSE
Model				
Ridge	0.9900	0.9505	679.2758	2311.8932
Lasso	0.9938	0.9234	534.2514	2874.9289

After adding polynomial features and scaling, we get higher scores with Ridge and Lasso (default alpha 1.0 parameter).

After tuning alpha

Hyperparameter tuning is done by optuna.

https://optuna.org/

After tuning there is an increase in the score on both models.

The best model is Ridge with alpha 1.7323 with R-Squared on the test set 0.9558, and error (RMSE) on the test set 2184.24.

Recommendation

Recommendation

Suggestions for next steps in analyzing this data:

- This modeling was carried out with a train test split. To get better generalization results, it would be better to do it with cross validation.
- Perform different engineering features.
- Make modeling with different machine learning such as tree-based model or gradient boosting.