Project Report: Predicting Used Car Price

FEB 2022

ASHAD ALAM

Overview

Determining the selling price of a used car is a challenging task because there are several factors that affect its price. The focus of this project is on developing a machine learning model that can accurately predict the selling price of used cars using appropriate features.

Data Set

For this project we will be using Automobile Data Set available on UCI Machine Learning Repository. A brief description of attributes is provided below.

Attribute Information:

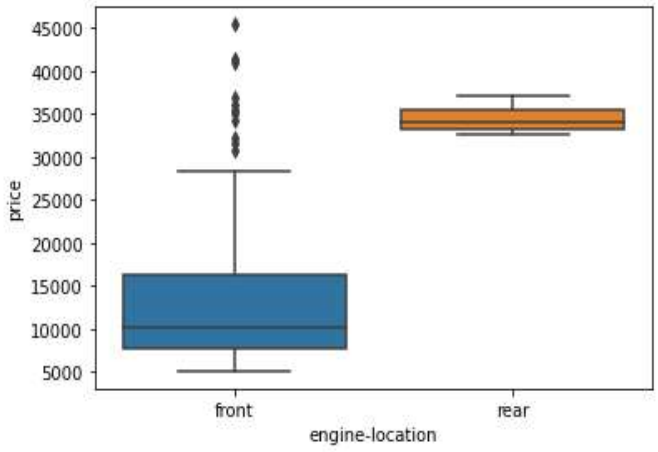
1. symboling: -3, -2, -1, 0, 1, 2, 3.
2. normalized-losses: continuous from 65 to 256.
3. make: alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo
4. fuel-type: diesel, gas.
5. aspiration: std, turbo.
6. num-of-doors: four, two.
7. body-style: hardtop, wagon, sedan, hatchback, convertible.
8. drive-wheels: 4wd, fwd, rwd.
9. engine-location: front, rear.
10. wheel-base: continuous from 86.6 120.9.
11. length: continuous from 141.1 to 208.1.
12. width: continuous from 60.3 to 72.3.
13. height: continuous from 47.8 to 59.8.
14. curb-weight: continuous from 1488 to 4066.
15. engine-type: dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
16. num-of-cylinders: eight, five, four, six, three, twelve, two.
17. engine-size: continuous from 61 to 326.
18. fuel-system: 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
19. bore: continuous from 2.54 to 3.94.
20. stroke: continuous from 2.07 to 4.17.
21. compression-ratio: continuous from 7 to 23.
22. horsepower: continuous from 48 to 288.
23. peak-rpm: continuous from 4150 to 6600.
24. city-mpg: continuous from 13 to 49.
25. highway-mpg: continuous from 16 to 54.
26. price: continuous from 5118 to 45400.

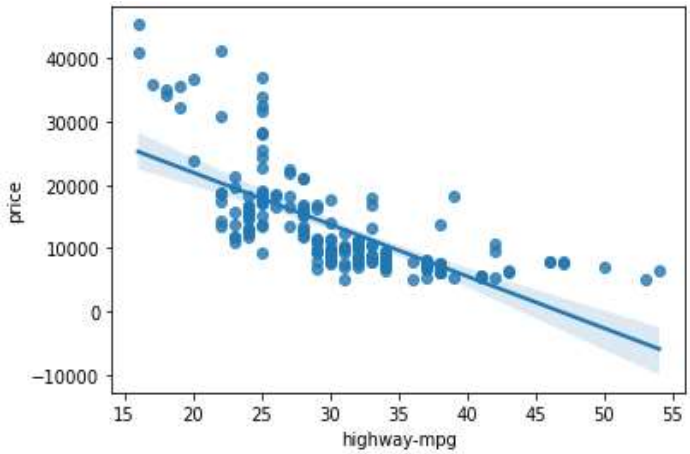
Data Exploration

Data Cleaning and Feature Engineering

We can see that column such as num-of-doors, normalized-loss, bore, stroke, peak-rpm, horsepower, price contains some missing values, therefore we have replaced missing values with appropriate values such as num-of-doors by four (most of vehicle contains four gates), normalized-loss, bore, stoke, peak-rpm, horsepower by average value, and dropped rows that contains missing value price (we cannot accurately guess the price of vehicle without knowing anything about it). We have also seen that some of the features are not in the correct data format hence we converted them into the appropriate data format.

Now our data set does not contain any missing value. To get a better understanding of our data we have plotted some scatterplots, and histograms. We have noticed that data contains many outliers, therefore we have normalized our data to remove outliers. We have also seen several columns skewed therefore we applied log transform on those columns. We have also applied one-hot-encoding on our dataset since our dataset contains several unique values.





Key Findings

We can see that vehicle whose engine is on the rear side are much costlier than front. We can also see that highway-mpg is negatively related to price as highway-mpg increases price decreases. And our data contains many columns which are not correlated to the price and will not improve the performance of our model hence we dropped those columns.

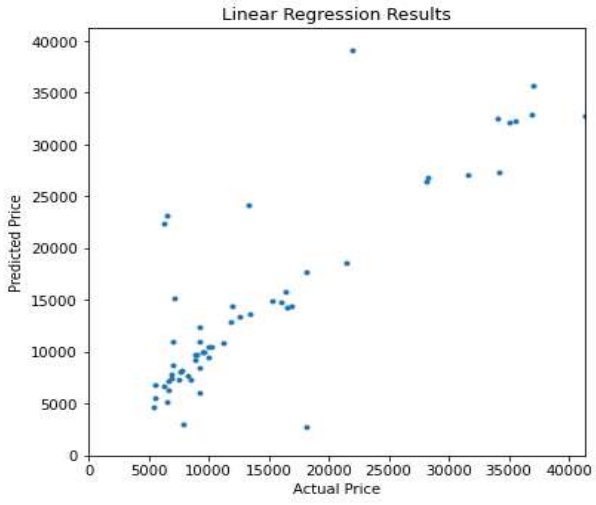
The given table shows the list of columns which are skewed we have set skew limit=0.75:

	Skew
compression-ratio	2.704644
price	1.992907
wheel-base	1.134976
width	0.843971
stroke	-0.943292

Methodology

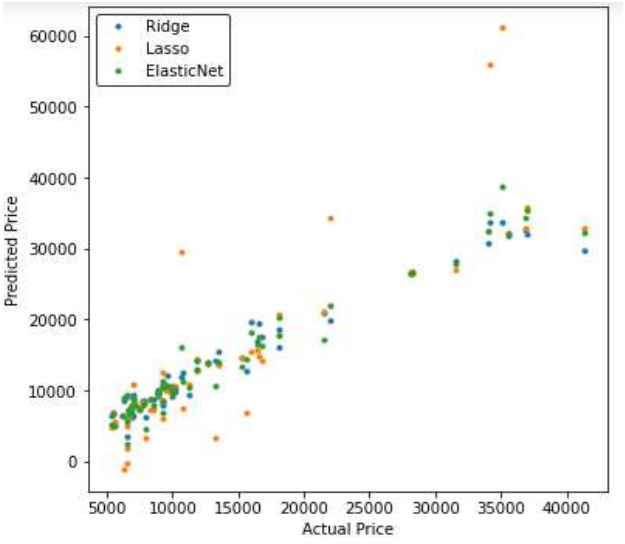
We used various methods and techniques, with a 70% - 30% splits for our train and test data. Simple Linear Regression, added polynomial affect, used regularization regression. For all these implications, scikit-learn packages were used.

1.Simple Linear Regression



2.Regularization

After applying regularization methods like lasso, ridge, elastic net the performance of our model improves. Figure shows Predicted Price vs Actual Price in lasso, ridge, elastic net.



Best Model

From the above figures we can see that ridge, elastic net performs much better than linear and lasso regression. For getting the best model we have calculated the RMSE values for all models. The results is shown below:

	RMSE
Linear	37313.093582
Ridge	2337.273140
Lasso	6004.287245
ElasticNet	2189.280262

We have also calculated the R2 score which is a measure of how close our predicted value is to the real value. Below Table shows the corresponding R2 score for our models.

	R2
Linear	-13.687653
Ridge	0.942370
Lasso	0.619677
ElasticNet	0.949437

We can see that R2 score of Ridge and Elastic Net is almost similar and equals 0.94 therefore we can say that 94 percent of our data is predicted by our model.

Future Work

For better performance, we plan to judiciously design deep learning network structures, use adaptive learning rates and train on clusters of data rather than the whole dataset. We can also use Random Forest, XG Boost algorithms to enhance performance of our model. To correct for overfitting in different selections of features and number of trees will be tested to check for change in performance.