Predicting the price of used cars

July 21, 2018

Made by-
Shivam Sachan

Overview

Predicting the price of used cars in both an important and interesting problem. With difficult economic conditions, it is likely that sales of second-hand imported (reconditioned) cars and used cars will increase. In many developed countries, it is common to lease a car rather than buying it outright. A lease is a binding contract between a buyer and a seller (or a third party –usually a bank, insurance firm or other financial institutions) in which the buyer must pay fixed instalments for a pre-defined number of months/years to the seller/financer.

Predicting the resale value of a car is not a simple task. It is trite knowledge that the value of used cars depends on a number of factors. The most important ones are usually the age of the car, its make (model and variant), its mileage (the number of kilometers it has run), color, on road price, and many other.

Methodology

Data plays very important role in any machine learning problem. Scraping is done from different sites in order to have good amount of data. Making the good quantity of data into good quality data is again the tedious task to do in order to increase accuracy.

Techniques used to remove outliers-

1. Mean and standard deviation technique
2. Clustering technique

We are ready with the quality data. Obviously it is a supervised-regression problem and can be depicted same by looking into the data. Now it's time to implement different machine learning models. Different regression models used are-

1. Linear regression
2. Decision tree
3. Support vector regressor
4. Lasso
5. Elastic net
6. SGD regressor
7. Random forest
8. Gradient boosting
9. MLP regressor

It is all about finding the best model for our data. Tested with 10% of total data and found that Gradient boosting tops the list in accuracy amongst all models.

Evaluation and conclusion

Model is ready to give the output with the accuracy of 90% but there is a problem as some of the models and variants are not in our trained model. To overcome from this problem, clustering is used. Within a cluster, there are some models and variants which are in the model means for them, depreciation can be calculated. For remaining data points in that particular cluster, depreciation would be approximately same or may be average of all the data points for which we can directly predict.

Milestones

The main limitation of this problem was not having every types of models and variants in the trained model but this will always be a problem because every day number of cars are produced having different models and variants. As future work, we intend to collect more data and to use more advanced techniques like artificial neural networks, or any other technique.