FLIP ROBO

Car Price Prediction

Submitted by:

Sanuj P O

ACKNOWLEDGEMENT

I thank Data Trained team and the faculty at FlipRobo to get an exploration on a real-life dataset which helped in exploration of the concepts that I use. I have also taken few references from Kaggle and also some YouTube videos to find apt resolution for the dataset.

Business Problem:

With the covid 19 impact in the market, we have seen lot of changes in the car market. Now some cars are in demand hence making them costly and some are not in demand hence cheaper.

One of our clients works with small traders, who sell used cars. With the change in market due to covid 19 impact, our client is facing problems with their previous car price valuation machine learning models. So, they are looking for new machine learning models from new data. We have to make car price valuation model.

We have 2 Phases here,

1st : Data scrapping. We have to scrape atleast 5000 data. In this section we will try to get a variety of websites and the prices from the websites to see the fluctuations in the market. We will scrape data from different locations and then we can also get relevant data like model, price, kms driven, fuel, etc.,

2nd: Need to build a machine learning model. Before model building we have to do all data pre-processing steps. Trying different models with different hyper parameters and selecting the best model.

Working on data:

Initially we have to go through the websites, I've tried many visited where in the varieties are less, so planned to move in with OLX.

Had set up multiple locations like Kerala, Maharashtra, Delhi, Kolkata etc.,

I have scraped details like, Price, Kms driven, Brand, Model, year of manufacture, Transmission and variant.

Here we have are using webchrome driver and Selenium to get our data, and created lists to store them, I've used 5 locations and 5 ipynb files which will keep the data segregated for reference and later on we can merge them to continue with the learning.

And with the load of data that we have to take all programs run for 2-3 hours to get the details from every location and the details of car available.

We had created the file however the Variant column does not have much information hence we will have to drop that column in the later part.

Had to terminate in between as it was taking a longer time to scrape more data.

The shape of the data is now 11203, 12.

As there were no values, we are dropping variant and Unnamed: 0, Unnamed: 0.1 as a part of cleaning.

In the kms column, we were having Km, ",", "-", etc, hence changed them to get just the kilometers run.

For price as well, we did the same, replaced the extra data that we have apart from the actual price.

For Owner, we had changed 1st to 1, 2nd to 2 and like that..

For unknown values in fuel we substituted with Mode.

Same with Transmission as well.

For brand, changed blank values, that is "-" with others.

Done the same with Model as well.

Got other brands into a new section as it is a noise in the data which will not infer much details.

Based on the details that we fetched here are the inferences made from visualisation:

1. In 2012 most number of cards were purchased.

2. With the number of year of manufacture increases, the prices increase as well.
3. With the number of Kilomometers increase, there is a decline in the price.
4. Mostly people prefer petrol compared to automatic.
5. Delhi is having the highest number of sellers and they have the range higher compared to rest of the cities.
6. Single owner vehicles tends to get more price that others.
7. Price is higher for diesel vehicles then comes petrol and others.
8. Kilometers and Price are highly correlated, Year of manufacture and price as well are highly correlated in these.
9. After removing the outliers, moved on to complete the scaling process as the price is in big range and we have to get them all together.

After completing the same we can move ahead with modelling and after running all the models, we found that Random forest was giving best prediction compared to rest and also performing good on cross validation.

We have saved the model and finalized Random forest.

Conclusion

We had a good accuracy on predicting the price which is upcoming based on the data that we have, if we put in new cards in an affordable range, this could do wonders.