

Project Report on Car Price Prediction



BY:

MAYANK SHUKLA

A solid orange horizontal bar at the bottom of the page.

Background & Introduction

We are about to deploy an ML model for car selling price prediction and analysis. This kind of system becomes handy for many people.

Imagine a situation where you have an old car and want to sell it. You may of course approach an agent for this and find the market price, but later may have to pay pocket money for his service in selling your car. But what if you can know your car selling price without the intervention of an agent. Or if you are an agent, definitely this will make your work easier. Yes, this system has already learned about previous selling prices over years of various cars.

So, to be clear, this deployed web application will provide you will the approximate selling price for your car based on the fuel type, years of service, showroom price, the number of previous owners, kilometres driven, if dealer/individual, and finally if the transmission type is manual/automatic. And that's a brownie point

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

Approach to Data Cleaning


- * First we check the information of the given dataset because it tells that how many rows and columns are present in our dataset and data type of the columns whether they are object, integer or float.
- Drop duplicates rows if present in dataset.+Then we check for the null values present in our dataset.
- If null values are present then fill it via mean, median or mode. Or also you can remove that rows but kindly check it properly.
- After that we check the summary statistics of our dataset. This part tells about the statistics of our dataset i.e. mean, median, max value ,min values and also it tell whether outliers are present in our dataset or not.

Approach to Data Cleaning (Cntd...)

- * We also check the correlation of our dataset to check the correlation of the columns with each other. If columns are highly correlated with each other let's say 90% or above then remove those columns to avoid multi coli-nearity problem.
- * We extract data from date column and make new columns like day, month and year to see the outcomes with our target column that is label.
- * We delete the pcircle column because it has only one unique value that tells that collected data is only for one circle.
- * We cannot remove outliers because more than 20% of our data are removed.



Visualization.

- * We plot correlation matrix via heatmap to see the correlation of the columns with other columns.
 - * We also visualize the correlation of columns with target column via bar graph to see which column is highly correlated with target column.
 - * We see the number of defaulter and non defaulter customers with the help of count plot.
 - * We plot histogram to displays the shape and spread of continuous sample data.
 - * We also see the customers labels i.e defaluter/Non-defaulter according to date and month with count plot.
 - * We also see the distribution of the data with the help of distribution plot whether it is left skewed or right skewed.
- 

Modelling part

- * We know that this is classification problem so we use accuracy score, classification report and confusion matrix as our evaluation matrix. We also see the AUC score and also plot the AUC_ROC curve for our final model.
- * As we know this dataset is imbalance so we don't too much focus on accuracy score . We see the precision and recall value along with f1_score.
- * First we see the result without doing any sampling technique and for that I use Decision Tree Regressor.
- * We also use Random Forest Classifier as our evaluation model without using hyper-parameter tuning because our dataset is too large and it takes more than hour to give the result.
- * At the end we use Voting Regressor, which is an ensemble meta-estimator that fits several base regressors, each on the whole dataset to average the individual predictions to form a final prediction.

Conclusion

- So here 'Voting Regressor Model' is the best model out of all model tested above and by looking this we can conclude that our model is predicting around 87.35% of correct results for Label '0' indicates that the loan has not been payed i.e. defaulter.