CAR PRICE PREDICTION

Submitted by:

GOURAV KUMAR

ACKNOWLEDGMENT

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

INTRODUCTION

This notebook is going to be focused on solving the problem of Cars Price Prediction for Cars sellers.

- A Car value is simply more than location and Kilometers Driven.
- We are going to take advantage of all of the feature variables available to use and use it to analyze and predict Cars prices.
- We are going to break everything into logical steps that allow us to ensure the cleanest, most realistic data for our model to make accurate predictions from.
- Load Data and Packages
- Analyzing the Test Variable (Car Price)
- Multivariable Analysis
- Impute Missing Data and Clean Data
- Feature Transformation/Engineering
- Modeling and Predictions

Analytical Problem Framing

- A benefit to this study is that we can have two clients at the same time! (Think of being a divorce lawyer for both interested parties) However, in this case, we can have both clients with no conflict of interest!
- Client car buyer: This client wants to find their next car with a reasonable price tag. They have their locations of interest ready. Now, they want to know if the car price matches the car value. With this study, they can understand which features (ex. Kilometers, location, etc.) influence the final price of the car. If all matches, they can ensure that they are getting a fair price.
- Client car seller: Think of the average car-flipper. This client wants to take advantage of the features that influence a car price the most. They typically want to buy a car at a low price and invest on the features that will give the highest return. For example, buying a car at a good location but small Kilometers Driven. The client will invest on Kilometers Driven at a small cost to get a large return.

Model/s Development and Evaluation

- Data for training and validation (Measure MSE)

  To select a set of training data that will be input in the Machine Learning algorithm, to ensure that the classification algorithm training can be generalized well to new data. For this study using a sample size of 15%, assumed it ideal ratio between training and validation

- GridSearchCV: RandomForestRegressor

  GridSearchCV is a library function that is a member of sklearn's model_selection package. It helps to loop through predefined hyperparameters and fit your estimator (model) on your training set. So, in the end, you can select the best parameters from the listed hyperparameters.

  A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees. Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap

- k-fold

  k-fold is a popular kind of cross-validation technique, in which, say k=10 for example, 9 folds for training and 1 fold for testing purpose and this repeats unless all folds get a chance to be the test set one by one. This way, it provides a good idea of the generalization ability of the model, especially when we have limited data and can't afford to split into test an d training data.

CONCLUSION

In this research paper, we have used machine learning algorithms to predict the car prices. We have mentioned the step by step procedure to analyze the dataset and finding the correlation between the parameters. Thus we can select the parameters which are not correlated to each other and are independent in nature. These feature set were then given as an input to four algorithms and a csv file was generated consisting of predicted house prices. Hence we calculated the performance of each model using different performance metrics and compared them based on these metrics. We found that Decision Tree overfits our dataset and gives the highest accuracy of 70%. KFold gives the least accuracy of 70%.