

CAR PRICE PREDICTION

Submitted by:

VINOD ADEP

**ACKNOWLEDGMENT**

I Would likely to express my special thanks to Flip Robo technology for given me a wonderful opportunity to explore on Large Dataset and Data Science Project

I have to say special thanks to Data Trainded Mentor’s had guided me to Solving logical problems and explained project concept in deeply.

Link that helps me to complete this project:

<https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.RandomForesrRegressor.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.DecisionTreeReRegressor.html>

https://scikit-learn.org/stable/modules/generated/sklearn.feature\_extraction.text.CountVectorizer.html

**INTRODUCTION**

* Business Problem Framing :-

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. This project contains two phase

* Review of Literature :-

Type of Car Names This is a comprehensive summary of the research done on the topic. The review should enumerate, describe, summarize, evaluate and clarify the research done.

* Motivation for the Problem Undertaken :-

Motivation of this project is to determine which feature are affecting much to predict the car price. Here, there are number of feature and motive is to extract a those feature’s which is more usefull for to predict the car price of the market.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem :-

In this sample data have some features numerical columns and some are text columns, I had used nlp to covert text to vectors of array ,which will give the frequency of words in column, With CountVectorizer

* Data Sources and their formats :-

Format of the data set is csv. Pandas I used to read and access the data for understand the structure of the data. From the data set we can see that some columns are have more text which is likely to be documents type.

* Data Preprocessing Done :-

In this project I have used Get Dummies on some columns which are fuel, owner, insurance, etc. I had Splitted some columns with numerical data.

* Data Inputs- Logic- Output Relationships :-

Aim of this project is to identify the best input’s. Based on this inputs sale price will predict accurately. Here some of the inputs which are highly affecting to predict the sale price like Overall quality of the house which is most important feature to predict the price old house. In this columns have 10 number which are ordinal. 10 means the quality of house is soo rich having well neighbours quality of material used while building the house all the thing’s are consider based on this sale price will increase or decrease. Suppose the quality of houses rated out of 10 have only 3 or 4 this type of houses price have less compared to rating have more than 7. Such that other features also acts over target column i.e sale price and help to predict the price of house.

* State the set of assumptions (if any) related to the problem under consideration :-

House Price Prediction is one of the main role in Real Estate buissness , I think we have to add another column in this dataset for different area have different sale price, So that will be very usefull when we deploy project by area location. Also House buyers have get idea about which area is reasonable for purchasing house.

* Hardware and Software Requirements and Tools Used

Listing Tools and libraries that I have used to solve project:

* Software: Anaconda, Jupyter Notebook, Python3
* Libraries: Numpy, Pandas, Matplotlib,NLTK
* Sklearn

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods) :-

While performing EDA we have got insite information from dataset to predict Sale price. Some of the columns are interdependent like garage area and garege can have number of cars space. Number of cars are directly correlated to area of garage. Multicorrelation will effect some of the models.

Most of the columns are ordinal type which giving the grades, rating and ranks to the house’s which have directly correlation to our target column.

Filling null values have major role in this dataset beacuse dataset have lots of null values in both categorical and continues data. Based on the type of column i have repalce the null values to some other values. I have handle null values in continuos column after encoding entire data set into numberical then after based on all the columns null values are imported bye usinf KNN Imputer.

* Testing of Identified Approaches (Algorithms) :-

I have used some of the model of machine learning like Linear regression, Decision Tree Regressor, Random Forest Regressor. I have used for loop for random state so it can give best random state with accuracy. It used to select best random state to split the data to resepective model and also find best cv(cross validation). Each model have differect random state and cv. I have tested by applying before scaling, after scaling the data. And also before feature selection and after feature selection. How models are giving there output also observed that after scaling the data linear regression model’s accuracy has increase heigh. To handle the problem of over fitting and under fitting the model. Each model will give there out put after tuning the model.

* Run and Evaluate selected models :-

In this project I got good accuracies from Linear regression and Random Forest Regressor algorithms.

Linear Regression is used in suppervised regression model. Linear regression is used to fit the best fit line on the basis of that line sale prices would predicted. Best fit line is a line where the difference between actual and predicted values is less that is called best fit line.   
Linear regression contain two major values based on this values line would drawn i.e intercept and the slope of the line. Best slope would be calculated by using the gradient decent method with the small learning rate.

Randome Foresr is a tree based algorithm it cab be used on both classification and regression problems.

Now we have a regression problem so used Random Forest regressor. In this algorithm first it is calculte the impurity of each column and check the information gain. Which gives high information gain that columns beacomes a root node then their child would be their features. Then again calcualte the entropy or gini and information gain like this tree will build.

Interpretation of the Results:-

To see the accuracy of each model I have used r2 score. R2 score is the score how good slope has genreate for the best fit line. Linear regression and random foresr regression is giving the r2 score near to 87.05 %.

**CONCLUSION**

* Learning Outcomes of the Study in respect of Data Science:-

I have learn some inbuild funtion of python at the time of data cleaning. Object columns to Integer are done by using python map function over on lambda function and Mean encoder and ordinary encoder. The main challenging is that to run all model by hypertuning each model and fetching each model best random state and best cv once and get the best model name as a result.