Car price prediction project

▶   ***Problem statement:-***

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

Data Collection Phase

You have to scrape at least 5000 used cars data. You can scrape more data as well, it's up to you. more the data better the model

In this section You need to scrape the data of used cars from websites (Olx, cardekho, Cars24 etc.) You need web scraping for this. You have to fetch data for different locations. The number of columns for data doesn't have limit, it's up to you and your creativity. Generally, these columns are Brand, model, variant, manufacturing year, driven kilometers, fuel, number of owners, location and at last target variable Price of the car. This data is to give you a hint about important variables in used car model. You can make changes to it, you can add or you can remove some columns, it completely depends on the website from which you are fetching the data. Try to include all types of cars in your data for example- SUV, Sedans, Coupe, minivan, Hatchback.

Note – The data which you are collecting is important to us. Kindly don't share it on any public platforms.

Model Building Phase

After collecting the data, you need to build a machine learning model. Before model building do all data pre-processing steps. Try different models with different hyper parameters and select the best model.

Follow the complete life cycle of data science. Include all the steps like.

1.   Data Cleaning

2.   Exploratory Data Analysis

3.   Data Pre-processing

4.   Model Building

5.   Model Evaluation

6.   Selecting the best modEL

Data scrapping

❑ First we will scrap the data for model building , for that purpose import all the required libraries like:-selenium,webdriver,pandas,time etc.

❑ Then using driver import the chromedriver.exe and then import the webpage from where we have to scrap the data

❑ Then create some empty list to store the scrapped data

❑ Afte this using xpath scrap all the reuired data and make data frame of it so that we can use that data frame for model building

Model building

.Import All the important Libraries

2.Load the data set which is to be used for analysis

3.Check for the null values in the dataset

  ➤ If there are null values fill them using fillna method

4. Check data description

  ❑ data description show full description of every column like:- std,mean,median,25% of the column and some other important information about every columns

Visualization steps

❖ Check for correlation in the columns using heatmap or other techniques

  ❑ Using correlation we can check that if any columns having high correlation with each other we can remove one of them

❖ After checking correlation we plot the graph for data distribution among all the columns

❖ With the help of data distribution graph we can see that data is distributed well or not , if not distributed well we can check skewness using data.skew() method

❖ If the value of any column is in between +0.5 and -0.5 that means data is distributed well otherwise there is skewness in the data

Data preprocessing

- ✓ After that check for outliers

- ✓ After checking for outlier scaled the data using StandardScaler

- ✓ Now check for vif score for every column, which column having vif score more than 5 remove that column

Training process

- ❑ In this project we are going to use linear regression for predicting the sales prices of houses.

- ❑ First import all the important libraries i.e. LinearRegression, r2_score, confusion matrix, classification report and train_test_split for the training and testing of the model.

- ❑ After getting the random_state split the data into training and testing.

- ❑ Then fit the training and testing data into the model and check the accuracy score that how well it fits our data.

- ❑ Then perform the cross validation and hyper parameter tunning on the model and plot the ROC AUC curve for the final accuracy.

- ❑ Finally save tho model for the later prediction of the model.