Project on Car Price Prediction

Title : To predict the price of a car on the basis of car name and its features.

Objective: Develop a robust machine learning model to predict car prices based on a dataset containing various features such as car specifications and pricing information. This project aims to conduct thorough data exploration and cleaning, perform feature engineering to enhance model performance, train multiple algorithms, and deploy the final model using Streamlit for practical usage.

Project Implimentation:

- ❖Importing libraries and reading file

1) Importing some important libraries like pandas , numpy, stats etc.

2) Reading the csv file from local machine. The dataset contains features like Name, Location, Year, Kilometers, Fuel type, Transmission, Owner type, Mileage, Engine, Power, Seats, New price, Price.

3) By applying info function I observed that some columns contains null values and also observed the data

type of each column, also describe method tells me count,min, max, std etc of numerical columns.

❖ Data Cleaning

4) First of all mileage column contains string data ‘kmpl’ and we have to remove it and therefore split the data by ‘k’ and take first index. And after that change the datatype of column to float. Mileage column also contains null values so replace that null values with mean.

5) After that Engine column contains string data ‘cc’ and we have to remove it and therefore split the data by ‘c’ and take first index. And after that change the datatype of column to int. Engine column also contains null values so replace that null values with median because it contains some outliers.

6) After that Power column contains string data ‘bhp’ and we have to remove it and therefore split the data by ‘b’ and take first index. And after that change the datatype of column to int , But some of the values in power column are not numeric therefore we first convert it to numeric and then Power column also contains null values so replace that null values with mean and change its datatype to float.

7) Seats column also contains nan values which are replaced by mode. And then change the datatype to int.

Drop the unnecessary columns. We are done with a data cleaning.

❖Exploratory Data Analysis

8) Some of the columns are not numeric so will covert then to numeric. First will take Owner type and I observed that there is sort of order therefore will do ordinal encoding .Instead of sklearn I did maping and gives the values to the categories. After that changed the datatype to int.

9) In fuel_type column ‘Electric’ and ‘LPG’ categories are very less so I deleted that rows which contains ‘Electric’ and ‘LPG’. Also drop the new price column. From a dataset we observed that location column is not important and that’s why we will delete that column.

10) From a Name column we can extract one column which is company, so we simply apply split function and take first index i.e company name. But in car name we have to remove the company name and to do it so we apply function to remove company name from car name column. And we got two columns which are company contains car’s company and rename column contains only name of cars.

11) From company column we can only take top 10 categories and other data we can give it to others category.

❖ Data Visualization

12) Plot a barplot of company vs price and from that I observed Audi, Mercedies, BMW are the companies whose price is too high.

13) Plot a barplot of Owner_Type vs price and from that I observed owner no. 1 will have the maximum price for a car and that is very much obvious.

14) Plot a barplot of Seats vs price and from that I observed the car which has two seats that cars price is high. Also plot of Transmission vs Price in which car which have automatic transmission have high price.

15) To chek for the outliers we will plot boxplots. So we plot boxplot for year, Kilometers_Driven , Mileage, Engine and observed that all the columns have outliers.

16) For year column all the outliers are below IQR range and therefore by quantile function we fetched all values below 1 percent and then remove it. We did same for mileage column in which we apply lambda function but we replace the outliers with mean of that cokumn.Simultaneously we are ploting a boxplot.

17) For Engine and Power column we replaced the outliers with mean by same trick of taking percentile and then applying lambda function. In seats column also

there are outliers so remove that categories by using not in operator.

18) Now to plot a distribution to check columns are normally distributed or not and for that I used histplot and Q-Q plot. First power column which was not normally distributed so I had applied various transformation to make it normally distributed and at the end box_cox gives me proper result so I applied box_cox on that column.

19) Then I checked for year column and after that checked mileage and mileage column littile bit deviated at the tail and therefore I checked for various transformation and finally select sqrt transformation. I kept Engine column as it is.

20) Price column was right sqewed and therefore we directly apply log transformation on it. After that I goes with Kilometers_Driven in which I observed some outliers and firstly I removed it and then apply transformation. Box_cox transformation gives me result that I want.

- ❖ Splitting a Data

21) After doing EDA and visualization we will split the data and in x independent columns are there and in y dependent column which is price. Split the data into training and testing and test_size taken is 0.2.

❖Making a Pipelines

22) For pipelines import all related libraries. I used column transformer in which I gives One hot encoder and standard scaler. Firstly encoded that column which are categorical and after that apply standardization to that column which are numerical and remainder data keep as it is.

❖Model Evaluation

- Linear Regression

23) Make the object of linear regression and then passed the column transformer and linear regression to pipeline and do fit, predict. I got r2 score of 89 percent also calculate mse and mae which are 0.0738 and 0.1830 respectively and then plot scatterplot of true values vs predicted values.

- Support Vector Machine

24) Make the object of Support Vector Regressor and then passed the column transformer and Support Vector Regressor to pipeline and do fit, predict. I got r2 score of 90 percent also calculate mse and mae which are 0.066 and 0.172 respectively and then plot scatterplot of true values vs predicted values.

- Random Forest

25) Make the object of Random Forest Regressor and then passed the column transformer and Random Forest Regressor to pipeline and do fit, predict. I got r2 score of 92 percent also calculate mse and mae which are 0.0585 and 0.162 respectively and then plot scatterplot of true values vs predicted values.

- AdaBoostRegressor

26) Make the object of AdaBoost Regressor and then passed the column transformer and AdaBoost Regressor to pipeline and do fit, predict. I got r2 score of 84 percent also calculate mse and mae which are 0.109 and 0.255 respectively and then plot scatterplot of true values vs predicted values.

- KNN

27) Make the object of KNN and then passed the column transformer and KNN to pipeline and do fit, predict. I got r2 score of 90 percent also calculate mse and mae which are 0.065 and 0.176 respectively and then plot scatterplot of true values vs predicted values.

28) Make a pickel file to make a small web application. And for that I selected Random Forest.

Conclusion: This project showcases my proficiency in end-to-end machine learning development, from data preprocessing and feature engineering to model training, evaluation, and deployment. The Streamlit interface provides a practical demonstration of how the developed model can be utilized in a real-world scenario.