CAR PRICE PREDICTION

Submitted by:

ASHISH KUMAR SAMAL

ACKNOWLEDGMENT

It is a matter of great pleasure to express my profound feeling of reference to all the people who helped and supported me during the project. I would like to convey my sincere thanks to FLIPROBO TECHNOLOGIES AND DATATRAINED EDUCATION for constantly helping me with the valuable inputs during the project duration. Their inspiring guidance and everlasting enthusiasm have been valuable assets during the tenure of my project.

ASHISH KUMAR SAMAL

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.

- Review of Literature

There are two primary phases in the system: 1. Training phase: The system is trained by using the data in the data set and fits a model (line/curve) based on the algorithm chosen accordingly. 2. Testing phase: the system is provided with the inputs and is tested for its working. The accuracy is checked. And therefore, the data that is used to train the model or test it, has to be appropriate. The system is designed to detect and predict price of used car and hence appropriate algorithms must be used to do the two different tasks. Before the algorithms are selected for further use, different algorithms were compared for its accuracy. The well-suited one for the task was chosen.

- Motivation for the Problem Undertaken

To develop a efficient and effective model which predicts the price of a used car according to user's inputs. To achieve good accuracy. To develop a User Interface(UI) which is user-friendly and takes input from the user and predicts the price.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

	price	year	brand	model	kms	owner	Fuel	location
count	5.011000e+03	5011.000000	5011.000000	5011.000000	5011.000000	5011.000000	5011.000000	5011.000000
mean	5.764094e+05	2014.618639	12.964678	99.911794	54080.125723	1.268409	1.732788	3.286170
std	4.278102e+05	3.742143	5.816021	62.499020	42577.189396	0.531256	1.510126	2.532252
min	3.350000e+04	1998.000000	0.000000	0.000000	100.000000	1.000000	0.000000	0.000000
25%	3.250000e+05	2012.000000	8.000000	43.000000	26407.500000	1.000000	0.000000	1.000000
50%	4.826500e+05	2015.000000	15.000000	90.000000	51864.000000	1.000000	3.000000	3.000000
75%	6.849245e+05	2018.000000	15.000000	157.000000	74000.000000	1.000000	3.000000	6.000000
max	5.000000e+06	2021.000000	24.000000	201.000000	850002.000000	4.000000	7.000000	8.000000

- Data Sources and their formats

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.

- Data Preprocessing Done

Dataset has no null values.We had to convert the datatypes of object values into float and int values.Outliers are removed and scaling of data was done.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

-Data Cleansing
-Visualization and EDA
-Outlier Removal
-Standard Scaling
-Train Test Split
-Model Building and evaluation
-Cross Validation
-Hyperparameter tuning of best model
-Saving the best model

- Testing of Identified Approaches (Algorithms)

```
from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import AdaBoostRegressor
from sklearn.neighbors import KNeighborsRegressor
from sklearn.metrics import mean_squared_error,mean_absolute_error
from sklearn.metrics import r2_score,accuracy_score
from sklearn.model_selection import train_test_split
```

Run and Evaluate selected models

```
model=[DecisionTreeRegressor(),KNeighborsRegressor(),AdaBoostRegressor(),LinearRegression()]
max_r2_score=0
for r_state in range(40,90):
    x_train,x_test,y_train,y_test=train_test_split(x_std,y,random_state=r_state,test_size=0.33)
    for i in model:
        i.fit(x_train,y_train)
        pred=i.predict(x_test)
        r2_sc=r2_score(y_test,pred)
        if r2_sc>max_r2_score:
            max_r2_score=r2_sc
            final_state=r_state
            final_model=i

print("max r2 score correspond to random state",final_state,"is",max_r2_score,"and model is",final_model)
```

```
r2 score correspond to random state 76 is 0.7179849012515411 and model is DecisionTreeRegressor()
```

- Key Metrics for success in solving problem under consideration

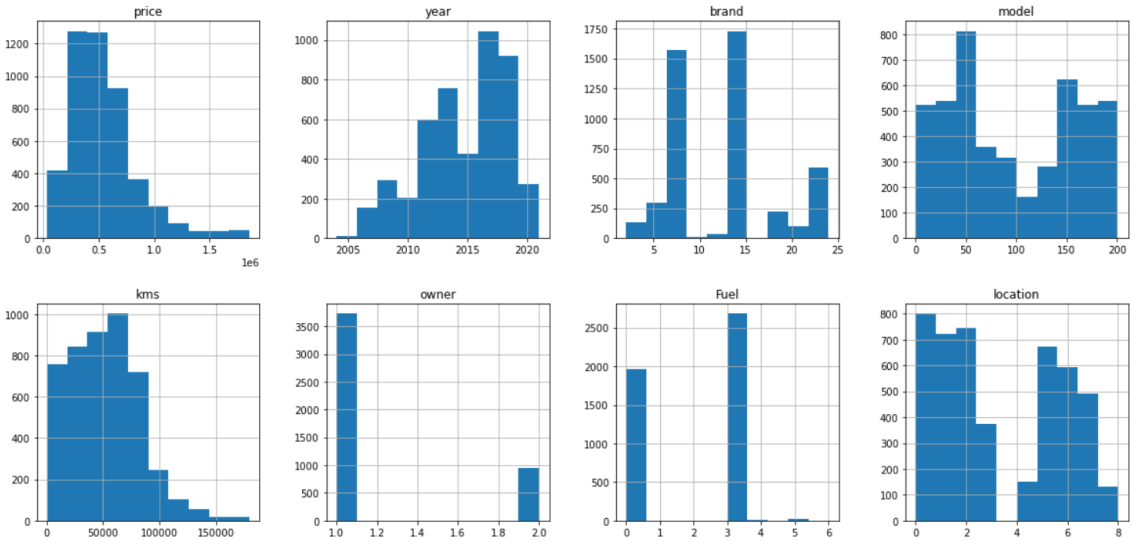
```
parameter={'criterion':["squared_error", "absolute_error"],
           'max_depth':range(1,9),
           'max_features':["auto", "sqrt", "log2"],
           'min_samples_leaf':(1,2),
           'splitter':['best','random']}
GCV=GridSearchCV(DecisionTreeRegressor(),parameter,cv=5)
GCV.fit(x_train,y_train)

GCV.best_params_
```

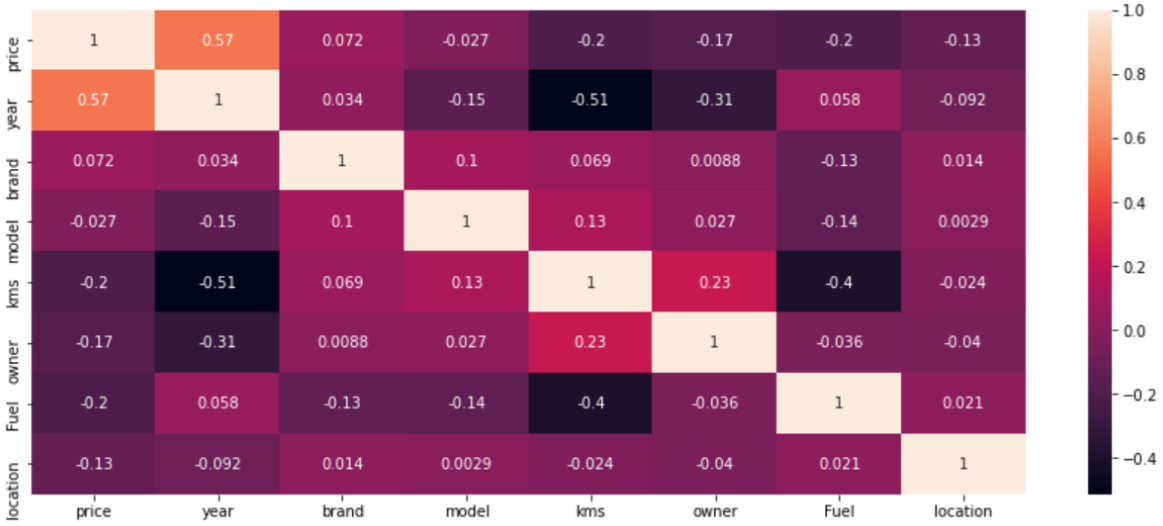
```
{'criterion': 'squared_error',
 'max_depth': 8,
 'max_features': 'auto',
 'min_samples_leaf': 2,
 'splitter': 'best'}
```

```
final_mod=DecisionTreeRegressor(criterion='squared_error',max_depth=8,
                                max_features='auto',min_samples_leaf=2,splitter='best')
final_mod.fit(x_train,y_train)
pred=final_mod.predict(x_test)
acc=r2_score(y_test,pred)
print(acc)
```

```
0.6702326124420339
```

- Visualizations





CONCLUSION

- ## Key Findings and Conclusions of the Study

  The increased prices of new cars and the financial incapability of the customers to buy them, Used Car sales are on a global increase. Therefore, there is an urgent need for a Used Car Price Prediction system which effectively determines the worthiness of the car using a variety of features. The proposed system will help to determine the accurate price of used car price prediction. This paper compares 4 different algorithms for machine learning : Linear Regression, AdaBoost Regression,K-Neighbors Regressor and decision Tree Regression.

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

  Price depends on no of kms and year of registration.
  Price of car depends on the number of past owners.

- ## Limitations of this work and Scope for Future Work

In future this machine learning model may bind with various website which can provide real time data for price prediction. Also we may add large historical data of car price which can help to improve accuracy of the machine learning

model. We can build an android app as user interface for interacting with user. For better performance, we plan to judiciously design deep learning network structures, use adaptive learning rates and train on clusters of data rather than the whole dataset.