CAR PRICE PREDICTION

Submitted by:

YASH JAISWAL

ACKNOWLEDGMENT

I would like to thank you Flip Robo Technologies for providing me with the opportunity to work on this project from which I have learned a lot

INTRODUCTION

- Business Problem Framing

car are one of the necessary need of each and every person around the globe and therefore CAR PRICE Focusing on changing trends in cars sales and purchases predictive modelling market mix modelling Using machine learning in order to predict the actual values of the prospective and decide whether to amount to the CAR PRICE.

- Conceptual Background of the Domain Problem

CAR PRICE Prediction model and problem patterns making sure last-minute purchases are cars Keeping the price as full as they want it raising prices on a rate of car price which is filling up in order to reduce rate of cars price and hold back inventory for those expensive lastminute expensive purchases This usually happens as an attempt to maximize revenue based on CAR PRICE Prediction

- Review of Literature

This is a comprehensive summary of the research done on the behalf You have to data set at least 5638 rows of data. You can data set more data as well, it’s up to you, More the data better the model In this section you have to Cars price of the data Car price Prediction of from different websites

- Motivation for the Problem Undertaken

Car price Prediction model problem are likely to rate of less car price of the data set purchases are loan Keeping the car price as full as they want it raising prices on a car price prediction which is filling up

in order to reduce car sale price and hold back inventory for those expensive last-minute expensive purchases loan of the rate of intrest

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

```
#Lets check the data type of dataset
df.dtypes
```

```
Unnamed: 0                int64
Brand Name               object
Car model                object
Car variant              object
Running kilometers       object
Number of owners         object
fuel System              object
manufacturing year        int64
Price                    object
dtype: object
```

so here we have

int64 type - Unnamed: 0 , manufacturing year columns

Object type - Than All columns are object columns

- Data Sources and their formats

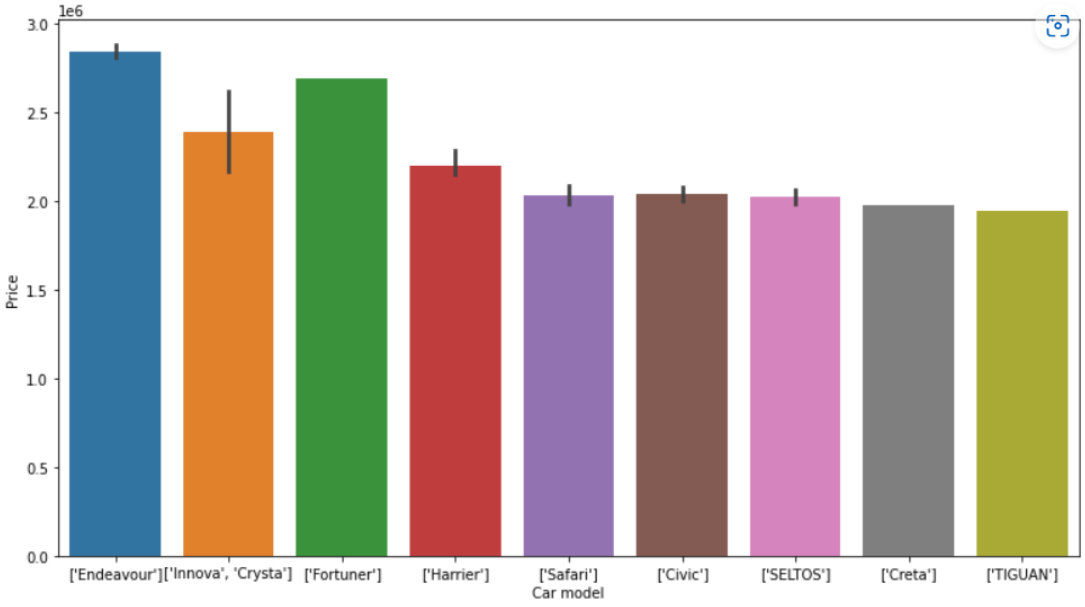
```
df.describe()
```

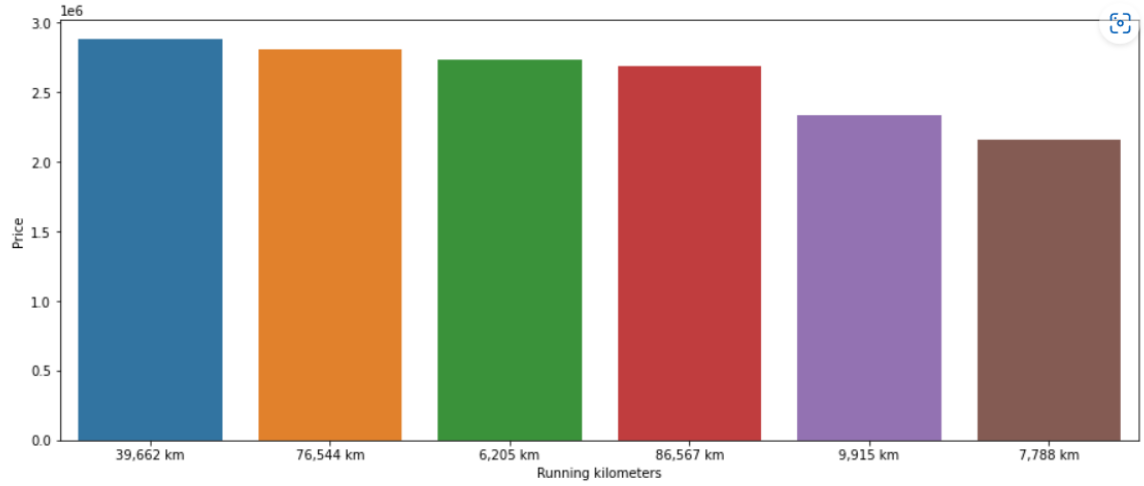
	manufacturing year	Price
count	5638.000000	5.638000e+03
mean	2017.600213	6.753235e+05
std	2.733470	3.366062e+05
min	2009.000000	1.480000e+05
25%	2016.000000	4.470000e+05
50%	2018.000000	5.930000e+05
75%	2020.000000	8.210000e+05
max	2022.000000	2.880000e+06

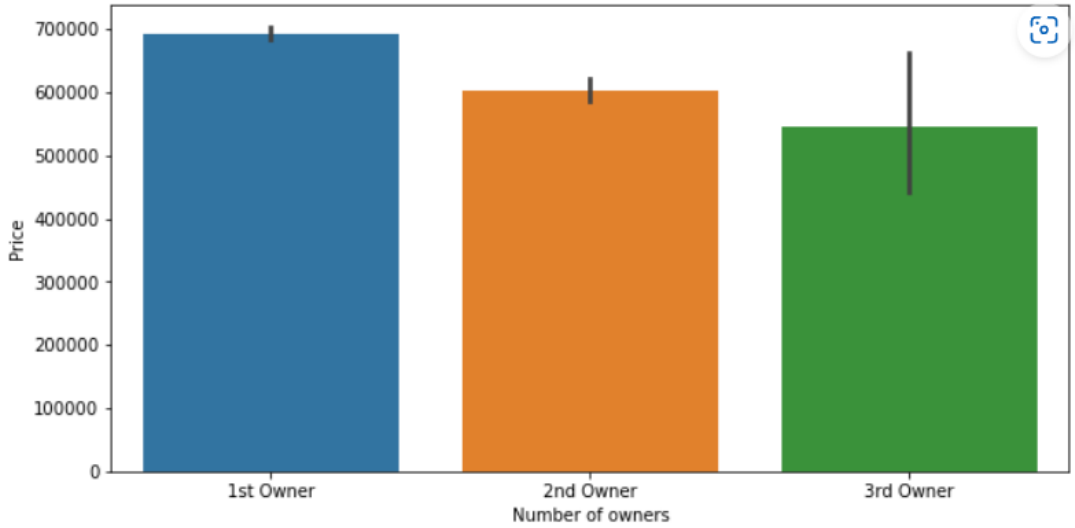
- Data Preprocessing Done

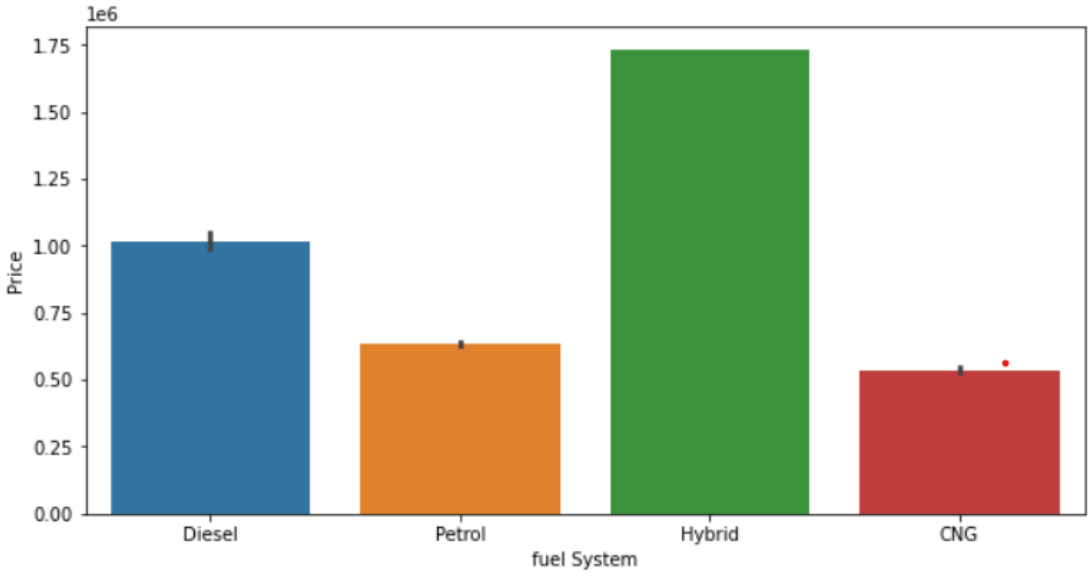
Loading the training data set as a dataframe , used pandas to set display I ensuring we do not see any truncated information , checked the number of rows and columns present in our training data set , checked for missing data and the number of rows with null values , verified the percentage of missing data in each columns are decide to dicard the once that value more than , dropped all the unwanted columns are duplicated data present in our data frame, separated categorical columns and numeric columns name in separate list variable for ease in visulazation , checked the unique values information in each column to get a gist for categorical data. Used pandas profiling during the visulazing phase along with pie plot count plot scatter plot and the other , with the help of label encoding technique converted all object data type columns to numeric data types.

- Data Inputs- Logic- Output Relationships



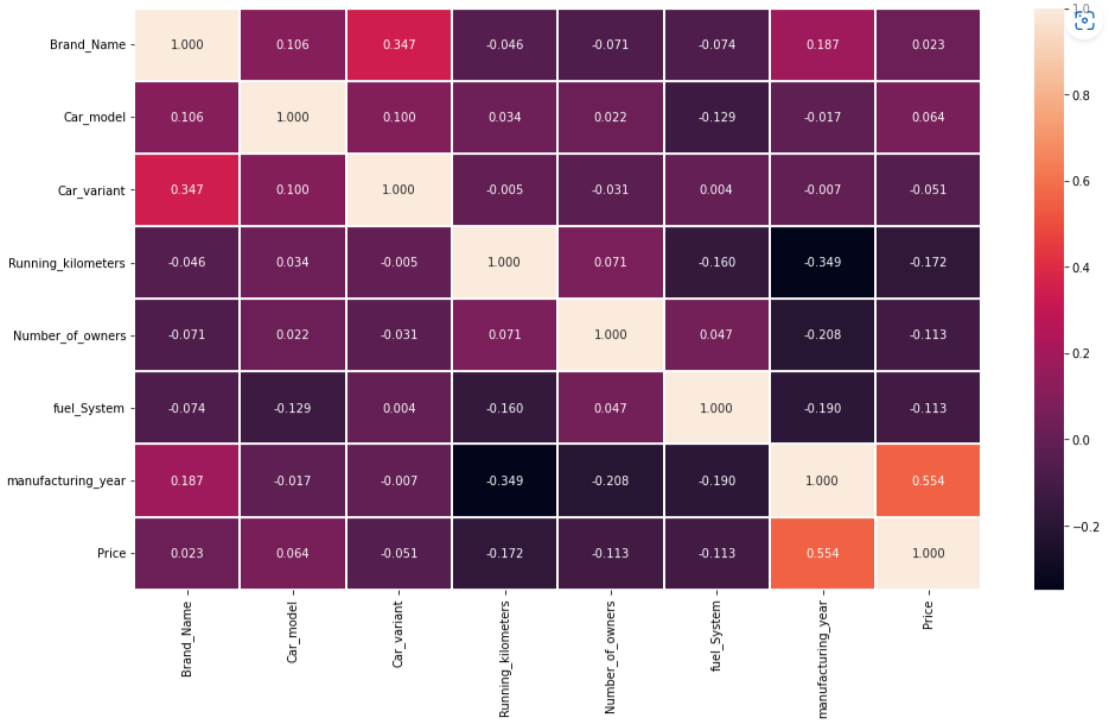

1e6
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Price
39,662 km
76,544 km
6,205 km
86,567 km
9,915 km
7,788 km
Running kilometers

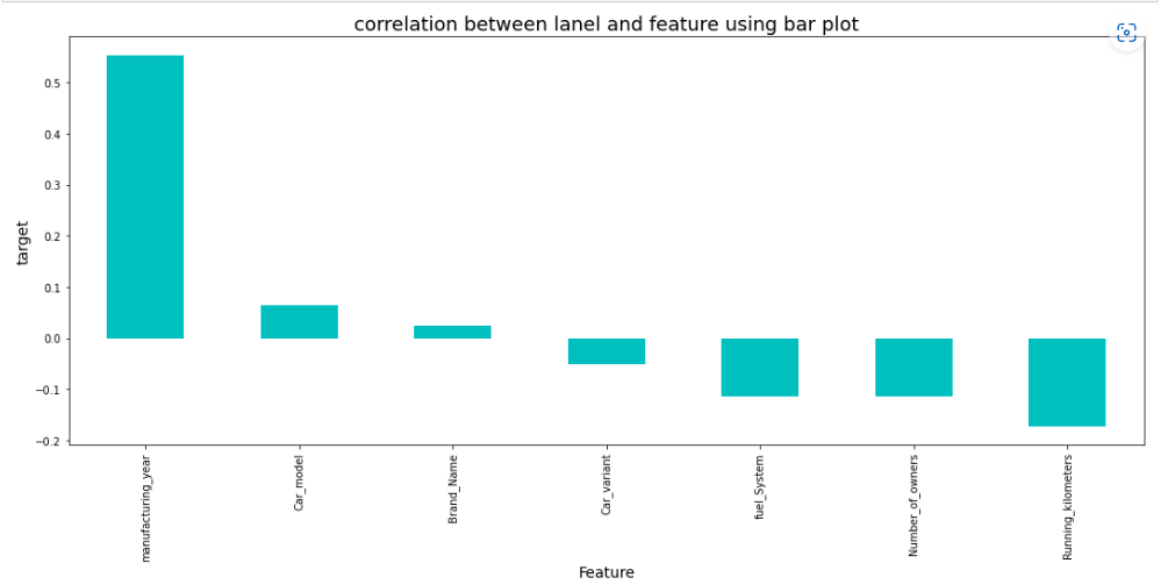

700000
600000
500000
400000
300000
200000
100000
0
Price
1st Owner
2nd Owner
3rd Owner
Number of owners


1e6
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Price
Diesel
Petrol
Hybrid
CNG
fuel System

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

	Brand_Name	Car_model	Car_variant	Running_kilometers	Number_of_owners	fuel_System	manufacturing_year	Price
Brand_Name	1.000000	0.106175	0.346606	-0.046180	-0.070661	-0.073750	0.187023	0.023357
Car_model	0.106175	1.000000	0.100085	0.033821	0.022206	-0.128968	-0.017095	0.064362
Car_variant	0.346606	0.100085	1.000000	-0.005185	-0.030502	0.003665	-0.007186	-0.050945
Running_kilometers	-0.046180	0.033821	-0.005185	1.000000	0.071287	-0.160023	-0.348569	-0.171950
Number_of_owners	-0.070661	0.022206	-0.030502	0.071287	1.000000	0.047103	-0.207799	-0.113367
fuel_System	-0.073750	-0.128968	0.003665	-0.160023	0.047103	1.000000	-0.190241	-0.113269
manufacturing_year	0.187023	-0.017095	-0.007186	-0.348569	-0.207799	-0.190241	1.000000	0.553882
Price	0.023357	0.064362	-0.050945	-0.171950	-0.113367	-0.113269	0.553882	1.000000





- Hardware and Software Requirements and Tools Used

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

import warnings
warnings.filterwarnings('ignore')
```

```
data1 = pd.read_csv("car_price_2 ")
data1
```

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

I have used both statistical and analytical approaches to solve the problem which mainly inculd the preprocessing of the data and EDA to check the correlation of independent and dependent features also before building the model I made sure that the point data was cleaned and claed before it was fed into machine learing models for this project we need to predict the loan label meaning our target columns is continue so this is a Logistic problem I have used various

regression algorthim I have sekect random forest regression as the best suitalble algorithm for our final models as it is giving a good and least different in and cv-score among all the algorithm used other regression algorithm are also giving accuracy but some are over fitting and some are under fitting the result which may be because of less label performance as well as accuracy and to check my model from overfitting and under fitting I have made use of the k fold and then hyper tuning the final model once I was able to get my desired final model I ensured to save that models i before i loaded the testing data and stared performance the data as training data set and obtaining the predicted label out of the regression machine learning model.

- Testing of Identified Approaches (Algorithms)

1)logistic regression

2)decision tree

3)xg boost

4)random forest

Here we select Random Forest Tree Regression for the model building.

```
parameter={'criterion':['mse','mae'],
           'max_features':['auto','sqrt','log2'],
           'max_depth':range(10,15),
           'min_samples_split':range(9,10),
           'min_samples_leaf':range(5,6)}
```

```
rf = RandomForestRegressor()
clf=GridSearchCV(rf,parameter)
clf.fit(x_train,y_train)
print(clf.best_params_)
```

```
{'criterion': 'mse', 'max_depth': 14, 'max_features': 'auto', 'min_samples_leaf': 5, 'min_samples_split': 9}
```

```
R2 Score 88.93042064431853
Cross Val Score 89.0572293722042
```

CONCLUSION

- Key Findings and Conclusions of the Study

Post models building and choosing the appropriate model I want ahead and scrape the data and join the dataset. After applying all the data pre processing steps as the dataset I was then able to get the predicted label result. Once the dataset with feature columns are predicted label was format I exported the value in a comma separated value file to be accessed as needed.

Conclusion

```
loaded_model=pickle.load(open('CAR_PRICE_PREDICTION','rb'))
result=loaded_model.score(x_test,y_test)
print(result*100)
```

```
88.93042064431853
```