Project Car Prediction

Submitted By:

Aman Saxena

ACKNOWLEDGMENT

- https://scikit-learn.org/stable/ - For the libraries used in the project.
- Rest project is done by myself only.

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. 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.

- Conceptual Background of the Domain Problem

For more understanding we can simply correlate it with the business in which we buy lots of cars at the lower price and then sale it with the large profit.

- Motivation for the Problem Undertaken

In this project we have to build the model that will predict the Car price of the Car using the independent variables so that the company can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

First of all we Scrape the dataset from the differenet websites then we see the shape of the dataset i.e 5075 rows and 7 columns and we

notice that there are columns cantaining the ‘-’ values. So we replace these values with the Nan values and later dealt those nan values with the mode of the columns as all the columns are of object type as theywere scraped from the website . After that we check for the outliers using the boxplot we found lots of outliers in the columns but those were the categorical columns in which the outliers were present and some outliers are the possible values so we did not remove the outliers from the dataset.

Then we convert the datatype of two columns i.e Price and Km driven so that we can predict the price of the car.

Since the categorical values were present so we have to encode them to make the prediction for that we have used ordinal encoder.

- Data Sources and their formats

We load the dtatset and check their datatypes so we observe that all the columns are of object datatype and then we converted the datatype of two columns so that we can predict the car price.

Data Preprocessing Done

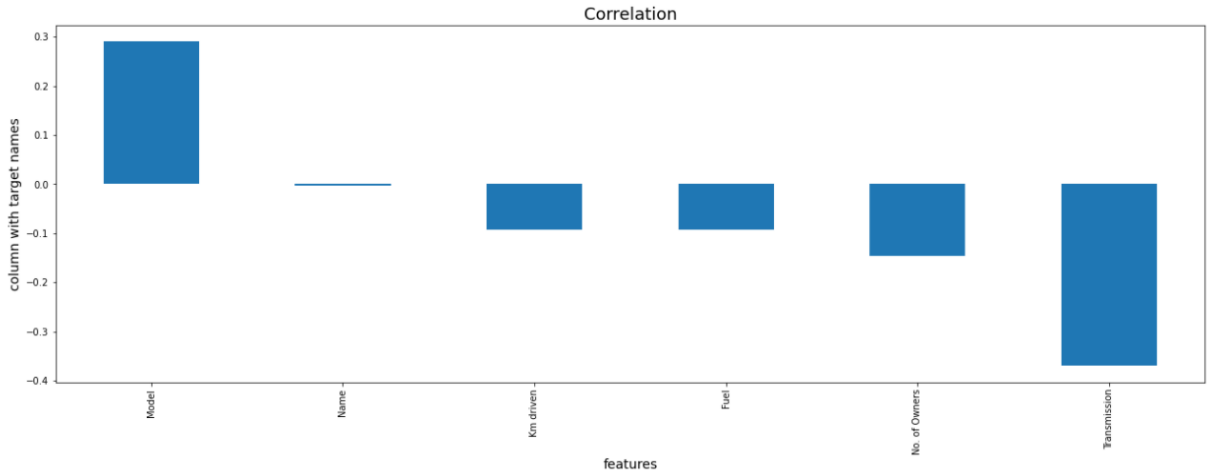
First of all we Scrape the dataset from the differenet websites then we see the shape of the dataset i.e 5075 rows and 7 columns and we notice that there are columns cantaining the ‘-’ values. So we replace these values with the Nan values and later dealt those nan values with the mode of the columns as all the columns are of object type as theywere scraped from the website . After that we check for the outliers using the boxplot we found lots of outliers in the columns but

those were the categorical columns in which the outliers were present and some outliers are the possible values so we did not remove the outliers from the dataset.

Then we convert the datatype of two columns i.e Price and Km driven so that we can predict the price of the car.

Since the categorical values were present so we have to encode them to make the prediction for that we have used ordinal encoder.

- Data Inputs- Logic- Output Relationships



In the above picture we can clearly see that what features are positively related with the target variable and what features are negatively related with the target variable.

- Hardware and Software Requirements and Tools Used

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

- from sklearn.linear_model import LinearRegression
- from sklearn.preprocessing import StandardScaler
- import warnings
- warnings.filterwarnings('ignore')
- from sklearn.preprocessing import OrdinalEncoder
- from sklearn.metrics import r2_score
- from sklearn.model_selection import train_test_split,GridSearchCV
- from sklearn.linear_model import Lasso,Ridge

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

Since the problem is of regression and we have to predict the continuous output so the approaches which we can use are Linear Regression, Lasso Regression, Ridge Regression, Random forest Regressor but the approaches which we have used in our project are Linear Regression, Lasso Regression, Ridge Regression.

- Testing of Identified Approaches (Algorithms)
  - Lasso Regression
  - Ridge Regression
  - Linear Regression

- Run and Evaluate selected models
  - Linear Regression

Linear Regression

```
LR.fit(x_train,y_train)
y_pred=LR.predict(x_test)
print(r2_score(y_test,y_pred))
```

```
0.3920226269041658
```

Cross Validation

```
from sklearn.model_selection import cross_val_score
cv_score=cross_val_score(LR,x_new,y,cv=5)
print(cv_score.mean())
```

```
0.19900522923299566
```

➢ Lasso Regression

```
]:  1 parameters={'alpha':[.0001,.001,.01,.1,1,10],'random_state':list(range(0,10)),'max_iter':[100,200,500,1000],'normalize':[Tru
    2 ls=Lasso()
    3 clf=GridSearchCV(ls,parameters)
    4 clf.fit(x_new,y)
    5 print(clf.best_params_)

{'alpha': 10, 'max_iter': 100, 'normalize': True, 'random_state': 0}

]:  1 ls=Lasso(alpha=10,random_state=0,max_iter=100,normalize=True)
    2 ls.fit(x_new,y)

]: Lasso(alpha=10, max_iter=100, normalize=True, random_state=0)

]:  1 ls_pred=ls.predict(x_test)
    2 lss=r2_score(y_test,ls_pred)
    3 lss

]: 0.39329133932272353
```

➢ Ridge Regression

Ridge

```
[145]:  1 parameters={'alpha':[.0001,.001,.01,.1,1,10],'random_state':list(range(0,10)),'max_iter':[100,200,500,1000],'normalize':[Tru
        2 Rg=Ridge()
        3 clfR=GridSearchCV(Rg,parameters)
        4 clfR.fit(x_new,y)
        5 print(clfR.best_params_)

{'alpha': 0.1, 'max_iter': 100, 'normalize': True, 'random_state': 0}

[146]:  1 Rgg=Ridge(alpha=0.1,random_state=0,max_iter=100,normalize=True)
        2 Rgg.fit(x_new,y)

it[146]: Ridge(alpha=0.1, max_iter=100, normalize=True, random_state=0)

[149]:  1 Rg_pred=Rgg.predict(x_test)
        2 Rgs=r2_score(y_test,Rg_pred)
        3 Rgs

it[149]: 0.26200961643311327
```

- Key Metrics for success in solving problem under consideration

  R2 Score → for calculating the accuracy
  Cross Val Score → For cross validation

CONCLUSION

- Key Findings and Conclusions of the Study

  First of all we Scrape the dataset from the differenet websites then we see the shape of the dataset i.e 5075 rows and 7 columns and we notice that there are columns cantaining the '-' values. So we replace these values with the Nan values and later dealt those nan values with the mode of the columns as all the columns are of object type as theywere scraped from the website . After that we check for the outliers using the boxplot we found lots of outliers in the columns but those were the categorical columns in which the outliers were present and some outliers are the possible values so we did not remove the outliers from the dataset.
  Then we convert the datatype of two columns i.e Price and Km driven so that we can predict the price of the car.
  Since the categorical values were present so we have to encode them to make the prediction for that we have used ordinal encoder.