FLIP ROBO

CAR PRICE PREDICTION

Submitted by:

ANEESHA B SOMAN

ACKNOWLEDGMENT

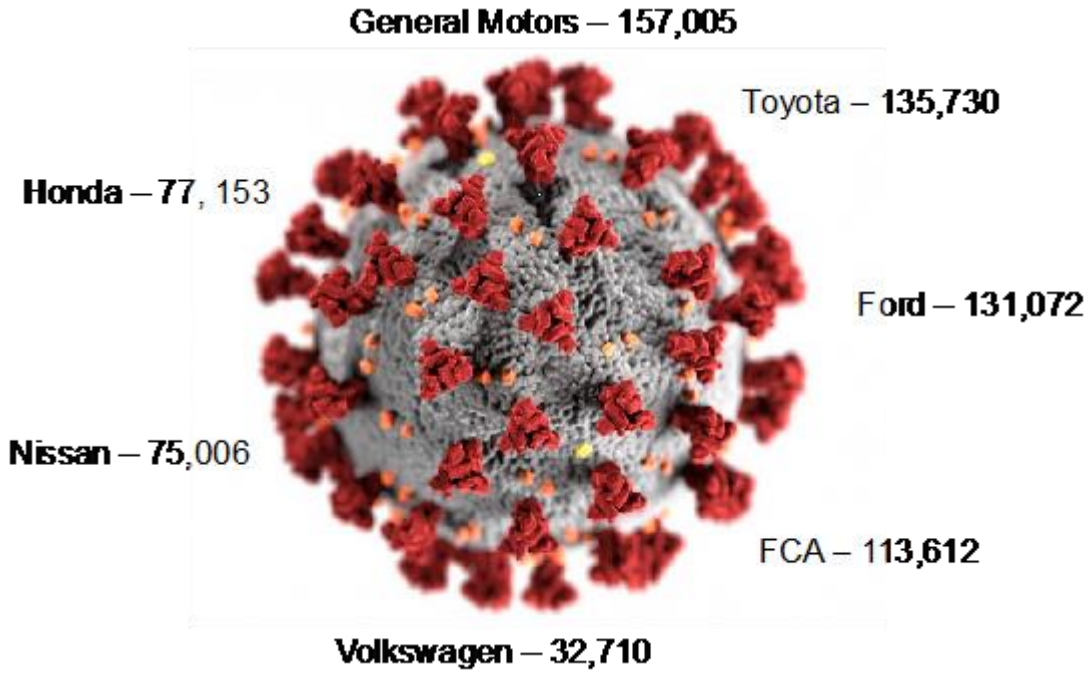
The success and final outcome of this project required a lot of guidance and assistance from Keshav Bansal Sir and I am Extremely fortunate to have got this all along the completion of my project work Whatever I have done is only due to such guidance and assistance and I would not forget to thank him. I respect and thank Keshav Bansal Sir, for giving me an opportunity to do the project work in Data mining,Data Modelling and Analytics and providing me all the support and guidance which made me complete the project on time . I am extremely grateful to him for providing such a nice support and guidance though he had busy schedule managing the company affairs.

I have also referred to various articles in Stackoverflow while cleaning the data, referred algorithms in sklearn and various pictures for the projects are obtained from google.

INTRODUCTION

- Business Problem Framing

  Due to Covid 19 Pandemic 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. I have made a car price valuation model.



- Conceptual Background of the Domain Problem

  There are various companies such as olx,cardekho,car24,etc which performs Car price prediction for used cars. These predictions are done through various data models. This data model allows the person who is in the buyer and the seller side to understand the current market value of the used car

- ## Review of Literature

  Data was collected from various websites such as olx,cardekho,car24. This data was cleaned and analysed. It showed the impact that various factors had on the price of the car.Then a model is created using the data by splitting the data as dependent and independent variable. These datas are further split into test and train. The train data is trained through various regression algorithms.

  The algorithm having the lease difference between r2 score and cross val score will be used for hyperparameter tuning. The best parameters are used to tune the model. This model is given to the client in further using to visualise data for car price prediction.

- ## Motivation for the Problem Undertaken

  Car has become a significant part of most of the households, specially where the public transport is not advanced. Hence Used car plays the pivotal role among cars as it expands the market of cars to a wider populations.

  Further Driving a car is important for people in general because it provides status and the opportunity for personal control and autonomy

  Also In sparsely populated areas, owning a car is even more important, since it provides the only opportunity for travelling long distances due to a lack of public transport

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

a.**Statistical models used:**

- Linear Regression
- DecisionTreeClassifier
- Random forest regressor
- Gradient Boosting Regressor
- Ada Boost Regressor
- PCA
- Standard Scalar

b.Analytical models:

Descriptive Analytics	**Diagnostic Analytics**	**Predictive Analytics**	**Prescriptive Analytics**
Data visualization done through between the Price and independent variables through **scatter plots**	The reason for change is understood through correlation techniques	Prediction is done through Gradient Boosting Regressor	Prescritive analysis is done through the Model created

- Data Sources and their formats

a.Data sources are

522 data is collected from OLX

5292 data from Cardekho

2289 data from cars24

b.format of data collected: Excel sheet

c.Data description

The features are Price,Brand,distance,year,model,owner and gas

there are outliers and null values

there are both categorical and numerical values

	Price	year	randMode	distance	owner	Gas	Brand	Model
2	4,11,699	2016	Maruti Sw	29,893 km	1st Owner	Petrol	Maruti	Swift
3	2,90,699	2014	Maruti Al	12,535 km	1st Owner	Petrol	Maruti	Alto
4	2,63,599	2017	Maruti Al	18,113 km	1st Owner	Petrol	Maruti	Alto
5	3,87,599	2019	Maruti Al	16,224 km	1st Owner	Petrol	Maruti	Alto K
6	1,94,999	2016	Tata Nan	20,039 km	1st Owner	Petrol	Tata	Nano
7	4,62,899	2018	Hyundai (	15,754 km	2nd Owne	Petrol	Hyundai	Grand i
8	2,68,599	2017	Maruti Al	23,227 km	1st Owner	Petrol	Maruti	Alto
9	5,82,999	2018	Ford Figo	5,915 km	1st Owner	Petrol	Ford	Figo Aspire
10	3,18,299	2011	Toyota Et	49,654 km	2nd Owne	Petrol	Toyota	Etios
11	2,95,999	2015	Maruti Al	26,183 km	1st Owner	Petrol	Maruti	Alto K
12	3,22,999	2017	Maruti Ee	21,722 km	1st Owner	Petrol	Maruti	Eeco
13	4,06,999	2014	Maruti Sw	13,509 km	2nd Owne	Petrol	Maruti	Swift
14	4,19,699	2014	Hyundai i	39,731 km	1st Owner	Petrol	Hyundai	i
15	3,50,799	2018	Maruti Al	5,198 km	1st Owner	Petrol	Maruti	Alto K
16	1,90,499	2013	Maruti Al	47,071 km	2nd Owne	Petrol	Maruti	Alto
17	3,50,699	2016	Maruti Al	15,471 km	1st Owner	Petrol	Maruti	Alto K

- Data Preprocessing Done
  a. removing ₹ from Price
  b. removing km from distance
  c. converting price and distance into integer format
  d. label encoding year column
  e. one hot encoding the rest of the columns and applying PCA
  f. Filling null values of Model with mode

- Data Inputs- Logic- Output Relationships

  Price is the dependent variable and year,distance,owner,Gas,Brand and Model are the dependent variable

  Highest relationship is between price and distance covered

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

  Assumed that the above stated features are only required for predicting the price of the car

- Hardware and Software Requirements and Tools Used

  Hardware:Windows 10

  Software:Jupyter notebook

  Libraries: seaborn, matplotlib, statsmodels, numpy, pandas, sklearn, statsmodels, scipy, pickle

Model/s Development and Evaluation

- **<u>Identification of possible problem-solving approaches (methods)</u>**

  EDA performed on the data.

  Performed one hot encoding on the categorical data and PCA was done to obtain optimum number of columns

  The label was numerical data hence the approaches that can be applied are: LinearRegression, DecisionTreeRegressor RandomForestRegressor, Gradient Boosting Regressor and Ada Boost Regressor

  The data is split and the best random state is found. Then the data is split again with the best random state

- Testing of Identified Approaches (Algorithms)
  a. Linear Regression
  b. DecisionTreeClassifier
  c. Random forest regressor
  d. Gradient Boosting Regressor
  e. Ada Boost Regressor

- Run and Evaluate selected models

R2 score and cross val score

		r2 score	cv score	difference
	DecisionTreeRegressor	75.28	61.65	13.63
	RandomForestRegressor	75.4	77.05	-1.65
	GradientBoostingRegressor	77.7	76.09	1.61
	AdaBoostRegressor	52.44	34.27	18.17

- Key Metrics for success in solving problem under consideration

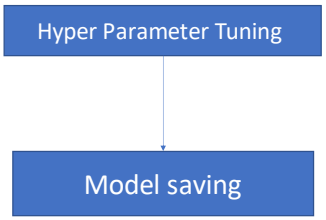
  The least difference between r2score and cv score was seen in gradient boosting regressor hence provinf least fitting problem. Hence it is considered for modelling the data under consideration(test data).

  The **r2 score** was used.
  R-squared is a metric of correlation. Correlation is measured by "r" and it **tells us how strongly two variables can be related**.

  A correlation closer to **+1 means a strong relationship in the positive direction, while -1 means a stronger relationship in the opposite direction**.

  A value closer to 0 means that there is not much of a relationship between the variables. R-squared is closely related to correlation

Hyperparameter tuning



Hyper parameter tuning

```
#GradientBoostingRegressor
parameters={'loss':['squared_error', 'ls', 'absolute_error'],
            'learning_rate':[1,2,3],
            'n_estimators':[1,2,3],
            'subsample':[1,2,3],
            'criterion':['friedman_mse', 'squared_error'],
            'min_samples_split':[1,2,3],
            'min_samples_leaf':[1,2,3]
}
```

```
GCV=GridSearchCV(GradientBoostingRegressor(),parameters,cv=4)
```

```
GCV.fit(x_train,y_train)
```

```
33]: GridSearchCV(cv=4, estimator=GradientBoostingRegressor(),
                   param_grid={'criterion': ['friedman_mse', 'squared_error'],
                               'learning_rate': [1, 2, 3],
                               'loss': ['squared_error', 'ls', 'absolute_error'],
                               'min_samples_leaf': [1, 2, 3],
                               'min_samples_split': [1, 2, 3],
                               'n_estimators': [1, 2, 3], 'subsample': [1, 2, 3]})
```

```
GCV.best_estimator_
```

```
4]: GradientBoostingRegressor(learning_rate=1, min_samples_split=3, n_estimators=3,
                              subsample=1)
```

```
mod=GradientBoostingRegressor(learning_rate=1,n_estimators=3,subsample=1,min_samples_split=3)
```

```
regression2=GradientBoostingRegressor()
regression2.fit(x_train,y_train)
```

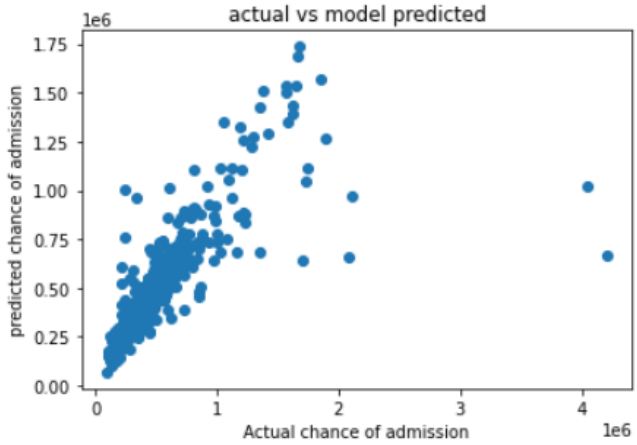
```
9]: GradientBoostingRegressor()
```

Saving of model

```
#saving GradientBoostingRegressor model
GradientBoostingRegressor_model2=GradientBoostingRegressor()
GradientBoostingRegressor_model2.fit(x_train,y_train)

filename='finalized_model.pickle'
pickle.dump(GradientBoostingRegressor_model2,open(filename,'wb'))

#Adjusted R2
GradientBoostingRegressor_model2.score(x_train,y_train)
```

```
t[41]: 0.9515383018681569
```

```
y_pred=regression2.predict(x_test)

plt.scatter(y_test,y_pred)
plt.xlabel('Actual chance of admission')
plt.ylabel('predicted chance of admission')
plt.title('actual vs model predicted')
plt.show()
```



Model scores:

Model evaluation ¶

```
#Adjusted R2
GradientBoostingRegressor_model2.score(x_train,y_train)
```

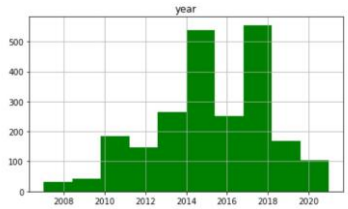
```
4]: 0.9515383018681569
```

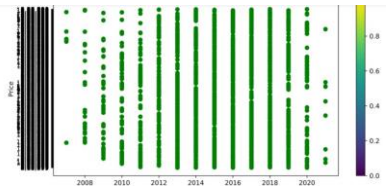
- Visualizations

visualizations

There are cars from 2007 to 2021

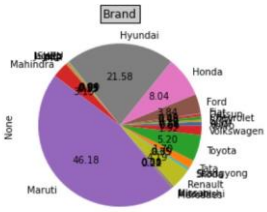
Highest number of cars are between 2013 and 2019

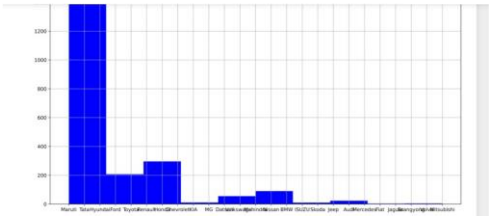




Brands

Highest number of cars are of maruthi brand and then hyundai

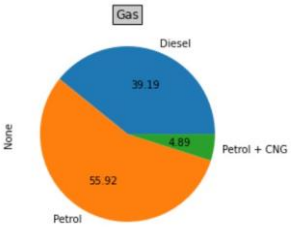


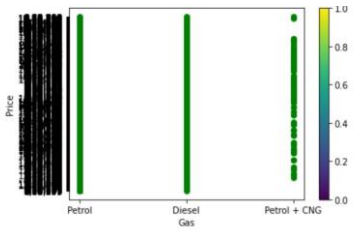


Gas

Highest number for cars are petrol driven

Price of car is lower for petrol+CNG

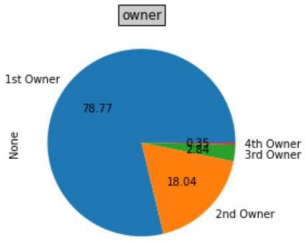


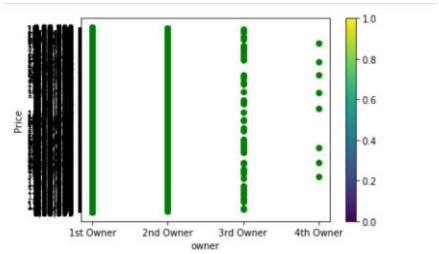


Owner

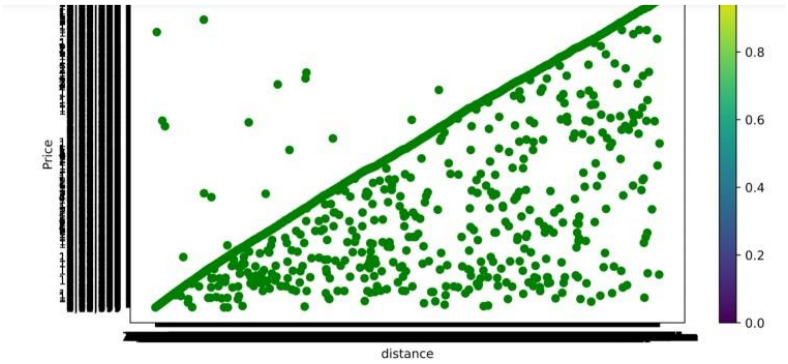
Maximum number for cars are of first owners

Maximum price is for 1st and 2nd owner cars





Distance



.

CONCLUSION

- Key Findings and Conclusions of the Study

  The project has build a model to predict the car prices of used cars.This can be used by sellers or buyers in the car selling platforms to analyse their car price

  The car price depends on factors like Distance travelled, ownership,gas,year of manufacture, brand and model

- Learning Outcomes of the Study in respect of Data Science

  The dependent variable was a numerical data hence regression algorithms were used. The algorithms used were Linear Regression ,DecisionTreeClassifier, Random forest regressor ,Gradient Boosting Regressor and Ada Boost Regressor

- Limitations of this work and Scope for Future Work

  Some of the limitations are that The number of features are less. Further the place is only for delhi. Various other parameters are needed in the dataset. Also since a large number of data was present the processing power of the computer was less to perform hyperparameter tuning with large amount of choices.

  These limitations can be imporved by mining more data for the project.