Car Price Prediction Project

Submitted by:

Ambika Saraf

ACKNOWLEDGMENT

I have taken efforts in this project however it would not have been completed without guidance from other people. I warmly acknowledge the invaluable supervision and an inspired guidance by our SME Mr. Keshav Bansal, FlipRobo Technology.

I would also like to express my sincere thanks to Data trained Education and FlipRobo Technology for giving me an opportunity to work on this project.

I also want to express my gratitude towards my friends and family who have patiently extended all sorts of help for accomplishing this.

I am grateful to one and all who are directly or indirectly involved in successful completion of this project.

INTRODUCTION

From necessities cars have turned into a luxury these days and with this change the sales in car market was really high before the pandemic hit the world. Covid-19 has impacted many businesses we have seen a lot of changes in car market too. 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.

The project consists of two phase :

1. **Data Collecting Phase**
2. **Model Building Phase**

In first phase we have to collect data of used cars from online websites. Here data is collected from "www.olx.in" and "www.cars24.com" website using Selenium technique for web scraping.

Next, we have built a regression model to predict prices of car. In the long term, this would allow people to better explain and review their purchase with each other in this increasingly digital world.

ANALYTICAL FRAMING

First thing to do was to collect data from different websites which is done using Selenium. We have tried to scrap data from all locations and include all types of cars in our data for example- SUV, Sedans, Coupe, minivan, Hatchback. Below is the data description of each variable:

Target Variable:

Price: Car Prices given on website

Features:

1. **CarName:** (Categorical) Name of Car
2. **Owner:** (Numeric Categorical) Number of owners
3. **Model:** (Categorical) Model Name
4. **Fuel:** (Categorical)Fuel used in car
5. **YearofPurchase:** (Continuous)car purchased in which year
6. **Transmission:** (Categorical) Automatic/ Manual
7. **Distance:** (Continuous) Distance covered by car
8. **Website:** (Categorical) Website from which data is scrapped
9. **Product_url:** url for each product

Hardware and Software Requirements and Tools Used

- Laptop with stable internet connection (Project done in jupyter notebook)
- scikit-learn
- matplotlib
- pandas
- numpy
- Google Chrome Web-driver
- Selenium

The project consists of Data Cleaning, Exploratory Data Analysis, Data Pre-processing, Model Building, Model Evaluation and selecting the best model.

Data Cleaning

```
In [3]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1800 entries, 0 to 1799
Data columns (total 11 columns):
Unnamed: 0        1800 non-null int64
CarName           1779 non-null object
Owner             1779 non-null object
YearofPurchase    1799 non-null float64
Transmission      1757 non-null object
Model             1741 non-null object
Distance          1779 non-null object
Fuel              1779 non-null object
Price             1779 non-null object
website           1800 non-null object
Product_url       1800 non-null object
dtypes: float64(1), int64(1), object(9)
memory usage: 154.8+ KB
```

Our data set consists of 1800 rows and 11 columns. We also observe missing data in features and thus we will be first dealing with null values.

- We have first removed rows having all important features as null values and then observe Transmission and Model columns having missing data and replaced it with highest weighted data.
- Next, we have removed words and "," from object data and convert them into int data type.
- Removed unnecessary features from dataset.

Statistical Description of data:

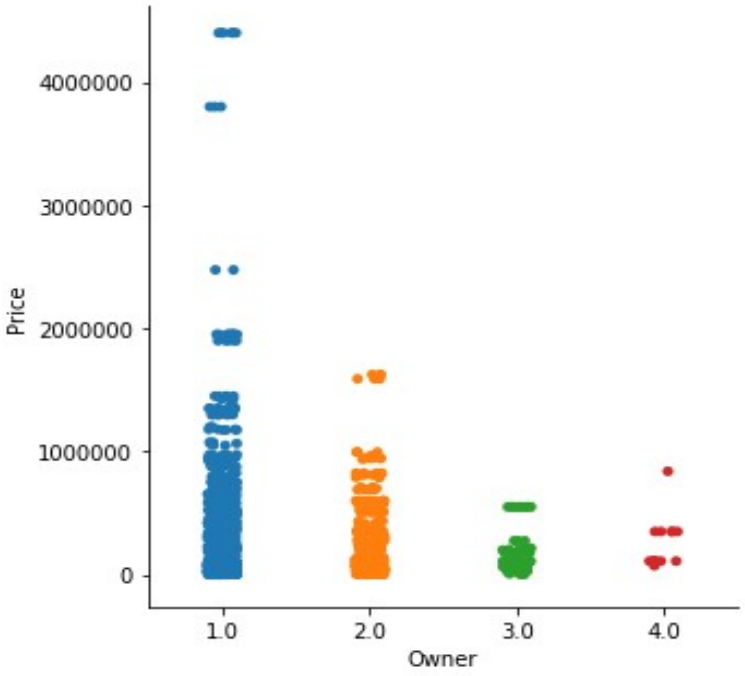
	index	Unnamed: 0	Owner	YearofPurchase	Distance	Price
count	1779.000000	1779.000000	1779.000000	1779.000000	1779.000000	1.779000e+03
mean	905.418212	905.418212	1.307476	2014.310287	59428.679595	2.776836e+05
std	519.330474	519.330474	0.581303	3.551701	65715.863432	4.480986e+05
min	0.000000	0.000000	1.000000	2003.000000	0.000000	0.000000e+00
25%	456.500000	456.500000	1.000000	2012.000000	33598.500000	4.189950e+04
50%	910.000000	910.000000	1.000000	2015.000000	54000.000000	8.839900e+04
75%	1354.500000	1354.500000	2.000000	2017.000000	74844.500000	3.750000e+05
max	1799.000000	1799.000000	4.000000	2021.000000	999999.000000	4.400000e+06

Conclusions made from EDA

By Univariate Analysis we observed the value counts and distribution of variable. By looking at visualizations we made sure how to treat data in an appropriate manner.

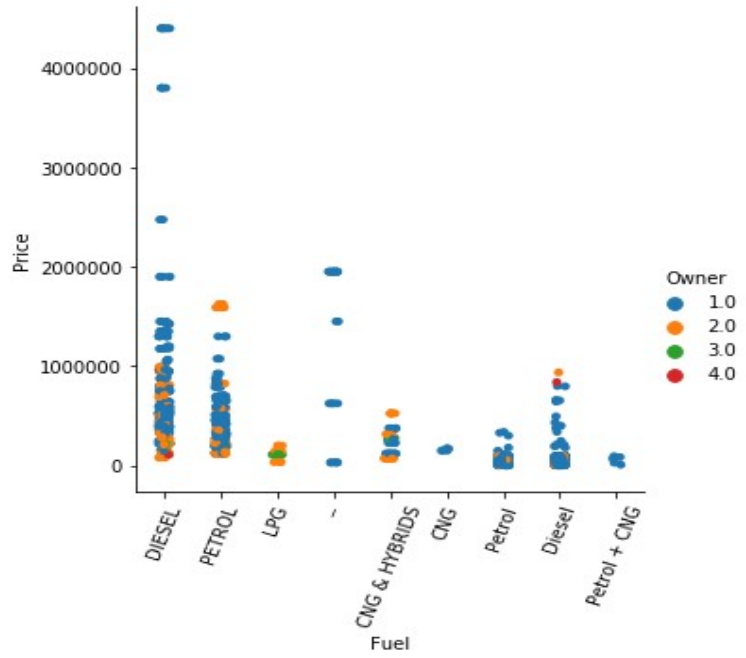
By Bivariate analysis we observed how each feature is affecting our target variable and also how more than one feature is effecting target variable. Below are the variables that show direct relationship with target variable.

1) Owner



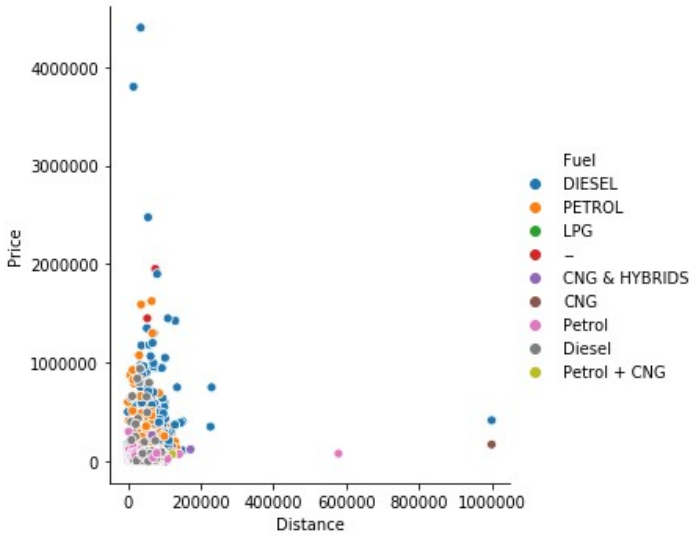
Less the number of owners greater the price of car

2) Fuel



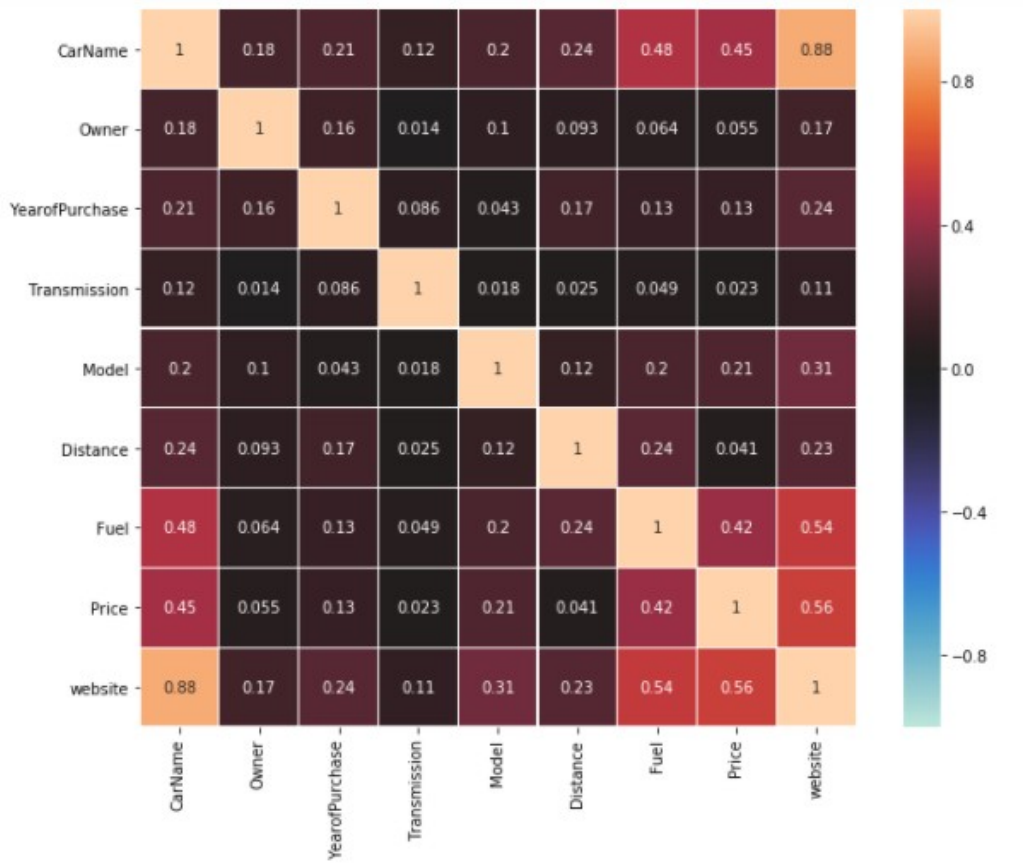
Diesel car have higher price as compared to those with other fuels.

3) Distance



Less the distance covered greater is the price also diesel cars are among those with high prices.

Next we have encoded all object type data to numerical data using Label Encoder. Ad moved further to check for correlation. We observe maximum correlation of target variable with “website” and minimum with “Transmission”. Below is the heatmap for correlation.



Data Pre-Processing

Since we had observed outliers in “Distance” column in this step we have used quantile method to remove outliers in data and also checked for skewness in data.

In this step we make our data ready for model and thus splitting and scaling of data is done using standard scalar.

MODEL BUILDING AND EVALUATION

First we found out the best random state for our linear regression model and then run each model on this random state.

Algorithms used are:

- Linear Regression
- Model Regularization using Lasso CV
- Decision Tree Regressor
- KNN Regressor
- Random Forest Regressor
- Gradient Boosting Regressor

Since range of target variable is too high we are getting a high value of mse therefore we will only look for R2 score and CV Score to determine our best model.

Linear regression:

Accuracy = 0.4012052494349163
Mean Absolute Error= 178701.69151004375
Mean Squared Error= 95296282075.04192

Lasso CV:

Accuracy= 0.4012274634080786
Mean Absolute Error= 178694.66568731022
Mean Squared Error= 95292746791.79088

KNN regression:

Accuracy = 0.5797470302191235
Mean Absolute Error= 121890.51494252874
Mean Squared Error= 66881924922.21848

Decision Tree regression:

Accuracy = 0.9574653116106683
Mean Absolute Error= 30138.885057471263
Mean Squared Error= 6769260516.894253

Random Forest regression:

Accuracy = 0.9570128070199103
Mean Absolute Error= 40700.53509578544

Mean Squared Error= 6841275184.825837

Gradient Boosting regression:

Accuracy = 0.9256661522224467
Mean Absolute Error= 64980.73303329423
Mean Squared Error= 11829995701.9463

Model Evaluation using CV Score:

	Model	R2_Score	CV Score	Difference
0	LinearRegression	0.401205	0.363671	0.037534
1	Lasso	0.401227	0.363673	0.037555
2	KNN	0.579747	0.520360	0.059387
3	DecsionTree	0.964691	0.849362	0.115328
4	RandomForest	0.961145	0.859417	0.101728
5	GradientBoosting	0.925546	0.880601	0.044945

And hence, we observe Gradient Boosting regression model as our best model and now we will move further to tune it using GridSearchCV.

Hyper Parametric Tuning

After parametric tuning of random forest model using Grid Search CV we obtain following results:

Gradient Boosting Regression: Accuracy = 0.9187764506451738

Mean Squared Error= 12926469818.162355

Root Mean Squared Error= 113694.6340781409

Mean Absolute Error= 71002.34225913782

In Last step we have compared original prices with the predicted prices and also saved our model to for future use.

CONCLUSIONS

KEY FINDINGS AND CONCLUSIONS OF THE STUDY

First, we collected the used cars data from different websites like olx, car24 and it was done by using Web scraping. The framework used for web scraping was Selenium. Then the scrapped data was saved in a csv file to use it for modeling purpose.

From the extensive EDA performed in this project we observed year of purchase, car name, fuel and owner directly influencing prices. We did data cleaning, data-preprocessing steps like finding and handling null values, removing words from numbers, converting object to int type, data visualization, handling outliers, etc. We have used multiple algorithms and chose gradient boosting model as our best model and thus tuned it using Grid Search CV.

The model build after hyper-parametric tuning gives an accuracy for 91.87%.

LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE

After the completion of this project, we got an insight of how to collect data, pre-processing the data, analyzing the data and building a model. It helped me to gain conclusions from graphs. Also it helped me in exploring multiple algorithms and metrics to get the best output.

LIMITATIONS OF THIS WORK AND SCOPE FOR FUTURE WORK

Since the data keeps changing we cannot fully rely on this project in the distant future we need to update it with updation in data. Also website was poorly designed because the scrapping took a lot of time and there were many issues in accessing to next page and thus less data was fetched.
This project is done with limited resources and can be made more efficient in future.