

**CAR PRICE PREDICTION**

**Submitted by:**

Pratiksha Avinash Potghan

**ACKNOWLEDGMENT**

I would like to convey my heartfelt gratitude to Flip Robo Technologies for providing me with this wonderful opportunity to work on a Machine Learning project “Car Price Prediction Model” and I also want to thank my SME “Mohd Kashif” for providing the dataset to complete this project.

This project would not have been accomplished without their help and insights.

Working on this project was a great experience.

**INTRODUCTION**

* **Business Problem Framing**

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. This project contains two phases:

a. Data Collection Phase

b. Model Building Phase

* **Conceptual Background of the Domain Problem**

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. So, one of our clients works with small traders, who sell used cars 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.

* **Review of Literature**

We have to made car price valuation model. This project contains two

phases:

1. Data Collection Phase:

We have scraped more than 6000 used cars data from websites: Olx and cardekho. We have fetched data for different locations. All types of cars are present in data for

example- SUV, Sedans, Coupe, minivan, Hatchback.

1. Model Building Phase:

After collecting the data, built a machine learning model. Before model building have done all data pre processing steps. Tried different models with different hyper parameters and selected the best model. Followed the complete life cycle of data science. Include all the steps like.

1. Data Cleaning

2. Exploratory Data Analysis

3. Data Pre-processing

4. Model Building

5. Model Evaluation

6. Selecting the best model

* **Motivation for the Problem Undertaken**

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.

**Analytical Problem Framing**

* **Mathematical/ Analytical Modeling of the Problem**

1) Scrapped Data from websites: Cardekho, OLX and Cars24

2) Used Panda’s Library to save data into csv file

3) Descriptive Statistics

4) Analysed correlation

5) Detected Outliers and removed

6) Detected Skewness and removed

7) Scaled data using Standard Scaler

8) Removed Multicollinearity

* **Data Sources and their formats**

Scraped Data from websites: Cardekho, OLX and Cars24and used Panda’s Library to save data into csv file: car\_price.csv. Target and Features variables of this dataset are:

*Target:*

• Car\_Price: Price of the used cars

*Features:*

• Brand

• Model

• Variant

• Manufacturing\_Year

• Driven\_KiloMeters

• Fuel

• Number\_of\_Owners

• Location

* **Data Preprocessing Done**

Checked Total Numbers of Rows and Column

Checked All Column Name

Checked Data Type of All Data

Checked for Null Values

There is null value in the dataset.

Information about Data

Checked total number of unique values

* **Data cleaning**

Column "Variant" contains missing value (127) which isvery less, less than 5%. So, we will keep this column and will drop those rows.

Checked all values of each column and replaced irrelevant value or data.

Converted Data Type of column "Car\_Price" as this column contains numeric values but it's datatype is object.

After conversion of Datatype, we seen Car\_Price column contains null values. So, dropped those rows only as total missing value is 6.

Checked again total no of rows and column

* **Data Visualization**

*Univariate Analysis*

➢ Used Countplot

*Bivariate Analysis*

(For comparison between each feature with target)

➢ Used Catplot and Scatterplot

*Multivariate Analysis*

(For comparison between all feature with target)

➢ Used Pairplot

* **Data Inputs- Logic- Output Relationships**
* Checking Correlation
* Checking skewness
* Checking skewness after outlier removal
* **State the set of assumptions (if any) related to the problem under consideration**
* By observing Target Variable “label” it is already assumed that it is a Regression Problem and to understand it have to use Regression model.
* Have to convert datatype of “Car\_Price” column.
* **Hardware and Software Requirements and Tools Used**

*Hardware used:*

Processor: 11th Gen Intel(R) Core(TM) i3-1125G4 @

2.00GHz 2.00 GHz

System Type: 64-bit OS

*Software used:*

Anaconda for 64-bit OS

Jupyter notebook

**Model/s Development and Evaluation**

* **Identification of possible problem-solving approaches (methods)**

In this project, we want to predict the micro-credit defaulter and for this we have used these approaches:

• Checked Total Numbers of Rows and Column

• Checked All Column Name

• Checked Data Type of All Data

• Checked for Null Values

• Checked for special character present in dataset or not

• Checked total number of unique values

• Dropped irrelevant Features

• Replaced duplicate values, special characters and irrelevant data

• Checked all features through visualization.

• Information about Data

• Checked correlation of features with target

• Detected Outliers and removed

• Checked skewness and removed

• Scaled data using Standard Scaler

• Checked Multicollinearity

• Used Feature Selection Method: Variance threshold method.

* **Testing of Identified Approaches (Algorithms)**

1. Linear Regression

2. Random Forest Regressor

3. KNN Regressor

4. Gradient Boosting Regressor

5. Decision Tree Regressor

R2 Score, Mean abs error, Mean squared error, Root Mean Squared Error, CV score are used for success.

* **Key Metrics for success in solving problem under consideration**

R2 Score, Mean abs error, Mean squared error, Root Mean Squared Error, CV score are used for success.

* **Visualizations**

**Univariate Analysis**

➢ Using Countplot

**Bivariate Analysis**

(For comparison between each feature with target)

➢ Using Catplot

➢ Using Scatterplot

**Multivariate Analysis**

(For comparison between all feature with target)

➢ Using Pairplot

* **Interpretation of the Results**

Through Visualization it is interpretated that Data is skewed due to presence of outliers in Dataset.Through Pre-processing it is interpretated that outliers & skewness was present in dataset, data was improper scaled, multicollinearity was present. By creating/building model we get best model: Gradient Boosting Regressor

**CONCLUSION**

* **Key Findings and Conclusions of the Study**

Here we have made a new car price valuation model as due to covid 19 impact previous car price valuation machine learning models is not working well because some cars are in demand hence making them costly and some are not in demand hence cheaper.

For new car price valuation model, we have done prediction on basis of Data using EDA, Data Cleaning, Data Visualization, Data Preprocessing, Checked Correlation, removed irrelevant features, Removed Outliers, Removed Skewness and at last train our data by splitting our data through train-test split process. Built our model using 5 models and finally selected best model which was giving best accuracy that is Gradient Boosting Regressor. Then tunned our model through Hyper Tunning using GridSearchCV. And at last compared our predicted and Actual Price of Car. Thus, our project is completed.

* **Learning Outcomes of the Study in respect of Data Science**

▪ This project has demonstrated the importance of scrapping data then converting that data of those data into csv and then using that csv file built a model to predict on data.

▪ Through different powerful tools of visualization, we were able to analyse and interpret different hidden insights about the data.

▪ Through data cleaning we were able to remove unnecessary columns and outliers from our dataset due to which our model would have suffered from overfitting or underfitting.

The few challenges while working on this project were: -

▪ Improper scaling: scaled it to a single scale using Standard Scaler

▪ Too many features: 10 features were present in the dataset, after data cleaning 2 features were reduced due to no relation with target variable and last after removing multicollinearity, we were able to reduce 1 more feature.

▪ Converted datatype of target column

▪ Replaced irrelevant values or data from features

▪ Skewed data due to outliers: Removed using power transformer

'yeo-johnson' method and outliers was removed through zscore.

* **Limitations of this work and Scope for Future Work**

While we couldn’t reach out goal of 100% accuracy in detecting defaulter but created a system that made data get very close to that goal. This project allows multiple algorithms to be integrated together to combine modules and their results to increase the accuracy of the final result. This model can further be improved with the addition of more algorithms into it. However, the output of these algorithms needs to be in the same format as the others which will make modules easy to add as done in the code.