Car Price Prediction

Submitted by:

ASHISH KUMAR DEWANGAN

INTRODUCTION

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. 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 phaseData Collection Phase

Objective

You have to scrape at least 5000 used cars data. You can scrape more data as well, it's up to you. more the data better the model. In this section You need to scrape the data of used cars from websites (Olx, cardekho, Cars24 etc.) You need web scraping for this. You have to fetch data for different locations. The number of columns for data doesn't have limit, it's up to you and your creativity. Generally, these columns are Brand, model, variant, manufacturing year, driven kilometers, fuel, number of owners, location and at last target variable Price of the car. This data is to give you a hint about important variables in used car model. You can make changes to it, you can add or you can remove some columns, it completely depends on the website from which you are fetching the data.

Try to include all types of cars in your data for example- SUV, Sedans, Coupe, minivan, Hatchback.

Exploratory Data Analysis

- Data Pre-processing and Visualizations

The data was collected from 3 different sources Cars24, Olx and CarDekho. The amount of data from each of these websites is as follows:

Source	Number of Records Extracted	Number of Features
Cars24	4163	12
Olx	486	15
CarDekho	2570	17
Total	7219	11

The features which have been selected for price prediction are:

```
['Brand', 'Model', 'Variant', 'Make_Year', 'Fuel', 'Km_driven',
       'Transmission', 'Number_of_Owners', 'Location', 'Price', 'Source
']
```

The inference from the analysis of missing values are as follows:

1. Variant: It 1247 missing values out of 7219 records. We have dropped the missing values.
2. Transmission: It has 219 missing value out of 7219. We have dropped the missing values

Univariate Analysis

1. We have a feature which contains the manufactured year of cars. Using this info we have engineered another feature 'Age'.
2. There are huge number of second hand cars of Maruti Suzuki, followed by Hyundai and Honda.
3. The price of the cars from the brand which is abundantly available lies in the same range.
4. Chevrolet is available at a cheap price.
5. The price of Honda and Tata are a little bit higher.
6. The price of the cars which are less than a year old are highest. At the same time the cars which are 2 or 3 years old are sold at a higher price than a car which has age 1 to 2 years.
7. Most of the cars available for sale are from 2015 to 2018.
8. There are very less number of cars which run on CNG or LPG.

Data Pre-Processing

First of all the name of the car is strip from its variant from the column model. Some of the values in the Km_driven feature and Price feature were in different formats. It was handled to avoid any ambiguity at a later stage. The features were encoded using Binary Encoder.

After data pre-processing labels and features were separated, the features were split into train and test dataset in the ratio of 3:1.

The model which were tested in this data are-

1. Logistic Regression
2. Lasso
3. Ridge
4. KNN Regressor
5. Decision Tree Regressor
6. Random Forest Regressor
7. Ada Boost Regressor

8. Gradient Boost Regressor
9. XG Boost Regressor

The parameters of the model were tuned were ever applicable to improve model accuracy and precision. The parameters precision, recall and F1 score were also compared.

The model performance is tabulated below for reference:

S. No.	Model Name	MSE
1.	Linear Regression	156314
2.	Lasso	156308
3.	Ridge	156239
4.	KNN	220941
5.	Decision Tree	163920
6.	Random Forest	117846
7.	Ada Boost	167328
8.	Gradient Boost	138487
9.	XG Boost	121599

The accuracy scores of model after hyper parameter tuning are:

S. No.	Model Name	Score
1.	Decision Tree	0.56
2.	XG Boost	0.712
3.	Random Forest	0.700
4.	Gradient Boost	0.641

The performance of XG Boost Regressor is better than any other model. Hence it is selected as the final model for prediction of property price.

The prediction was performed on the test data using the final model.

Conclusion

We cleaned the data, performed EDA and successfully trained a model and tuned the hyper parameters to predict the price of properties based on the features and dataset available with us. The XG Boost Regressor proved to be the best model in prediction due to its high

- **Score – 0.712**
- **MAE** – 64535

- **RMSE** – 118563

Hyper Parameters –
```
{'min_child_weight': 9,
 'max_depth': 20,
 'learning_rate': 0.2,
 'gamma': 0.3,
 'colsample_bytree': 0.5}
```

The MAE and MSME are high may be because the model assumes that all the cars have fair condition. If we can get the condition scores of the cars, there is a chance that the performance will further improve.