

# *CAR PRICE PREDICTION PROJECT*

**SUBMITTED BY:**

**RITU RANI**

# BUSINESS 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. 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 HAS 2 PHASES**

**First,** is to collect data from website. we have collected used data from Cars24 website in one python file.

### **second phase is:**

In this section we will upload our collected data and build a machine learning model to predict the Sale price of used Cars.

we have collected data of used cars from Cars24 website from 5 different popular cities in one Python file. After that I have uploaded that data in 2nd Python file to predict the price of used cars.

We will build a machine learning model to predict the Sale price of used Cars.

**PHASE 1:**

First, I have collected dataset from Cars24 website among popular cities i.e. New Delhi, Pune, Bangalore, Chennai, Kolkata. I have scrapped the details of 3543 cars details using Selenium library. I will use this dataset to build a machine learning model which will predict the prices of used cars.

Click on the below link to find my juypter notebook in GitHub for web-scrapping:

<https://github.com/riturani2403/car_price_predictions/blob/main/USED_CAR_SCRAPPING.ipynb>

**PHASE 2:**

In this notebook, I'll build a model to predict the value of used cars. Machine Learning techniques are very useful in predicting outcomes for large amount of data. In this paper six machine learning algorithms, Linear Regression (LR), Support Vector Regressor (SVR), Decision Tree Regressor, K Nearest Neighbour (KNN), AdaBoost Regressor and Random Forest regressor (RF) are applied to predict the prices of used cars. The experimental results conclude that the accuracy of Random Forest Classifier machine learning algorithm is better as compared other machine learning approaches.

Our task is to predict the price fo used cars by using various regression algorithms. Exploratory data analysis is done on the dataset to achieve insights and the pre-processing pipeline is done to get the data ready for the training. 80% of the data is used for training purpose and 20% for the testing purposes. Six regression models are trained and their performances are compared with various performance metrics like r2 score, cross validation score.

**LOADING DATASET**

The data set is of shape (3543, 9) i.e. It has 9 attributes and 3543 rows.

The dataset provides 8 input variables and 1 target variable that are a mixture of ordinal, categorical and numerical data types. Following are the variables is our dataset:

We have three kinds of data types:

1. **car\_year** : Car manufacturing year.
2. **car\_name** : Name of car
3. **car\_model** : model of that particular car
4. **zero\_payment per month** : Instalments per month
5. **distance covered in km** : Car covered distance
6. **petrol/diesel** : whether Car is of petrol, diesel, cng, electric.
7. **owner**: how many owners
8. **City** : current city
9. **car\_price\_INR** : current valuation of used car ( Our Target Variable)

**Object:** It means variables are categorical.

Following are the Categorical variables in our dataset: car\_name, car\_model, petrol/diesel, city.

**int64:** It represents the integer variables.

Following are the integer variables in our dataset: car-year, zero\_payment per month, distance covered, car\_price\_INR.

**Exploratory Data Analysis**

Now, we will do exploratory data analysis to get the insight about the data and how target variable depends on various attributes.

First, we are analyzing our target variable i.e. “car\_price\_INR”.

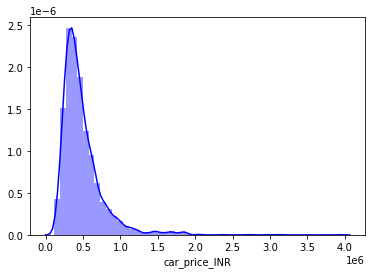


FIGURE. 1

This figure shows the analysis of our target variable.

We can visualize that car prices are left skewed.

**UNIVARIATE ANALYSIS OF INDEPENDENT VARIABLES:**

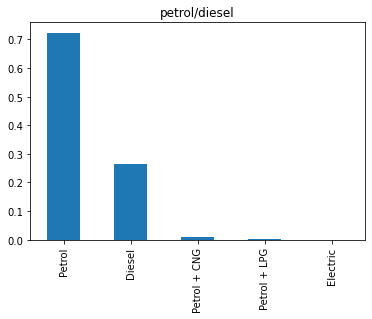
***PETROL/DIESEL:***  


FIGURE. 2

* We can visualize that from our dataset 70% of the cars are of petrol.
* Around 26% are diesel
* Rest are from CNG, LPG and Electric.

***CITY:***

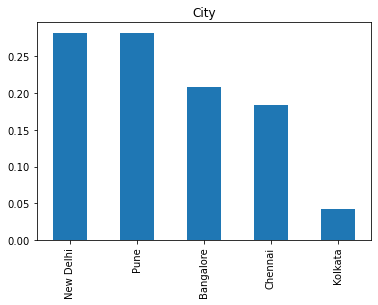


FIGURE. 3

* We have taken data mostly from New Delhi and Pune

***CAR YEAR***

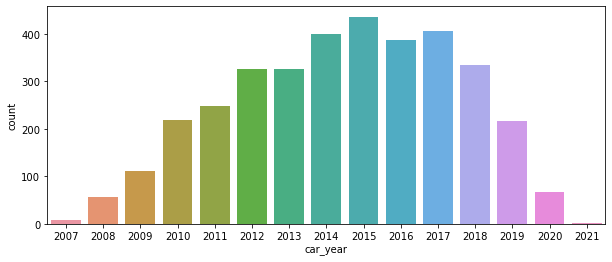
******

FIGURE. 4

Large no of cars is from 2015.

**DATA PRE-PROCESSING:**

**Data pre-processing** is very essential step in any **data** mining process. It directly impacts the predictions of the model. If data is unclean, have missing vales, missing attributes or contains outliers, if skewness is present, then all these factors degrade the quality of our results and our predictions will be biased.

First, we will check for missing or NaN values through heat-map:

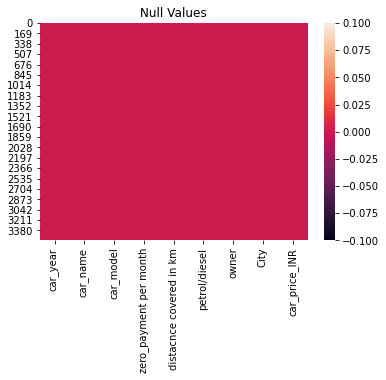


Figure.5

We can see that there is no NaN or missing values present in our dataset.

**Correlation matrix heatmap:**

Checking correlations is very important to analyse data. A heatmap has been plotted to check the correlation between the attributes, if there is positive or negative relationship. This is one of the methods to decide which attributes affect the target variable the most.



FIGURE. 6

OBSERVATIONS:

* zero payment per month is positively correlated with price of used cars.

**OUTLIERS:**

An outlier means an observation that falls outside the overall pattern or we can say an abnormal distance from other values in a random sample from a population.

We have outliers present in all attributes. From scipy.stats we imported Z-score and drop all the rows in which threshold value is greater than 3. By dropping these rows we lost our 6% percent data and the results came from our predictions will be unbiased. Therefore, we will drop these outlier values so that our predictions will be accurate.

**TREATING SKEWNESS:**

If the skewness is between -0.5 and 0.5 then the data is fairly symmetrical and represent normal distribution. If the skewness is between 0.5 and 1 or -1 and -0.5 then the data is moderately skewed. If the skewness is less than -1 or greater than 1then the data is highly skewed.

As we earlier analysed that skewness is present in all the attributes. Therefore, we will treat this with power transformation.

**SPLITTING DATASET:**

The shape of the dataset after dropping of the irrelevant columns is (3297,9). We split the dataset where 80% is used for training the model and 20% for testing the model. Hence out of 3297 data entries, 2637 are used for training and 660 are used for testing the model.

**FINDING BEST RANDOM STATE**

Our model best performs at random state 1 and we are achieving 99.9% accuracy score.

Six Regression Algorithms are used.

1. Linear Regression
2. RandomForestRegressor
3. SupportVectorRegressor
4. K Nearest Neighbour
5. Decision Tree Regressor
6. AdaBoostRegressor

We are achieving more than 95% accuracy score with each Algorithm.

We will proceed with Random Forest Regressor as it is giving 99.9% accuracy score

Therefore, we proceed with Random Forest Classifier as it is giving highest accuracy score.

**Cross Validation Score:**

Imported cross validation score to check the over-fitting and under fitting in our predictions.

**from** **sklearn.model\_selection** **import** cross\_val\_score

#### **We proceed with RandomForestClassifier as it is giving highest accuracy\_score and there is minimum difference between accuracy\_score & cross\_validation\_score**

We get best accuracy\_score from RandomForestClassifier.

**TUNNING WITH BEST PARAMETERS:**

Imported RandomizedSearchCV from sklearn.model\_selection and find out the best parameters of RandomForestClassifier which performed best on our model.

Following are the best parameters for our model:

{'criterion': 'mae', 'max\_features': 'auto'}

R2 score is 99.9% of our model.

**SAVE THE MODEL**

We finally save our best model by importing pickle. The use of pickle is widespread as they allow us to easily transfer data from one server or system to another and then store it in a file or database.

**LOAD THE MODEL:**

Loading model through pickle and checking accuracy.

**CONCLUSION**

We successfully predicted the prices of used cars with 100% accuracy score by using various regression algorithms. Exploratory data analysis is done on the dataset to achieve insights and the pre-processing pipeline is done to get the data ready for the training.80% of the data is used for training purpose and 20% for the testing purposes. Six regression models are trained and their performances are compared with various performance metrics like r2 accuracy score, cross validation score. The RandomForestClassifier comes out to be the best performing algorithm above all other models with an accuracy of 100% and over all performing best.

**DOWNLOAD JUPTYER NOTEBOOK**

Click on the below link to find my juypter notebook in GitHub:

<https://github.com/riturani2403/car_price_predictions/blob/main/USED_CAR_PricePredicions_Machine%20learning.ipynb>