A Project Report

On

**“Used Car Price Prediction”**

**TECHNEX’21**

**Submitted To:**

Mr. Mayur Dev Sewak

General Manager, Operations

Eisystems Services

Ms. Mallika Srivastava

Trainer, Data Science & Analytics Domain

Eisystems Services

**Submitted By**:

Rushik Patel

**Index**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Contents** | **Page No.** |
| **1.** | **List of Figures** | **3** |
| **2.** | **Abstract** | **4** |
| **3.** | **Summary** | **4** |
| **4.** | **Objectives** | **4** |
| **5.** | **System Requirements** | **5** |
| **6.** | **Data Flow Diagram/ Algorithm** | **5** |
| **7.** | **Code/Program** | **11** |
| **8.** | **Reference** | **14** |

1. **List of Figures**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Name of Figures** | **Page No.** |
| **1.** | **Flowchart of Project** | **6** |
| **2.** | **Head of Dataset** | **7** |
| **3.** | **Tail of Dataset** | **7** |
| **4.** | **Car Price Histogram** | **8** |
| **5.** | **Fuel Type vs Price (Box Plot)** | **8** |
| **6.** | **Owner Type vs Price (Box Plot)** | **9** |
| **7.** | **Company vs Count (Bar Graph)** | **10** |
| **8.** | **Dataset** | **11** |
| **9.** | **Removed Null Values** | **12** |
| **10.** | **Converting Features into Float Datatype** | **12** |
| **11.** | **Categorical Data to Features** | **13** |
| **12.** | **Predicted Output and Accuracy** | **14** |

1. **Abstract**

In this report, we look at how supervised machine learning technique can be used to forecast used car prices in India. Data from online website kaggle was used to make the predictions. The prediction was made using multiple linear regression analysis. The predictions are then analysed and compared to determine how much accuracy does the algorithm provides. A seemingly simple problem turned out to be difficult, as dataset has to be thoroughly clean for better accuracy. In future, we can use more advance algorithm to predict car prices with increase in accuracy.

1. **Summary**

The Price of new car is fixed by manufacturer with additional taxes of state as well as central government. Therefore, customer knows exact amount of money for buying new car. But, due to increase prices of new car and financial incapability of customer to buy them, second hand car market has been growing in popularity. These results in providing opportunity for both buyers and sellers. Buying the used car is best option for customer because the price is fair and affordable.

However, many factors affect the price of used car, including its age, current condition, mileage, engine efficiency and many more. In most of cases, the price of used car on the market fluctuates. As a result, model for evaluating car price is needed.

To overcome this, I developed a model based on multiple linear regression that can predict car price. Predicting used car’s resale value is not an easy job as it depends on number of other variables. The most significant are car’s age, model, engine, power (bhp) and mileage(kmpl). As this model produces continuous value as an output and not categorical, giving all these variable as input model predicts actual price of car.

1. **Objectives**

Deciding whether a used car is worth the posted price when listed online can be difficult. Several factors like mileage, year, model, etc. can influence car price. From seller’s perspective, it is also a dilemma to price a used car appropriately. Based on existing data, the goal is to use machine learning algorithm for developing model to predict used car price. In addition, to consider each effective factors while predicting car price. Therefore, to develop an efficient and effective model which predicts the price of a used car according to user’s inputs with good accuracy.

1. **System Requirements**

Hardware Requirements

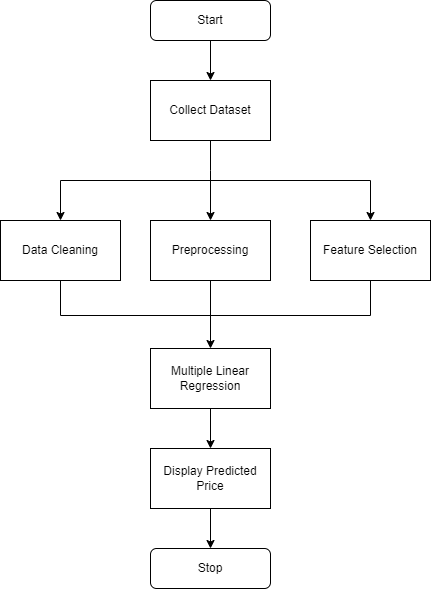
* Operating System - Windows 7, 8, 10
* Processor - dual core 2.4 GHz (i3, i5 series intel processor or equivalent other processor)
* RAM – 4 GB

Software Requirement

* Python
* Jupyter Notebook (recommended) or Pycharm
* Chrome (web browser)
* PIP 2.7

1. **Data Flow Diagram/ Algorithm**

The project starts with collecting dataset. The next step is to do Data cleaning, Data Preprocessing and feature selection. Data Preprocessing includes Data reduction, Data Transformation. Then, using machine learning algorithm we will predict the price. The algorithm used is multiple Linear Regression. Then model predicts the most accurate price. After that predicted price is displayed to the user according to user’s inputs.

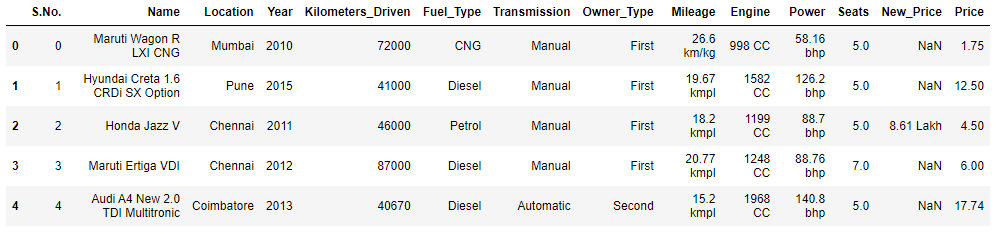


**Fig 5.1 Flowchart of Project**

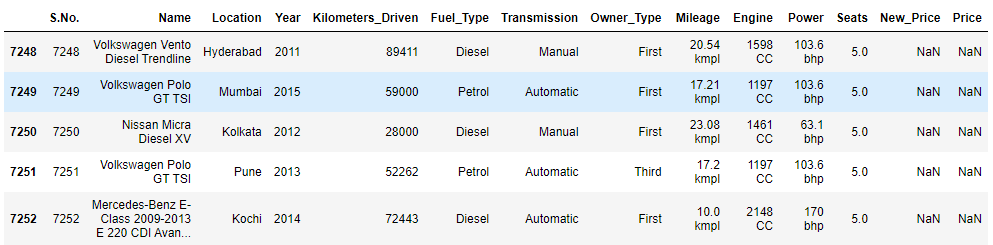
**6.1 Dataset**

For this project, I have used dataset on used car sales available on kaggle. The features available in this dataset are Name, Location, Year, Kilometers\_Driven, Fuel\_Type, Transmission, Owner\_Type, Mileage (kmpl), Engine (CC), Power (Bhp), Seats, New\_Price and Price.

A Short glimpse of first five row and last five rows of dataset is shown below



**Fig 6.1.1 Head of Dataset**



**Fig 6.1.2 Tail of Dataset**

The link for the dataset is <https://www.kaggle.com/yogidsba/predict-used-car-prices-linearregression/data.>

**6.2 Data Cleaning, Preprocessing and Feature Selection**

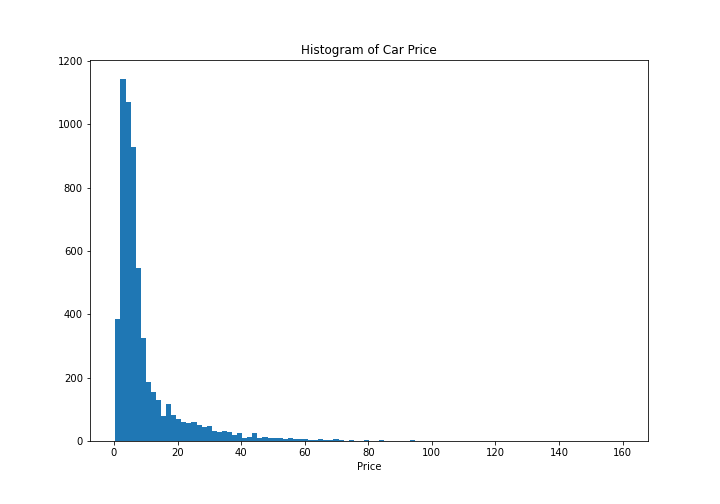
**Data Cleaning**

In order to get better understanding of data, I first dropped all the data having null values. Then I converted some features into float values for processing. For example Mileage, Engine and Power where trailed by kmpl, CC and Bhp respectively that makes them object datatype. After converting them into float values they are ready for further processing.

**Data Preprocessing**

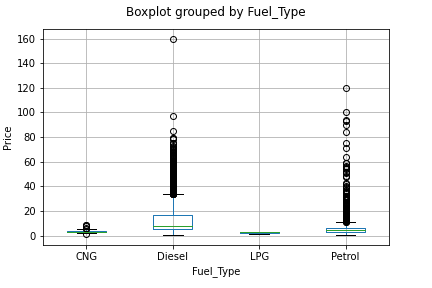
After cleaning data and making, them ready to visualize. Data visualization is the best way to find out how data looks like. I plotted various graph to get overview of our target column that is price vs various other column.

1. First I created histogram of car prices. This figure shows how car prices are distributed over the data. For example, majority of data is distributed between 0-20.



**Fig 6.2.1 Car Price Histogram**

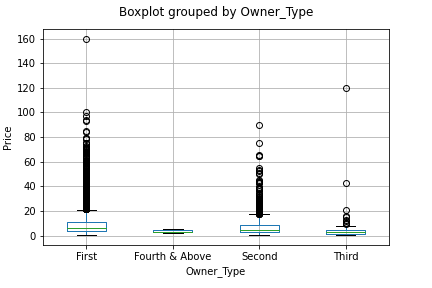
1. Then I plotted box plot of Fuel type vs Price to get insight on which type of vehicle is priced more and which is priced less.



**Fig 6.2.2 Fuel\_Type vs Price**

Above is the box plotted figure of Fuel type vs Price. As, we thought diesel car would cost followed Petrol and then CNG and LPG.

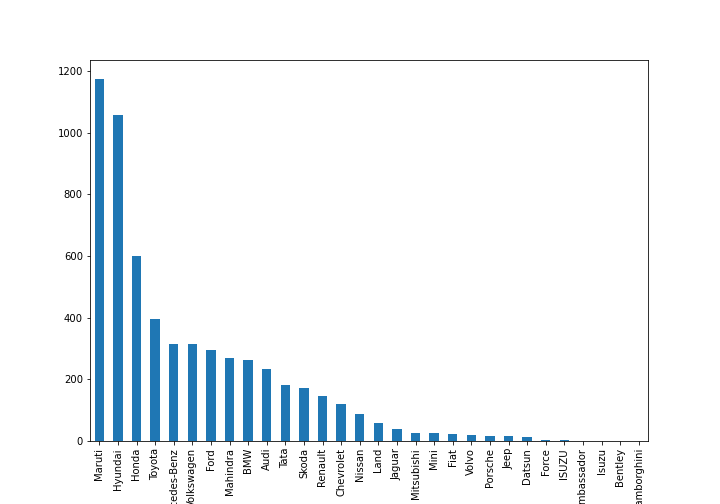
1. After Fuel\_Type vs Price boxplot I plotted box plot of Owner type vs Price. This helped me to gain knowledge about which car owner is willing to pay more for the car.



**Fig 6.2.3 Owner\_Type vs Price**

As we see in figure below first hand owner is willing to pay more followed by second hand and then third and fourth and above.

1. At last, I plotted bar graph of company vs its products count. This graph shows which car company is common and is more in dataset compared to other car company. For example, Maruti is most common brand followed by Hyundai.



**Fig 6.2.4 Company vs Count**

**Feature Selection**

Now it is more important to select appropriate feature from the dataset that would be input to the model which predicts output. We convert some categorical features into categories such as location, Fuel\_Type, Transmission which is used as an input into the model.

**6.3 Multiple Linear Regression**

Linear Regression attempts to model the relationship between two variables by fitting a linear equation to observed data. It performs the task to predict a dependent variable value (y) based on a given independent variable(x).

Similarly, multiple linear regression is a statistical technique that is used to analyse the relationship between a single dependent variable and several independent variable. The objective of multiple regression analysis is to use the independent variables whose values are known to predict the value of the single dependent variable. Each predictor value is weighted, the weights denoting their relative contribution to the overall prediction.

C:\Users\Rushik\Desktop\Capture.PNG

Here Y is the dependent variable, and X1, …. ,Xn are n dependent variables. In calculating the weights, a, b1,…,bn, regression analysis ensures maximal prediction of the dependent variable from the set of independent variables. This is usually done by least squares estimation.

**6.4 Display predict price**

After that, independent variables are given as an input to model. Model built with multiple linear regression predicts output as a dependent variable i.e. in our case it is car price.

1. **Code/ Program**

**# Importing necessary libraries and Dataset**

import numpy as np

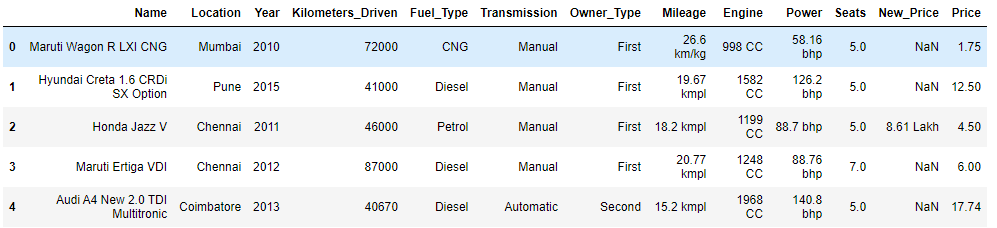
import pandas as pd

import matplotlib.pyplot as plt

**# Importing dataset**

df = pd.read\_csv('used\_cars\_data.csv')

df.head()



**Fig 7.1 Data Set**

**# finding null values, droping them and reseting index.**

print("Shape of df Before dropping any Row: ",df.shape)

df = df[df['Mileage'].notna()]

print("Shape of df After dropping Rows with NULL values in Mileage: ",df.shape)

df = df[df['Engine'].notna()]

print("Shape of df After dropping Rows with NULL values in Engine : ",df.shape)

df = df[df['Power'].notna()]

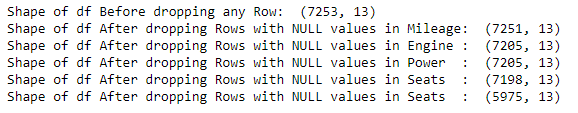
print("Shape of df After dropping Rows with NULL values in Power : ",df.shape)

df = df[df['Seats'].notna()]

print("Shape of df After dropping Rows with NULL values in Seats : ",df.shape)

df = df[df['Price'].notna()]

print("Shape of df After dropping Rows with NULL values in Seats : ",df.shape)



**Fig 7.2 Removed Null Values**

**# getting company name from name column and converting 'Mileage', 'Engine' and 'Power' column into float datatype.**

for i in range(df.shape[0]):

df.at[i, 'Company'] = df['Name'][i].split()[0]

df.at[i, 'Mileage(kmpl)'] = df['Mileage'][i].split()[0]

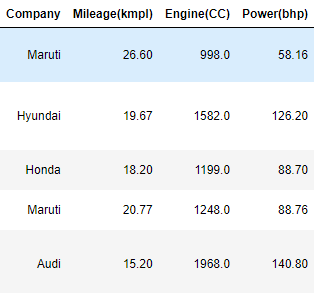
df.at[i, 'Engine(CC)'] = df['Engine'][i].split()[0]

df.at[i, 'Power(bhp)'] = df['Power'][i].split()[0]

df['Mileage(kmpl)'] = df['Mileage(kmpl)'].astype(float)

df['Engine(CC)'] = df['Engine(CC)'].astype(float)

df['Power(bhp)'] = df['Power(bhp)'].astype(float)



**Fig 7.3 Converting Features into float datatype**

**# As for now there are five categorical features.**

**#1.Location**

**#2.Fuel\_Type**

**#3.Transmission**

**#4.Owner\_Type**

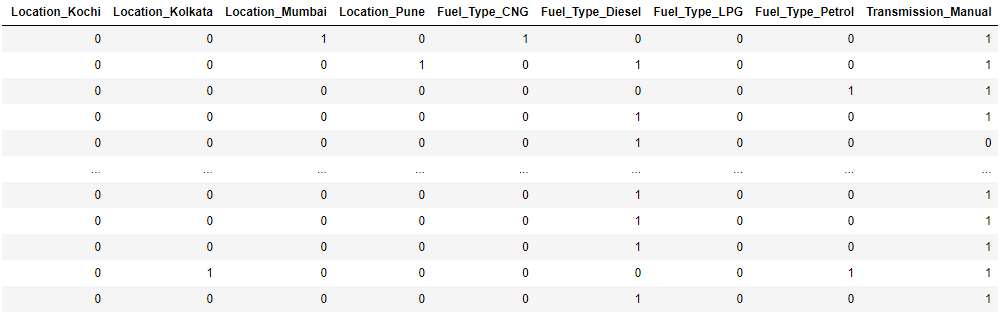
**#5.Company**

**# Dividing these each features into categories and generating new columns.**

df=pd.get\_dummies(df, columns=['Location', 'Fuel\_Type'], drop\_first=False)

df=pd.get\_dummies(df, columns=['Transmission'], drop\_first=True)

df.replace({"First":1,"Second":2,"Third": 3,"Fourth & Above":4},inplace=True)



**Fig 7.4 Categorical Data to Features**

**# Feature Selection**

**# Selecting final features that will be used for model building and droping all other useless feature.**

df.drop(["Company"],axis=1,inplace=True)

df.drop(['New\_car\_Price'],axis=1, inplace=True)

y=df['Price']

df.drop(['Price'],axis=1, inplace=True)

x=df

**# model building**

**# Building model using sklearn library.**

**# First splitting data to train and test into 80:20 ratio for the model.**

from sklearn.model\_selection import train\_test\_split

x\_train, x\_test, y\_train, y\_test = train\_test\_split(x, y, test\_size = 0.2)

**# Applying Linear Regression algorithm using sklearn library.**

from sklearn.linear\_model import LinearRegression

multi\_model = LinearRegression()

multi\_model.fit(x\_train, y\_train)

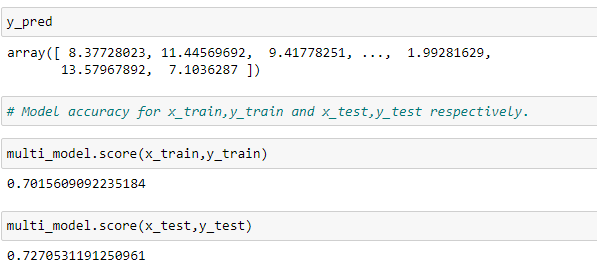
**# Predicting y\_test by giving x\_test as an input to the model**

y\_pred=multi\_model.predict(x\_test)

**# Model accuracy for x\_train,y\_train and x\_test,y\_test respectively.**

multi\_model.score(x\_train,y\_train)

multi\_model.score(x\_test,y\_test)



**Fig 7.5 Predicted Output and Accuracy**

(For full code/program please refer to following link:

<https://github.com/Rushik2900/Used-Car-Price-Prediction/blob/main/Car_Price_Prediction_Linear_Reg.ipynb>)

1. **References**

[1] <https://www.kaggle.com/yogidsba/predict-used-car-prices-linearregression/data>

[2] “Used Car Prediction” by Praful Rane, Deep Pandya, Dhawal Kotak (IRJET 2021)

[3] “Prediction of Car Price using Linear Regression” by Ravi Shastri (IJTSRD 2021)

[4] <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html> Scikit-learn - machine learning in python

[5] <https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.html> Matplotlib - Pyplot in python