A Project Report on

“**Car Price Prediction Using ML”**

Submitted in partial fulfillment of the requirements for

the award of the degree of

### Master of Science in Data Science and Big Data Analytics

in

### Data Science & Big Data Analytics

By

**Dhananjay S. Malusare**

4722655

Under the Guidance of

**Esmita Gupta**

****

#### Department of Information Technology

B. K. Birla College of Arts, Science and Commerce(Autonomous), Kalyan

B. K. Birla College Road, Near RTO, Kalyan

UNIVERSITY OFMUMBAI

#### Academic Year2021-2022

**Acknowledgement**

This Project Report entitled ***“Car Price Prediction using Ml”*** Submitted by ***“Dhananjay S. Malusare”(*** ***4722655)***is approved for the partial fulfillment of the requirement for the award of the degree of ***Master of Science*** in ***Data Sacience*** from ***University of Mumbai***.

Dhananjay S. Malusare.

Prof. Esmita Gupta

Head, Department of Information Technology

Place: B. K. Birla College (Autonomous), Kalyan

Date:

### CERTIFICATE

This is to certify that the project en titled ***“Car Price Prediction using ML”*** submitted by***“Dhananjay Malusare” (4722655)*** for the partial fulfillment of the requirement for award of a degree ***Master of Science*** in ***Data Science***, to the University of Mumbai,is a bonafide work carried out during academic year2021-2022.

Dhananjay S. Malusare

Prof.Esmita Gupt Dr.AvinashPatil

Head Department of IT Principal

ExternalExaminer(s)

1.

2.

**Place: B. K. Birla College (Autonomous), Kalyan**

Date:

**Declaration**

**I declare that this written submission represents my ideas in my own words, and where ideas or words of others have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the Institute and can also evoke penal action from the sources that have thus not been properly cited, or from whom proper permission has not been taken when needed.**

**--------------------------------**

**(signature)**

**Date: ----------------------------------------**

(Name of Student and Student ID)

**Abstract**

A fair car value prediction has made it so easy for the buyers to get a car home, as it just requires few efforts and brains of field experts. Day to day, there are many brands that bring new models in the market with lavish prices. Customers not being capable to buy a new car financially due to the higher market price, there is a need of used car value prediction globally which effectively determines the worthiness of a car that can be bought without much thinking. To train a model for predicting the price of used cars we applied machine learning techniques i.e., Regression Algorithms because it provides us continuous value as an output and not a categorized value such as Random Forest, linear regression and other algorithms for getting better accuracy. Then after processing on the data of dataset collected from Kaggle, we will be comparing the performance of different algorithms to get a chosen output. Further it would be available in GUI as a Web-application developed using Python-flask making it user friendly so that users could give input and get the price of a car according to it

**Index**

|  |  |  |
| --- | --- | --- |
| S.No. | Title | P.no |
| 1. | Introduction | 7 |
| **2.** | PROBLEM STATEMENT AND OBJECTIVE | 8 |
| **3.** | LITERATURE SURVEY | 9 |
| **4.** | Used Algorithm | 10 |
| **5.** | METHODOLOGY | 13 |
| **6.** | Source Code | 19 |
| **7.** | Tools Used Hardware | 24 |
| **8.** | Conclusion | 25 |
| **9.** | References | 26 |

1. **INTRODUCTION**

According to data obtained from the National Transport Authority, the number of cars registered in recent years has witnessed a spectacular increase. With difficult economic conditions, it is likely that sales of second-hand imported (reconditioned) cars and used cars will increase. It is reported that the sales of new cars have registered a decrease in 2020 & 2021 due to pandemic conditions. Predicting the resale value of a car is not a simple task. It is trite knowledge that the value of used cars depends on several factors. As we can see, the price depends on many factors. Unfortunately, information about all these factors is not always available and the buyer must make the decision to purchase at a certain price based on a few factors only. The focus of this project is on developing machine learning models that can accurately predict the price of a used car based on its features, in order to make informed purchases. We implement and evaluate various learning methods on a dataset and will compare the performance of various machine learning algorithms like Linear Regression, Random Forest Regression, Lasso Regression, Decision Tree Regressor, XGBoost and choose the best out of it. Depending on various parameters we will determine the price of the car.

**2.PROBLEM STATEMENT AND OBJECTIVE**

Imagine a situation where you have an old car and want to sell it. You may of course approach an agent for this and find the market price, but later may have to pay pocket money for his service in selling your car. But what if you can know your car selling price without the intervention of an agent. Or if you are an agent, definitely this will make your work easier. Yes, this system has already learned about previous selling prices over years of various cars.

So, to be clear, we will provide you will the approximate selling price for your car based on the fuel type, years of service, showroom price, the number of previous owners, kilometers driven, if dealer/individual, and finally if the transmission type is manual/automatic.

The main objective of this this project are:

1.To develop an efficient and effective model which predicts the price of a used car according to user’s inputs.

2.To achieve good accuracy.

3.To develop a User Interface (UI) which is user-friendly and takes input from the user and predicts the price.

1. **LITERATURE SURVEY**

* **2.1 Survey of Existing system-1**

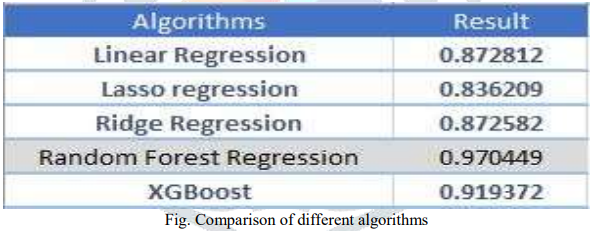
In the first existing survey Ref- TEM Journal. Volume 8, Issue 1, Pages 113-118, ISSN 2217-8309, DOI: 10.18421/TEM81- 16, February 2019 ‘Car Price Prediction Using Machine Learning Techniques’ according to authors Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric, in this paper they mainly concentrate on collecting various data from web portal by using web scrap techniques. And those have been compared with the help of different machine learning algorithms to predict the vehicle price in easy manner. They classified the price according to different ranges of price that is already given. Artificial neural network, support vector machine, random forest algorithms were used on different datasets to build classifiers model.

* **2.2 Survey of Existing system-2**

In the second existing survey Ref: International Journal of Information & Computation Technology. ISSN 0974-2239 Volume 4, Number 7 (2014) ‘Predicting the Price of Used’ Cars Using Machine Learning Techniques’ according to author Sameerchand Pudaruth they have done the predictions of car price from the historical data that has been collected from daily newspapers. For predicting the price of vehicles, they have used supervised machine learning techniques. Other algorithms were also used to predict such as multiple linear regression, some decision tree algorithms. All these algorithms were compared and found the best algorithm for prediction. They have faced some difficulties in comparing the algorithms, somehow they have managed.

**4.USED ALGORITHM**

There are several algorithms a person can use to predict a value as every algorithm itself has different significance and its own equation to predict. Algorithms like linear regression, Lasso regression, Ridge regression, Random forest regression, XGBoost, etc. have different equation through which we get a predicted value using past data. Let us compare all the equation and see which result is more accurate in this project:

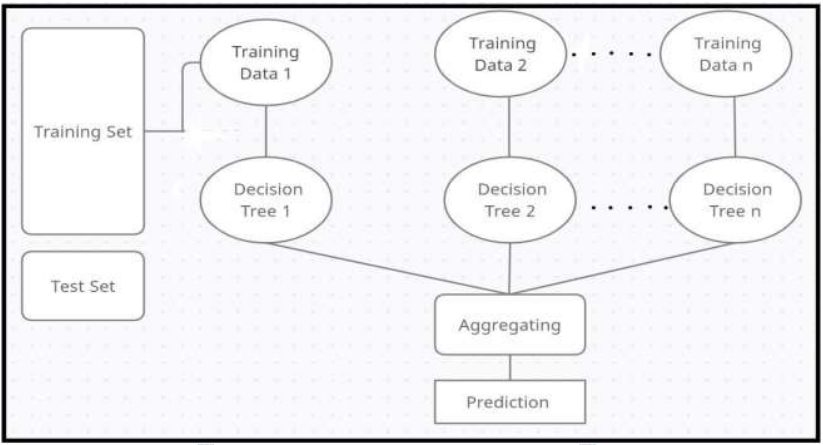


Here we can see Random forest regression is getting more accuracy than rest of the algorithms after comparing with our data which we are using in this project to predict the car value. Our point here showing you this comparison between different algorithms is not that Random Forest regression is the best of all the algorithms, but to tell you that using the data of cars for this model we are getting Random Forest regression as most accurate algorithm. This does not mean that all the other algorithms are not accurate much, they all have their different roles where they get more accurate value. Like if we say that Linear regression algorithm, it is an algorithm which is a statistical regression method used for predictive analysis and solving regression problem showing relationship between the continuous variable. Similarly, Ridge regression is one of the robust versions of linear regression. A general linear regression would fail if there is high collinearity between independent variables so Ridge regression is used here.

Likewise, every algorithm has its own specified role where they are performing more accurately than other algorithms. And this is why in this project we are using Random Forest regression. Let us know go in detail to understand our regression;

**Random Forest Regression:-**

We have previously understood that why we are using this algorithm in this model. Now we will see steps that has helped us to build this model for car value prediction; Firstly, we have import dataset of different cars and their year, kms driven, fuel type, transmission, etc. These are the attributes which we are going to need to predict. Then we will be using feature engineering where we will correlate between attributes in our dataset,



**Advantages of Random Forest Regression:**

1) Irrespective of size of the database; large or small it efficiently runs and gives more accurate value.

2) Generally, it takes lesser time to train the model and there is no overfitting.

3) Having extensive amount of missing data but still it maintains higher accuracy and estimates missing values.

**Hyper parameters Random forest:**

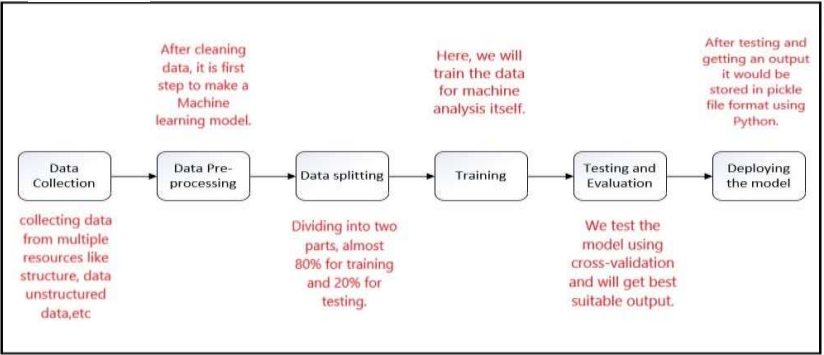
* Hyper parameters help us to increase the speed of processing which are provided in Sklearn library of python; max\_depth: Using this parameter, we can limit up to what depth I want every tree in my random forest to grow.
* min\_sample\_split: It gives us minimum required number while sampling in every decision tree of random forests.
* max\_leaf\_nodes: If we are getting more terminal nodes, this hyper parameter sets condition to restricts the growth of the
* tree, and stops the splitting. min\_sample\_leaf: In this it sets a condition to the terminal(leaf) nodes after splitting and we only get the wanted leaf
* nodes. n\_estimators: Taking more number of decision tree will increase time complexity so, it limits the number of decision tree.
* max\_samples: It is not necessary to give each decision tree the full data. Fraction of the original dataset is given to any
* individual tree; this helps us to reduce the training time in random forest. max\_features: It sets the number of maximum features which is provided to each decision tree. It should be set because
* in testing, problem of overfitting comes into place.

**Extra-tree Regressor: -**

An extra-tree Regressor, this class implements a meta estimator that fits a number of randomized decision trees on various subsamples of the data sets and uses averaging to improve the predictive accuracy and control overfitting. It works by creating a large number of unpruned decision trees from the training dataset. Extra-tree Regressor adds randomization but still has optimization.

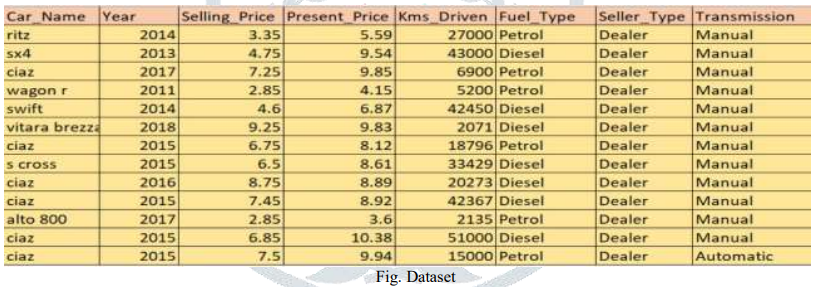
1. **METHODOLOGY**

* **Proposed Model:-**

****

* **Data collection:**

Solving machine learning problems firstly we require raw data because without raw data we cannot do machine learning problems. raw data we get from further discussion of the problem with client and data scientist team we focus on data that is a data integration and data integration is a very difficult task because we collect data from multiple resources like structure data unstructured data, web scraping, etc. collected data stored in data warehouse and we get data from a data warehouse.

****

**Fig.Dataset**

* **Data Pre-processing:**

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not 0always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data pre-processing task.

* **Data splitting:**

The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model. In this step data is split for training and testing almost 80% of data is for training and 20% for testing is a basic rule in the machine learning.

* **Training:**

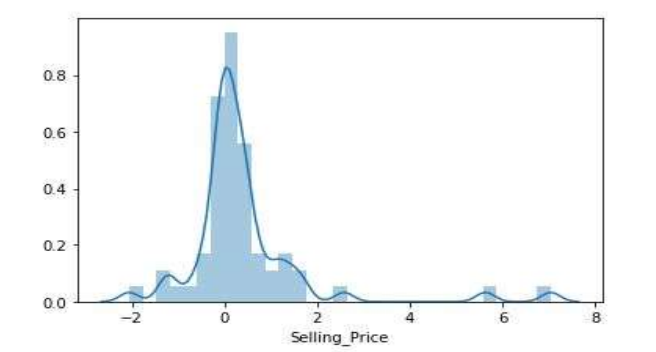
In this step, we do training data for machine analysis itself and we do another step is to validate training data because training data set will produce either overfitting or under the fitting problem that means false positive output or true negative output that means overfitting means when you go new area and 1st person give disrespect and you considering all people are same this is.

* **Testing and Evaluation:**

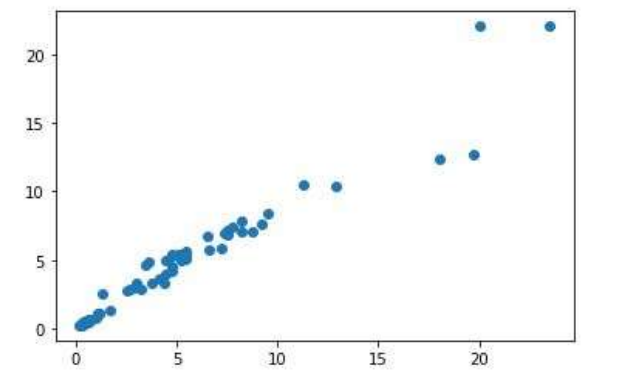
In the testing phase, we test the model using cross-validation, we check the model is well or not and going is right or not, there are some technique of cross-validation and we use confusion matrix for checking model performance. We will test in all algorithms and will get best suitable output.

Seaborn is a Python-based data visualization library based on matplotlib. Provides an advanced interface for drawing attractive and informative mathematical graphics.

Below the graph representation tells us about how much the prediction is accurate. The graph, basically, shows that the model which we have created gives us a very good result as it is a normal distribution representation. The close range on the graph gives us the idea that it is predicting more accurately.

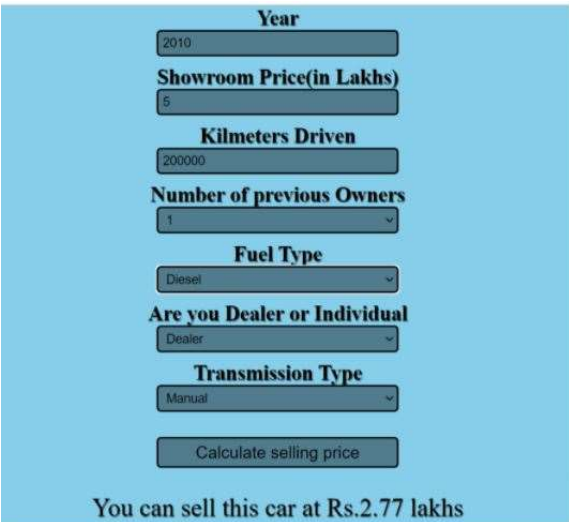


Here, the plotting of the dots below in the graph is linearly seen, that means the model which we are running is giving us much better prediction. Therefore, the linearity of the dots in the graph should be there which says us how much your model is pretty accurate. More the linearity of the dots, more accurate the prediction we will get.



* Deploying the model:

In this we will be deploying the model on web. After testing and getting an output it would be stored in pickle file format using Python. We have used version controlled system as GitHub and containerized our application which has been uploaded on Docker hub and further deployed on Azure web app service making the whole process more dynamic.



1. **Source Code:**

\*App.py:-

from flask import Flask, render\_template, request

import pickle

import numpy as np

import os

from sklearn.preprocessing import StandardScaler

app = Flask(\_\_name\_\_,static\_url\_path='/static')

model = pickle.load(open('random\_forest\_regression\_model.pkl', 'rb'))

@app.route('/',methods=['GET'])

def Home():

return render\_template('index.html')

standard\_to = StandardScaler()

@app.route("/predict", methods=['POST'])

def predict():

Fuel\_Type\_Diesel=0

if request.method == 'POST':

Year = int(request.form['Year'])

Present\_Price=float(request.form['Present\_Price'])

Kms\_Driven=int(request.form['Kms\_Driven'])

Kms\_Driven2=np.log(Kms\_Driven)

Owner=int(request.form['Owner'])

Fuel\_Type\_Petrol=request.form['Fuel\_Type\_Petrol']

if(Fuel\_Type\_Petrol=='Petrol'):

Fuel\_Type\_Petrol=1

Fuel\_Type\_Diesel=0

elif(Fuel\_Type\_Petrol=='Diesel'):

Fuel\_Type\_Petrol=0

Fuel\_Type\_Diesel=1

else:

Fuel\_Type\_Petrol=0

Fuel\_Type\_Diesel=0

Year=2020-Year

Seller\_Type\_Individual=request.form['Seller\_Type\_Individual']

if(Seller\_Type\_Individual=='Individual'):

Seller\_Type\_Individual=1

else:

Seller\_Type\_Individual=0

Transmission\_Mannual=request.form['Transmission\_Mannual']

if(Transmission\_Mannual=='Mannual'):

Transmission\_Mannual=1

else:

Transmission\_Mannual=0

prediction=model.predict([[Present\_Price,Kms\_Driven2,Owner,Year,Fuel\_Type\_Diesel,Fuel\_Type\_Petrol,Seller\_Type\_Individual,Transmission\_Mannual]])

output=round(prediction[0],2)

if output<0:

return render\_template('index.html',prediction\_texts="Sorry you cannot sell this car")

else:

return render\_template('index.html',prediction\_text="You Can Sell The Car at {}".format(output))

else:

return render\_template('index.html')

if \_\_name\_\_=="\_\_main\_\_":

port = int(os.environ.get('PORT', 33507))

app.run(debug=True,port=port)

\*index.html:

<!DOCTYPE html>

<html lang="en">

<head>

<meta charset="UTF-8">

<meta name="viewport" content="width=device-width, initial-scale=1.0">

<title>Document</title>

<link rel="stylesheet" href="/static/main.css">

<style>@import url('https://fonts.googleapis.com/css2?family=Poppins:wght@300;400&display=swap');</style>

</head>

<body>

<div class="container">

<div class="container-wrapper">

<form action="{{ url\_for('predict')}}" method="post">

<h2 class="text\_center" >Predictive analysis</h2>

<h3>Year</h3>

<input id="first" name="Year" type="number ">

<h3>What is the Showroom Price?(In lakhs)</h3><input id="second" name="Present\_Price" required="required">

<h3>How Many Kilometers Drived?</h3><input id="third" name="Kms\_Driven" required="required">

<h3>How much owners previously had the car(0 or 1 or 3) ?</h3><input id="fourth" name="Owner" required="required">

<h3>What Is the Fuel type?</h3><select name="Fuel\_Type\_Petrol" id="fuel" required="required">

<option value="Petrol">Petrol</option>

<option value="Diesel">Diesel</option>

<option value="CNG">CNG</option>

</select>

<h3>Are you A Dealer or Individual</h3><select name="Seller\_Type\_Individual" id="resea" required="required">

<option value="Dealer">Dealer</option>

<option value="Individual">Individual</option>

</select>

<h3>Transmission type</h3><select name="Transmission\_Mannual" id="research" required="required">

<option value="Mannual">Manual Car</option>

<option value="Automatic">Automatic Car</option>

</select>

<div class="submit-button">

<button id="sub" type="submit ">Calculate the Selling Price</button>

</div>

</form>

</div>

<h3 class="text\_center">{{ prediction\_text }}<h3></h3>

</div>

</body></html>

1. **Tools Used Hardware**

* **Hardware:-**

Minimum requirements needed to install and run are as followed:

1. Operating system- Windows 7,8,10
2. Processor- dual core 2.4 GHz (i5 or i7 series Intel processor
3. or equivalent AMD)
4. RAM-4GB

* **Software: -**

1. Python Flask

2. Pycharm/VSCode

3. PIP 2.7

4. Jupyter Notebook

5. Chrome

6. Docker Container

1. **Conclusion**

The model which we were making is to predict a value of second hand car using machine learning techniques. We have collected the data of cars from Kaggle having attributes like different cars and their year, kms driven, fuel type, transmission, etc. The data is then processed using different algorithms where results of each algorithms is compared, getting Random Forest algorithm as the most accurate amongst them, so we have used Random Forest because irrespective of size it runs efficiently and gives more accuracy than any other algorithm. Further it would be available in GUI as a Web-application developed using Pythonflask making it user friendly so that users could give input and get the price of a car according to it.

1. **References**
2. Sameerchand Pudaruth, “Predicting the Price of Used Cars using Machine Learning Techniques” ;(IJICT 2014).
3. Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric,” Car Price Prediction Using Machine Learning”; (TEM Journal 2019).
4. S.E. Viswapriya, Durbaka Sai Sandeep Sharma, Gandavarapu Sathya kiran, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-8, June 2020.
5. Pattabiraman Venkatasubbu, Mukkesh Ganesh, ‘Used Cars Price Prediction using Supervised Learning Techniques’ (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-1S3, December 2019.
6. Praful Rane, Deep Pandya, Dhawal Kotak, ‘Used Car Prediction’, (IRJET), Volume: 08 Issue: 04, Apr 2021