SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

A Major Project Report on

USED CAR PRICE PREDICTION USING MACHINE LEARNING

Submitted in fulfillment of the requirements for the award of the Degree of

Bachelor of Technology
In
Computer Science and Engineering

Submitted by

D S Aumkaareshwar	(R18CS103)
Blesson E	(R18CS085)
C Y Revanth Raju	(R18CS099)
David G B	(R18CS106)

Under the guidance of

Dr. Meenakshi S A
School of CSE
REVA UNIVERSITY

May 2022

Rukmini Knowledge Park, Kattigenahalli, Yelahanka, Bengaluru-560064
www.reva.edu.in

DECLARATION

We, **D S Aumkaareshwar (R18CS103), Blesson E (R18CS085), C Y Revanth Raju (R18CS099), David G B (R18CS106)** students of **B.Tech, belong in to School of CSE, REVA University, declare that this Project Report / Dissertation entitled "USED CAR PRICE PREDICTION USING MACHINE LEARNING" is the result the of project / dissertation work done by us under the supervision of Dr.** Meenakshi S A **at CSE**

We are submitting this Project Report / Dissertation in partial fulfillment of the requirements for the award of the degree of **Bachelor of technology** in **Computer Science and Engineering** by the REVA University, Bangalore during the academic year 2021-22.

We declare that this project report has been tested for plagiarism and has passed the plagiarism test with the similarity score less than 25% and it satisfies the academic requirements in respect of Project work prescribed for the said Degree.

We further declare that this project / dissertation report or any part of it has not been submitted for award of any other Degree/ Diploma of this University or any other University/ Institution.

1.
2.
3.
4.

(Signature of the candidate) Signed by me on 25TH May 2022

*Certified that this project work was submitted by **D S Aumkaareshwar (R18CS103), Blesson E (R18CS085), C Y Revanth Raju (R18CS099), David G B (R18CS106)** has been carried out under my / our guidance and the declaration made by the candidate is true to the best of my knowledge.*

Signature of Guide
Date :

Signature of Director of School
Date :
With Official Seal of the School

REVA UNIVERSITY
Bengaluru, India

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

Certified that the project work entitled **"USED CAR PRICE PREDICTION USING MACHINE LEARNING"** carried out under my / our guidance by **D S Aumkaareshwar (R18CS103), Blesson E (R18CS085), C Y Revanth Raju (R18CS099), David G B(R18CS106),** bonafide students of REVA University during the academic year 2021-22, are submitting the project report in partial fulfillment for the award of **Bachelor of Technology** in **Computer Science and Engineering** during the academic year **2021-22**. The project report has been tested for plagiarism and has passed the plagiarism test with the similarity score less than 20%. The project report has been approved as it satisfies the academic requirements in respect of Project work prescribed for the said Degree.

Signature with date

Dr. Meenakshi S A
Guide

Signature with date

Dr. Ashwinkumar U M

Deputy Director

Signature with date

Dr. M Dhanamjaya
Vice Chancellor

External Examiner
Name of the Examiner with affiliation **Signature with Date**

1.

2.

ACKNOWLEDGEMENT

Any given task achieved is never the result of efforts of a single individual. There are always a bunch of people who play an instrumental role is leading a task to its completion. Our joy at having successfully finished our mini project work would be incomplete without thanking everyone who helped us out along the way. We would like to express our sense of gratitude to our REVA University for providing us the means of attaining our most cherished goal.

We would like to thank our Hon'ble Chancellor, **Dr. P. Shyama Raju**, Hon'ble Vice-Chancellor, **Dr. M. Dhanamjaya** for their immense support towards students to showcase innovative ideas.

We cannot express enough thanks to our respected Deputy Director, **Dr. Ashwin Kumar U. M.** for providing us with a highly conducive environment and encouraging the growth and creativity of each student. We would also like to offer our sincere gratitude to our Major Project Coordinators, **Prof. Kiran M.** for the numerous learning opportunities that have been provided.

We would like to take this opportunity to express our gratitude to our Major Project Guide, **Dr. Meenakshi S A, Associate Professor, School of CSE** for continuously supporting and guiding us in our every endeavor as well for taking a keen and active interest in the progress of every phase of our Major Project. Thank you for providing us with the necessary inputs and suggestions for advancing with our project work. We deeply appreciate the wise guidance that sir/madam has provided.

Finally, we would like to extend our sincere thanks to all the faculty members, staff from School of Computer Science and Engineering.

D S Aumkaareshwar (R18CS103)
Blesson E (R18CS085)
C Y Revanth Raju (R18CS099)
David G B (R18CS106)

TABLE OF CONTENTS

Chapter No.	Chapter Name	Page No.
1	ABSTRACT	6-7
2	INTRODUCTION	7-9
3	LITERATURE SURVEY: **3.1** EXISTING SYSTEM **3.2** PROPOSED SYSTEM	9-12
4	WORKFLOW OF THE PROJECT PROBLEM DEFINITION	12-14
5	OBJECTIVES COST ESTIMATION	14-17
6	METHODOLOGY: 1. CREATING ENVIRONMENT 2. INSTALLING LIBRARIES 3. DATA COLLECTION 4. READING DATA 5. DATA PREPROCESSING 6. TEST TRAIN AND SPLIT	17-23
7	RESULTS CONCLUSION REFERENCES	24-27

CHAPTER 1

ABSTRACT:

Car price prediction has always been an important and popular research topic because it requires the significant struggling effort and tremendous amount of knowledge from the field expert. For a reliable and accurate prediction, a large number of distinct attributes are examined. We used three machine learning techniques to create a model for predicting the price of used cars the SVN also known as the Support Vector Machine, the ANN also known as Artificial Neural Network and the popular one the RF which was abbreviated for Random Forest. However, the technical methods mentioned were used as a core in the group settings. The data for the prediction was acquired from some of the well-known websites using a web scraper coded in the Hypertext Preprocessor programming language.

A car's value begins to depreciate the moment it is purchased, and it continues to depreciate with each passing year. A car's value drops by 20% in the first year. The car's make and model, total kilometers driven, overall condition, and a variety of other factors all influence its resale value.

The relative performances of various algorithms were then compared in order to determine which one best suited the available data set. The final prediction model was implemented in a Java programme. Vehicle manufacturers set the price of new cars and vehicles in the society, with the governmenting bodies improvising with some additional taxes. As a result, a customer looking to purchase a car or motorcycle can rest assured that their money will be well spent because many variables are in play. However, many people cannot afford a car or a bike due to a variety of factors such as a lack of funds or rising prices. So, these people's next thought is to buy a used vehicle, such as used cars or motorcycles, but it's important to understand the actual market value of both buying and selling. As a result, a second-hand vehicle prediction model is required to accurately predict the price of used cars and motorcycles.

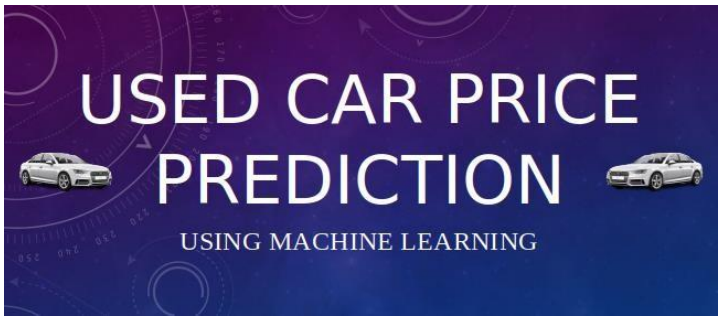
It makes money based on cost for impression, cost for view and cost for click. The e-marketplace has emerged as an efficient and important vehicle for transactions in the e-commerce industry, and both academia and industry have identified trust as a key factor

in enabling e-commerce, which is why this study looked into how trust can be established in Nigerian e-markets. A group of 32 students taking an E-business course posted online an item for sale and were then interviewed about what happened between them and their potential buyers. According to the findings, a major barrier to e-commerce adoption in Nigeria is a lack of trust. The practical implication is that e-markets, must design and implement a system that will screen both buyers and sellers, ensuring that both parties are protected. When transacting, have faith in one another. On the other hand, the Nigerian government must create an environment that allows trust to flourish in the e-market. The object of this project is to find the impact of our project on consumers decisions and how the goal of this project is to determine the extent to which our project influences consumer decisions. The project's goal is to be accomplished using a questionnaire that will aid in understanding the impact of our project on consumers. To determine the impact of our project on consumers, the data collected from the questionnaire will be statistically condensed. This will be the project's ultimate goal.

CHAPTER 2

INTRODUCTION:

This project entails the need for a module for a car price prediction using machine learning techniques. Without any interference from outer entities such as brokers and mediators. The chances of getting trapped in the greed of some brokers and some malfunctions are also more in the existing system. The transparency of module is also compromised by the existing module as there is no promised transparent methodology or the way in which the machine is predicting the value of our vehicle. The user should be aware of how his car value is detected that the way how the process of the price detection. Another main factor is the security which is not promised through the old or existing way of car price prediction.



Vehicle manufacturers set the price value amount of new cars and vehicles in the society, with the government bodies improvising with some increased taxes. As a result, a customer looking to purchase a car or motorcycle can rest assured that their money will be well spent because many variables are in play. However, many people cannot afford a car or a bike due to a variety of factors such as a lack of funds or rising prices. So, these people's next thought is to buy a used vehicle, such as used cars or motorcycles, but it's important to understand the actual market value of both buying and selling. As a result, a second-hand vehicle prediction model is required to accurately predict the price of used cars and motorcycles.

In proposed system we provide such an environment where the system decides the value through machine learning techniques and earlier prices of sold cars. Since there is no involvement of any such brokers the unbiased machine is wholly responsible for the predicted car value. There are more than one mechanism to predict the price of the car based on the previous sold car valuation. As the existing systems are totally unsatisfactory to the consumers, so we as a group of 4 members have come together to a new type of module which predicts the estimated car selling value. But the speciality of our module is the prediction using the machine learning techniques. This property makes the proposed module different from the existed system, as the old modules are mainly dependent from the rates fixed by the brokers and some modules have the functions as the buyer and seller have direct contact with each other, which causes the ruckus among the users. So, the module which we are proposing is much more efficient than the modules which are existing. The system which we are proposing here is mainly dependent on the machine learning techniques which takes the values from the database which is been stored in an excel sheet. The date consists of the previous experiences which when some users shared their details about the selling and the experiences they

had during the whole process. The whole process is system generated as the data is only thing which is been provided the coder and the system takes the values provided and then calculates and measures using the inbuilt algorithms and then returns the webpage which asks the user for some inputs and the same should be provided by the user and then the input is dynamically taken and the same measurements are collected and output is resulted and that is the rate at which the user can sell his vehicle and the value resulted is the minimum amount of the price at which user can sell the vehicle.

CHAPTER 3

LITERATURE SURVEY:

For the literature survey we searched for some existing modules and managed for some great changes and created a new different type of system which can learn from the drawbacks of the older systems and adapting those in the new system. The literature survey for our project was basically dependent on the comparison and the updation between the existing modules and the proposed module. Below is the detailed response and the comparisons between the existing modules of the used car price prediction through the traditional way and the new one proposed model based on the machine learning techniques. This will help for the better understanding about the traditional and the new machine learning way of price prediction.

EXISTING SYSTEM:

Table 1. Existing vs Proposed System.

Existing System:	Proposed System:
Existing websites are only designed in such a way that only the dealer can predict the value of the vehicle.	In proposed system we provide such an environment where the system decides the value through machine learning techniques and earlier prices of sold cars.
The brokers may manipulate the value of the vehicles.	Since there is no involvement of any such brokers the unbiased machine is wholly responsible for the predicted car value.
There is no such earlier mechanism which predicts the price of vehicles with respect to its valuation.	There are more than one mechanisms to predict the price of the car based on the previous sold car valuation.

Existing websites are only designed in such a way that only the dealer can predict the value of the vehicle. The brokers may manipulate the value of the vehicles. There is no such earlier mechanism which predicts the price of vehicles with respect to its valuation. During the process of project when went through all the related modules. Came to know that the existing modules are not that feasible, user friendly, reliable, transparent, and mainly the estimated cost would be calculated through some untrusted sources. The main defect was the cost estimation by some untrusted sources made us to think for some alternative modules for the betterment of the field of estimation modules. And another drawback of the exiting module is that the brokers who operate the modules and applications may manipulate the values and data being biased towards any one of the seller and buyer in greed of some greater commissions. So, this creates a great trust deficit within the customer. Now when we come to some other drawbacks there comes the great thing of valuation. The main benefit that makes the customers attracted towards such modules are the valuation which are less than or matching to the market value of the estimations. But the valuations provided by the earlier modules are depended on the buyer's choice or the choice of the mediator. The chances of getting trapped in the greed of some brokers and some malfunctions are also more in the existing system. The transparency of module is also compromised by the existing module as there is no promised transparent methodology or the way in which the machine is predicting the value of our vehicle. The user should be aware of how his car value is detected that the way how the process of the price detection. Another main factor is the security which is not promised through the old or existing way of car price prediction.

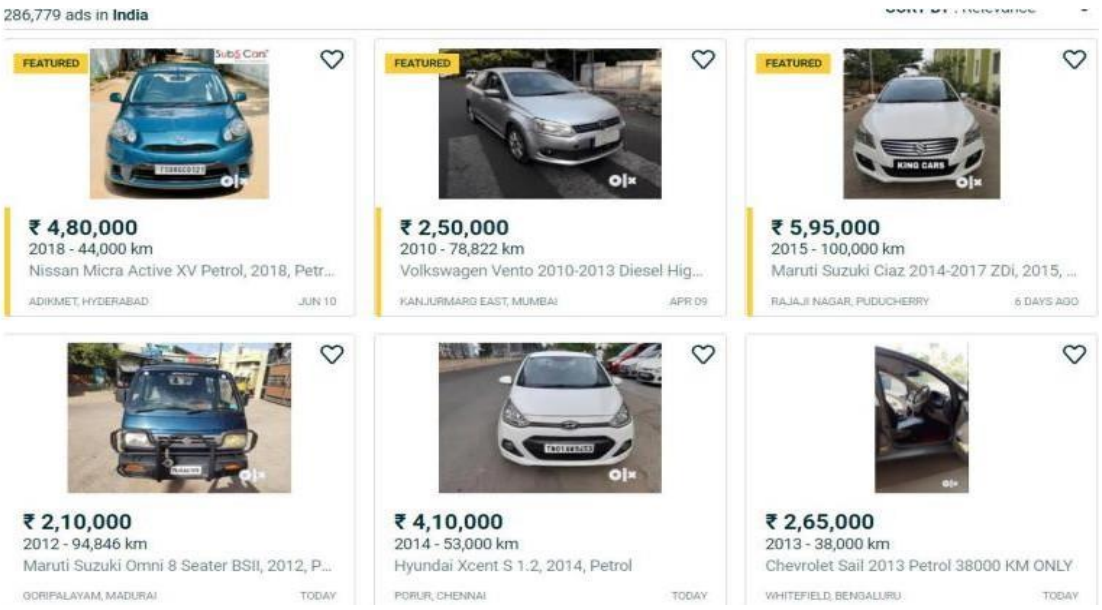


Fig. 1. Existing systems

PROPOSED SYSTEM:

In proposed system we provide such an environment where the system decides the value through machine learning techniques and earlier prices of sold cars. Since there is no involvement of any such brokers the unbiased machine is wholly responsible for the predicted car value. There are more than one mechanism to predict the price of the car based on the previous sold car valuation. As the existing systems are totally unsatisfactory to the consumers, so we as a group of 4 members have come together to a new type of module which predicts the estimated car selling value. But the speciality of our module is the prediction using the machine learning techniques. This property makes the proposed module different from the existed system, as the old modules are mainly dependent from the rates fixed by the brokers and some modules have the functions as the buyer and seller have direct contact with each other, which causes the ruckus among the users. So, the module which we are proposing is much more efficient than the modules which are existing. The system which we are proposing here is mainly dependent on the machine learning techniques which takes the values from the database which is been stored in an excel sheet.

The date consists of the previous experiences which when some users shared their details about the selling and the experiences they had during the whole process. The whole process is system generated as the data is only thing which is been provided the coder and the system takes the values provided and then calculates and measures using the inbuilt algorithms and then returns the webpage which asks the user for some inputs and the same should be provided by the user and then the input is dynamically taken and the same measurements are collected and output is resulted and that is the rate at which the user can sell his vehicle and the value resulted is the minimum amount of the price at which user can sell the vehicle.

CHAPTER 4

WORKFLOW OF THE PROJECT:

We choose the best model in terms of both train and test data in this project and save it in a pickle file using the pickle function. We write a piece of code that connects the pickle and python files so that they can be deployed. We will create a website using HTML, CSS, and Bootstrap to accept user input and accurately predict the price of a used vehicle.

This is a machine learning project in which we use various modules to predict the price of used cars and bikes. We save the file with the pickle function and then deploy it with flask. For the front end of this website, HTML [Hyper Text Markup Language], CSS [Cascaded Style Sheet], and Bootstrap were used. The system which we are proposing here is mainly dependent on the machine learning techniques which takes the values from the database which is been stored in an excel sheet. The date consists of the previous experiences which when some users shared their details about the selling and the experiences they had during the whole process. The whole process is system generated as the data is only thing which is been provided the coder and the system takes the values provided and then calculates and measures using the inbuilt algorithms and then returns the webpage which asks the user for some inputs and the same should

be provided by the user and then the input is dynamically taken and the same measurements are collected and output is resulted and that is the rate at which the user can sell his vehicle and the value resulted is the minimum amount of the price at which user can sell the vehicle. The code below contains of the basic features of python codes and needs to be

```
from flask import Flask, render_template, request
import jsonify
import requests
import pickle
import numpy as np
import sklearn
from sklearn.preprocessing import StandardScaler
app = Flask(__name__)
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
        else:
            Fuel_Type_Petrol=0
            Fuel_Type_Diesel=1
        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__":
    app.run(debug=True)
```

PROBLEM DEFINITION:

There are many websites and applications which display the used car price prediction, but those values usually depend on the dealer or the individual, who are not trustworthy.

The lack of knowledge in cars and vehicles make the users to trapped into a fraud. People easily get fooled as through the existing systems allows the brokers and dealers to increase the prices of the predicting cars and vehicles.

CHAPTER 5

OBJECTIVES:

To create an effective and efficient model that can predict the price of a used vehicle based on the inputs of the user. To ensure that the models used are accurate. Develop a good User-Interface that takes the user's inputs into account. To gather information from the user to make more accurate predictions. To take the car brand into account so that the prediction is accurate. All the models' accuracy is displayed. The UI

is clean thanks to Bootstrap. Predicting various data visualizations so that we can better understand the data. Using a variety of machine learning models to better understand accuracy and which model data works best. Pickle is used to save the model so that it can be reused.

This study tested a single machine learning classifier approach that had been used in all previous studies. Artificial Neural Network, Support Vector Machine, and Random Forest classifier models were built using the entire data set collected in this study, which was split into training (90%) and testing (10%) subsets. The random forest (RF), also known as the random decision forest, is an ensemble method. For classification and regression problems, RF can be used. Ho created the algorithm as a solution to the problem of overfitting in decision tree algorithms. Artificial Neural Networks (ANNs) are a type of machine learning model that attempts to solve problems in the same way that the human brain does. Artificial neurons are also used by the ANN instead of neurons known as perceptron. Neurons in the human brain are connected by axons, whereas in ANN, weighted matrices are used to connect artificial neurons. Information travels between neurons via connections, and information from one neuron travels to all neurons connected to it. The system for adjusting the weights between neurons can be trained using input examples. The Support Vector Machine can be used to solve problems involving classification and regression. The SVM can make a binary decision for the input data set and determine which of the two categories the input sample belongs to. The SVM algorithm is trained to divide input data into two categories with the largest possible area between them. The SVM algorithm cannot be used when the input data is not labelled. Unsupervised learning is required for unlabeled data, and SVM has an implementation called Support Vector Clustering (SVC).

COST ESTIMATION:

The project can be implemented with little financial aid but needs good knowledge about the creating webpages and coding. The systems required for the smooth run of the module are the basic things that can be used to run any operating system.

Software requirements:

- Jupyter notebook
- Spyder notebook
- Anaconda prompt [to run the python codes of the jupyter]
- Operating system
- Pre-installed python modules:

  certifi==2020.6.20

  chardet==3.0.4

  click==7.1.2

  Flask==1.1.2

  idna==2.10

  itsdangerous==1.1.0

  Jinja2==2.11.2

  joblib==0.15.1

  jsonify==0.5

  MarkupSafe==1.1.1

  numpy==1.19.0

  requests==2.24.0

  scikit-learn==0.23.1

  scipy==1.5.0

  sklearn==0.0

  threadpoolctl==2.1.0

  urllib3==1.25.9

Werkzeug==1.0.1

wincertstore==0.2

Hardware requirements:

- Laptop [any laptop or PC]
- RAM – 2GB and above
- Mouse and Keyboard
- Processor – [Intel Pentium or intel core or Amd]
- Hard disk – 500 MB and above

CHAPTER 6

METHODOLOGY:

1. CREATING ENVIRONMENT:

As we form different applications in our structure, and it needs different python versions so we design an environment for each project and work the same project in that environment so that there would be no compound of versions with each other.

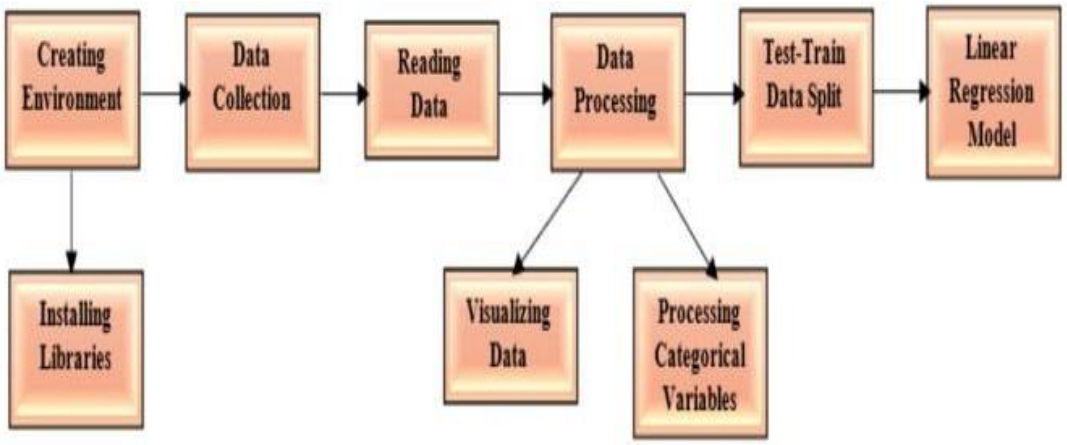


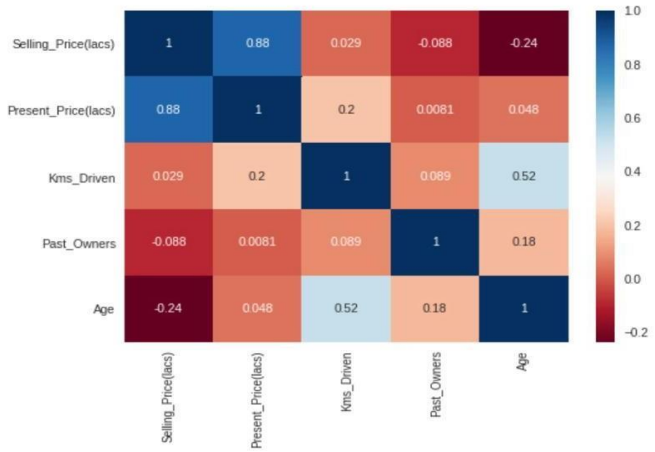
Fig. 2. Methodology

2. INSTALLING LIBRAIRES:

After designing an environment, we can stop the libraries using the function called as $ pip stop requirement.txt. So, we get a file of all need libraries and its versions to download. We can download by opening anaconda prompt and running the version as pip install library name version of that library.

Chart 1. Showing datatypes and their libraries

```
sns.heatmap(df_main.corr(), annot=
True, cmap="RdBu")
plt.show()
```



3. DATA COLLECTION:

The important and the only one task that the user must work on is to collect the data. And the only job that the programmer has to do, if once the user provides the data or the dataset the data should be dynamic and random because the data is the main part of the whole process and once the dataset is given the machine learning method they take the parameters as the core unit and make the further transitions for the calculations that take place during the input of the user provides. This is the only task that must be done by the programmer so that the system is automatic.

The data is nothing but the CSV (comma separated values) which consists of rows and columns which consists of a lot of data approximately 15 MB.

Table 2. Showing the type of dataset.

	Car_Name	Year	Selling_Price	Present_Price	Kms_Driven	Fuel_Type	Seller_Type	Transmission	Owner
0	ritz	2014	3.35	5.59	27000	Petrol	Dealer	Manual	0
1	sx4	2013	4.75	9.54	43000	Diesel	Dealer	Manual	0
2	ciaz	2017	7.25	9.85	6900	Petrol	Dealer	Manual	0
3	wagon r	2011	2.85	4.15	5200	Petrol	Dealer	Manual	0
4	swift	2014	4.60	6.87	42450	Diesel	Dealer	Manual	0

- **About dataset:**

This dataset contains information about used motorcycles. This data can be used for a lot of purposes such as price prediction to verify the use of linear regression in machine learning.

The columns in the given dataset are as follows:

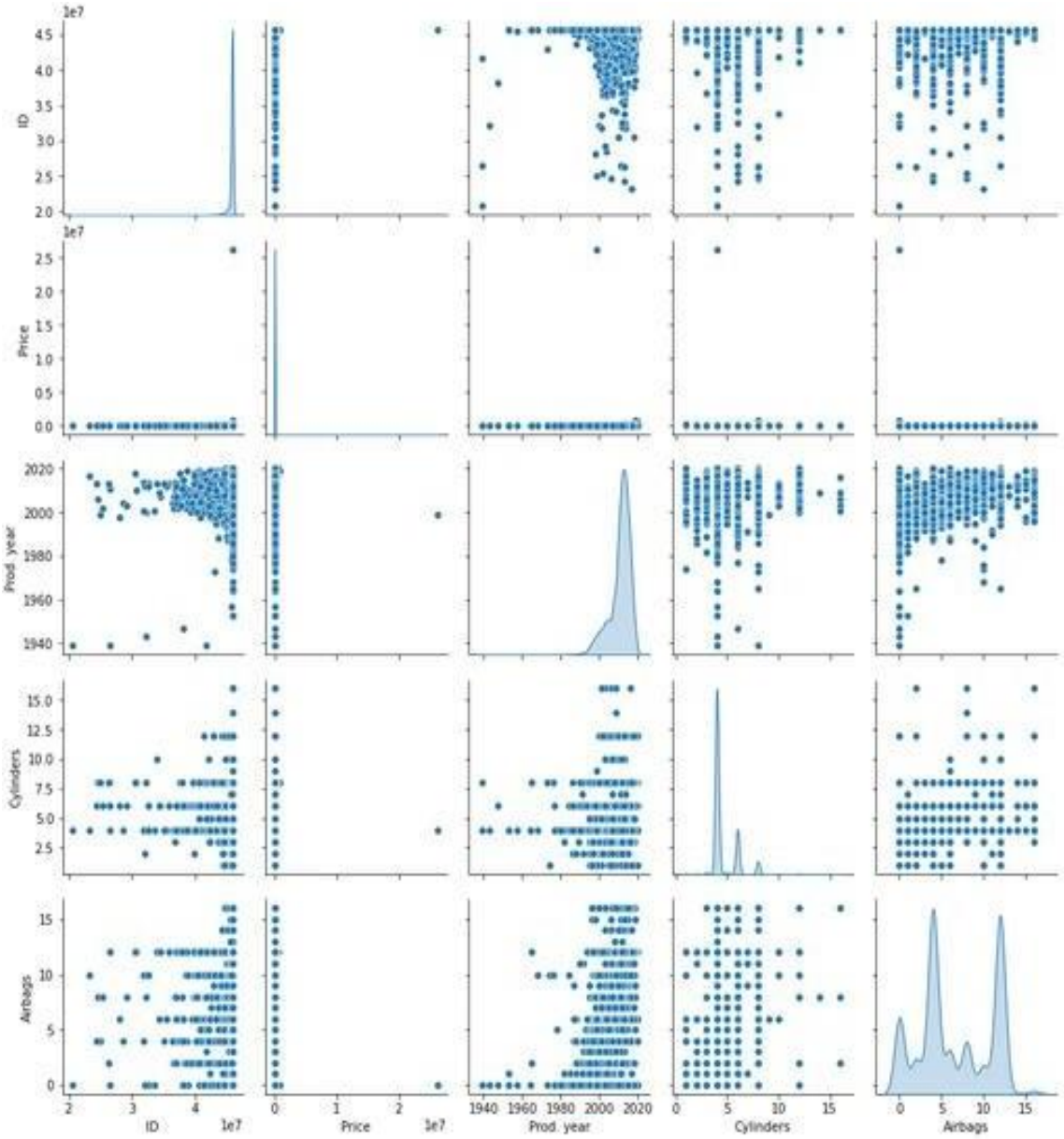
1. Name
2. Selling price
3. Year
4. Selling type
5. Number of previous Owner
6. Kilometer driven
7. Ex showroom price
8. Present price in the market

Table 3. Collection of data

	Car_Name	Year	Selling_Pri	Present_Pr	Kms_Drive	Fuel_Type	Seller_Typ	Transmissi	Owner
2	ritz	2014	3.35	5.59	27000	Petrol	Dealer	Manual	0
3	sx4	2013	4.75	9.54	43000	Diesel	Dealer	Manual	0
4	ciaz	2017	7.25	9.85	6900	Petrol	Dealer	Manual	0
5	wagon r	2011	2.85	4.15	5200	Petrol	Dealer	Manual	0
6	swift	2014	4.6	6.87	42450	Diesel	Dealer	Manual	0
7	vitara brez	2018	9.25	9.83	2071	Diesel	Dealer	Manual	0
8	ciaz	2015	6.75	8.12	18796	Petrol	Dealer	Manual	0
9	s cross	2015	6.5	8.61	33429	Diesel	Dealer	Manual	0
10	ciaz	2016	8.75	8.89	20273	Diesel	Dealer	Manual	0
11	ciaz	2015	7.45	8.92	42367	Diesel	Dealer	Manual	0
12	alto 800	2017	2.85	3.6	2135	Petrol	Dealer	Manual	0
13	ciaz	2015	6.85	10.38	51000	Diesel	Dealer	Manual	0
14	ciaz	2015	7.5	9.94	15000	Petrol	Dealer	Automatic	0
15	ertiga	2015	6.1	7.71	26000	Petrol	Dealer	Manual	0

4. READING DATA:

We read the information called CSV (comma separated values) which is in the form of rows and columns. The comma separated values which are collected is provided to the python code/algorithms that use the provided data for the further code that is to the calculations of the rate of the vehicle value.



Graph 1. Distribution and varieties of data

5. DATA PREPROCESSING:

At first, we should have to preprocess the data before using it so that it doesn't contain any null or invalid values which affects the perfection of the program. That is why we preprocess the null and the invalid values by discarding the row that contains the value or handling the erroneous values. The preprocessing of the data is also the main part of the whole process. As the data which is been used as the core of

our whole process, contains some errors that will collapse the whole system with some threats to the outputs.

Missing values before data cleaning process

id	0
region	0
year	1527
manufacturer	22764
model	7989
condition	231934
cylinders	199683
fuel	3985
odometer	92324
title_status	3062
transmission	3719
drive	144143
size	342003
type	141531
paint_color	164706
lat	10292
long	10292
price	0

Grey color shows the distribution of null values.

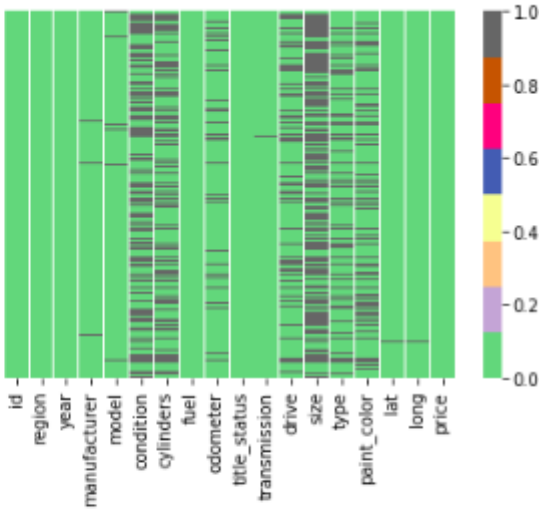


Table 4. Processing of data (removal of null and invalid data)

7 Visualizing data:

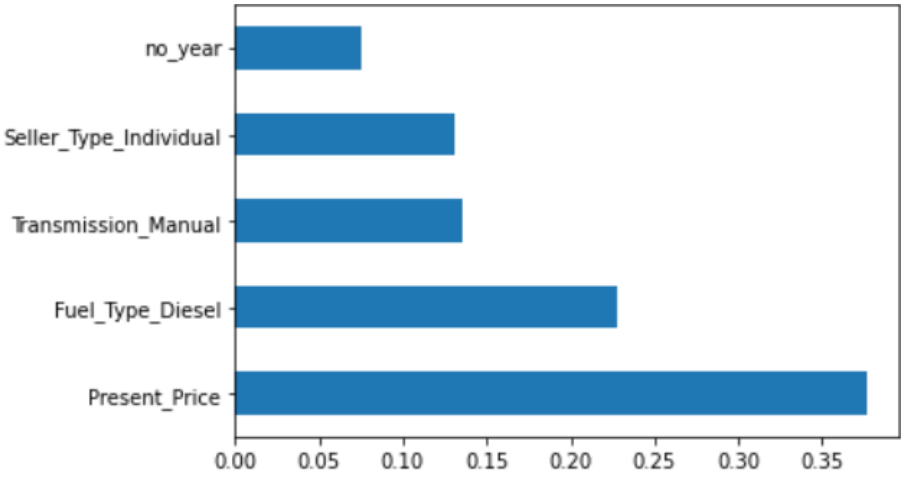
We can have a vision the data of different variables so that we can understand the data more clearly and use them according to the need of it. We can also envision the different important features of all the variables so that we can know how important the variable depends on model. The goal of this study is to create a good regression model that can accurately predict car prices. To do so, we'll need some previous used car data, for which we'll use price and other standard attributes. Other attributes are considered independent variables, while car price is considered the dependent variable.

Random Forest is a regression model based on ensemble learning. It employs a decision tree model, specifically multiple decision trees to generate the ensemble model, which generates a prediction collectively. The benefit of this model is that the trees are created

in parallel and are relatively uncorrelated, resulting in good results because each tree is not susceptible to individual tree errors.

7 Processing categorical features:

We should have to process the categorical features of the data by encoding them from 0 to 1 so that it can be used later in the process and helps in the further completion of the process and model growth successfully. The use of Bootstrap Aggregation or bagging, which provides the randomness required to produce robust and uncorrelated trees, helps to ensure this uncorrelated behaviour. This model was chosen to account for the dataset's large number of features and to compare a bagging technique to the following gradient boosting methods.



Graph 2. Processing categorical features

6.TEST TRAIN AND SPLIT:

We need to split the columns after encoding it to categorical features into two variables know as X and Y (consider) and then perform the instruct and test operation on those variables. So that we get the instruct and test of both X and Y. Later we use the test and instruct variables to develop a model.

CHAPTER 7

RESULTS:

First of all, the when the html code is also executed then the below page is displayed and then the data should be given as the input then the put will be given in the figure will be shown below. The figure contains the snap of a web page that is been created with the help of html, CSS and JavaScript modules which create the below page which offers the user to input the values. Once the values are input in the blank spaces and need to fill the values and click on the button which further enables inner systems which calculates the values which needed for the desired output. Then comes the buying price of the vehicle in the showroom, then comes the blank where the input should be as the number of kilometers the vehicle has ran. Then the part of input is for the number of owners the vehicle has been owned. Then comes the type of fuel used by the vehicle as diesel or petrol or CNG. Then if you are a dealer or a individual who is verifying the values. Then the gear or transmission type of manual or automatic should be selected from dropbox. Then click on the calculate the selling price of the car button. Then your output is ready to be displayed.

Fig. 2. Snap of the page created without any input

The figure shown below shows the values given as input as shown below, the first input is the year in which the vehicle was bought. Then comes the buying price of the vehicle in the showroom, then comes the blank where the input should be as the number of kilometers the vehicle has ran. Then the part of input is for the number of owners the vehicle has been owned. Then comes the type of fuel used by the vehicle as diesel or petrol or CNG. Then if you are a dealer or a individual who is verifying the values. Then the gear or transmission type of manual or automatic should be selected from dropbox. Then click on the calculate the selling price of the car button. Then your output is ready to be displayed.

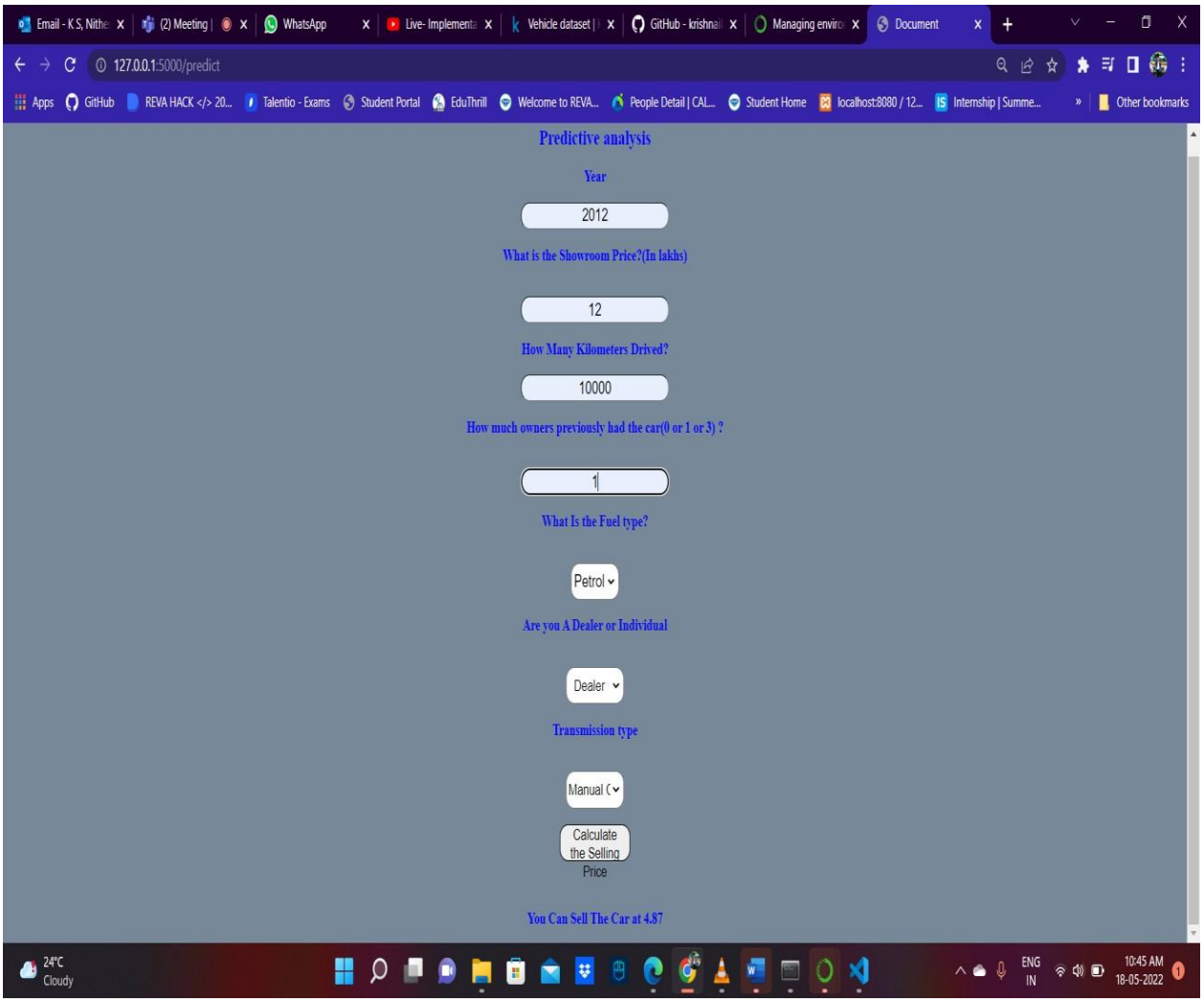


Fig. 3. Snap of the page with all filled data and the result.

CONCLUSIONS:

Due to the large number of attributes that must be considered for an accurate forecast, car price forecasting can be a difficult task. The collection and preprocessing of data is a crucial step in the forecasting process. In this study, PHP scripts were developed to integrate, standardize, and clean data in order to reduce noise in machine learning algorithms.

Although data cleaning is one of the processes that improves forecast performance, it is insufficient for complex data sets like the one used in this study. Initially when the process of the project was started the machine learning algorithms only returned the accuracy of less than 50%. So later on when some changes were made to algorithms referring to some of renowned literatures regarding the concept, and also when we tried to the data set from initially some kb's to more than 10 MB which also helped our algorithm to be much more accurate than it at the start of the project, so by all these improvements we were able to increase the accuracy to more than 80%. This is a significant improvement over using a single machine learning method. However, the proposed system has the disadvantage of consuming significantly more computational resources than a single machine learning algorithm. Although this system has shown remarkable results in the car price forecasting problem, our goal for future research is to see how well it works with different data sets. We'll compare our test data to used car data from eBay and OLX to confirm the proposed approach.

REFERENCES

1. Agencija za statistiku BiH. (n.d.), retrieved from: http://www.bhas.ba . [accessed July 18, 2018.]
2. Listiani, M. (2009). Support vector regression analysis for price prediction in a car leasing application (Doctoral dissertation, Master thesis, TU Hamburg-Harburg).
3. Richardson, M. S. (2009). Determinants of used car resale value.Retrieved from: shttps://digitalcc.coloradocollege.edu/islandora/object
4. Wu, J. D., Hsu, C. C., & Chen, H. C. (2009). An expert system of price forecasting for used cars using adaptive neuro- fuzzy inference. Expert Systems with Applications, 36(4), 7809-7817.
5. Du, J., Xie, L., & Schroeder, S. (2009). Practice Prize Paper—PIN Optimal Distribution of Auction Vehicles System: Applying Price Forecasting, Elasticity Estimation, and Genetic Algorithms to Used-Vehicle Distribution. Marketing Science, 28(4), 637-644.
6. Gongqi, S., Yansong, W., & Qiang, Z. (2011, January). New Model for Residual Value Prediction of the Used Car Based on BP Neural Network and Nonlinear Curve Fit. In Measuring Technology and Mechatronics Automation (ICMTMA), 2011 Third International Conference on (Vol. 2, pp. 682-685). IEEE.
7. Pudaruth, S. (2014). Predicting the price of used cars using machine learning techniques. Int. J. Inf. Comput. Technol, 4(7), 753-764.

Used car Price Prediction

ORIGINALITY REPORT

13%	6%	3%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.researchgate.net Internet Source	2%
2	Submitted to BITS, Pilani-Dubai Student Paper	2%
3	github.com Internet Source	2%
4	Submitted to Coventry University Student Paper	1%
5	Submitted to The University of Wolverhampton Student Paper	1%
6	Submitted to Istanbul Aydin University Student Paper	1%
7	Samed Jukic, Muzafer Saracevic, Abdulhamit Subasi, Jasmin Kevric. "Comparison of Ensemble Machine Learning Methods for Automated Classification of Focal and Non- Focal Epileptic EEG Signals", Mathematics, 2020 Publication	1%

8	Submitted to University of Greenwich Student Paper	1%
9	ijsrcseit.com Internet Source	1%
10	Submitted to The University of the South Pacific Student Paper	<1%
11	Submitted to Harrisburg University of Science and Technology Student Paper	<1%
12	zeen.com Internet Source	<1%
13	nerd2geek.blogspot.com Internet Source	<1%
14	www.ijaerd.co.in Internet Source	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

Used Cars Price Prediction using Machine Learning Models

Meenakshi S A
Associate Professor
School of CSE
meenakshi.sa@reva.edu.in

D S Aumkaareshwar
School of CSE
r18cs103@cit.reva.edu.in

Blesson E
School of CSE
r18cs085@cit.reva.edu.in

C Y Revanth Raju
School of CSE
r18cs099@cit.reva.edu.in

David G B
School of CSE
r18cs106@cit.reva.edu.in

Abstract: In a car price prediction there is a substantial number of distinct properties are examined for accurate predictions. The data was available in Kaggle which is updated regularly, we use that within the sort of CSV as input for the project. to make a model for predicting the worth of used cars, we apply three machine learning methods Random Forest Regressor, simple regression, and Decision tree. However, the mentioned methods were applied to figure as together.

I. INTRODUCTION

The cost of a new car in the industry is given by vehicle manufacturers with some extra taxes added by government. So, a customer who needs to buy a car or bike because it's controlled by many factors. But due to many factors such as lack of funds, and increased prices many people cannot afford a car or bike. So, the next thought of these people is buying a used vehicle i.e., the used cars and bikes, but it's important to know the actual market value of both buying and selling. So, there is a necessity for a second-hand vehicle prediction model that predicts the price of used cars and bikes appropriately. We select the finest model in terms of both train and test data and save the same model in the pickle file using the pickle function. We write a snippet of code to connect both pickle and python files so that it could be used to deploy. In this, we create a website using HTML, CSS, and Bootstrap to take the user input and predict the price of the used vehicle appropriately.

The following are the variables used:

Year: The year in which car is bought.
Selling price: Latest selling price of used car.
Present Price: The present value of car
Kms Driven: Total kms driven by the car.
Fuel type: The kind of fuel used in car.
Transmission: Manual or Automatic transmission.
Owners: Number of owners who used the cars.

II. LITERATURE SURVEY

Underfitting and overfitting picture after we create us required models. The models can be very biased to the training data and do not perform well on the test dataset. this is called overfitting. Likewise, the models do not take into consideration all the variances present in the population and perform poorly on a test data set, it is called underfitting. a balance must be achieved, which gives in the concept called Bias-Variance tradeoff. the choice of attribute plays a important role in influencing both the bias and variance to the statistical model. Robert Tibshirani [3] came up with a idea of replacement method which he called Lasso, which reduces the residual sum of squares. This returns a subset of attributes which is included in multivariate analysis to remove the minimal error rate. Decision trees has a problem, which is overfitting if they're not shrunk. However, hypothesis is tested using ANOVA is used to check whether the various types of errors really differ from each another.

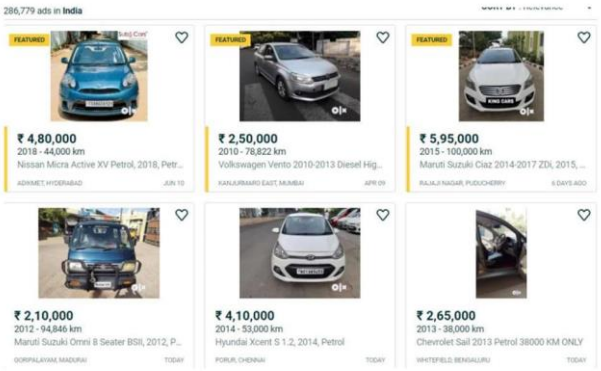


Figure 1: Existing websites of selling cars

III. PROPOSED MODEL

Data was collected from the website called Kaggle.com where a vehicle dataset from car Dekho is provided for buying and selling cars.
The attributes were used for each car: Selling Price, Kilometers are driven, Car Name, Year, Present or the Current Price, Fuel Type: petrol, diesel, or Compressed Natural Gas, Transmission Type: Automatic or Manual, Owners (No. of previous owners).

The cars which are available or mentioned in the dataset are taken under consideration in this project. Further, some modifications of the data were done with null entities to maintain homogeneity in the dataset so it could not affect the prediction model.

	Car_Name	Year	Selling_Pri	Present_Pr	Kms_Drive	Fuel_Type	Seller_Typ	Transmissi	Owner
2	ritz	2014	3.35	5.59	27000	Petrol	Dealer	Manual	0
3	sx4	2013	4.75	9.54	43000	Diesel	Dealer	Manual	0
4	ciaz	2017	7.25	9.85	6900	Petrol	Dealer	Manual	0
5	wagon r	2011	2.85	4.15	5200	Petrol	Dealer	Manual	0
6	swift	2014	4.6	6.87	42450	Diesel	Dealer	Manual	0
7	vitara brez	2018	9.25	9.83	2071	Diesel	Dealer	Manual	0
8	ciaz	2015	6.75	8.12	18796	Petrol	Dealer	Manual	0
9	s cross	2015	6.5	8.61	33429	Diesel	Dealer	Manual	0
10	ciaz	2016	8.75	8.89	20273	Diesel	Dealer	Manual	0
11	ciaz	2015	7.45	8.92	42367	Diesel	Dealer	Manual	0
12	alto 800	2017	2.85	3.6	2135	Petrol	Dealer	Manual	0
13	ciaz	2015	6.85	10.38	51000	Diesel	Dealer	Manual	0
14	ciaz	2015	7.5	9.94	15000	Petrol	Dealer	Automatic	0
15	ertiga	2015	6.1	7.71	26000	Petrol	Dealer	Manual	0

Figure 2: Dataset used in the project.

Dataset used in the project. As we use this data set for creating the model, we first do some data visualizations so to understand the columns properly. Later we divide the columns into X and Y variables i.e., according to what we need to predict from what values. Later we train and test this, so we get the trained and

tested variables for further analysis.
We import the modules and create a dictionary file of imported modules along with their parameters and call each of them and fit it with test and training data which is done by RandomSearchCv.

Workflow of the project:

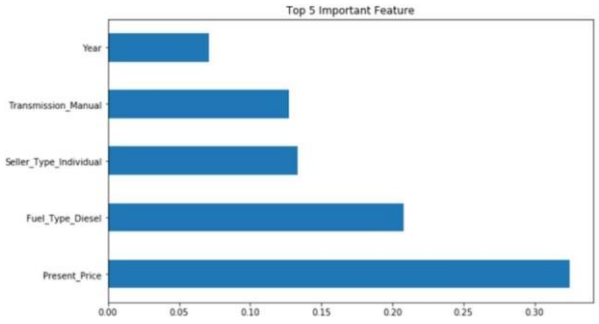


Figure 3: Important features in the project.

Key features in the project. As we need to know the important features in the data set so that we can find the variable We can save the model using pickle function. In this, we connect the file with HTML using a Python language and we deploy this using a Flask module. Data was collected from the website called Kaggle.com where a vehicle dataset from car Dekho is provided for buying and selling cars. The attributes were used for each car: Selling Price of the car, Kilometers driven, Car Name, Year, Present or the Current Price, Fuel Type: Diesel, Petrol, or Compressed Natural Gas, Transmission Type: Automatic or Manual, Owners (No. of previous owners). The cars which are available or mentioned in the dataset are taken under consideration in this project. Further, some modifications of the data were done with null entities to maintain homogeneity in the dataset so it could not affect the prediction model. Dataset used in the project. As we use this data set for creating the model, we first do some data visualizations so to understand the columns properly. Later we divide the columns into X and Y variables i.e., according to what we need to predict from what values. Later we train and test this, so we get the trained and tested variables for further analysis. We import the modules and create a dictionary file of imported modules along with their parameters and call each of them and fit it with test and training data which is done by RandomSearchCv.

Workflow of the project Important features in the project. As we need to know the important features in the data set so that we can find the variable We can save the model using pickle function.

In this, we connect the file with HTML using a Python language and we deploy this using a Flask module. Data was collected from the website called Kaggle.com where a vehicle dataset from car Dekho is provided for buying and selling cars.

IV. SYSTEM DESIGN FOR PROPOSED MODEL

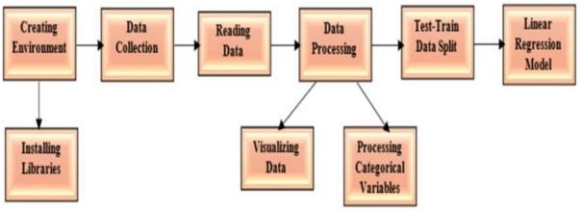


Figure 4: System Design Architecture

THE ARCHITECTURE COMPRISES OF THE FOLLOWING METHODS:

CREATING ENVIRONMENT

As we build different applications in our system, and it needs different python versions so we create an environment for each project and work on the same project in that environment so that there would be no colloidal of versions with each other.
After creating an environment, we can freeze the libraries using the function called pip freeze requirement.txt so we get a file of all required libraries and their versions to download we can download by opening the anaconda prompt and running the version as pip install library name version of that library

DATA COLLECTION

we can collect the dataset from the website Kaggle.com and we use the same data for our project to predict the price of vehicles this data is taken from car Dekho which is popular for selling used cars

READING DATA

we read the data called csv comma separated values which is in the form of rows and columns we read this dataset using read.csv function

DATA REPROCESSING

We need to pre-process the data before use so that it doesn't contain any null values which may affect the program accuracy. So, we pre-process the null values by removing the row or handling the null values by visualizing data. We can visualize the data of different variables so that we can understand the data more clearly and use them according to the need. Further we can also visualize the different important features of all the variables so that we can know how important the variable depends on the model processing and categorical attributes. We need to process the categorical attributes by encoding them to 0 and 1 so that they can be used later in the model development.

TEST TRAIN DATA BUILT

We need to divide the columns after encoding them to categorical features into two variables known as x and y consider and then perform the train and test operation on those variables so that we get the train and test of both x and y later we use the test and train variables to develop a model.

CREATING A MODEL

We need to import all the necessary models so that we can test and train on that model first we will create a dictionary where we import all the models such as linear regression random forest regressor and decision tree later we store all the models and their parameters accordingly in the dictionary and use them to fit one by one accordingly and we can find the test and train score of all the models.

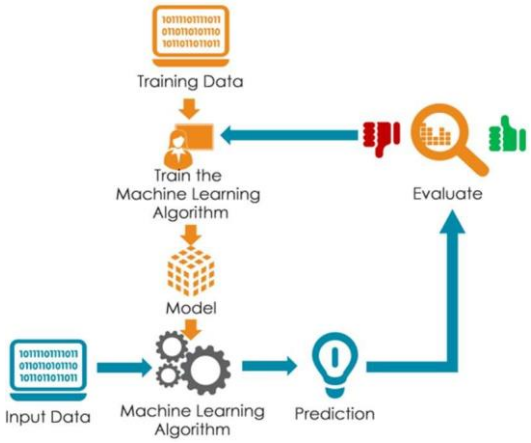


Figure 5: Workflow of model

V. ALGORITHM

Basic working algorithm of Used Vehicle Price prediction is described below where the other things are discussed deeply further.

Basic functionality of website:

STEP 1: Start
STEP 2: Onboarding [display the user interface].
STEP 3: select the car or bike you need to predict.
STEP 4: It displays the car image.
STEP 4: Enter the required text fields.

- ➢ Purchased Year
- ➢ Show room price (in lakhs).
- ➢ How many kilometers driven.
- ➢ How many owners previously owned the cars.
- ➢ What is the fuel type?
- ➢ Are you a dealer or Individual?
- ➢ Transmission Type.
- ➢ Estimated Price.

STEP 5: Enter predict button.
STEP 6: you get the predicted price of the car in lakhs.
STEP 7: Stop.

Model creation algorithm:

STEP 1: Start
STEP 2: Read the csv file.
STEP 3: Pre-process the data.
STEP 4: Visualizing the dataset using data analysis.
STEP 5: Encode the categorical features.
STEP 6: Dividing the dataset accordingly to test and train.
STEP 7: Train and test the columns divided.
STEP 8: Import all the models.
STEP 9: Fit using RandomForestCv.
STEP 10: Display the accuracy of train and test of models.
STEP 11: Select the highest accuracy model.
STEP 12: Save the model using Pickle.
STEP 13: STOP

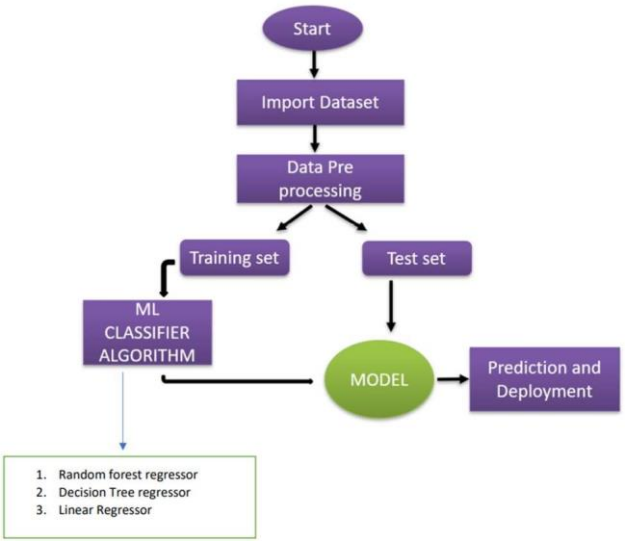


Figure 6: Implementation of Flowchart

Linear Regression

Linear regression is a regression model where we find the least-square errored values of all the points, and we find the rsquare value of that and we find the probability of r square value. So, that we can fit the train and test score value of the dataset. The objective is to explain variation to the response variables that are attributed to variation in explanatory variables. Linear regression analysis is applied to quantify the strength of the relationship between explanatory variables and response. it is also used to determine whether explanatory variables may have no linear relationship with the response or to identify the subsets of explanatory variables contain useless information about the response.

Random forest regressor

Random forest regressor is of multiple decision trees where we aggregate all the rows and columns such that we find the decision tree values of each decision tree. So, we bootstrap all the values and find the mean of all, and we get the predicted value from the regressor.

VI. CONCLUSION

Car price prediction is challenging task due to the many numbers of properties should be considered for the approximate prediction. The important step in the process is collection, pre-processing of data. In this PHP scripts were built to normalize and clean the data to remove noise for machine learning algorithms. Data cleaning process increases the prediction performance, still insufficient for the complex datasets. Applying a single machine model on the dataset there will be a accuracy was less than 40%. The ensemble of multiple machine learning models has been proposed and this combination of Machine learning methods has accuracy of 90%. This is significant improvement compared to single machine learning

model approach. However, the disadvantage of the proposed system is that it consumes more resources than single machine learning model. However, this system has achieved high performance in price prediction our aim for the future research is to train this system to work successfully with all datasets.

VII. REFERENCES

[1] Agencija za statistiku BiH. (n.d.), retrieved from: http://www.bhas.ba . [accessed July 18, 2018.]

[2] Listiani, M. (2009). Support vector regression analysis for price prediction in a car leasing application (Doctoral dissertation, Master thesis, TU Hamburg-Harburg).

[3] Richardson, M. S. (2009). Determinants of used car resale value. Retrieved from: shttps://digitalcc.coloradocollege.edu/islandora/object /coccc%3A1346 [accessed: August 1, 2018.]

[4] Wu, J. D., Hsu, C. C., & Chen, H. C. (2009). An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference. Expert Systems with Applications, 36(4), 7809-7817.

[5] Du, J., Xie, L., & Schroeder, S. (2009). Practice Prize Paper—PIN Optimal Distribution of Auction Vehicles System: Applying Price Forecasting, Elasticity Estimation, and Genetic Algorithms to Used-Vehicle Distribution. Marketing Science, 28(4), 637-644.

[6] Gongqi, S., Yansong, W., & Qiang, Z. (2011, January). New Model for Residual Value Prediction of the Used Car Based on BP Neural Network and Nonlinear Curve Fit. In Measuring Technology and Mechatronics Automation (ICMTMA), 2011 Third International Conference on (Vol. 2, pp. 682-685). IEEE.

[7] Pudaruth, S. (2014). Predicting the price of used cars using machine learning techniques. Int. J. Inf. Comput. Technol, 4(7), 753-764.

[8] Noor, K., & Jan, S. (2017). Vehicle Price Prediction System using Machine Learning Techniques. International Journal of Computer Applications, 167(9), 27-31. [9] Auto pijaca BiH. (n.d.), Retrieved from: https://www.autopijaca.ba. [accessed August 10, 2018].

[10] Weka 3 - Data Mining with Open Source Machine Learning Software in Java. (n.d.), Retrieved from: https://www.cs.waikato.ac.nz/ml/weka/. [August 04, 2018].

[11] Ho, T. K. (1995, August). Random decision forests. In Document analysis and recognition, 1995., proceedings of the third international conference on (Vol. 1, pp. 278-282). IEEE.

[12] Russell, S. (2015). Artificial Intelligence: A Modern Approach (3rd edition). PE.

[13] Ben-Hur, A., Horn, D., Siegelmann, H. T., & Vapnik, V. (2001). Support vector clustering. Journal of machine learning research, 2(Dec), 125-137.

[14] Aizerman, M. A. (1964). Theoretical foundations of the potential function method in pattern recognition learning. Automation and remote control, 25, 821- 837.

[15] 3.2.4.3.1. sklearn.ensemble.RandomForestClassifier — scikit-learn 0.19.2 documentation. (n.d.). Retrieved from: http://scikitlearn.org/stable/modules/generated/sklearn.ensemble .RandomForestClassifier.html [accessed: August 30, 2018].

[16] Used cars database. (n.d.) Retrieved from: https://www.kaggle.com/orgesleka/used-carsdatabase. [accessed: June 04, 2018].

[17] OLX. (n.d.), Retrieved from: https://olx.ba. [accessed August 05,2018]