A MINI PROJECT REPORT

ON

"SECOND-HAND CAR PRICE PREDICTOR"

Submitted to

SAVITRIBAI PHULE PUNE UNIVERSITY

in completion of
SKILL DEVELOPMENT LABORATORY
(T.E Computer Engineering)

BY

Smit Mahajan	Exam No:
Neha Goundadkar	Exam No:
Mohit Patil	Exam No:
Amit Raj	Exam No:

Sinhgad Institutes

Department of Computer Engineering
Sinhgad College of Engineering, Pune-41
Accredited by NAAC with grade 'A'

YEAR 2019-20

CERTIFICATE

Sinhgad Technical Education Society,
Department of Computer Engineering
Sinhgad College of Engineering, Pune-41
Accredited by NAAC with grade 'A'

Sinhgad Institutes

"SECOND-HAND CAR PRICE PREDICTOR"

Submitted to

SAVITRIBAI PHULE PUNE UNIVERSITY

in completion of

SKILL DEVELOPMENT LABORATORY
(T.E Computer Engineering)

BY

Smit Mahajan	Exam No:
Neha Goundadkar	Exam No:
Mohit Patil	Exam No:
Amit Raj	Exam No:

Prof. N. A. Mhetre.
Internal Guide
Department of Computer Engineering
Engineering

Prof. M. P. Wankhade
Head of Dept.
Department of Computer

Dr. S.D. Lokhande
Principal
SCOE, Pune

CONTENTS

2.Acknowledgement

Every orientation works has an imprint of many people and it becomes our duty to express deep gratitude for the same.

During the entire duration of preparation for this Dissertation, we received endless help from a number of people and feel that this report would be incomplete if we do not convey graceful thanks to them. This acknowledgement is a humble attempt to thanks all those who were involved in the project work and were of immense help to us.

First and foremost we take the opportunity to extend our deep heartfelt gratitude to our guide **Prof Nalini Mhetre** for guiding us throughout the entire project and for her kind and valuable suggestions, without which this idea won't have executed. We also humbly thank **Prof M. P Wankhede, Head of Department of Computer Engineering SCOE, Pune** for his indispensable support, his priceless suggestions and for his valuable time

3.Abstract

Second-hand car price predictor focuses on predicting a fair price for one's used car based on the car's condition.It is difficult for an owner to get a good deal for one's used car due to unavailability of a car dealership company or lack of information about such existing dealer.

This issue is resolved using machine learning for predicting best price of a car according to it's current condition which includes it's power,year of registration,brand,model,etc.Necessary information regarding factors affecting the price is provided to seller using appropriate graphs.

The end user is notified with an email message which includes his/her car price and which is further useful for contact support and booking an inspection for one's car. Along with the present car price, price for upcoming 3^{rd} and 5^{th} year is also predicted and shown to the seller with the help of GUI.

4. List of figures

5. INTRODUCTION

5.1.Background and basics

5.1.1. Second-hand car price predictor– Introduction

It is a project developed in order to help sellers contact car dealers and vice versa for getting the best deal for one's car. The price is predicted on the basis of details entered by the user/seller on the GUI provided using machine learning algorithm. The user is also notified with an email message for booking a free car inspection.

5.1.2 Machine Learning

Machine learning is a field of artificial intelligence that uses statistical techniques to give computer systems the ability to "learn" (e.g., progressively improve performance on a specific task) from data, without being explicitly programmed. Within the field of data analytics, machine learning is a method used to devise complex models and algorithms that lend themselves to prediction; in commercial use, this is known as predictive analytics. These analytical models allow researchers, data scientists, engineers, and analysts to "produce reliable, repeatable decisions and results" and uncover "hidden insights" through learning from historical relationships and trends in the data.

Types of Machine Learning Techniques:

- Supervised Learning
  - Regression
    - RidgeCV
    - Linear Regression
    - Random Forest Regression
- Unsupervised Learning
  - Clustering
    - K-means
    - Heirarchical.
    - Gaussian Mixture
  - Classification
    - Support Vector Machines
    - Naïve Bayes
    - Nearest Neighbour

5.1.3 Data Analysis

Data analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, while being used in different business, science, and social science domains. Analysis refers to breaking a whole into its separate components for individual examination. Data analysis is a process for obtaining raw data and converting it into information useful for decision-making by users. Data is collected and analyzed to answer questions, test hypotheses or disprove theories.

5.2.Problem statement

To predict price of a car on the basis of following details taken from user:

1. Kilometers run.
2. Power.
3. Year of registration.
4. Month of registration.
5. Damage.
6. Brand.
7. Model.

5.2.1.Scope statement:

Price of a used car is calculated using machine learning algorithm for given details of a car.An email message is sent to the car seller consisting of predicted price of his/her car along with car inspector’s name.

6. PROJECT PLANNING AND MANAGEMENT

6.1 Software requirements

6.1.1 Python 3.7 64 bit – For preparing estimator object and training the model.

6.1.2 Django 2.5 – Framework for deploying the model on web page.

6.1.3 Plotly – For constructing graphs using javascript.

6.1.4 Pandas – For reading and refining the dataset and construct a dataframe..

6.1.5 Sklearn – For importing machine learning algorithms and metrics.

6.1.6 Editor-HTML, javascript – For creating GUI.

6.1.7 Browser – GUI platform.

6.2 Hardware requirements

6.2.1 4GB RAM

6.2.2 Intel core i5-7th gen

6.2.3 Node component(Laptop).

7. ANALYSIS AND DESIGN

7.1. Use Case Diagram

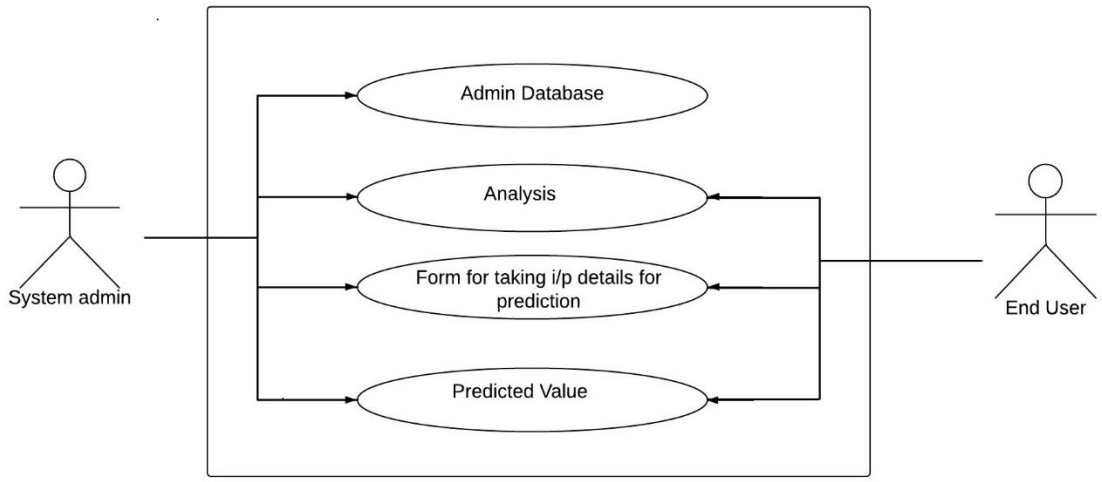

Fig 5.1.1 Use case diagram

A use case diagram is a **dynamic** or behavior diagram in UML. Use case diagrams model the functionality of a system using actors and use cases. Use cases are a set of actions, services, and functions that the system needs to perform

7.2. Activity diagram

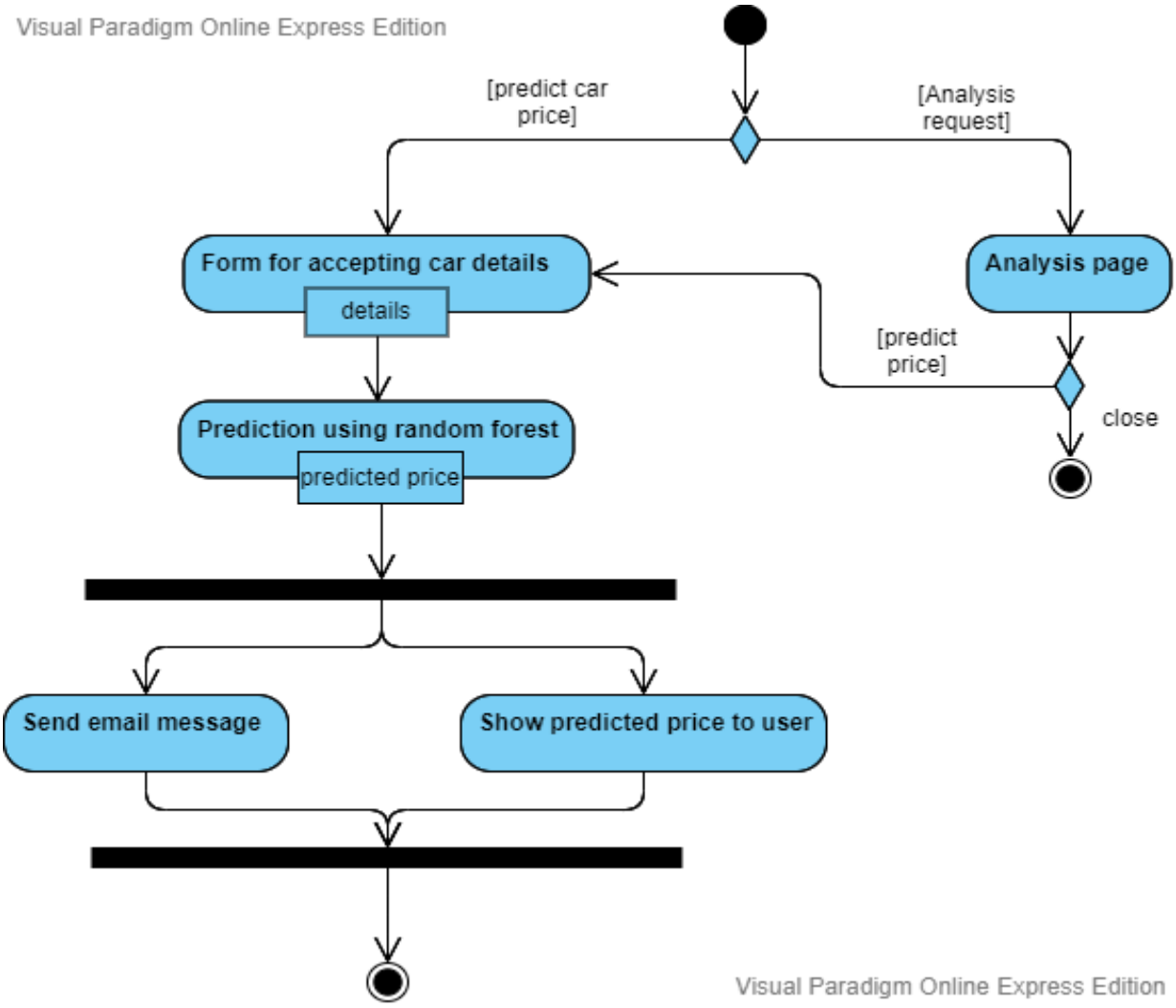

Fig 7.2.1 Activity diagram

The above activity diagram shows the entire project flow. The input and output parameter to an action if any, is displayed just at the top and bottom of action respectively.

7.3. System architecture

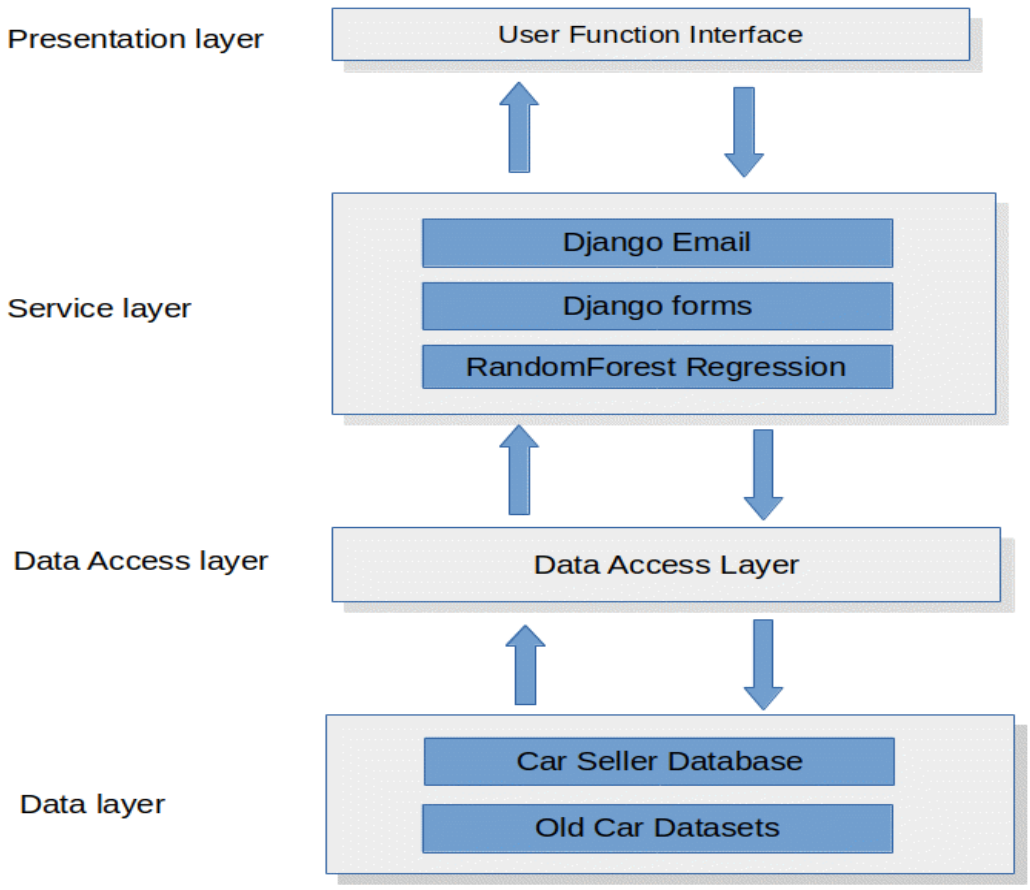

Fig 7.3.1 System architecture

The above shown system architecture shows the hierarchical representation of the project flow. Various levels of abstraction are shown where data layer is having lowest abstraction level and GUI is having highest abstraction level.

8. IMPLEMENTATION & CODING

8.1 METHODOLOGY

8.1.1 Gathering data

The dataset is taken from Kaggle which consists of data of about 370,000 sold cars in Germany. The dataset contains 20 attributes and is in the form of CSV.

8.1.2 Data Preparation

Unnecessary attributes which do not contribute to car's price were found by plotting correlation with price. Then such useless columns were dropped.

The data consisted of Null values in many attributes which were handled using 'fillna' function in pandas. Moreover, the dataset contained string format attribute values. As the regression algorithms has only numeric domain, these string values were encoded to numeric values using 'ColumnTransformer' along with 'OneHotEncoder'.

8.1.3 Choosing the model

Among the multiple choices available in the sklearn library for choosing the model, we tested Random Forest Regressor, Linear Regression, RidgeCV, Decision Tree Regression, Gradient Boost Regression.

The data was split into training and testing sets. Then the respective model was trained on the training dataset using 'fit' function. The models were evaluated using 'r2_score' metrics in sklearn for accuracy.

The accuracies of the above models were as follows:

Random Forest Regressor – 86.4
Linear Regression – 57.2
RidgeCV – 69.8

Since the highest accuracy was achieved by Random Forest, this was chosen as the preferred model.

8.1.4 Training

The data was split into train test split like in the previous step into training and test data in ratio 80-20. The model is fit to the training data.

8.1.5 Evaluation

To check whether the used model is accurate, predict function is used on the test data. The accuracy score is measured with the function r2 score.
The accuracy on the test data came out to be : 86.4

8.1.6 Prediction

The values required as input from users are Kilometers run, power, brand of car, model of car, damage(yes/no), year of registration, month of registration which are accepted in the GUI. The price of car is predicted from the pretrained model.

8.2 DATASET

The dataset consists of data of over 370,000 sold cars in Germany.

Every country's dataset contains the following colums –

1. dateCrawled.
2. name
3. seller
4. OfferType
5. price
6. abtest
7. vehiclePrice
8. yearOfRegistration
9. gearbox
10. powerPS
11. model
12. kilometer
13. monthOfRegistration
14. fuelType
15. brand
16. notRepairedDamage
17. dateCreated
18. nrOfPictures
19. postalCode
20. lastSeenOnline

8.3 GUI DESIGN AND SCREENSHOTS

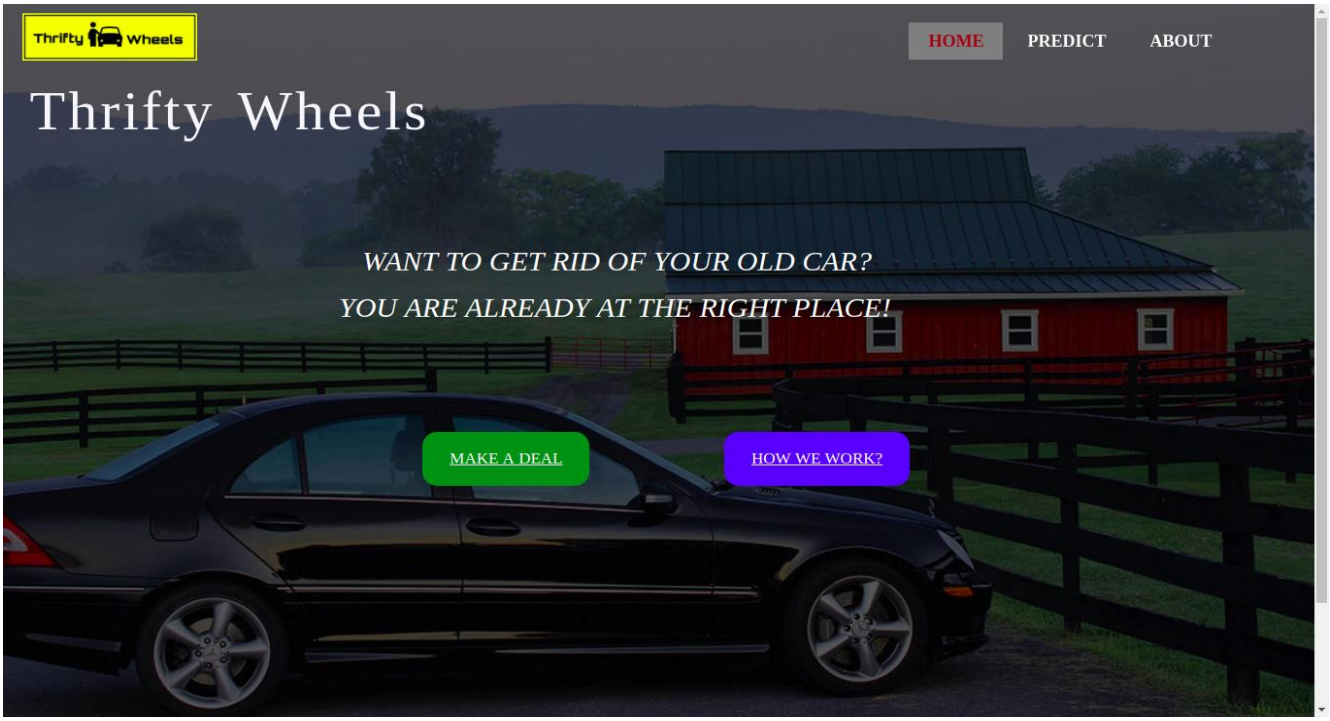

Fig 8.4.1 Homegage

The below image shows the homepage of the project GUI. The prediction button takes the user to the prediction page. The 'HOW WE WORK' button takes the user to the analysis page.The above GUI is created using Django framework and HTML. All the html files are located in a directory named static. The Django views.py contains functions for respective web page.

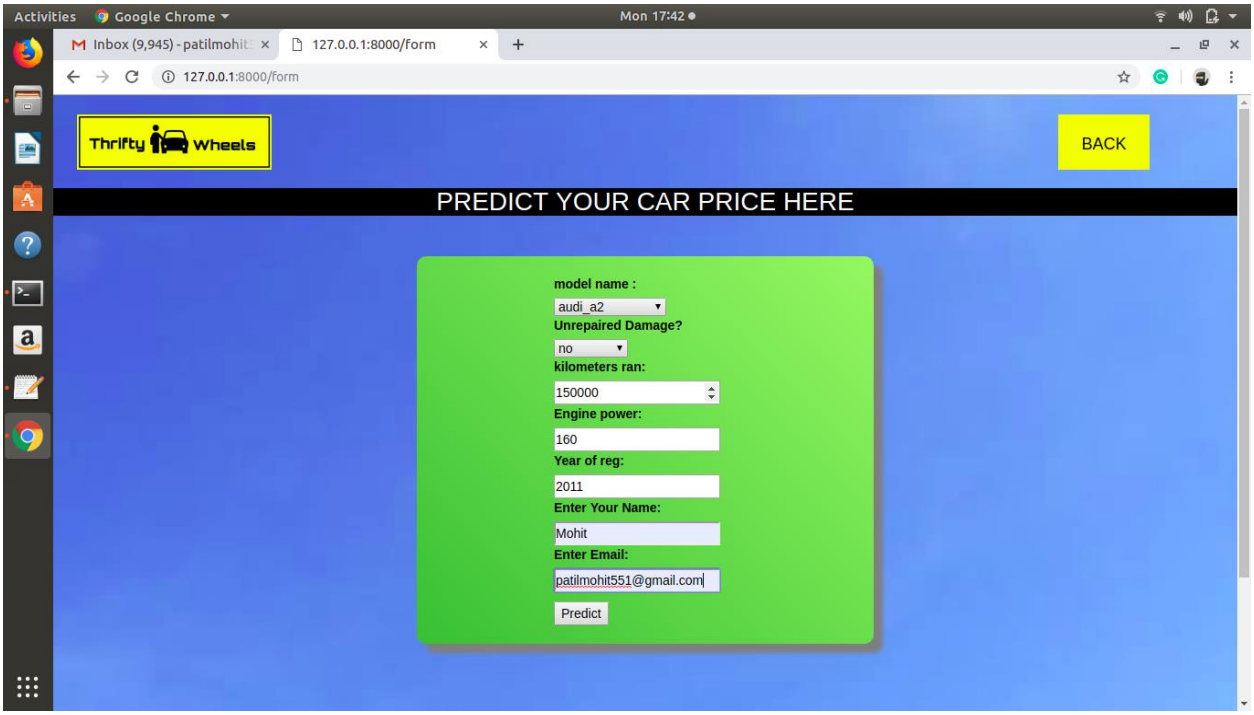

Fig. 8.4.2 GUI-prediction form

The above screenshot shows the ML prediction page. The user should input the required fields in the form. After clicking on the “Predict ” button, the predicted price of the car is shown on the next page.

9. RESULTS & DISCUSSIONS

9.1 VISUALIZATION OF RESULTS

The below shown graph shows the variation of price of car on the basis of car power. The graph is constructed using plotly and javascript.

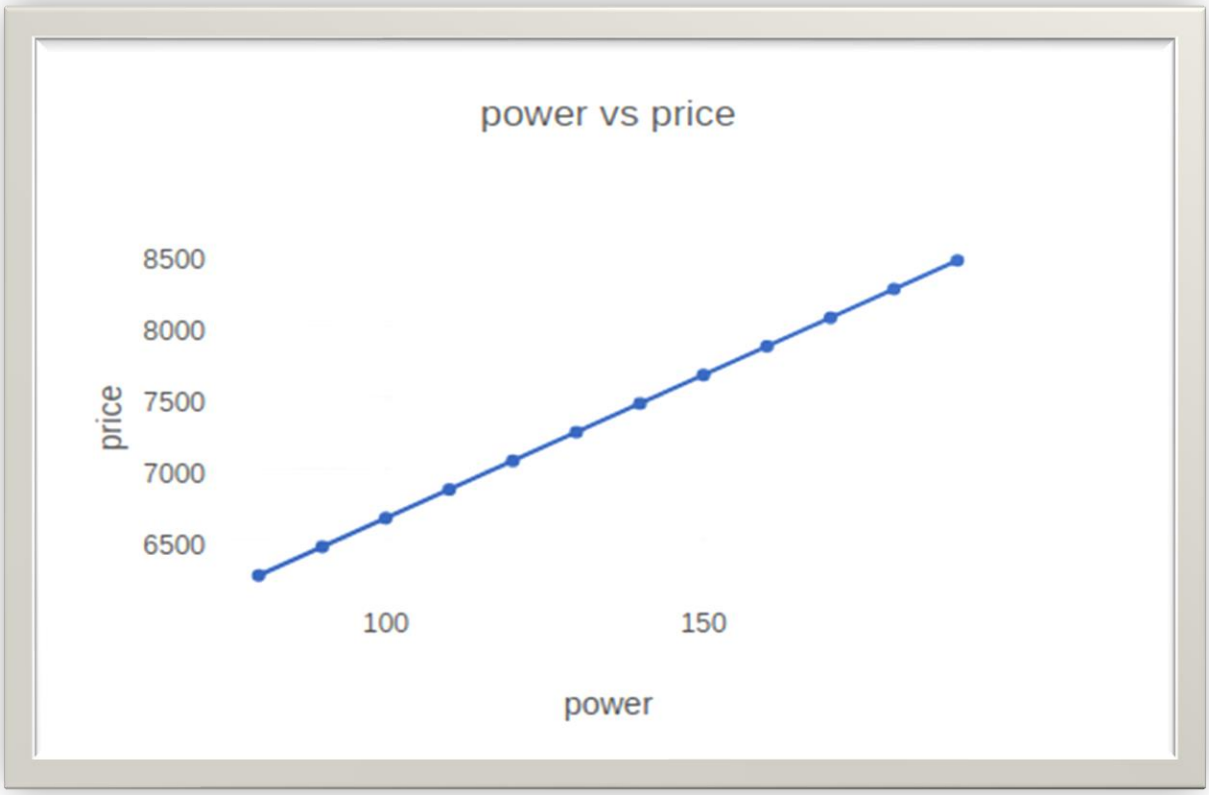

Fig 9.1.1 Power vs Price

The above graph depicts that with an increase in power of car which is measured in horsepower, the price of car increases linearly.

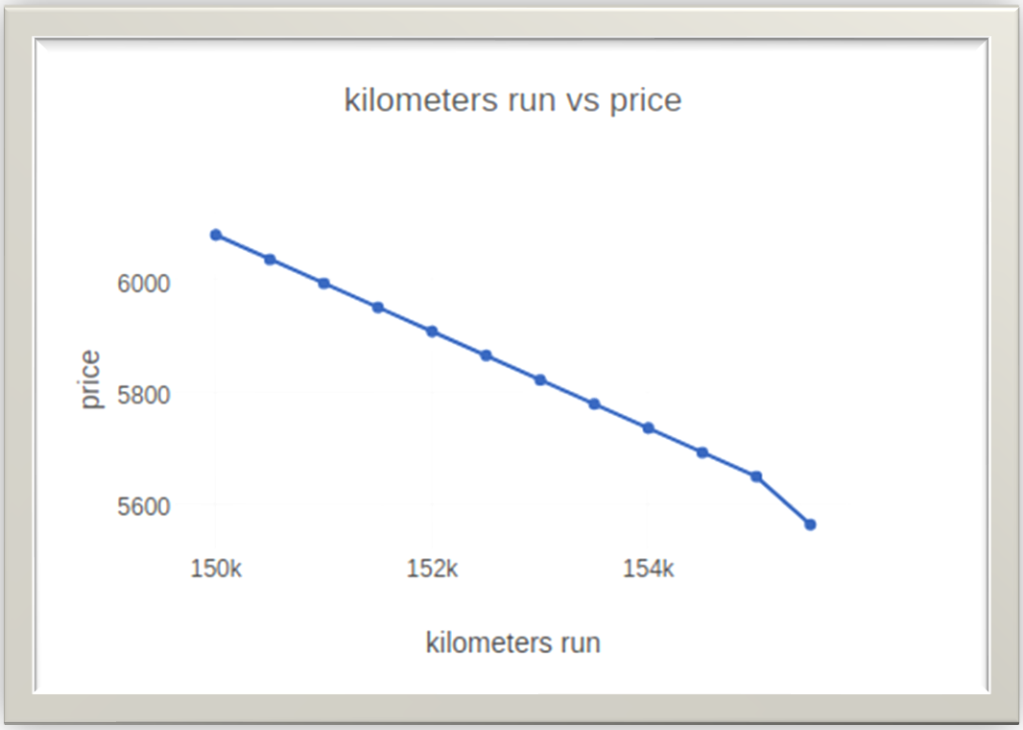

Fig 9.1.2 Kilometers vs price

The above plotted graph shows the variation of price of car with kilometers run parameter. It can be seen that with an increase in kilometer count, the price of car decreases.

Fig 9.1.3 YearofRegistration vs Price

The above graph show variation of price of car with year of registration of car. As the car gets older it's price gradually decrease. Year of Registration affects the overall price by 20%.

Fig 9.1.4 Predicted value page.

The above page shows the GUI used to represent the predicted value to user.

10. CONCLUSION

Using the data from dataset along with random forest regressor, the price of used car is predicted. The seller is notified with an email about the car's price. The Seller is also provided with statistical visualization in the form of graphs. Seller's data is stored in database for future use. The database access is controlled using user-admin password. The user/seller data which is saved for future purpose can be used to notify the users about deals on festivals.

The analysis page in project GUI helped user to understand about the variation of price of one's car with respect to kilometers run, power, brand, model, year of registration and damage if any. The analysis page GUI contained statistical visualization in the form of graphs for easy understanding for the user. The seller is also provided with the predicted price of his/her car in the next 3 and 5 years respectively.

11. REFERENCES

1. https://www.kaggle.com/orgesleka/used-cars-database.(Dataset)
2. https://docs.djangoproject.com
3. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
4. https://www.w3schools.com/html/
5. https://www.tutorialspoint.com/managerial_economics/regression_technique.htm
6. https://tutorial.djangogirls.org/en/django_start_project/
7. https://www.freepik.com/
8. http://web4.cs.ucl.ac.uk/staff/D.Barber/textbook/091117.pdf, by D.Barber.
9. Breiman. L. (1998b). Randomizing outputs to increase prediction accuracy. Technical Report 518, May 1, 1998, Statistics Department, UCB (in press, Machine Learning).
10. https://w3layouts.com/s