PROJECT REPORT
ON

CAR RESALE VALUE PREDICTION

AT

ज्ञानेन प्रकाशते जगत्
INDUS
UNIVERSITY

In the partial fulfillment of the requirement

for the degree of

Bachelor of Technology

in

Information Technology

PREPARED BY

DHWANI RAJEEV NIMBARK (IU1741220018)

AKSHAT NILESHBHAI PATEL (IU1741220021)

UNDER THE GUIDANCE OF

Internal Guide
Prof. Sejal Thakkar
Assistant Professor,
Department of Computer Engineering,
I.T.E, Indus University,
Ahmedabad

SUBMITTED TO
INSTITUTE OF TECHNOLOGY AND ENGINEERING
INDUS UNIVERSITY CAMPUS, RANCHARDA, VIA-THALTEJ
AHMEDABAD-382115, GUJARAT, INDIA,
WEB: www.indusuni.ac.in
MAY 2021

CANDIDATE’S DECLARATION

I declare that final semester report entitled “**CAR RESALE VALUE PREDICTION**” is my own work conducted under the supervision of the guide **MRS.SEJAL THAKKAR.**

I further declare that to the best of my knowledge, the report for B. Tech final semester does not contain part of the work which has been submitted for the award of B. Tech Degree either in this university or any other university without proper citation.

Candidate’s Signature

DHWANI RAJEEV NIMBARK (IU1741220018)
AKSHAT NILESHBHAI PATEL (IU1741220021)

Guide: Sejal Thakkar
Assistant Professor,
Department of Computer Engineering,
Indus Institute of Technology and Engineering
INDUS UNIVERSITY– Ahmedabad,
State: Gujarat

INDUS INSTITUTE OF TECHNOLOGY AND ENGINEERING
COMPUTER ENGINEERING
2020 -2021

CERTIFICATE

Date: 17/05/2021

This is to certify that the project work entitled "**CAR RESALE VALUE PREDICTION"** has been carried out by **DHWANI RAJEEV NIMBARK & AKSHAT NILESHBHAI PATEL** under my guidance in partial fulfillment of degree of Bachelor of Technology in **INFORMATION TECHNOLOGY ENGINEERING (Final Year)** of Indus University, Ahmedabad during the academic year 2020 - 2021.

Sejal Thakkar
Assistant Professor,
Department of Computer Engineering,
I.T.E, Indus University
Ahmedabad

Seema Mahajan
Head of the Department,
Department of Computer Engineering,
I.T.E, Indus University
Ahmedabad

TABLE OF CONTENT

Title **Page No**

ABSTRACT

Used car resale market in India was marked at 24.2 billion US dollars in 2019. Due to huge requirement of used cars and lack of experts who can determine the correct valuation, there is an utmost need of bridging this gap between sellers and buyers. This project focuses on building a system that can accurately predict a resale value of the car based on minimal features like kms driven, year of purchase etc. without manual or human interference and hence it remains unbiased.

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

Abbreviation	Full-Form
SVR	Support Vector Regression
ML	Machine Learning
GBR	Gradient Boosting Regression
API	Application Programming Interface
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
SVM	Support Vector Machine

CHAPTER 1

INTRODUCTION

- **PROJECT SUMMARY**
- **PROJECT PURPOSE**
- **PROJECT SCOPE**
- **OBJECTIVE**
- **TECHNOLOGY AND LITERATURE OVERVIEW**

1.1 PROJECT SUMMARY

In this project we have used different algorithm with different techniques for developing Car resale value prediction system considering different features of the car. In a nutshell, car resale value prediction helps the user to predict the resale value of the car depending upon various features like kilometers driven, fuel type, etc.

1.2 PROJECT PURPOSE

The main idea of making a car resale value prediction system is to get hands-on practice for python using Data Science. Car resale value prediction is the system to predict the amount of resale value based on the parameters provided by the user. User enters the details of the car into the form given and accordingly the car resale value is predicted.

1.3 PROJECT SCOPE

The system currently can predict the resale value of a car with an RMSE (Root Mean Squared Error) of 42,000 INR. The system can only predict resale value for single model i.e., Maruti Swift Dzire as currently system is in development stage. Also, current system takes only few important features such as kms driven, fuel type, year of purchase and city to predict the price. Since the system is being developed for educational purposes, scope of the project is minimal. However, it can be extended to further models and features of a car as and when required.

1.4 OBJECTIVE

Car resale value prediction system is made with the purpose of predicting the correct valuation of used cars that helps users to sell the car remotely with perfect valuation and without human intervention in the process to eliminate biased valuation.

1.5 TECHNOLOGY AND LITERATURE OVERVIEW

1.5.1 LIBRARIES

Scikit-learn:

Scikit-learn is an open-source python library that implements a range of machine learning, preprocessing, cross-validation and visualization algorithms using a unified interface.

Important feature of Scikit-learn:

- Simple and efficient tool for data analysis and data mining. It features various regression, clustering and classification and algorithm including gradient boosting, SVM, k-means.
- Built on NumPy, SciPy and matplotlib.
- Reusable in various contents and everybody can access.

Pandas:

Pandas is one of the most popular Python libraries for data manipulation and analysis. Pandas is designed for quick and easy data manipulation, aggregation, and visualization. There are two data structure in pandas:

- ➢ **Series** – It handles and store data in one-dimensional data.
- ➢ **Data frame**– It handles and store data in two-dimensional data.

Important features of pandas:

- It provides large data structures and manipulating numerical tables and time series data.
- Pandas is a perfect tool for data wrangling.

1.5.2 TOOLS

- ➢ Jupyter Notebook
- ➢ PyCharm

1.5.3 LANGUAGE

Python:

Python is a high-level programming language and is widely being used among the developers' community. Python was mainly developed for emphasis on code readability, and its syntax allows programmers to express concepts in fewer lines of code. Python is a programming language that lets developers work quickly and integrate systems more efficiently.

1.5.4 LITERATURE OVERVIEW

A literature review of predicting used car prices
By Pranav Gadre

Determining whether the listed price of a used car is a challenging task, due to the many factors that drive a used vehicle's price on the market. The focus of this project is 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 consisting of the sale prices of different makes and models across cities in the United States. Our results show that Random Forest model and K-Means clustering with linear regression yield the best results, but are compute heavy.

Conventional linear regression also yielded satisfactory results, with the advantage of a significantly lower training time in comparison to the aforementioned methods.

Car Price Prediction using Machine Learning Techniques
By Enis Gegic, Becir Isakovic, Dino Keco

This paper provides knowledge to build a model for predicting the price of used cars in Bosnia and Herzegovina, we applied three machine learning techniques (Artificial Neural Network, Support Vector Machine and Random Forest). However, the mentioned techniques were applied to work as an ensemble. The data used for the prediction was collected from the web portal autopijaca.ba using web scraper that was written in PHP programming language. Respective performances of different algorithms were then compared to find one that best suits the available data set. The final prediction model was integrated into Java application. Furthermore, the model was evaluated using test data and the accuracy of 87.38% was obtained.

CHAPTER 2

LITERATURE SURVEY

- **INTRODUCTION OF SURVEY**
- **WHY SURVEY**

2.1 INTRODUCTION OF SURVEY

Survey is an efficient way of collecting information from a large number of respondents. Very large samplings are possible. Statistical techniques can be used to determine validity, reliability, and statistical significance. There is an economy in data collection due to the focus provided by standardized questions. Survey is flexible in the sense that a wide range of information can be collected. They can be used to study attributes, values, beliefs, and past behaviors. Only questions of interest to the researcher are asked, recorded, codified, and analyzed. Time and money is not spent on tangential questions. They can be administered from remote locations using mail, email or telephone.

FEATURES:

- Surveys are useful in describing the characteristics of a large dataset. No other method of observation can provide this general capability.
- Consequently, very large samples are feasible, making the results statistically significant even when analyzing multiple variables.
- A large area can be covered under survey in the available time and money.
- As in sample study few units are to be examined detailed study of the survey can be done.
- As few units are to be examined the survey work requires less time.
- Thus, in this way sample survey saves time.
- As few units are to be examined the survey work requires less money.
- Thus, in this way sample survey saves lots of money.

2.2 WHY SURVEY?

The literature review plays a very important role in the research process. It is a source from where research ideas are draw and developed into concepts and finally theories. It also provides the researcher a bird's eye view about the research done in that area so far. Depending on what is observed in the literature review, a researcher will understand where his/her research stands. Here in this literature survey, all primary, secondary and tertiary sources of information were searched the literature review plays a very important role in the research process.

- Place each work in the context of its contribution to understanding the research problem being studied.
- Describe the relationship of each work to the others under consideration.
- Identify new ways to interpret prior research.
- Reveal any gaps that exist in the literature.
- Resolve conflicts amongst seemingly contradictory previous studies.
- Identify areas of prior scholarship to prevent duplication of effort.
- Point the way in fulfilling a need for additional research.
- Locate your own research within the context of existing literature (very important)

CHAPTER 3

PROJECT MANAGEMENT

- **PROJECT PLANNING**
- **PROJECT SCHEDULING**
- **RISK MANAGEMENT**

3.1 PROJECT PLANNING OBJECTIVES

Project management objectives are the successful development of the project's procedures of initiation, planning, execution, regulation and closure as well as the guidance of the project team's operations towards achieving all the agreed upon goals within the set scope, time, quality and budget standards.

3.1.1 SOFTWARE SCOPE

- Due to limited data, system only takes into account limited features for predicting the resale value of the car.
- Since this is an online system, current system does not take into account any physical damage to the car body or engine while predicting the resale value.

3.1.2 PROJECT DEVELOPMENT APPROACH

3.1.2.1 SCRUM

Scrum is a framework for project management and development that emphasizes teamwork, accountability and iterative progress towards a well-defined goal. The three pillars of Scrum are transparency, inspection and adaptation. The framework, which is often part of agile software development, is named for a rugby formation.

3.1.2.2 SCRUM MODEL DESIGN

Scrum is an agile way to manage a project, usually software development. Agile software development with Scrum is often perceived as a methodology; but rather than viewing Scrum as methodology, think of it as a framework for managing a process.

Figure 3.1 Project Development Approach

3.2 PROJECT SCHEDULING

Project scheduling is a process to communicate what tasks need to get done and what organization resources will be allocated to complete this task in a time-frame. Project task scheduling is a project planning activity, for scheduling a project, it is necessary to -

- Break down the project tasks into smaller, manageable form
- Find out various tasks and correlate them
- Estimate time frame required for each task
- Divide time into work-units
- Assign adequate number of work-units for each task
- Calculate total time required for the project from start to finish

3.2.1 BASIC PRINCIPLE OF PROJECT SCHEDULING

- **Compartmentalization –** define distinct tasks
- **Interdependency –** parallel and sequential tasks
- **Time allocation –** assigned person days, start time, ending time
- **Effort validation –** be sure resources are available
- **Defined responsibilities –** people must be assigned
- **Defined Outcomes –** each task must have an output
- **Defined milestones –** review for quality

3.2.2 COMPARTMENTALIZATION

The development of the project is divided into the following set of activities/tasks:

- Data gathering
    - Data cleaning
    - Data sorting
- EDA – Exploratory Data Analysis
    - Understanding and implementing feature engineering
- Choosing appropriate machine learning model
- Training the model with appropriate data
    - Hyper parameter tuning of the model
- Testing and deployment

3.2.3 TIMELINE CHART

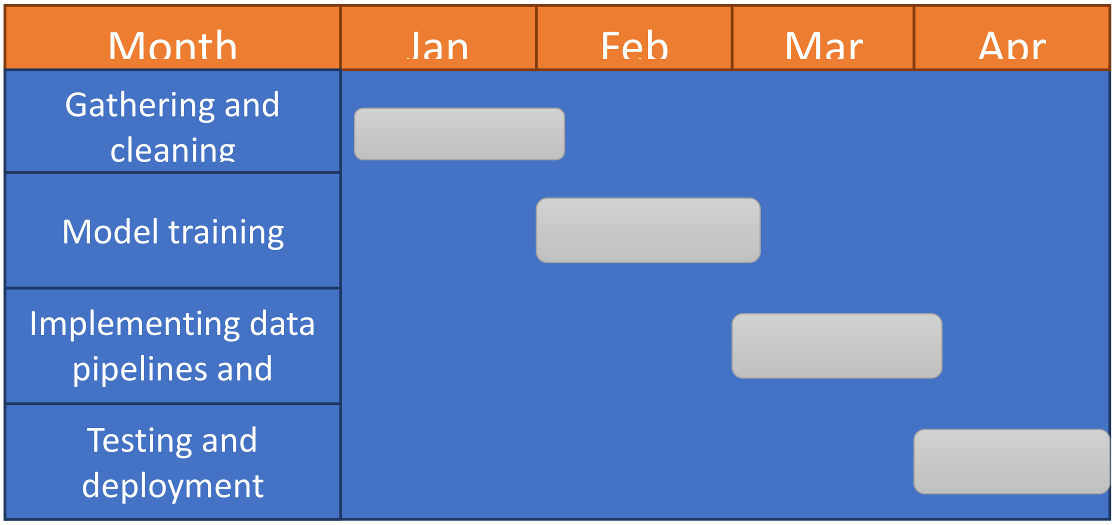


Figure 3.2 Timeline Chart

3.3 RISK MANAGEMENT

Risk management is the human activity which integrates recognition of risk (Identification), risk assessment (Analysis), developing strategies to manage it (Planning) and mitigation of risk using managerial resources. Some categories of risk include product size, business impact, and customer-related process, Technology, Development Environment, Staffing, Schedule, and cost.

3.3.1 RISK IDENTIFICATION

- Understanding the words “if you do not actively attack the risk, they will attack you”, we tried to find all possible risks.
- Risk means a danger to the project. By developing an application or a project there are many risks so the risk should have been calculated before the whole application can be accomplished. Risk must be calculated before it spoils the whole system.

- Risk Management is a process to determining risk and thinking about it before it comes so before risk, we have to think over the steps to overcome risks or to regain the original data which are affected by some or risks like electric power loss or internet connection failure
- The following table includes the risk for this system and risk type and its description

Id	Risk	Description
1	Failure of internet Connection	Internet connection is also one of the prime requirements for this project. If it fails then our Car data will not be updated
2	Electricity	Electricity is a prime requirement of any hardware. If electricity is not available to hardware, then the searching will not work

Table 3.3.1 Risk Identification Artifacts

3.3.2 RISK PROJECTION

Risk Projection is developing and documenting organized, comprehensive, and interactive strategies and methods for identifying risks. The risk planning process considers each of the risks which have been identified and identifies strategies to manage the risk. Again, there is no simple process which can be followed to established risk management plans. It relies on the judgment and experience of the project development team.

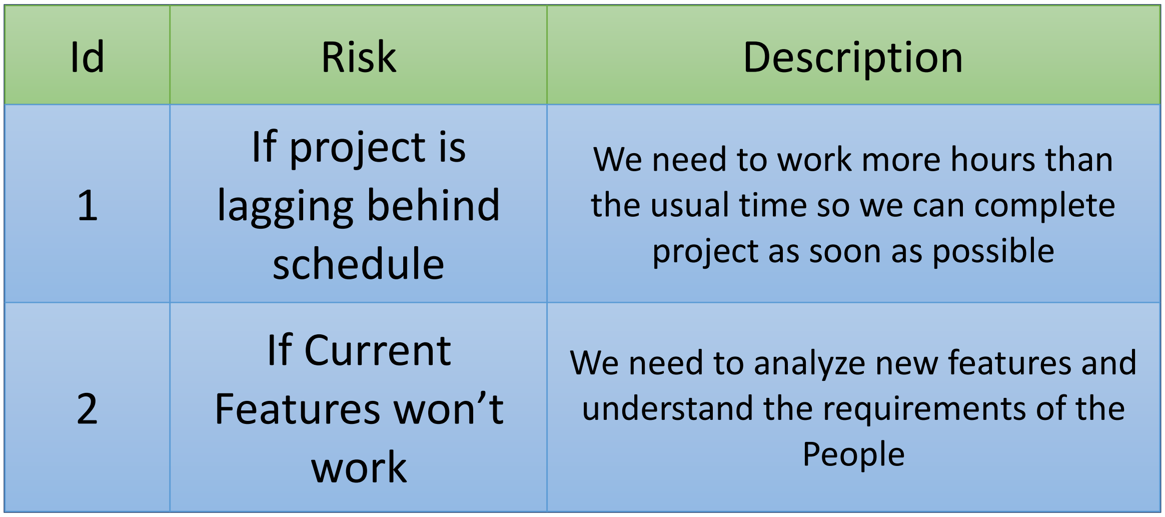
Below table describes the risk and the strategies identified. These strategies fall into three categories.

- **Avoidance strategies:** The probability that the risk will arise, will be reduced
- **Minimization strategies:** Impact of the risk will be reduced.
- **Contingency plans:** If the worst happens, you are prepared for it and have a strategy in place to deal with it.

Risk Projection/Identification:

- Since system is remote and works only on limited features, it might predict wrong valuation of the car which is physically damaged and might incur loss to the buyer.

- Since prediction is based on machine learning model, we have to keep adding new data time to time so as to avoid irrelevant predictions with new data



Id	Risk	Description
1	If project is lagging behind schedule	We need to work more hours than the usual time so we can complete project as soon as possible
2	If Current Features won't work	We need to analyze new features and understand the requirements of the People

Table 3.3.2 Risk Projection

CHAPTER 4

SYSTEM REQUIREMENT

- **USER CHARACTERISTIC**
- **FUNCTIONAL REQUIREMENT**
- **NON-FUNCTIONAL REQUIREMENT**
- **HARDWARE & SOFTWARE REQUIREMENT**

4.1 USER CHARACTERISTICS

- **User:** User will use the tool to get the prediction for the resolve value of car by providing the basic information about the car, the tool will predict the amount by machine learning model.
- **Admin:** Manages the data collection, data cleaning and data processing. Implementing best machine learning algorithm suitable for data and training the dataset.

4.2 FUNCTIONAL REQUIREMENT

4.2.1 Activity and Proposed System

- Only authorized users are allowed to access the system.
- Necessary car data is given to the system to make the future predictions.
- The system was developed to calculate the resale values from the received data.

4.3 NON-FUNCTIONAL REQUIREMENT

- User needs to fill up the form proposed by tool.

4.4 HARDWARE AND SOFTWARE REQUIREMENT

4.4.1 HARDWAREREQUIREMENT

- Processor (CPU) with 2GHz frequency or above
- A minimum of 8 GB RAM
- A minimum of 1 GB Graphics card

4.4.2 SOFTWARE REQUIREMENT

- Jupyter Notebook
- PyCharm
- Visual Studio 2019
- Operating Systems: Windows 10 and Ubuntu

4.4.3 SERVER HOSTING REQUIREMENT

- Since Hosting on cloud / VM is costly, so we can’t use VM for the long time. We can work on VM short term.
- We can’t apply ML Algorithms on Data Bricks, Data Factory, Data Lake because it costs too much. So, it’s better to use on local system.

CHAPTER 5

SYSTEM ANALYSIS

- **STUDY OF CURRENT SYSTEM**
- **PROBLEMS IN CURRENT SYSTEM**
- **REQUIREMENT OF NEW SYSTEM**
- **PROCESS MODEL**
- **FEASIBILITY STUDY**
- **FEATURES OF NEW SYSTEM**

5.1 STUDY OF CURRENT SYSTEM

In the current scenario, resale value of car is approximated by garage owners and car resellers through manual examination which usually is not correct due to various reasons.

Currently, there are few systems like cars24 which uses statistical data to analyze and predict resale value of the car. However statistical methods purely rely on the historical data and cannot accurately deal with new data all the time.

5.2 PROBLEMS IN CURRENT SYSTEM

- It is practically not possible to remotely predict the resale value of the car on the basis of limited information.
- Also due to monopoly and biased decisions, many times resale car customers feel cheated or fooled due to high price as compared to the condition of the car.
- There is a lack of experts who can examine correct resale value of the car in current times.
- There are few online car valuation platforms that predict car resale value but they are highly based on statistical methods. Hence the need arises to come up with a method that can accurately predict resale value of a car like experts without human intervention remotely.

5.3 REQUIREMENT OF NEW SYSTEM

In order to overcome all the challenges faced by the current system and to reduce the cost so as to make the system profitable, we have designed machine learning based model to effectively predict resale value of cars with higher accuracy so as to decrease the penalties and computational cost. This will make the business economically profitable and easier for customers to make a valid and profitable deal with these accurate predictions for resale value of cars.

5.4 PROCESS MODEL – AGILE

Phases of Agile Model:

Following are the phases in the Agile model are as follows:

- Requirements gathering
- Design the requirements
- Construction/ iteration
- Testing/ Quality assurance
- Deployment
- Feedback

1. **Requirements gathering**: In this phase, Data gathering is done through client communication such as different features, parameters etc.

2. **Design the requirements**: When you have identified the project, planned for different machine learning algorithm and deep learning neural network and applied hyperparameter optimization.

3. **Construction/ iteration**: With defined requirements, the work begins. We applied different machine learning algorithm such as KNN, SVR, GBR, and logistic.

4. **Testing**: In this phase, pilot testing of model for two weeks continuous.

5. **Deployment**: In this phase, we issued a product for the user's work environment and deployed.

6. **Feedback**: After releasing the product, the last step is feedback. In this, the team receives feedback about the product and works through the feedback.

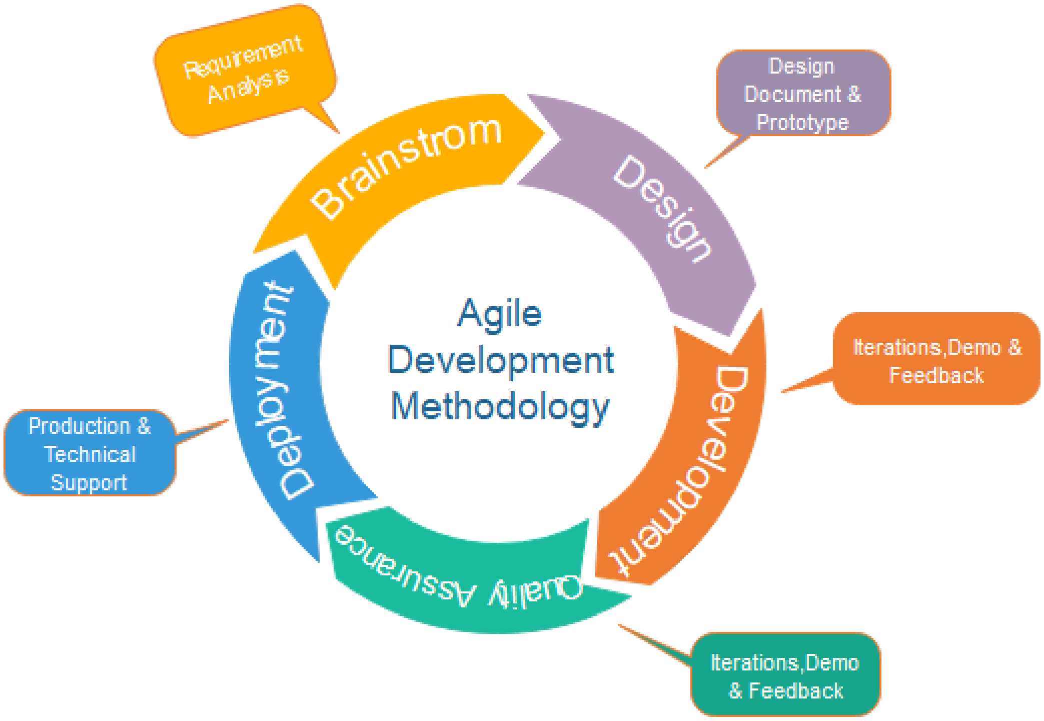

Requirement Analysis
Brainstrom
Design Document & Prototype
Design
Agile Development Methodology
Iterations,Demo & Feedback
Development
Quality Assurance
Iterations,Demo & Feedback
Deployment
Production & Technical Support

Figure 5.4.1 Agile Development model

5.5 FEASIBILITY STUDY

Feasibility is the measure of how beneficial the development of the system will be to an organization in comparison to the present system.

The feasibility analysis is categorized under four different types.

1. Technical Feasibility
2. Operational Feasibility
3. Economical Feasibility
4. Schedule Feasibility

5.5.1 TECHNICAL FEASIBILITY:

- Technical Feasibility is defined as the feasibility that is concerned with specifying equipment and software that will successfully satisfy the user requirement.
- Our system runs on a saved Machine Learning model which is very lightweight and hence there is no interruption on the user side due to technical reasons.

5.5.2 OPERATIONAL FEASIBILITY

Suppose due to some issues even if we get less amount of data, our model is still capable of getting the good prediction / accuracy.

5.5.3 ECONOMICAL FASIBLITY

Due to the use of Artificial Intelligence and its higher accuracy, seller and buyer both gets best price and thus it is economically feasible to both the parties.

5.5.4 SCHEDULE FEASIBILITY

The time period of our project is 4 months from January to April. In first month, we did data gathering and cleaning, in second month we did model training, in third month we implemented data pipelines and models, in fourth month testing and deployment.

5.6 FEATURES OF NEW SYSTEM

- The new system developed by us consists of two parts - data gathering and prediction using Machine Learning based algorithms

- We have used web scraping libraries to gather data from the webpages of cars24 website. The script runs and captures data from the HTML div mentioned in the code via URL. URL should be entered by the user. For now, we have captured data by entering URL for Swift Dezire cars for 5 cities.

- The second part is the web-based car resale value prediction. We have trained boosting algorithm-based ML model using data from the previous step after preprocessing and cleaning.

- The trained model is used for prediction. The front-end form asks user to fill values which are required for ML model to make prediction i.e. - city, kms driven, year of purchase and fuel type.

- Upon form submission, the data is sent to ML model via Flask API and the model responds with a predicted resale value of the car based on user input.

- This prediction is displayed on the webpage using render template. Thus, with minimal information and without human intervention or manual examination, user can predict the resale value of his/her car.

CHAPTER 6

DETAIL DESCRIPTION

- **PREDICTION REQUIREMENT**
- **PREDICTION APPROACH**
- **DATA FLOW**

6.1 Prediction Requirement

Due to huge number of factors affecting the resale value of a car and also monopoly of various resellers, predicting a realistic value of a car becomes difficult. Apart from that buyers feel cheated when price claimed is much higher as compared to the condition of the car. Also, resale car business is growing exponentially since last few years and it requires a lot of man power to manually examine and determine the resale value of the car. To overcome such issues, it is very necessary to come up with an intelligent system that can predict a perfect resale value of a car without human intervention to avoid biased results.

6.2 Prediction Approach

- For accurate prediction and better model training, huge dataset of resale cars of Swift Dezire of 5 cities is gathered via web scraping cars24 website. This dataset contains data of 5 main features i.e., fuel type, kms driven, city, car purchase year and resale value. Here resale value becomes our target column whereas other columns served as features for our model.

- Data scraped consists of many unwanted characters like comma, whitespaces etc. which has to be removed as model can only understand numbers. Moreover, fuel type was converted into numerical codes via one-hot encoding. A one hot encoding is a representation of categorical variables as binary vectors. This requires that the categorical values be mapped to integer values. After data preprocessing, all 5 files, each representing each city has to be merged for model training.

- Various different machine learning algorithms were implemented on the dataset along with hyperparameter tuning using GRID SEARCH CV. Below are the results of different machine learning algorithms implemented with parameter fine tuning:

- ➢ Logistic Regression - RMSE: 1,60,000
- ➢ Support Vector Regression - RMSE: 82,000
- ➢ Random Forest Regression - RMSE: 79,000
- ➢ Gradient Boosting Regression - RMSE: 50,000

- Reason behind GBR's good performance is because of its mathematical working.
- The reason why GBR could outcome all other regression algorithms is the mathematics behind it.
- Gradient boosting involves three elements:
  - ➢ A loss function to be optimized.
  - ➢ A weak learner to make predictions.
  - ➢ An additive model to add weak learners to minimize the loss function.

➢ **1. Loss Function**

The loss function used depends on the type of problem being solved.

It must be differentiable, but many standard loss functions are supported and you can define your own. For example, regression may use a squared error and classification may use logarithmic loss. A benefit of the gradient boosting framework is that a new boosting algorithm does not have to be derived for each loss function that may want to be used, instead, it is a generic enough framework that any differentiable loss function can be used.

➢ **2. Weak Learner**

Decision trees are used as the weak learner in gradient boosting.

Specifically, regression trees are used that output real values for splits and whose output can be added together, allowing subsequent models outputs to be added and “correct” the residuals in the predictions. Trees are constructed in a greedy manner, choosing the best split points based on purity scores like Gini or to minimize the loss. It is common to constrain the weak learners in specific ways, such as a maximum number of layers, nodes, splits or leaf nodes. This is to ensure that the learners remain weak, but can still be constructed in a greedy manner.

- **3. Additive Model**

Trees are added one at a time, and existing trees in the model are not changed.

A gradient descent procedure is used to minimize the loss when adding trees. Traditionally, gradient descent is used to minimize a set of parameters, such as the coefficients in a regression equation or weights in a neural network. After calculating error or loss, the weights are updated to minimize that error. Instead of parameters, we have weak learner sub-models or more specifically decision trees. After calculating the loss, to perform the gradient descent procedure, we must add a tree to the model that reduces the loss (i.e., follow the gradient). We do this by parameterizing the tree, then modify the parameters of the tree and move in the right direction by (reducing the residual loss.

6.3 DATA FLOW

- So, our flask-based web app consists of trained GBR model at the backend.

- GBR model was trained using the cleaned and preprocessed data. The trained GBR model was saved as a pkl file using pickle serialization.

- We have created a web-based form using html, CSS and JS that takes all the necessary input features that is year of purchase, kms driven, city and type of fuel.

- This data is then sent to the python script which performs preprocessing. Firstly, it removes all unwanted characters like comma, whitespaces etc. Then it converts city into numeric code.

- Then the data is passed to the saved model for prediction. The prediction of the model is then displayed on the webpage using flask render template.

CHAPTER 7

SYSTEM DESIGN

- **USE CASE DIAGRAM**
- **SUB USE CASE DIAGRAM**
- **DATA FLOW DIAGRAM**

7.1 USE-CASE DIAGRAM

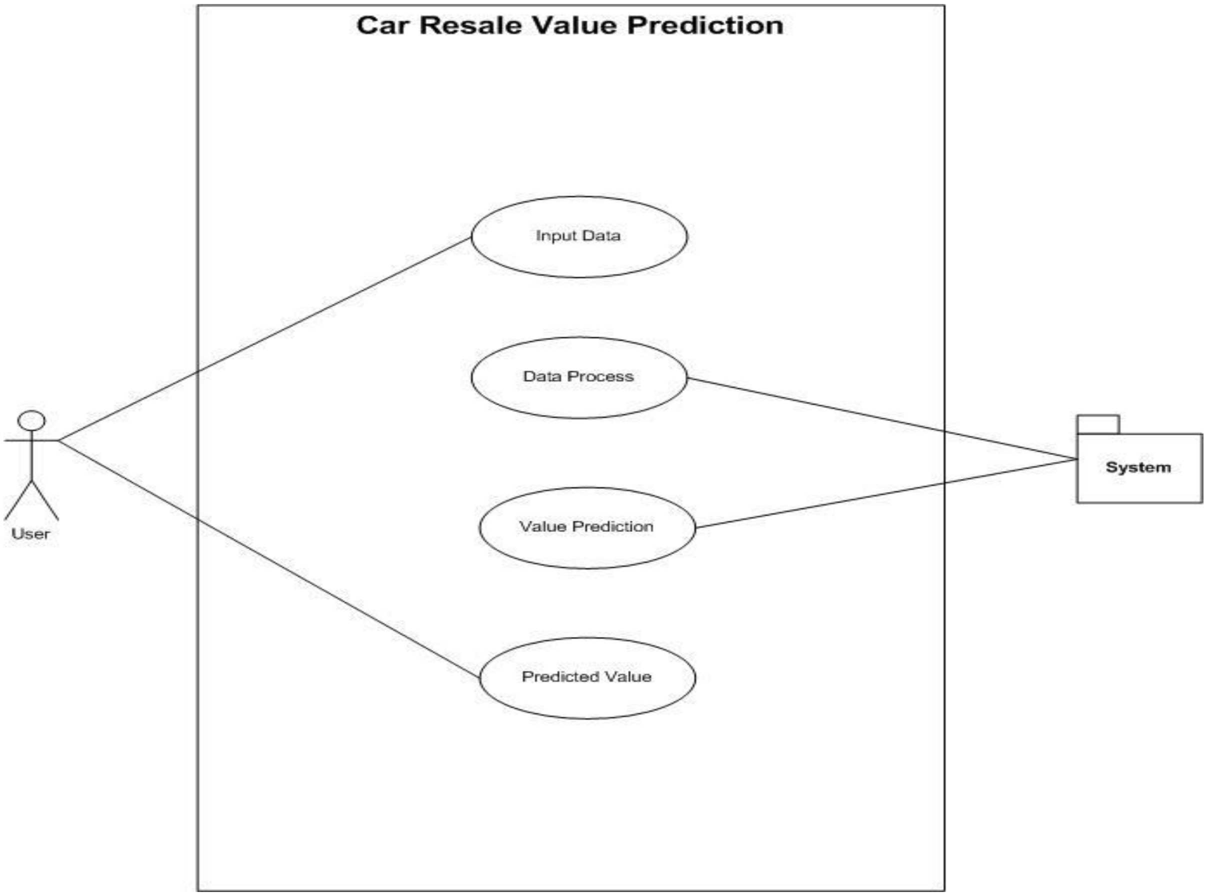


Figure 7.1.1 Use case Diagram

7.2 SUB USE-CASE DIAGRAM

DATA PROCESS

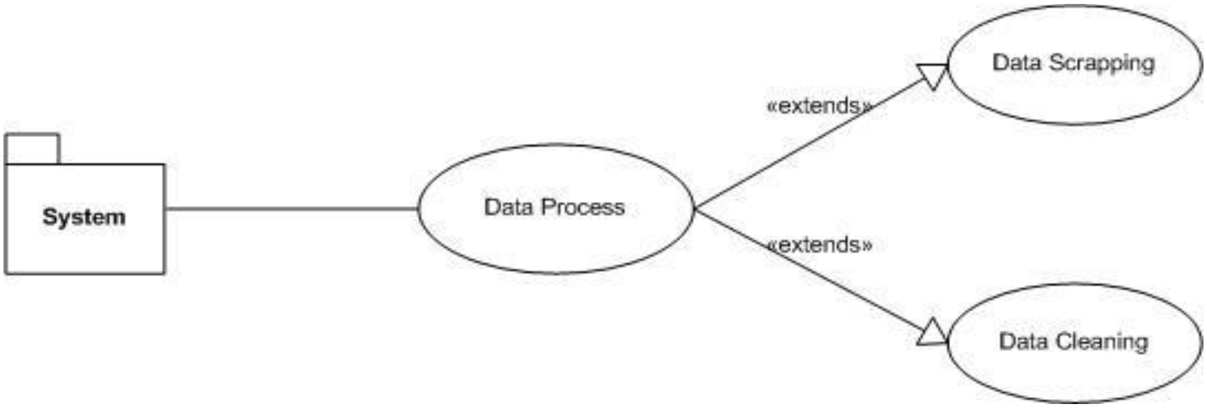


Figure 7.2.1 Sub use case Diagram

VALUE PREDICTION

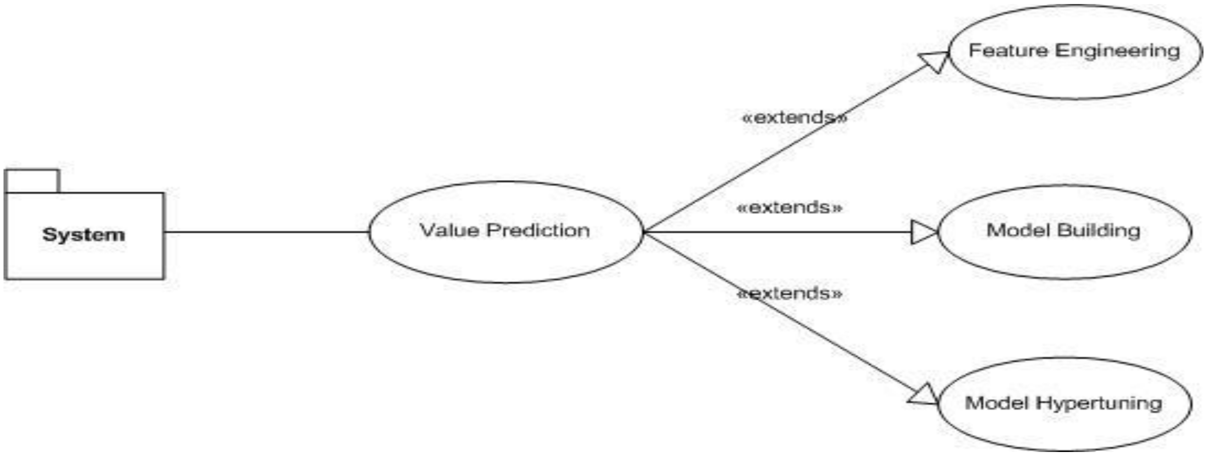


Figure 7.2.2 Sub use case Diagram

7.3 DATA-FLOW DIAGRAM

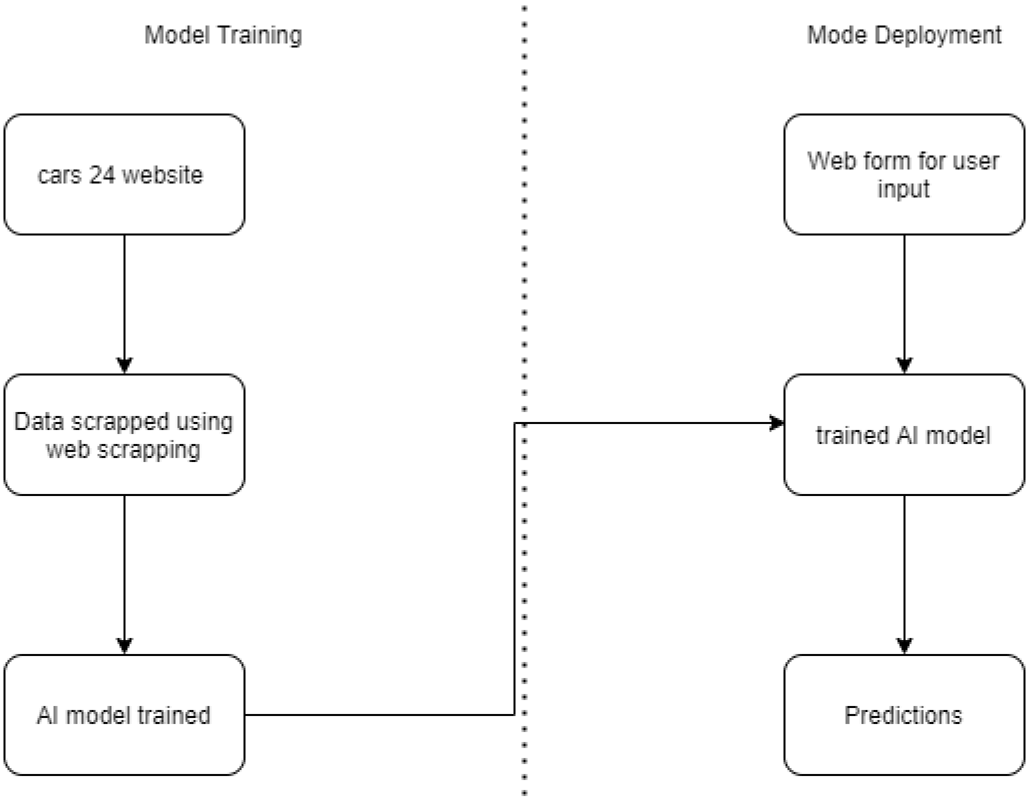


Figure 7.3.1 Data Flow Diagram

CHAPTER 8

LIMITATIONS AND FUTURE ENHANCEMENT

- **LIMITATIONS**
- **FUTURE ENHANCEMENT**

8.1 LIMITATIONS

Due to storage and database limitations, currently we have only trained model with data of Swift Dezire in 5 cities only. Thus, current system can only predict for single car and 5 cities. However, with more data collection and model training this can be extended.

Also, since the system is in initial stage and data available at the present moment is very less, current system can give accurate results up to Root Mean Squared error of 42,000 which means for instance, if resale value of a car is supposed to be 5 lacs, our system would predict somewhere in the range of 4,60,000 to 5,40,000.

The current data is scraped from cars24 website and thus the accuracy of the model is dependent on the authenticity of the data available on 3rd party website.

Current system only uses 4 features i.e., year of purchase, kms driven, city and fuel type. Due to computational limitations, more data could not be scraped and thus current system predicts resale value only based on these 4 features. However, if there are any other damage to the body of the car or engine, it cannot be accessed or taken into consideration while predicting the value of the car.

8.2 FUTURE ENHANCEMENT

Currently, system can only deal with Swift Dzire cars due to lack of data. Also, data has been collected of only 5 cities of India. This can be extended to multiple car models and cities so as to improve accuracy and usability.

Efficient use of deep learning such as LSTM (Long short-term memory) or RNN (Recurrent Neural networks) can be implemented once enough data is collected. This can improve accuracy and decrease RMSE drastically.

Currently, only few features are used to predict resale value of the car. This can be extended to more features. One can also implement CNN to determine physical condition of the car from images like identifying dents, scratches etc. and thus predicting more relevant resale valueofacar.

CHAPTER 9

TESTING

- **BLACK BOX TESTING**
- **WHITE BOX TESTING**
- **TEST CASES**

9.1 BLACK BOX TESTING

In Black-box testing, as we have applied ML algorithms and Neural Networks, we don’t have to know how the weights and biases are managed internally by NN. Default parameters given by the algorithms. We provide the inputs and required parameters and it gives us the predicted output.

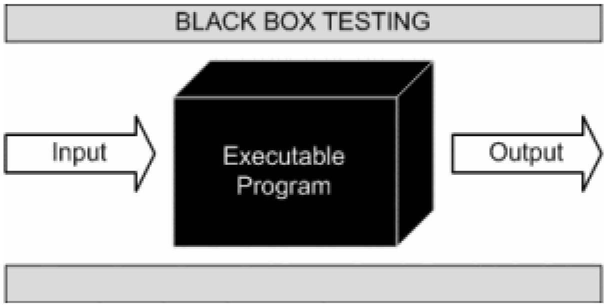


Figure 9.1.1 Black box testing

9.2 WHITE BOX TESTING

White-box testing is a testing technique which checks the internal functioning of the system. In this method, we know how the Loss function, activation functions, hyper-parameters work. We can control the under-fitting and over-fitting by hyper-tuning optimization.

In ML algorithm, for ex. We have applied decision tree algorithm, then we are aware of the depth of tree, leaf nodes and we can change them according to the needs.

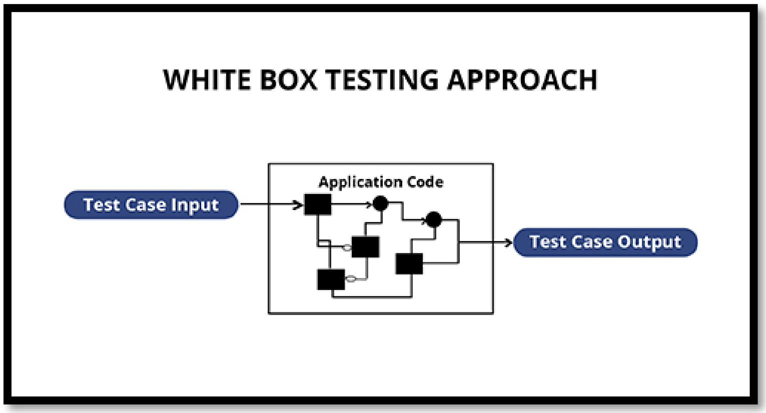


Figure 9.2.1 White box testing

9.3 TEST CASES

- **Missing values**

  The trained ML model requires 4 feature inputs for predicting the output. Failing which, the model throws invalid Input error. All the fields in the html form have been marked required using CSS and thus user must input all fields.

  - **Output**: User must input all the fields, failing which, form shows warning message "this field needs to be filled". Thus, there can be no errors in model prediction.

- **Invalid Input**

  The trained ML model requires only numerical input for all 4 features. Thus, if user uses symbols such as comma while input, model may throw error. To overcome the same, preprocessing script is deployed in backend which removes all unwanted characters like comma, whitespaces etc. so that model gets required input.

  - **Output**: Due to python preprocessing script, model will get the desired input and thus will give accurate prediction.

- **Unseen year of purchase**

  The model is trained with data from cars purchased since 2011 to 2020. If the user inputs details of car purchased after that i.e., 2021, model may get confused since that data is quite new and unseen to model.

  - **Output**: Model has been trained with boosting algorithm and thus it gives quite accurate results with around RMSE 65,000 INR.

CHAPTER 10

APPENDICES

- **BUSINESS MODEL**
- **PROJECT DEPLOYEMENT**

10.1 BUSINESS MODEL

A business model is a company's plan for making a profit. It identifies the products or services the business will sell, the target market it has identified, and the expenses it anticipates.

A new business in development has to have a business model, if only in order to attract investment, help it recruit talent, and motivate management and staff. Established businesses have to revisit and update their business plans often or they'll fail to anticipate trends and challenges ahead. Investors need to review and evaluate the business plans of companies that interest them.

The digital revolution is resulting in deep changes in consumer habits, caused, among other factors, by greater access to information and increasingly developing new technologies. All this invites us to take an in-depth look into the business models currently being used.

A fundamental driver of business model transformation is data science, which is based on the combined use of machine learning techniques, artificial intelligence, mathematics, statistics, databases and optimization.

This system aims to provide accurate car valuation for resale and thus can be sold to online car resale companies like cars, drool etc. The system can be charged on yearly basis as it includes continuous process of retraining model with more data which also includes data preprocessing and cleaning.

10.2 PROJECT DEPLOYEMENT

The system is currently deployed on local host using flask in python. However, the system can be deployed on a web hosting platform and can also be implemented via mobile app. The trained model file must be deployed on the database and the web form should send the data after preprocessing to the trained model for prediction. All this can be carried out through an API call.

CHAPTER 11

CONCLUSION

- **CONCLUSION**

11.1 CONCLUSION

Car resale market is growing exponentially and thus the need for accurate resale valuation system started growing, so as to reduce human intervention and make things remote and easy.

To test the system concept, we scrapped data of a particular car model from 5 different cities. With this limited data, we trained a machine learning model with several algorithms and found that GBR – Gradient Boosting Regression performs best out of all the algorithms tried.

However, once more data is collected and various different cars are included in the system, deep learning-based ANN or LSTM would perform better. But currently, GBR based car valuation system can predict resale value of a car with Root Mean Squared Error (RMSE) of 50,000 INR.

References

- https://www.researchgate.net/publication/319306871_Predicting_the_Price_of_Used_Cars_using_Machine_Learning_Techniques - Predicting the Price of Used Cars using Machine Learning Techniques.
- https://irjmets.com/rootaccess/forms/uploads/IRJMETS462275.pdf - Used car price prediction using linear regression model.
- http://cs229.stanford.edu/proj2019aut/data/assignment_308875_raw/26463410.pdf - Predicting used car prices.
- API Reference — scikit-learn 0.24.2 documentation
- ijictv4n7spl_17.pdf (ripublication.com) – Predicting the prices of the used cars using Machine Learning Techniques.
- sklearn.ensemble.GradientBoostingRegressor — scikit-learn 0.24.2 documentation