Savitribai Phule Pune University

A

FINAL PROJECT REPORT

ON

"House Price Prediction And Analysis"

SUBMITTED BY

Mr. Abhishek Thombare	**Exam No: 71725670H**
Mr. Sumit Chavan	**Exam No: 71725654F**
Mr. Gautam Mudaliar	**Exam No: 71725478L**
Mr. Nimish Vaidya	**Exam No: 71725682M**

UNDER THE GUIDANCE OF

PROF. P. V. Kasture

Sinhgad Institutes

DEPARTMENT OF COMPUTER ENGINEERING

Rasiklal Manikchand Dhariwal Sinhgad
School of Engineering, Pune - 411 058
[2019 - 2020]

Rasiklal Manikchand Dhariwal Sinhgad

School of Engineering, Pune

Sinhgad Institutes

C E R T I F I C A T E

This is to certify that Project Entitled

"House Price Prediction And Analysis"

SUBMITTED BY

Mr. Abhishek Thombare	**Exam No: 71725670H**
Mr. Sumit Chavan	**Exam No: 71725654F**
Mr. Gautam Mudaliar	**Exam No: 71725478L**
Mr. Nimish Vaidya	**Exam No: 71725682M**

is a bonafide work carried out by Students under the supervision of Prof. V. M. Lomte and it is submitted towards the partial fulfillment of the requirement of Bachelor of Engineering (Computer Engineering)

Prof. P. V. Kasture
Project Guide,
Department of Computer Engineering

Prof. Parth Sagar
Project Co-ordinator,
Department of Computer Engineering

Prof. V. M. Lomte
Head,
Department of Computer Engineering

Dr. V. V. Dixit
Principal,
RMDSSOE, Warje, Pune 58

Place: RMDSSOE, Pune.
Date: / /2019

ACKNOWLEDGEMENT

It gives me great pleasure and satisfaction in presenting this final report on "House Price Prediction And Analysis".

I am thankful to and fortunate enough to get constant encouragement, support and guidance from all Teaching staffs [of Computer Dept] which helped us in successfully completing our project work. Also, I would like to extend our sincere esteems to all staff in laboratory for their timely support.

I have furthermore to thank Computer Department HOD **Prof V. M. Lomte** *and* Guide **Prof. P. V. Kasture**

I would like to thank all those, who have directly or indirectly helped me for the completion of the work during this seminar report.

Abhishek Thombare
Sumit Chavan
Gautam Mudaliar
Nimish Vaidya
B.E. Computer

Abstract

Real Estate / Land is the least straight-forward industry in our society. House costs keep changing everyday consistently and from time to time are publicized instead of being established on valuation. House estimation process is a critical purpose of land which can be used to gain profits. The composition tries to get supportive gaining from valid data of property markets. Computer based intelligence strategies are applied to explore evident property trades to discover profitable models for house buyers and vendors. For evaluating the cost various regression techniques were attempted and one with most raised exactness is used.

Keywords - *Regression, ridge regression, lasso regression, random forest tree, prediction, parameters.*

TABLE OF CONTENTS

List of abbreviations

- ML : Machine Learning
- CNN : Convolutional Neural Network
- SRS : Software Requirement Specification
- GUI : Graphical User Interface
- IDE : Integrated Development Environment
- API : Application Program Interface

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

1.1 Overview

The aim of this report is to suggest best method that will be helpful in Predicting House Prices. After receiving input values those are passed to our model and appropriate price is predicted. After receiving inputs three intermediary models are generated which predicts the house price. Based on these predicted house prices, we build a meta model to predict our final house price. The dataset can be viewed in real-time in the adjoining map. New purchase can be submitted to the system for considering in future prediction. Apart from the web version, Android application is also available.

1.2 Motivation

The housing sector is the second largest employment provider after agriculture sector and it is estimated to grow at 30 percent over the next decade. Ambiguity among the prices of houses makes it difficult for the buyer to select their dream house. The interest of both buyers and sellers should be satisfied so that they do not overestimate or underestimate price.

1.3 Problem Definition and Objective

1.3.1 Problem Definition

We propose to use Ensembling method 'stacking' instead of using individual methods to predict price. Ensembling takes the combination of multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone.

The aim of the project is to develop the optimum machine learning model for predicting the price of property on a dataset of house sale prices in King County, Seattle from May 2014 to May 2015. The dataset provides various features houses have like locality, condition, size, age, etc, and the prices at which they were sold.

1.3.2 Objectives

1. To use Ensemble methods to predict house prices.
2. To compare different regression techniques with Ensemble method and choose the best approach.

1.4 Project Scope and Limitation

1.4.1 Project Scope

We can use House Price Prediction system to predict the price of any type of property. Banks need to predict house prices to estimate asset value and to determine grantable loan amount. House owners and customers can use this system to minimize middle man cost, Brokers can provide valuable advice regarding sales and investment to the customers using this system.

1.4.2 Limitation

1. Accuracy of the earlier models were low.
2. Availability of reliable test datasets was a major problem.
3. Earlier systems were not cost and time efficient.

1.5 Methodologies of Problem solving

Earlier there was no system to predict house prices accurately. Our dataset involves different basic parameters and information mining has been at the base of our framework. We at first tidied up our whole dataset and furthermore truncated the exception esteems. Further, we gauged every parameter dependent on its significance in deciding the estimating of the framework and this drove us to build the worth that every parameter retains in the framework. We shortlisted 5 diverse

AI calculations and tried our framework with various mixes that can ensure best potentially dependability of our outcomes.

CHAPTER 2

LITERATURE SURVEY

In “Predicting House Price With a Memristor-Based Artificial Neural Network” by J.J.Wang, Zhen Liu, T.P.Chen and Sumio Hosaka, ANN was used to learn a regression model of the house prices of several Boston towns in the USA and the predicted results are found to be close to the target data. In this paper, a 2-layer feed-forward neural network is designed which has the ability to learn to predict then house price under training mode, and then can successfully predict the house price in the predicting mode.

In “Housing Price Prediction using Machine Learning Algorithms: The Case of Melbourne City, Australia” by The Danh Phan, Macquarie University Sydney, Australia, a real historical transactional dataset was used to derive valuable insight into the housing market in Melbourne city. In this paper, the combination of Step-wise and SVM model proved to be a competitive approach. This research can also be applied for transactional datasets of the housing market.

In “House Price Prediction Using Machine Learning And Neural Networks ” by Ayush Varma , Abhijit Sarma , Sagar Doshi and Rohini Nair, real factors such as proximity to railway station etc. in addition to basic parameter were considered while determining the price. In this paper, Optimal combination of linear regression, forest regression boosted regression is found. The system warns user against investing in the wrong house.

In “A Multi Feature Based Housing Price Prediction for Indian Market Using Machine Learning” by Rushab Sawant, Saurabh Jain, Tushar Tiwari, Yashwant Jangid and Ms.Ankita Gupta, predicted the area to the developers which will maximize their investment returns for their future projects. The project developed was bound to predict housing prices based on features that do not change with time

In “Predicting House Sale Price Using Fuzzy Logic, Artificial Neural Network and K-Nearest Neighbour” by Muhammad Fahmi Mukhlishin, Ragil Saputra and

Adi Wibowo included sales value of taxable object land (NJOP-L) and sales value of taxable object building (NJOP-B) for prediction of prices. In this paper, results show that the fuzzy method is superior to neural networks as well as k-nearest neighbour for the house price prediction in limited data training.

In "Comparing Multilevel Modelling and Artificial Neural Networks in House Price Prediction" by Yingyu Feng and Kelvyn Jones, compared MLM and ANN approach to modelling housing prices with the widely accepted HPM approach in terms of goodness-of-fit. All performance measures show that MLM is superior to ANN and HPM in each scenario

CHAPTER 3

SOFTWARE REQUIREMENT SPECIFICATION

3.1 Introduction

Software requirement and specification or SRS is term used to describe the requirement of the product in depth. SRS deals with all the requirements mainly consisting of *Functional , Non-function, Project scope* requirements.

3.1.1 Project Scope

The aim of this project is to use House Price Prediction system to predict the price of any type of property. Banks can predict house prices to estimate asset value and to determine grantable loan amount using this system. House owners and customers can use this system to minimize middle man cost. Brokers can provide valuable advice regarding sales and investment to the customers using this system.

3.1.2 User Classes Characteristics

User classes

- User

  User should provide house parameters.

3.2 Assumption and Dependencies

- Assumption

  It is assumed that the user wants to predict the house price of a house located in Seattle city, USA only.

  The user must have a basic knowledge of mobile and handling web application.

- Dependencies

  Requires Internet connectivity.

Only Administrators will be able to edit main configurations.

3.3 Functional Requirements

1. Upload Dataset
2. Accept Parameters
3. Build model
4. Test model
5. Build meta model
6. Display result

3.4 External Interface Requirements

This section describes interface requirements for House Price Prediction system. This defines the interaction of user with system. These requirement also includes required hardware, software and communication requirements.

3.4.1 User Interfaces

1. Accept Inputs
2. Displays predicted price.
3. Show map overview

3.4.2 Hardware Interfaces

1. Processor - Intel i3/i5/i7
2. Speed - 1.1 GHz
3. RAM - 2 GB(min)
4. Hard Disk - 40 GB

5. Key Board - Standard Windows Keyboard
6. Mouse - Two or Three Button Mouse
7. Monitor - SVGA

3.4.3 Software Interfaces

1. Operating System - Windows 7/8/10
2. Application Server - Flask
3. Front End - Python
4. Programming Language - Python

3.5 Nonfunctional Requirements

3.5.1 Performance Requirements

1. The system should correctly execute the process.
2. The performance of the system lies in the way it is handled. Every user must be given proper guidance regarding how to use the system. The other factor which affects the performance is the absence of any of the suggested requirements.
3. The response time of the system should be deterministic at all times and very low. Thus, the system will work in real time.
4. The software should run on all browsers(mobile/laptop).

3.5.2 Safety Requirements

1. Any unauthorized user should be prevented from accessing the system.
2. Password authentication can be introduced.

3.5.3 Security Requirements

1. Website should be SSL certified for improved security.

3.5.4 Software Quality Attributes

1. Reliable
2. Maintainability
3. Testability
4. Correctness
5. Efficiency
6. Interoperability

3.6 System Requirements

3.6.1 Database Requirements

- The database is required to be created and maintained in MySQL Server. Stored procedures are also created to retrieve and operate on data.

3.6.2 Software Requirements

- Python

  Python is an interpreted, high-level and general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant whitespace.

- Flask

  Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions.

3.7 Analysis Models: SDLC Model to be applied

Agile model consists of incremental as well as iterative process models, which mainly focuses on customer satisfaction and perfect delivery of end product. The whole spyware development process is split into number of iterations for better quality product. Each iteration results in incremented functionality as compared to previous working product.

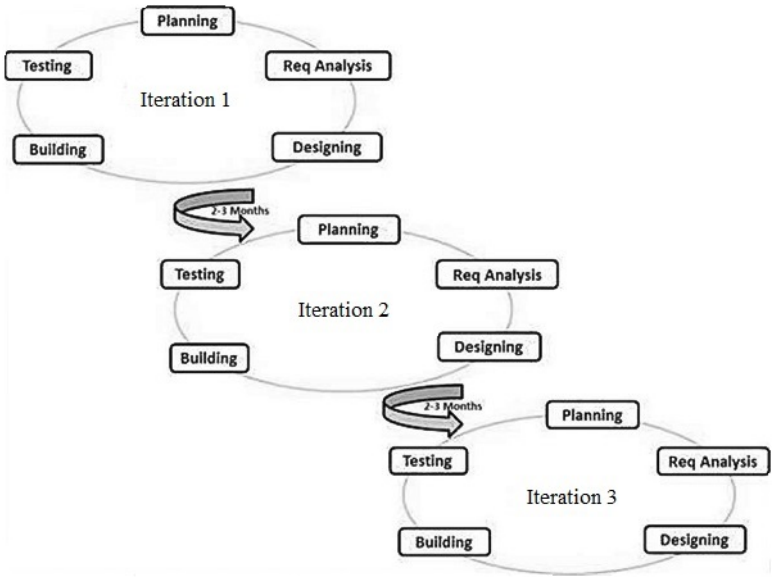


Figure 3.1: SDLC : Agile Model

In every iteration the system is redesigned from requirement specifications given by end user. From designing incremented product to testing it with new upgrades. Each iteration promises to deliver successive and improved version of previous end product. It is a real time approach in software development. It promotes alternative training and enhances teamwork. Rapid development and demonstration of functionality can be a great advantage of this model. But Agile practises have to carried out thoroughly and effectively in order to bring the outcome in less time. Delivery management follows strict rules which will decide the functionality to be given, scope, and tweaks in end product .

3.8 System Implementation Plan

3.8.1 Planning

1. Planning

   Planning is developing a strategic plan to execute a specific goal. Project planning develops the schedule starting from initialized phase to the final phase of the project.

   Understand the domain knowledge,evaluate the resources required for the projects through framing the problems and team required for the project.

   The team who are in the project are given a specific task to perform and deadline is given to complete the task.

   The meeting is conducted every week to determine the quality of the analytical model so that in every stages the error occurred in the project will be known and can be fixed.

   After the correct implementation of the project is done, then reports and technical documents are done.

2. Phases of Planning:
   1. Requirement Gathering.
   2. Literature Survey.
   3. Model Planning.
   4. Model Construction.
   5. Implementation of Code.
   6. Report Generation.

CHAPTER 4

SYSTEM DESIGN

4.1 System Architecture

A system architecture is a conceptual model that will define the structure or the behaviour of the system built. System model deals with the components and the connectors and how they communicate with each other. In House Price Prediction system, System, there are 4 main types of modules -

- Module 1 - Training Data
- Module 2 - Preprocessing
- Module 3 - Stacking
- Module 4 - Prediction

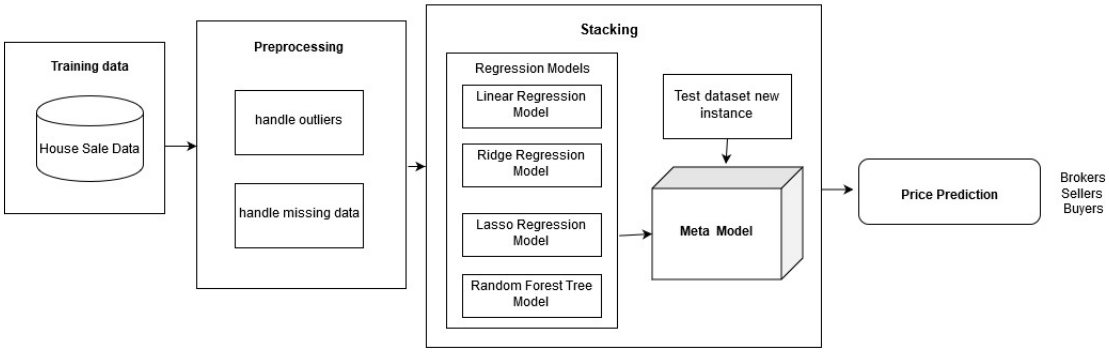


Figure 4.1: System Architecture

Training data is the data you use to train an algorithm or machine learning model to predict the outcome you design your model to predict.

Data preprocessing, where we load our data into a suitable place and prepare it for use in our machine learning training. In this step we prepare our data by removing outliers, handling missing data and normalizing the data

Stacking is an ensemble learning technique that combines multiple classification or regression models via a meta-classifier or a meta-regressor. The base level models are trained based on a complete training set, then the meta-model is trained on the outputs of the base level model as features. For our model we used

Linear, Ridge, Lasso regression and Random Forest for creating meta model. The meta model uses Linear Regression.

The trained model is used to predict house price using unknown data. This price can be used by Brokers, Sellers and Buyers.

4.2 Data Flow Diagrams

Data Flow Diagrams are the diagrams which helps to understand the flow of the system. These diagrams help in representing the flow of system or process.

4.2.1 Data flow Level 0

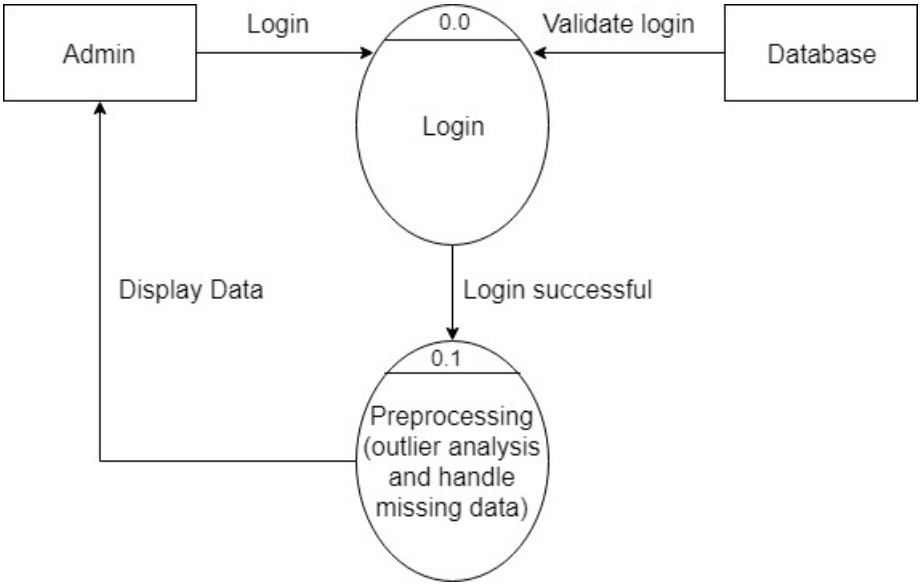


Figure 4.2: Level 0 : Data flow Diagram

DFD 0 is a context diagram. This DFD level 0 example shows how such a system will function with external entities such as Room, Unity and User.

4.2.2 Data flow Level 1

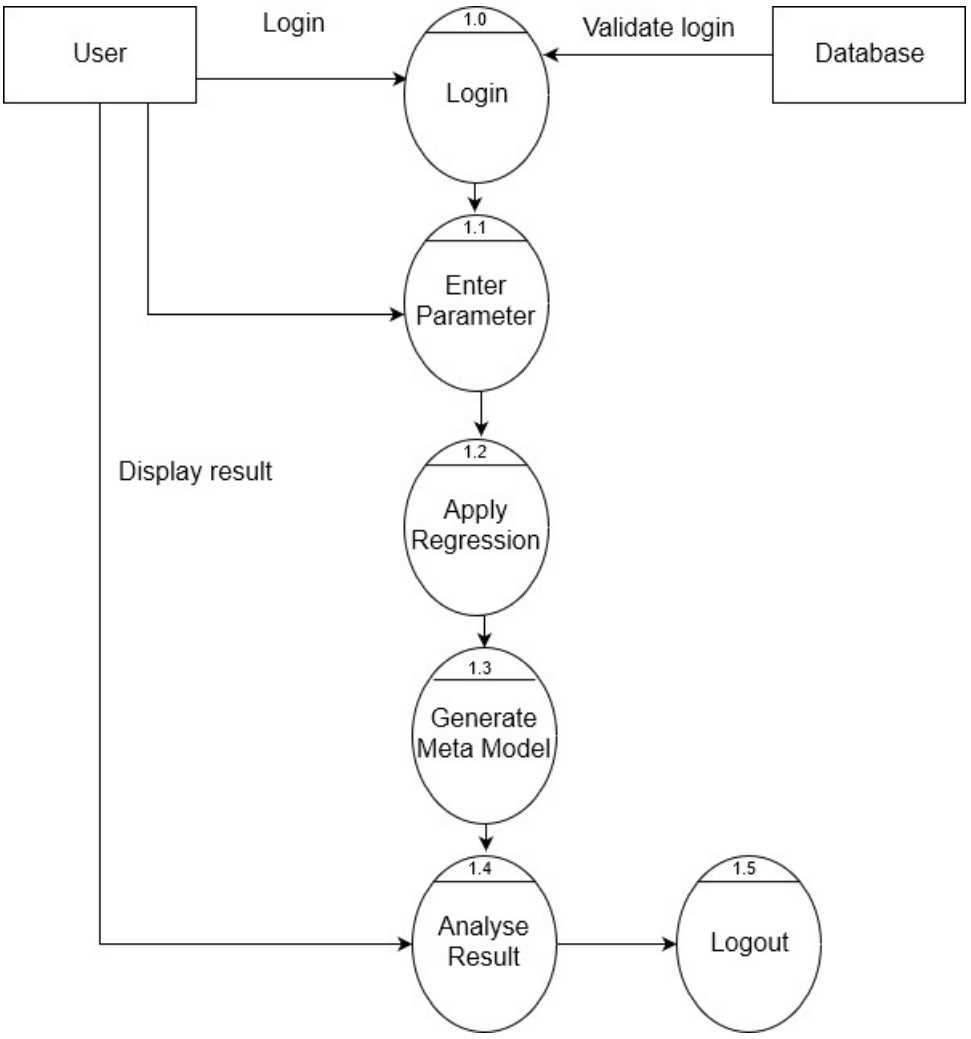


Figure 4.3: Level 1 : Dataflow Diagram

DFD 1 is a more detailed diagram than DFD 0. In this figure how the system will function when the Database is also connected can be seen. Authentication of user and inputs are supplied and stored in database.

4.3 Entity Relationship Diagrams

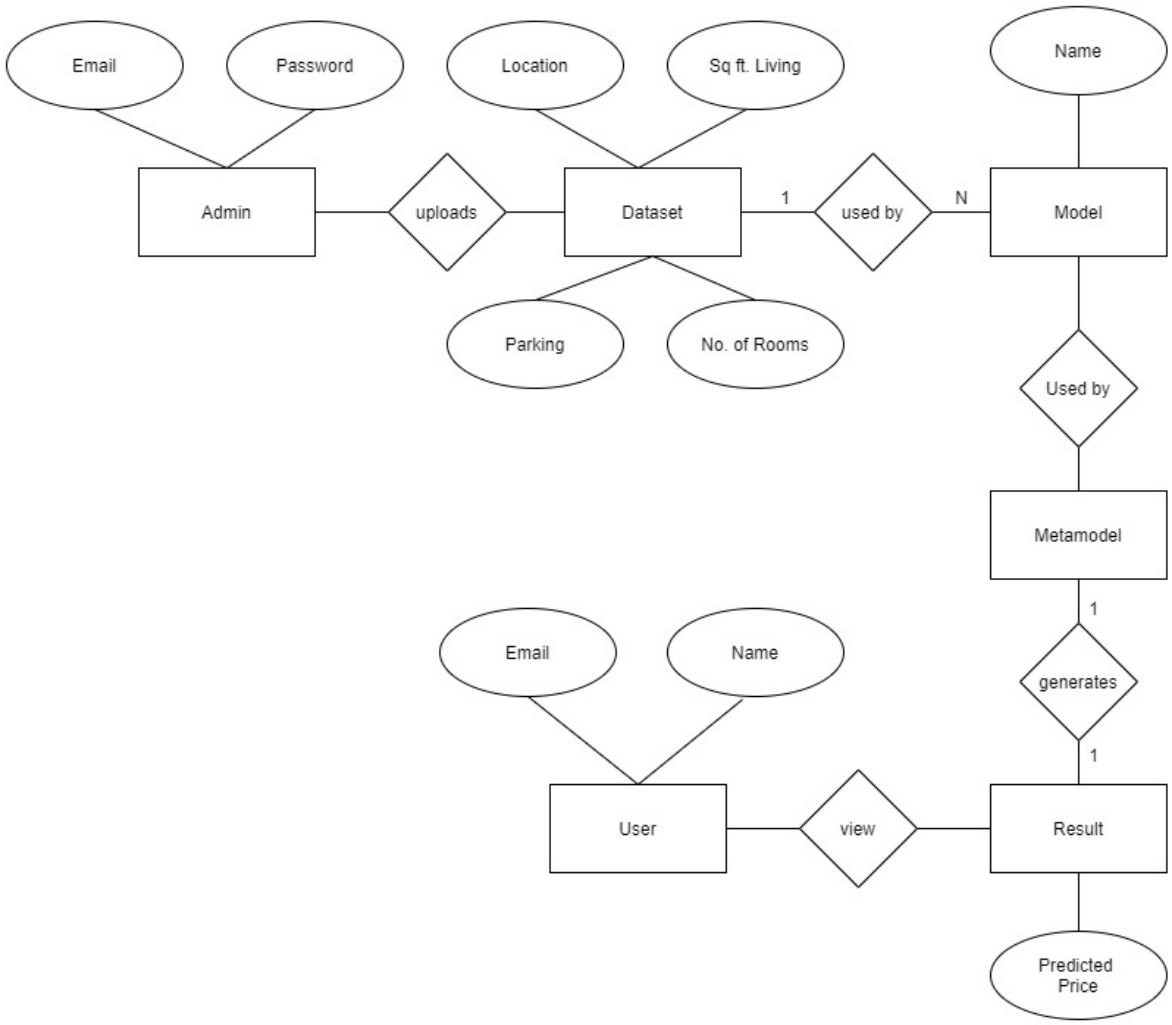


Figure 4.4: Entity Relationship Diagram

Entity Relationship Diagram shows entities within the system scope, and the inter-relationships among these entities. This diagram shows how the Robot, User and Unity are present in the system and how they are connected with each other.

4.4 UML Diagrams

A UML diagram is a diagram based on the UML (Unified Modeling Language) with the purpose of visually representing a system along with its main actors, roles, actions, artifacts or classes, in order to better understand, alter, maintain, or document information about the system.

4.4.1 Use case

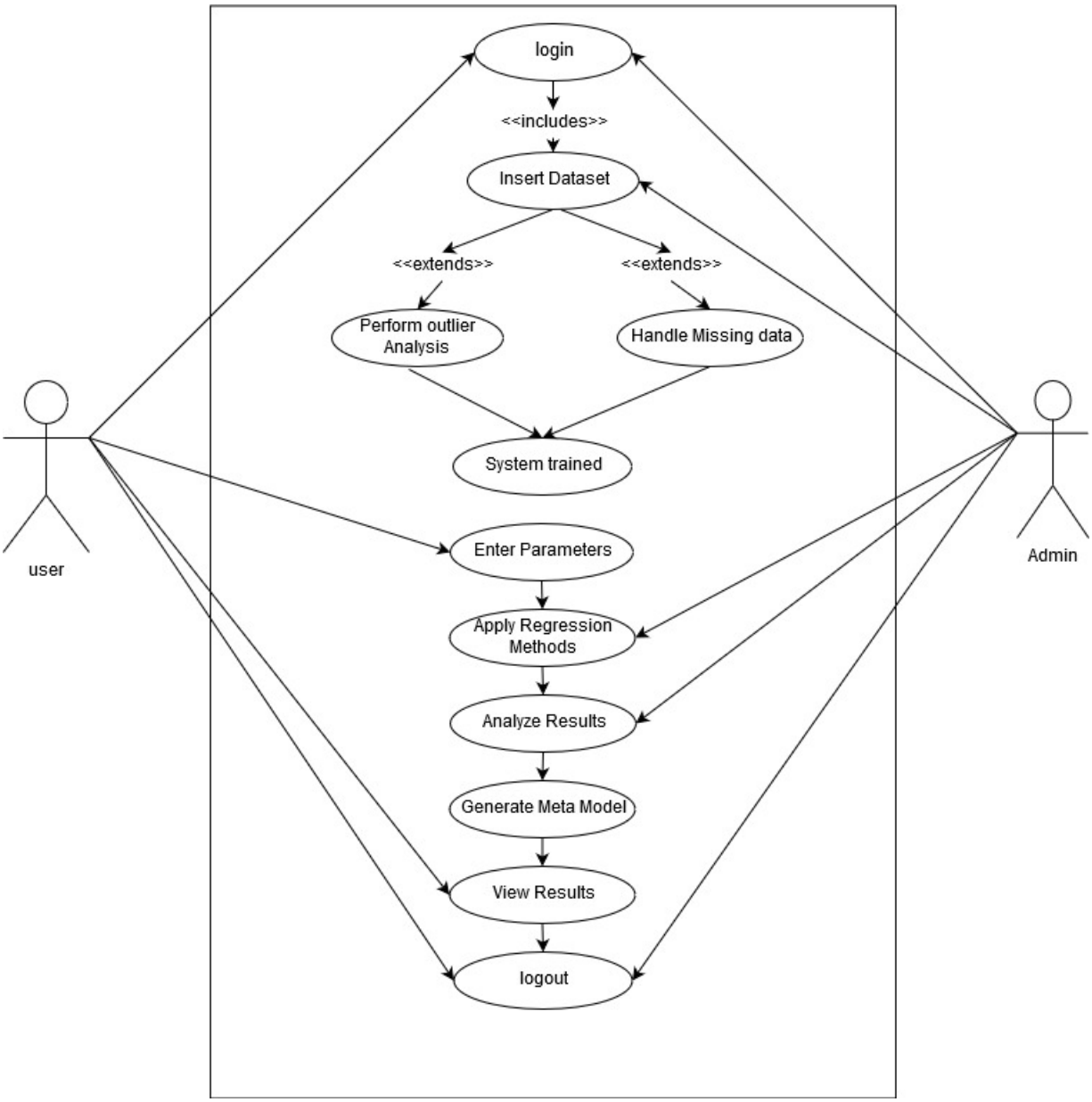


Figure 4.5: Use case Diagram

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. This diagram shows the set of actions, services and functions that the system needs to perform.

4.4.2 Sequence Diagram

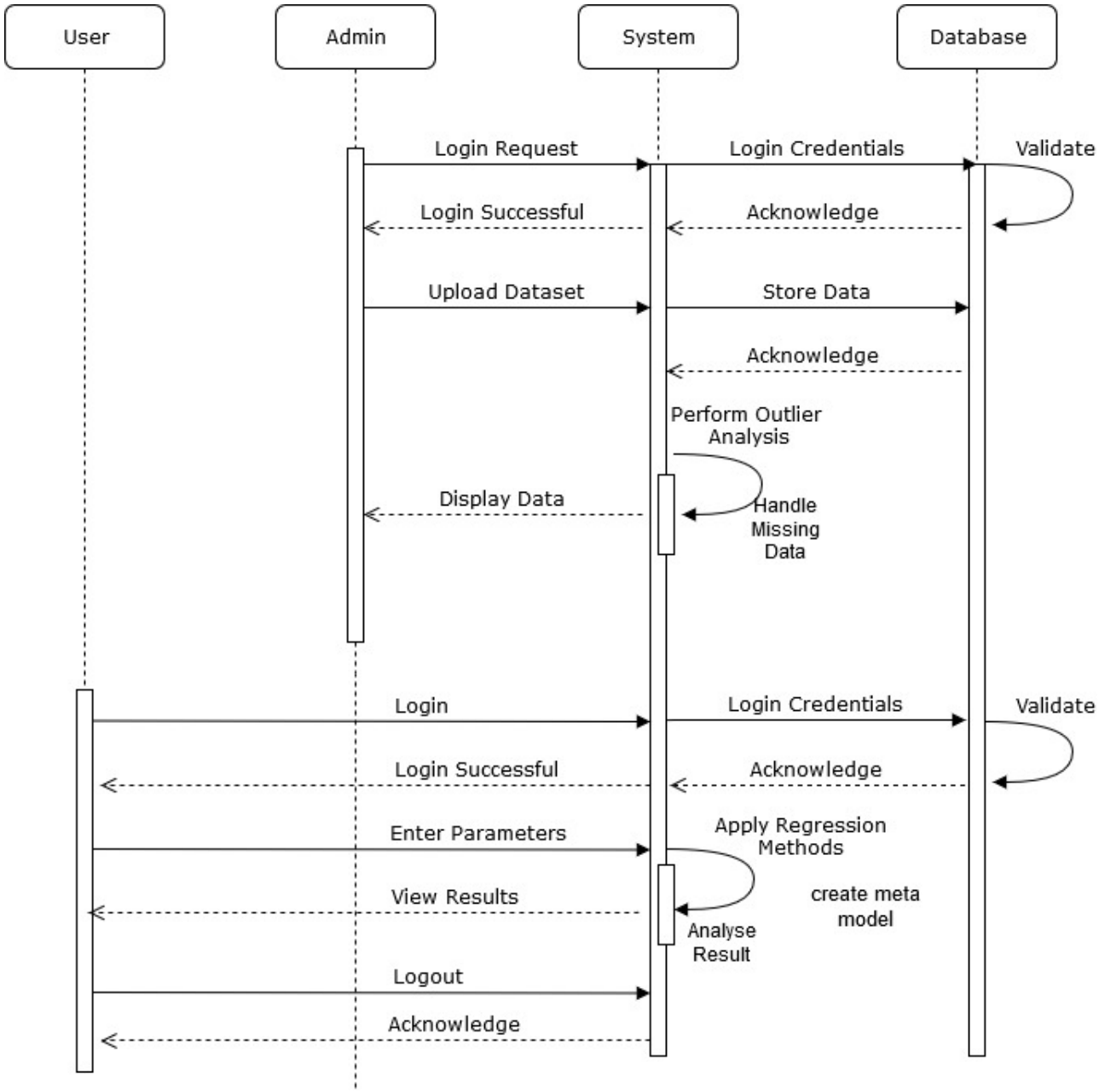


Figure 4.6: Sequence Diagram

A sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario. We have User, Admin, System and Database and they interact with each other.

4.4.3 Component Diagram

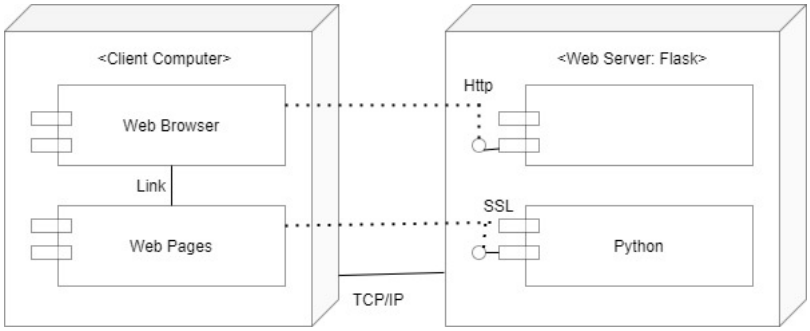


Figure 4.7: Component diagram

Component diagram is a special kind of diagram in UML. It does not describe the functionality of the system but it describes the components used to make those functionalities.

4.4.4 Class Diagram

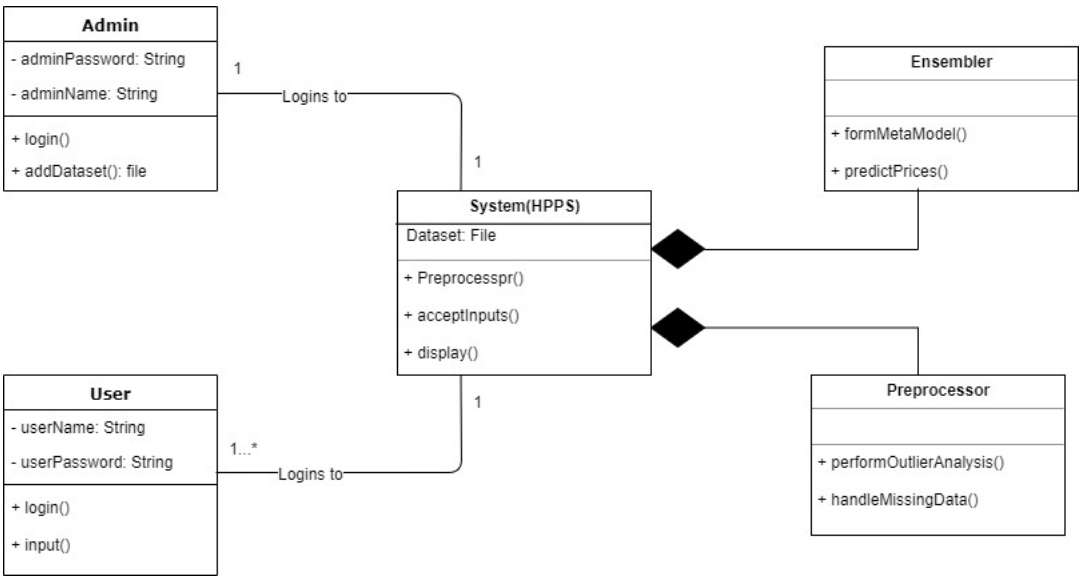


Figure 4.8: Class diagram

Class diagram is a static diagram. It represents the static view of an application. Class diagram is used for visualizing, describing, and documenting different aspects of a system. This diagram shows the User, Admin, System, Ensembler and Preprocessor.

CHAPTER 5

PROJECT PLAN

5.1 Project Estimate

Analysis is the process of cost evaluation of a project which includes labor cost, product development cost, resource cost as well as it is used for estimating profit of the project. For Cost analysis, we are using COCOMO model which is also called as constructive cost model. It is used to compute the efforts required for software development. It is a function of the program size in which the program size is nothing but the line of source code expressed in thousands.

As COCOMO model has three classes:

- Organic Projects
- Semi Detached Projects
- Embedded Projects

Here in our project, as we are a small team having good knowledge of the domain who can work efficiently with rigid requirements, we will choose the organic project class from COCOMO model. The basic COCOMO equations take the form

$$Effort\ Applied(E) = a(KLOC)^b$$

$$Development\ Time(D) = c(Effort\ Applied)^d$$

$$People\ required(P) = \frac{Effort\ Applied}{Development\ Time}$$

Here KLOC is the Line of Source code estimated in thousands. Efforts applied is the number of lines of code per hour by a person. Development time is estimated time required to develop the KLOC in months. People required is the man force which will write the KLOC for the development of the problem in the estimated time. Constant variables are a, b, c, d.

So, in our program our assumptions are as follows:

$Effort\ Applied(E) = a(KLOC)^b$

$= 2.4(12)^{1.05}$

$= 33.97$

Software Project	a	b	c	d
Organic Projects	2.4	1.05	2.5	0.38
semi detached Projects	3.0	1.12	2.5	0.35
Embedded Project	3.6	1.20	2.5	0.32

Table 5.1: COCOMO Constant Variables

$$Development\,Time(D) = c(EffortApplied)^d$$

$$= 2.5(33.97)^{0.38}$$

$$= 9.54$$

$$People\,required(P) = \frac{Effort\,Applied}{Development\,Time}$$

$$= \frac{33.97}{9.54}$$

$$= 3.56$$

$$\approx 4People$$

effort = a * KLOC^b, in person/months, with KLOC = lines of code, (in the thousands)
duration = c * effort^d
staffing = effort / duration

"A" variable	2.4
"B" variable	1.05
"C" variable	2.5
"D" variable	0.38
KLOC	18
EFFORT, (in person/months)	49.916873063396444
DURATION, (in months)	11.047754376548198
STAFFING, (recommended)	4.518282300822899

Figure 5.1: Cocomo Model

5.1.1 Project Resources

- Software Resources :

1. Platform : Flask, Android
2. Language : Python

5.2 Risk Management

Risk management is the process of identifying, assessing and controlling threats to an organization's capital and earnings. These threats, or risks, could stem from a wide variety of sources, including financial uncertainty, legal liabilities, strategic management errors, accidents and natural disasters.

5.2.1 Risk Identification

It is the process of determining potential threats which can later harm the performance of the project. One method of identifying risks is to create a risk item checklist. The checklist can be used for risk identification and focuses on some subset of known and predictable risks in the following generic subcategories:

- System Crash: System may crash due to overload.
- Storage : Risk Associated with storage of data.
- Response Time : First time model building may take long time.
- Security : Risk associated with access to each stakeholder.
- Connectivity : Stable Internet connectivity is required.

5.2.2 Risk Analysis

Risk analysis is concerned with analyzing the identified risks and by using Risk management drawing up plans to minimize their effect on a project. There might be a chance to happen or not.

Risk Analysis & Mitigation

Low Level	Medium level	High level
Dataset maybe missing	Dataset preprocessing may not done properly	System may crash due to overload.
Dataset attributes may be repeated	Dataset maybe in different format	Improper implementation of algorithms
Data maybe dirty	Preprocessing maybe time consuming process	_
Use of outlier function and proper handling of missing data	Use of sufficient hardware to reduce processing time	Regular backup of data and system logs

Figure 5.2: Risk Analysis

5.2.3 Overview of Risk Mitigation, Monitoring, Management

Risk mitigation planning is the process of developing options and actions to enhance opportunities and reduce threats to project objectives.

Risk : Computer Crash

- Mitigation - The cost associated with a computer crash resulting in a loss of data is crucial. A computer crash itself is not crucial, but rather the loss of data. A loss of data will result in not being able to deliver the product to the customer. This will result in a not receiving a letter of acceptance from the customer.
- Monitoring - When working on the product or documentation, the staff member should always be aware of the stability of the computing environment they're working in. Any changes in the stability of the environment should be recognized and taken seriously.

- Management - The lack of a stable-computing environment is extremely hazardous to a software development team. In the event that the computing environment is found unstable, the development team should cease work on that system until the environment is made stable again, or should move to a system that is stable and continue working there.

Risk : Technology not meeting requirement

- Mitigation - In order to prevent this from happening, meetings (formal and informal) will be held with the customer on a routine business. This insures that the product we are producing, and the specifications of the customer are equivalent.

- Monitoring - The meetings with the customer should ensure that the customer and our organization understand each other and the requirements for the product.

- Management - Should the development team come to the realization that their idea of the product specifications differs from those of the customer, the customer should be immediately notified and whatever steps necessary to rectify this problem should be done.

5.3 Project Schedule

5.3.1 Project Task Set

A task is accomplished by a set deadline, and must contribute toward work-related objectives. Just as project management is the coordination of individual tasks, a task can be broken down further into subtasks, which should also have clear start and end dates for completion.

	Project	**House Price Prediction System**		
	Starting Date	29-08-2019		
	Completion Date	22-02-2019		
Sr. No.	**Task Description**	**Starting Date**	**Ending Date**	**No. of Days**
1	Start of project	29-08-2019	-	-
2	Information Gathering	29-08-2019	29-09-2019	31
3	Surveying	30-09-2019	15-10-2019	16
4	Data Preprocessing	16-10-2019	30-10-2019	15
5	Modular Model building	31-10-2019	19-11-2019	20
6	Model Testing	20-11-2019	07-12-2019	18
7	Building UI	08-12-2019	01-01-2020	25
8	Integration	02-01-2020	29-01-2020	28
9	Testing	30-01-2019	22-02-2020	24
10	Completion	-		-

Figure 5.3: Project Task Set

5.3.2 Timeline Chart

A timeline chart is an effective way to visualize a process using chronological order. Since details are displayed graphically, important points in time can be easy seen and understood. Often used for managing a project's schedule, timeline charts function as a sort of calendar of events within a specific period of time. A Timeline for various modules in our system takes different period of time for complete it's operations.

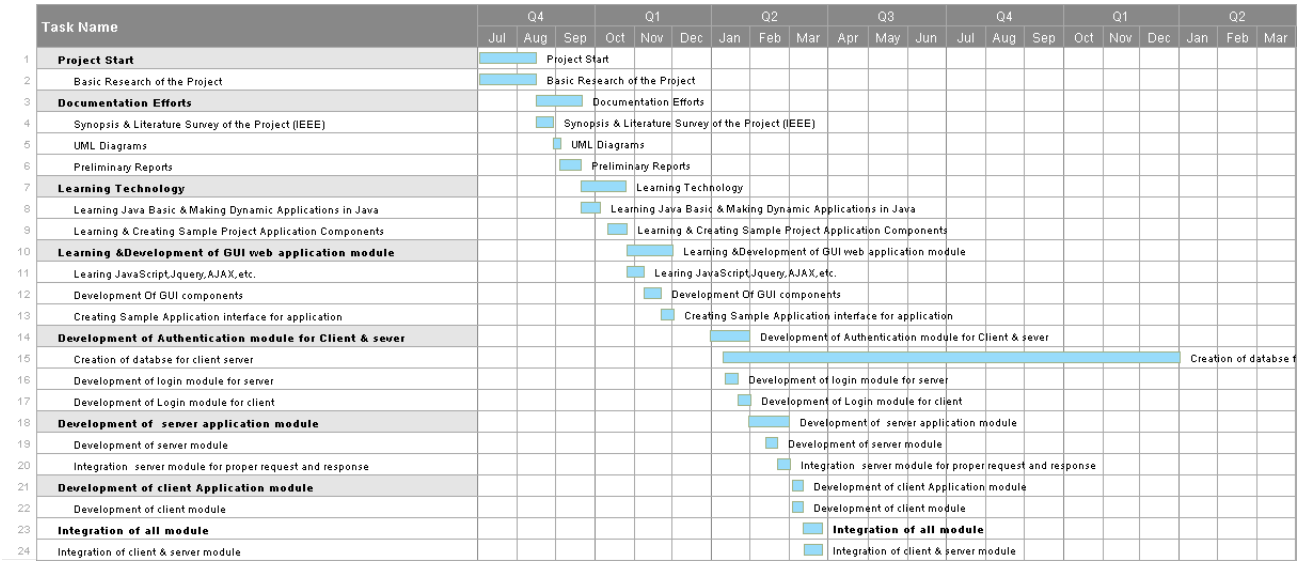


Figure 5.4: Timeline Chart

5.4 Team organization

5.4.1 Team Structure

Each and every member of the team is responsible for the identification of problems, proposing problem solving methodologies, identifying approaches for implementation and documentation.

Sr. No	Member	Responsibility
1	Abhishek Thombare	Project analysis,Developer and Design
2	Sumit Chavan	Requirement Gathering And Developer
3	Gautam Mudaliar	Requirement Gathering And Developer
4	Nimish Vaidya	Testing and Design

Table 5.2: Team Structure

5.4.2 Management reporting and communication

We report the progress of our project to our internal guide twice a week. We show our weekly status to our guide and incorporate the necessary changes. We

communicate among ourselves in case we want suggestions while executing our tasks.

CHAPTER 6

PROJECT IMPLEMENTATION

6.1 Overview of Project Modules

The proposed system has been implemented as a web application as well as android mobile application. The aim of project is to accurately predict prices of houses in Seattle city. The module involved in the application are explained below :

1. Upload Dataset
2. Preprocessing
3. Training submodel
4. Testing submodel
5. Training meta model
6. Prediction

6.2 Tools and Technologies used

1. Python:

Python is an interpreted, high-level and general-purpose programming language. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly, procedural), object-oriented, and functional programming.

2. FLASK :

Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions.

3. Spyder :

Spyder is a powerful scientific environment written in Python, for Python, and designed by and for scientists, engineers and data analysts. It offers a unique combination of the advanced editing, analysis, debugging, and profiling functionality of a comprehensive development tool with the data exploration, interactive execution, deep inspection, and beautiful visualization capabilities of a scientific package.

Algorithm For Predicting

Step 1: Use Linear regression:

It is a linear approach to model the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables).

Step 2: Use Ridge regression:

It is a way to create a model when the number of predictor variables in a set exceeds the number of observations, or when a data set has multicollinearity (correlations between predictor variables).

Step 3: Use Lasso regression:

It is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean.

Step 4: Use Random forest regression:

Random forest is a Supervised Learning algorithm which uses ensemble learning method for classification and regression. Random forest is a bagging technique and not a boosting technique. The trees in random forests are run in parallel.

Step 5: Stacking Based Ensemble

1. Learn first-level classifiers based on the original training data set.
2. Construct a new data set based on the output of base classifiers.
3. Learn a second-level classifier based on the newly constructed data set.

6.3 Mathematical Model

X: the matrix of input features (nrow: N, ncol: M+1) Y: the actual outcome variable (length:N) Y-hat: these are predicted values of Y (length:N) W: the weights or the coefficients (length: M+1) Here, N is the total number of data points available and M is the total number of features. X has M+1 columns because of M features and 1 intercept. The predicted outcome for any data point i is:

$$\hat{y}_i = \sum_{j=0}^{M} w_j * x_{ij}$$

Figure 6.1: Level 1 : Dataflow Diagram

1. Simple Linear Regression The objective function (also called as the cost) to be minimized is just the RSS (Residual Sum of Squares), i.e. the sum of squared errors of the predicted outcome as compared to the actual outcome. This can be depicted mathematically as:

$$Cost\ (W) = RSS(W) = \sum_{i=1}^{N} \{y_i - \hat{y}_i\}^2 = \sum_{i=1}^{N} \left\{ y_i - \sum_{j=0}^{M} w_j\, x_{ij} \right\}^2$$

2. Ridge Regression The objective function (also called the cost) to be minimized is the RSS plus the sum of square of the magnitude of weights. This can be depicted mathematically as:

$$Cost\ (W) = RSS(W) + \lambda * (sum\ of\ squares\ of\ weights)$$

$$= \sum_{i=1}^{N} \left\{ y_i - \sum_{j=0}^{M} w_j\, x_{ij} \right\}^2 + \lambda \sum_{j=0}^{M} w_j^2$$

3. Lasso Regression The objective function (also called the cost) to be minimized is the RSS plus the sum of absolute value of the magnitude of weights. This can be depicted mathematically as:

$$Cost\ (W) = RSS(W) + \lambda * (sum\ of\ absolute\ value\ of\ weights)$$

$$= \sum_{i=1}^{N} \left\{ y_i - \sum_{j=0}^{M} w_j\, x_{ij} \right\}^2 + \lambda \sum_{j=0}^{M} |w_j|$$

CHAPTER 7

SOFTWARE TESTING

Software testing is an activity aimed at evaluating an attribute or capability of a program or system and determining that it meets its required results. It is more than just running a program with the intention of finding faults. Every project is new with different parameters. No single yardstick maybe applicable in all circumstances. This is a unique and critical area with altogether different problems. Although critical to software quality and widely deployed by programs and testers.

Software testing still remains an art, due to limited understanding of principles of software. The difficulty systems from complexity of software. The purpose of software testing can be quality assurance, verification and validation or reliability estimation

7.1 Type of Testing

7.1.1 Unit Testing

Unit Testing is a method by which individual units of source code are tested to determine if they are fit to use.

1. First we check whether values entered by user are being accepted correctly.
2. Then we check whether first level models are predicting accurately.
3. Then we check if meta model is predicting prices accurately.

7.1.2 Integration Testing

It is a systematic approach for conducting tests to uncover interfacing errors. The main objective behind integration testing is to accept all the unit tested components and integrate into a program structure as given by the design.

Checking if accurate values are submitted to the first level models.

Checking if meta model is accepting accurate values from first level model.

7.1.3 System Testing

It is a level of software testing where a complete and integrated software is tested. The purpose of this test is to evaluate the system's compliance with the specified requirements.

7.1.4 Acceptance Testing

Cross validation can be used to simulate user inputs and gauge the accuracy of the system.

7.2 Test Cases and Test Results

Test ID	Test Case Name	Test Case Steps	Expected Result	Actual Result	Test Status
HPP_1	Check Web app accessibility	1. Open URL 2. Browser Compatibility	Website Opened	Website load without error	Pass
HPP_2	House Image visibility	Image visible after loading	Image visible	Image loaded	Pass
HPP_3	Button Clickability	1. Button visible 2. Click Button	Option toggled	Option toggled	Pass
HPP_4	Connection established	1. Click on button 2. Check dropdown	Able to select values from options	Values fetched and able to select	Pass
HPP_5	Add Sale data entry	1. Open tab 2. Enter required data	Data entered and submitted	Data edited, submitted and confirmation.	Pass
HPP_6	Predict price	1. Enter details 2. Click button 3. Internal calculation and result obtained	Able to predict price	Price predicted	Pass

Figure 7.1: Testcases

CHAPTER 8

RESULTS

8.1 Outcomes

With the help of proposed system We are getting accurate house price and also showing similar houses on google map along with price bands.

1. We have used Ensembling method Stacking
2. User can view house prices in nearby area and choose appropriate area for buying house.

8.2 Tables

	Train	Test
Linear Regression	75.03%	75.39%
Random Forest	95.04%	88.69%
Lasso Regression	75.04%	75.41%
Ridge Regression	75.03%	75.40%
Linear Regression(Stacking)	94.726%	88.993%

8.3 Graph

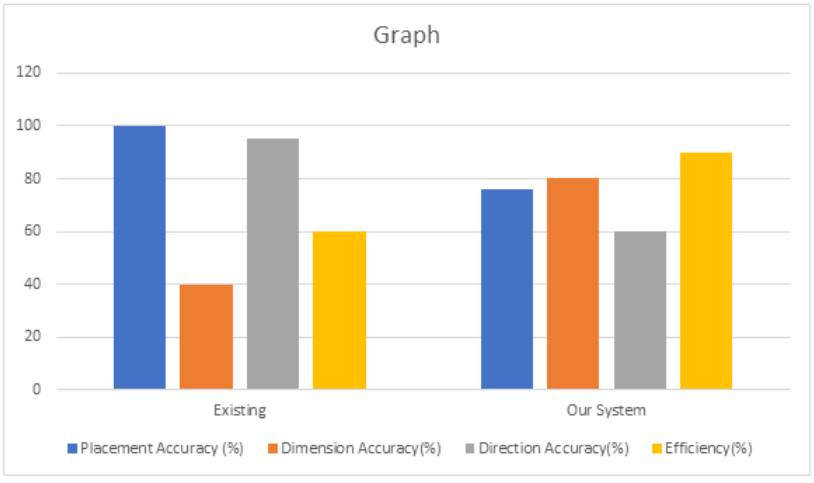


Figure 8.1: Graph

8.4 Screenshots

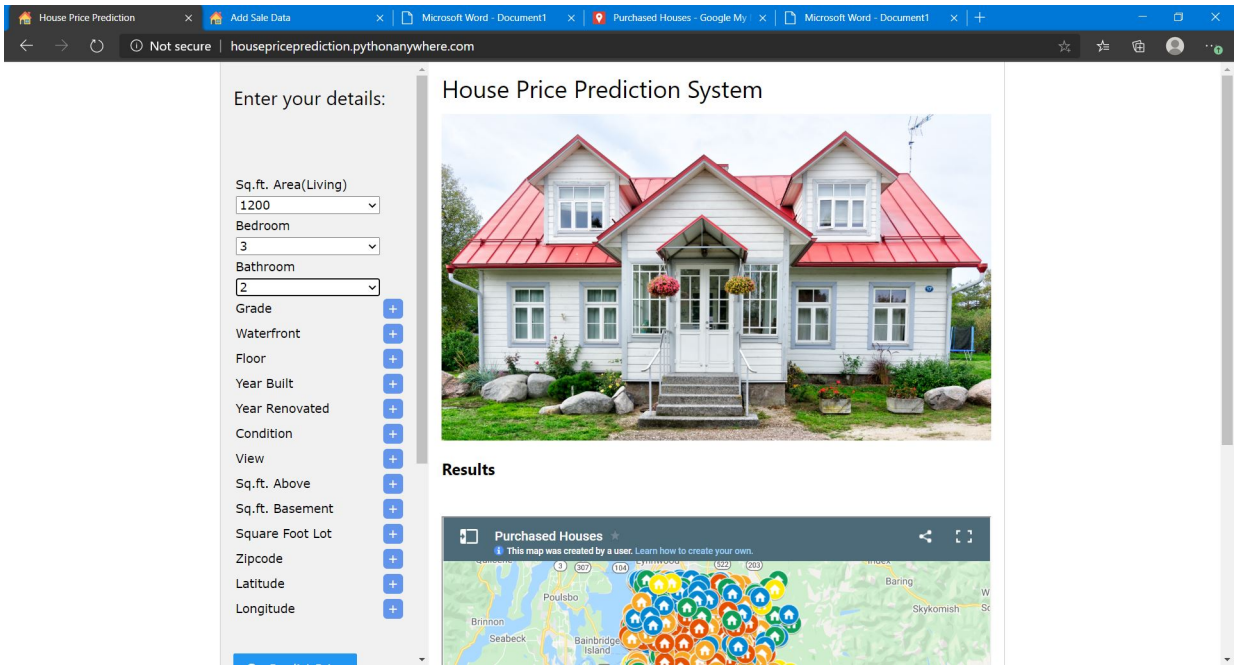


Figure 8.2: Website View

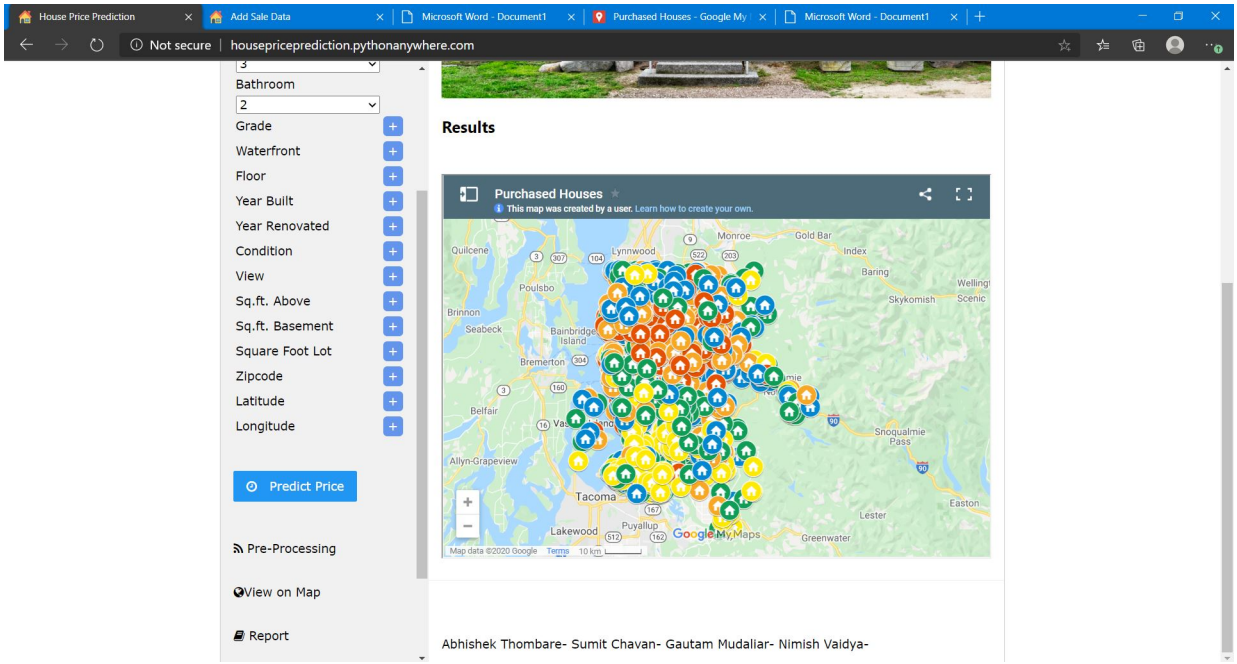


Figure 8.3: Google Map with price band

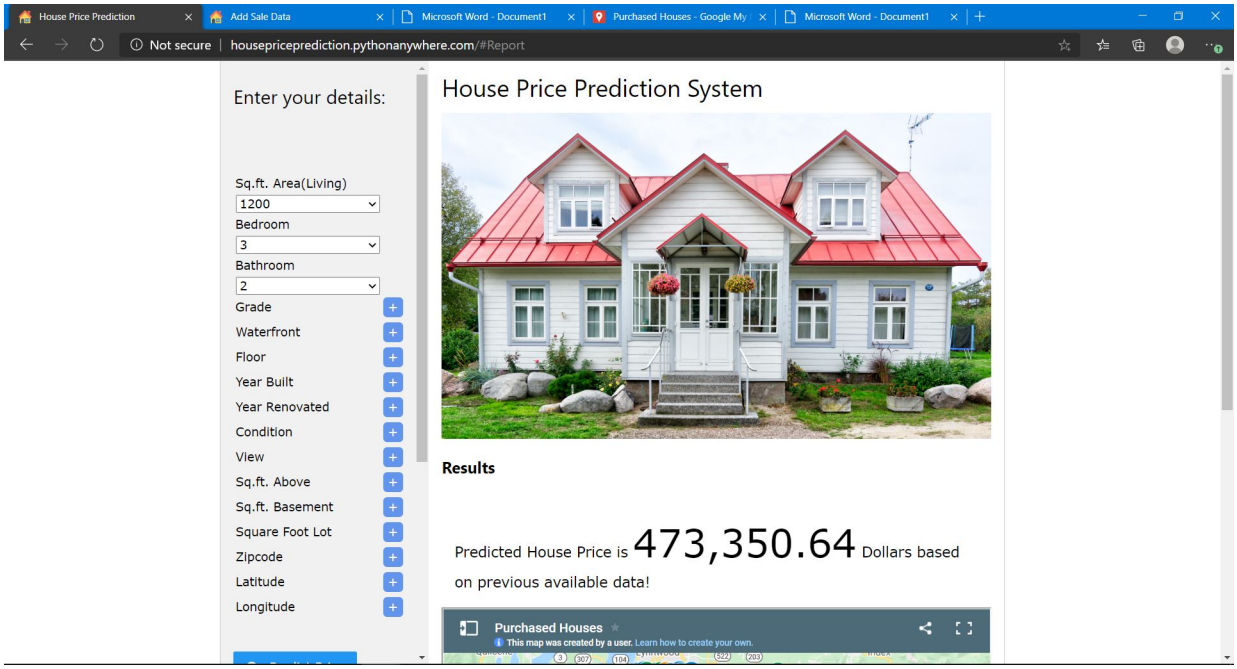


Figure 8.4: Calculated price

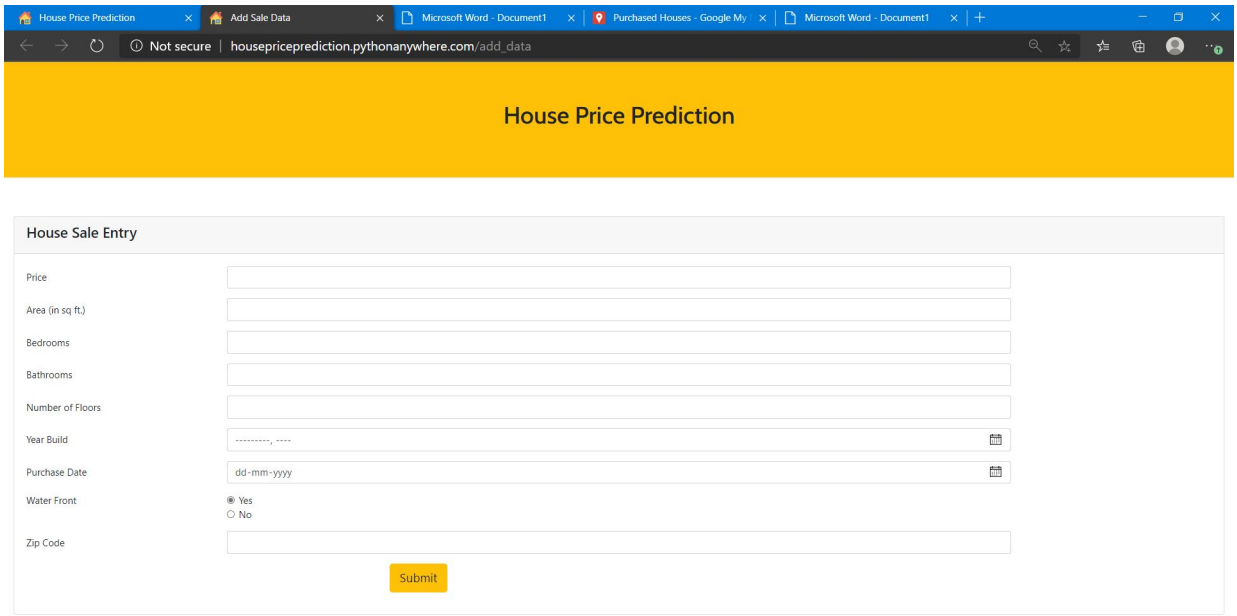


Figure 8.5: Form to submit new data

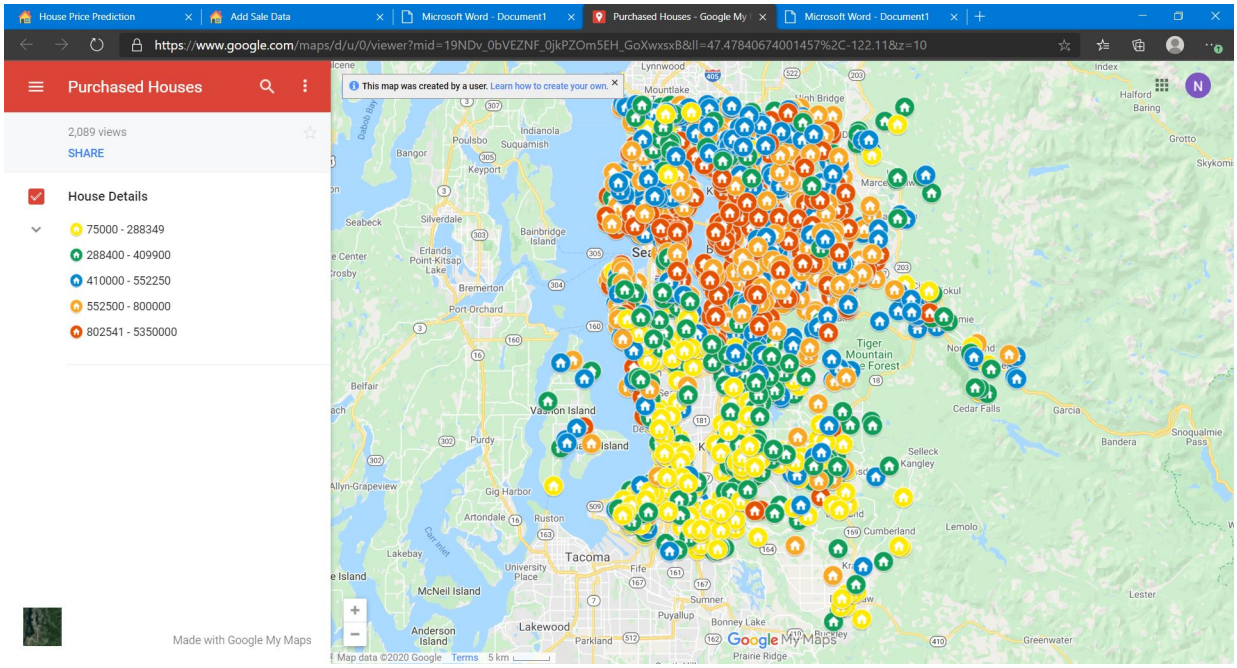


Figure 8.6: Google Maps with price band

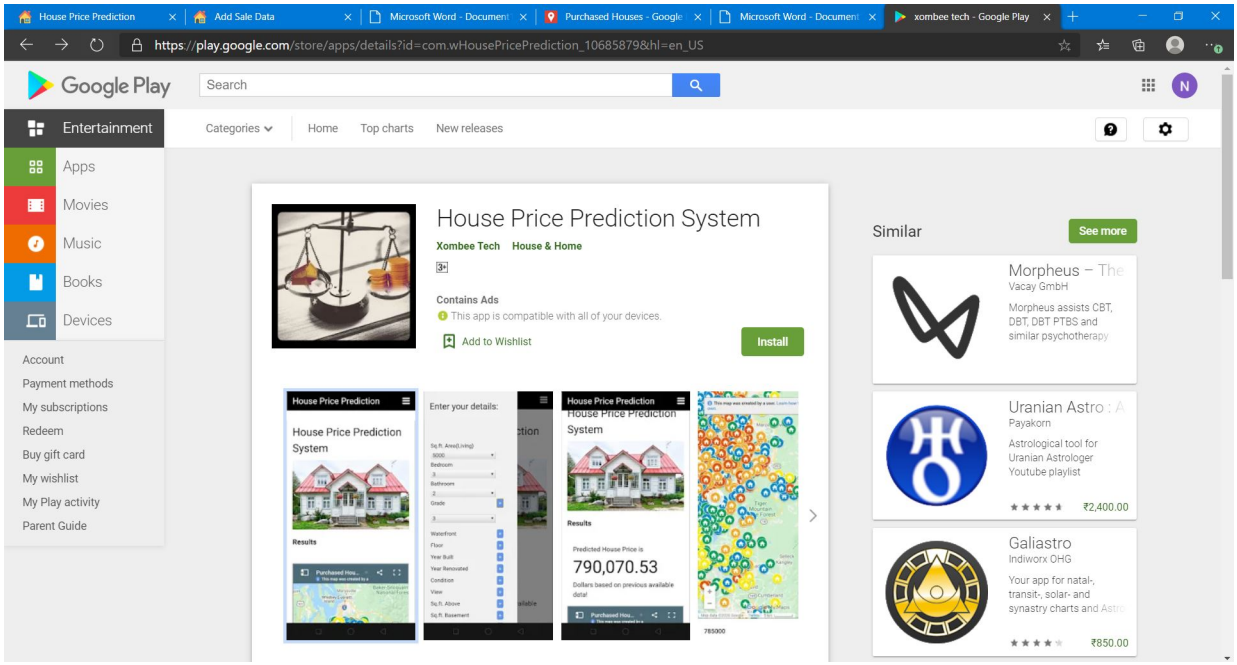


Figure 8.7: Application on Google Play Store

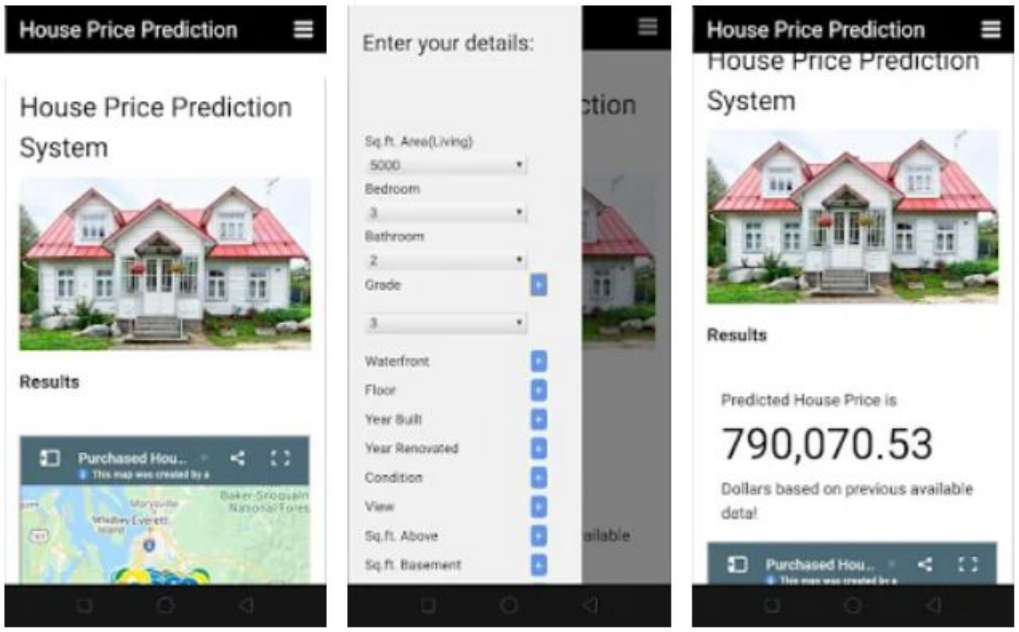


Figure 8.8: Screenshot of Android app

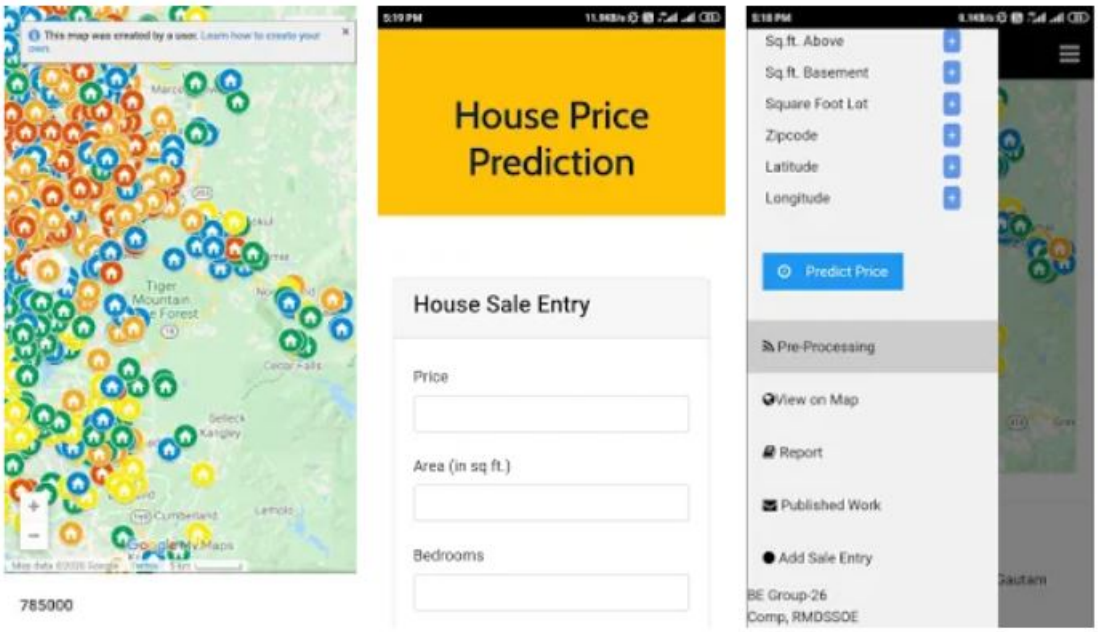


Figure 8.9: Screenshot of Android app

CHAPTER 9

OTHER SPECIFICATION

9.1 Advantages

- It is not mandatory to enter all parameters.
- Increased accuracy compared to earlier models.
- Ability to view houses on google maps with price bands.

9.2 Limitations

- Project limited to houses of Seattle city.
- Limited dataset availability.
- Input of less data parameters may lead to variance in predicted and actual price.

CHAPTER 10

CONCLUSION AND FUTURE WORK

10.1 Conclusion

We believe that greater accuracy in determining house prices can be achieved by using stacking of linear regression, random forest regression, lasso regression and ridge regression.

10.2 Future Work

We plan to extend the scope of this project beyond seattle city. We plan to use multi-layer stacking to improve the accuracy further. More parameters can be taken into account for better results.

10.3 Applications

- To predict the price of any type of property.
- Banks need to predict house prices to estimate asset value and to determine grantable loan amount.
- House owners and customers can use this system to minimize middle man cost
- Brokers can provide valuable advice regarding sales and investment to the customers using this system.

REFERENCES

1. J. J. WANG, S. G. Hu , X. T. Zhan , Q. Luo , Q. Yu , "Predicting House Price With a Memristor-Based Artificial Neural Network" IEEE Access DOI: 10.1109/ACCESS.2018.2814065

2. Ayush Varma, Abhijit Sarma, Sagar Doshi, Rohini Nair "House Price Prediction Using Machine Learning And Neural Networks" in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)

3. Rushab Sawant, Saurabh Jain, Tushar Tiwari, Yashwant Jangid, Ms.Ankita Gupta, "A Multi Feature Based Housing Price Prediction for Indian Market Using Machine Learning" International Journal of Computer Mathematical Sciences, Vol6 Issue 12

4. Muhammad Fahmi Mukhlishin, Ragil Saputra , Adi Wibowo, "Predicting House Sale Price Using Fuzzy Logic, Artificial Neural Network and K-Nearest Neighbour" in 2017 1st International Conference on Informatics and Computational Sciences (ICICoS)

5. Yingyu Feng, "Comparing Multilevel Modelling and Artificial Neural Networks in House Price Prediction" 2015 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM)

6. Yu, Jiafu Wu, "Real Estate Price Prediction with Regression and Classification" CS 229 Autumn 2016 Project Final Report, Stanford University

7. L.Li, K.-H. Chu, "Prediction of real estate price variation based on economic parameters" 2017 International Conference on Applied System Innovation (ICASI).

8. Xianghan Zheng, Hao Tian, "House Price Forecast Based on Dynamic Model Averaging Model Combined With Web Search Index" 2018 International Conference on Cloud Computing, Big Data and Blockchain (ICCBB)

9. Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, " A hybrid regression technique for house prices prediction" 2017 IEEE International Conference on

Industrial Engineering and Engineering Management (IEEM) [10] Ceyhun Abbasov, “The prediction of the chance of selling of houses as the factor of financial stability” 2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT)

10. Yang Li, Quan Pan, Tao Yang, Lantian Guo, “Reasonable price recommendation on Airbnb using Multi-Scale clustering” 2016 35th Chinese Control Conference (CCC)

11. CH. Raga Madhuri, VRSEC, Vijayawada, G Anuradha, M. Vani Pujitha, “House Price Prediction Using Regression Techniques: A Comparative Study” 2019 International Conference on Smart Structures and Systems (ICSSS)

12. Nehal N Ghosalkar, Sudhir N Dhage, “Real Estate Value Prediction Using Linear Regression” 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)

13. Parasich Andrey Viktorovich, Parasich Viktor Aleksandrovich, Kaftannikov Igor Leopoldovich, “ Predicting Sales Prices of the Houses Using Regression Methods of Machine Learning” 2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC)

APPENDIX A

Grenze International Journal of Engineering and Technology, Special Issue

Survey on House Price Prediction and Analysis

Nimish Vaidya [1], Gautam Mudaliar[2], Abhishek Thombare[3] and Sumit Chavan[4]
[1-4]RMD Sinhgad School of Engineering, Pune, India
Email: nimishvaidya99@gmail.com; gautammudliar@gmail.com

***Abstract*—Real Estate / Land is the least straight-forward industry in our society. House costs keep changing everyday consistently and from time to time are publicized instead of being established on valuation. House estimation process is a critical purpose of land which can be used to gain profits. The composition tries to get supportive gaining from valid data of property markets. Computer based intelligence strategies are applied to explore evident property trades to discover profitable models for house buyers and vendors. For evaluating the cost various regression techniques were attempted and one with most raised exactness is used.**

***Index Terms*— regression, ridge regression, lasso regression, random forest tree, prediction, parameters.**

I. Introduction

The objective of the project is to develop the optimum machine learning model for predicting the price of property on a dataset of house sale prices in King County, Seattle from May 2014 to May 2015. The dataset provides various features houses have like locality, condition, size, age, etc, and the prices at which they were sold.

II. Related Work

The land business has turned into an aggressive and nontransparent industry. The information mining process in such an industry gives a preferred position to the engineers by preparing those information, estimating future patterns and consequently helping them to settle on ideal learning driven choices. Our primary concentration here is to build up a model which predicts the property cost for a client as indicated by his\her interests. Our model examinations a lot of parameters chosen by the client to locate a perfect value as per their necessities.
For this it uses techniques called linear regression, ridge regression, lasso regression, random forest tree, for prediction and tries to gives an analysis of the results obtained. It helps to establishes the relationship between dependent variable and other changing independent variable also known as label attribute and regular attribute respectively.

III. Proposed System

Our dataset involves different basic parameters and information mining has been at the base of our framework. We at first tidied up our whole dataset and furthermore truncated the exception esteems. Further,

Grenze ID: 01.GIJET.6.2.502
© Grenze Scientific Society, 2020

we gauged every parameter dependent on its significance in deciding the estimating of the framework and this drove us to build the worth that every parameter retains in the framework. We shortlisted 5 diverse AI calculations and tried our framework with various mixes that can ensure best potentially dependability of our outcomes.

IV. VARIABLES

A total of 21 parameters were considered out of which price was dependent and rest all were independent parameters.
id (a notation for a house), date (Date house was sold), price (Price is prediction target), bedrooms (Number of Bedrooms/House), bathrooms (Number of bathrooms/House), sqft_living (square footage of the home), sqft_lot (square foot of the lot), floors (Total floors (levels) in house), waterfront (House which has a view to a waterfront), view (Has been viewed), condition (How good the condition), grade (overall grade given to the housing unit, based on King County grading system), sqft_above (square footage of house apart from basement), sqft_basement (square footage of the basement), yr_built , yr_renovated (Year when house was renovated), Zipcode, lat (Latitude coordinate), long (Longitude coordinate), sqft_living15 (Living room area in 2015(implies—some renovations) This might or might not have affected the lotsize area), sqft_lot15 (lotSize area in 2015(implies some renovations))

TABLE I. CO-RELATION

Parameter	Price	Parameter	Price
price	1	waterfront	0.26636943
sqft_living	0.70203505	floors	0.25679388
grade	0.66743425	yr_renovated	0.12643379
sqft_above	0.60556729	sqft_lot	0.08966086
sqft_living15	0.58537890	sqft_lot15	0.08244715
bathrooms	0.52513750	yr_built	0.05401153
view	0.39729348	zipcode	-0.0532028
sqft_basement	0.32381602	condition	0.03636178
bedrooms	0.30834959	long	0.02162624
lat	0.30700348	id	-0.0167621

This table shows co-relation values of different parameters with parameter price.

V. ALGORITHMS

A. Linear Regression

Linear regression is the most straight forward technique for forecast. It utilizes two things as factors which are the predictor variable and the other variable which is the most essential one. These regression evaluations are utilized to clarify the connection between one dependent variable and at least or more independent variables. The equation of the regression equation with one dependent and one independent variable is defined by the formula.

$$b = y + x*a \qquad (1)$$

where, b = estimated dependent variable score, y = constant, x = regression coefficient, and a = score on the independent variable.

B. Forest Regression

It uses technique called as Bagging of trees. The main idea here is to decorrelate the several trees. We then reduce the Variance in the Trees by averaging them. Using this approach, a large number of decision trees are created. Random forest training algorithm applies the technique of bootstrap aggregating, to tree learners.
Given a training set X = x1, ..., xn with responses Y = y1, ..., yn, bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples: For b = 1, ..., B:
1. Sample, with replacement, n training examples from X, Y; call these Xb, Yb.
2. Train a classification or regression tree fb on Xb, Yb.

C. Ridge Regression

It is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. Ridge regression equation is written in matrix form as

$$Y = XB + e \quad (2)$$

where Y is the dependent variable, X represents the independent variables, B is the regression coefficients to be estimated, and e represents the errors are residuals.

D. Lasso Regression

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, similar to the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of muticollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

VI. Conclusions

The best model performance is given by Linear Regression with .88 R-Squared value and minimum RMSE Values. Thus we can conclude that 88% of variations in dependent variable, ie price, are explained by the independent variables present in our model.

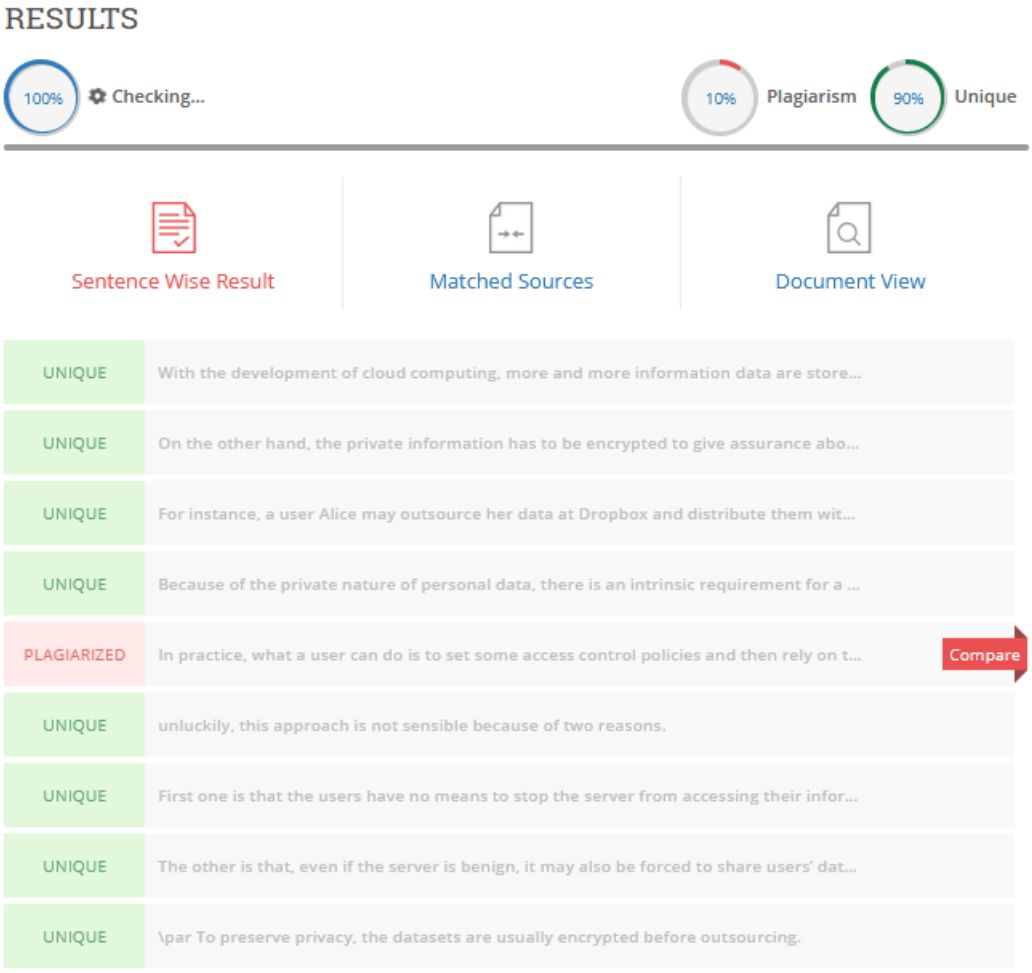
Name	rsquared	rmse
Linear Regression	0.88	127519.31
Ridge Regression	0.87	132052.04
Lasso Regression	0.85	142393.73
Decistion Tree	0.66	214681.94
Random Forest Tree	0.75	183825.07

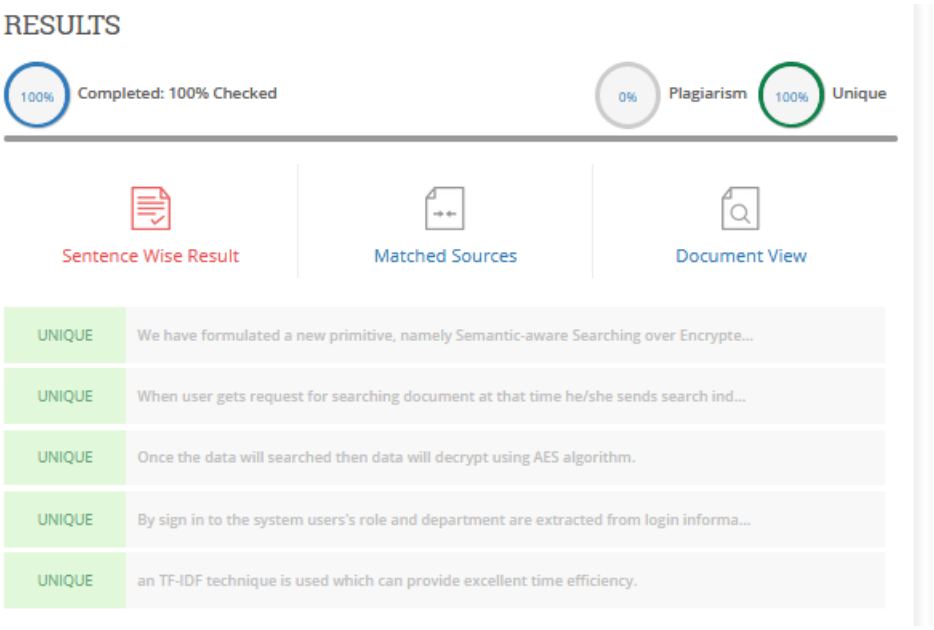
References

[1] J. J. WANG, "Predicting House Price With a Memristor-Based Artificial Neural Network"
[2] Ayush Varma, "House Price Prediction Using Machine Learning And Neural Networks"
[3] "Housing Price Prediction using Machine Learning Algorithms"
[4] Rushab Sawant, "A Multi Feature Based Housing Price Prediction for Indian Market Using Machine Learning"
[5] Muhammad Fahmi Mukhlishin, "Predicting House Sale Price Using Fuzzy Logic, Artificial Neural Network and K-Nearest Neighbour"
[6] Yingyu Feng, "Comparing Multilevel Modelling and Artificial Neural Networks in House Price Prediction"

APPENDIX B

Plagarism Report




RESULTS
100%
Completed: 100% Checked
0%
Plagiarism
100%
Unique
Sentence Wise Result
Matched Sources
Document View
UNIQUE
We have formulated a new primitive, namely Semantic-aware Searching over Encrypte...
UNIQUE
When user gets request for searching document at that time he/she sends search ind...
UNIQUE
Once the data will searched then data will decrypt using AES algorithm.
UNIQUE
By sign in to the system users's role and department are extracted from login informa...
UNIQUE
an TF-IDF technique is used which can provide excellent time efficiency.