A

Stage-1 Project Report

on

HOUSE PRICE PREDICTION USING MACHINE LEARNING

Submitted in Partial Fulfillment of
the Requirements for the Degree
of

Bachelor of Engineering

in

Computer Engineering

to

Kavayitri Bahinabai Chaudhari
North Maharashtra University, Jalgaon

Submitted by

Janhavi Kailas Chaudhari
Anup Sanjay Patil
Mayur Manohar Patil
Bhushan Ramkrushna Upasani
Priyakumari Kailassing Rajput

Under the Guidance of

Miss. Priti Sharma

DEPARTMENT OF COMPUTER ENGINEERING
SSBT's COLLEGE OF ENGINEERING AND TECHNOLOGY,
BAMBHORI, JALGAON - 425 001 (MS)
2020 - 2021

SSBT's COLLEGE OF ENGINEERING AND TECHNOLOGY, BAMBHORI, JALGAON - 425 001 (MS)

DEPARTMENT OF COMPUTER ENGINEERING

CERTIFICATE

This is to certify that the project entitled *House Price Prediction Using Machine Learning*, submitted by

Janhavi Kailas Chaudhari
Anup Sanjay Patil
Mayur Manohar Patil
Bhushan Ramkrushna Upasani
Priyakumari Kailassing Rajput

in partial fulfillment of the degree of *Bachelor of Engineering* in *Computer Engineering* has been satisfactorily carried out under my guidance as per the requirement of Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon.

Date: November 27, 2020
Place: Jalgaon

Miss. Priti Sharma
Guide

Prof. Dr. Girish K. Patnaik
Head

Prof. Dr. K. S. Wani
Principal

Acknowledgements

We take the opportunity to express our gratitude to all those who have rendered co-operation and guidance that supported us while analyzing and developing the design of this project right from beginning up till now. First of all we would like to express our deepest appreciation toward Dr. Girish K. Patnaik (H.O.D. of Computer Engineering) and Miss. Priti Sharma (Guide) whose valuable guidance supported us in completing the analysis and designing of our final year project. We would also like to thank and express our respect and gratitude to Dr. K. S. Wani (Principal) for providing us this opportunity. Last but not least we are profoundly grateful and obliged to Miss. Priti Sharma, for her guidance, motivation and continuous encouragement throughout to see that whole process is right on its target since its commencement.

Janhavi Kailas Chaudhari
Anup Sanjay Patil
Mayur Manohar Patil
Bhushan Ramkrushna Upasani
Priyakumari Kailassing Rajput

Contents

List of Tables

List of Figures

Abstract

Machine learning plays a major role from past years in image detection, spam reorganization, normal speech command, product recommendation and medical diagnosis. Present machine learning algorithm helps in enhancing security alerts, ensuring public safety and improve medical enhancements. Machine learning system also provides better customer service and safer automobile systems.

Designing an effective machine learning model for prediction of regression and classification problem is a tedious endeavor. Significant time and expertise are needed to customize the model for a specific problem. A significant way to reduce the complicated design is by using Automated Machine Learning (AML) that can intelligently optimize the best pipeline suitable for a problem or dataset. This study utilizes machine learning algorithms as a research method that develops housing price prediction models. In that point a housing cost prediction model to support a house vendor or a real estate agent for better information based on the valuation of house is recommended.

Chapter 1

Introduction

Proposed system aims to make a machine learning model which could predict the real estate house prices. Respective chapter has in view of the content focusing on scope, motivation, objectives and selection of life cycle model of the projected system. It also states the actual problem definition which is to be solved using machine learning models .
The Section 1.1 describes the background of the promised proposed system to be built. Motivation behind selecting particular problem statement is presented in section 1.2. The section 1.3 discuss about the problem definition. The scope of the system is discussed in section 1.4. The section 1.5 discuss about objectives of proposed system. The selection of life cycle model for development of system is discussed in section 1.6. The section 1.7 will discuss about organization of report. Finally, the summary is presented in the last section.

1.1 Background

Machine learning (ML) is the study of computer algorithms that improve automatically through experience. It is seen as a subset of artificial intelligence. Machine learning involves computers discovering how they can perform tasks without being explicitly programmed to do so. It involves computers learning from data provided so that they carry out certain tasks. For simple tasks assigned to computers, it is possible to program algorithms telling the machine how to execute all steps required to solve the problem at hand; on the computer's part, no learning is needed. For more advanced tasks, it can be challenging for a human to manually create the needed algorithms. In practice, it can turn out to be more effective to help the machine develop its own algorithm, rather than having human programmers specify every needed step. The term machine learning was coined in 1959 by Arthur Samuel, an American IBMer and pioneer in the field of computer gaming and artificial intelligence.Tom M. Mitchell provided a widely quoted, more formal definition of the algorithms studied in the machine learning field: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T,

as measured by P, improves with experience E." A core objective of a learner is to generalize from its experience. Generalization in this context is the ability of a learning machine to perform accurately on new, unseen examples/tasks after having experienced a learning data set. The training examples come from some generally unknown probability distribution (considered representative of the space of occurrences) and the learner has to build a general model about this space that enables it to produce sufficiently accurate predictions in new cases.

1.2 Motivation

Being extremely interested in everything having a relation with the Machine Learning, the final year project was a great occasion to devote time to learn and confirm the interest for this field. The fact that machine learning can make estimations, predictions and give the ability for machines to learn by themselves is both powerful and limitless in terms of application possibilities.
Machine learning can be used in finance, medicine, almost everywhere. Hence that is the pivotal motivation of carrying out the final year project around machine learning domain.

1.3 Problem Definition

In India, there are multiple real estate classified websites where properties are listed for sell/ buy/ rent purposes such as 99acres, housing, commonfloor, magicbricks and more. However, in each of these websites we can see lot of inconsistencies in terms of pricing of an apartment and there are some cases when similar apartments are priced differently and thus there is lot of in-transparency. Sometimes the consumers may feel the pricing is not justified for a particular listed apartment but there no way to confirm that either. Proper and justified prices of properties can bring in a lot of transparency and trust back to the real estate industry, which is very important as for most consumers especially in India the transaction prices are quite high and addressing this issue will help both the customers and the real estate industry in the long run. We propose to use machine learning and artificial intelligence techniques to develop an algorithm that can predict housing prices based on certain input features.
The business application of this algorithm is that classified websites can directly use this algorithm to predict prices of new properties that are going to be listed by taking some input variables and predicting the correct and justified price i.e. avoid taking price inputs from customers and thus not letting any error creeping in the system. This study on proactive pricing of houses in the Indian context has never been reported earlier in the literature to

the best of our knowledge.
However, the problem of house price prediction is quite old and there have been many studies and competitions addressing the same including the classic Boston housing price challenge on Kaggle. As far as housing price prediction in Indian context is concerned, use of machine learning algorithm can really get the job done.

1.4 Scope

Scope of Machine Learning (ML) is vast, and in the near future, it will deepen its reach into various fields like medical, finance, social media, facial and voice recognition, online fraud detection, and real estate. Gartner predicts that 30% of Government and large enterprise contracts will require AI-fueled solutions by 2025. ML shall also fuel areas that are highly dependent on data. For example, the real estate, machine learning can analyse data and evaluate it for better results.

1.5 Objectives

1. Predicting the house prices.
2. Minimizing the difference between predicted and actual house prices.
3. Machine learning in business therefore offers an important commercial benefit in being able to make the best use of your data.
4. Indeed, a key objective of the project is to enable you to keep up with those competitors already making best use of their data to maximise business opportunities.

1.6 Selection of Life Cycle Model for Development

In order to accomplish the stated objective, Agile model is selected as software development life cycle model. The incentive behind selecting Agile model is because requirements are stated clearly but the machine learning model needs to be trained and tested for better prediction on unseen data which makes the implementation and testing phase repetitive and to achieve the same Agile model for software development is taken in consideration.

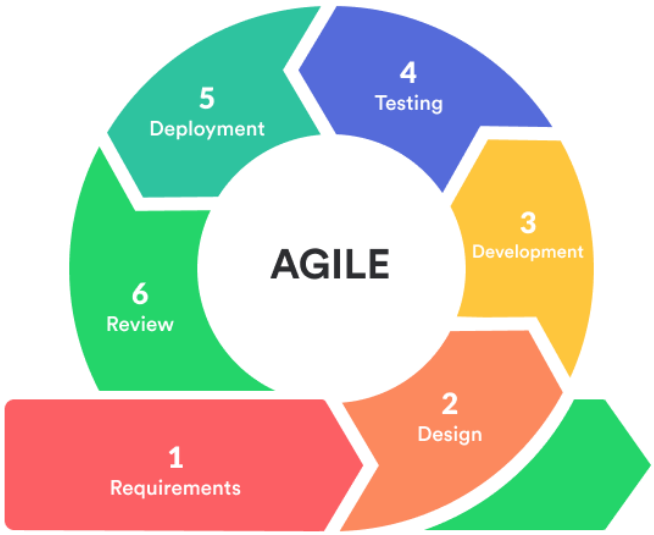

Figure 1.1: Software Development Life Cycle Model

1.7 Organization of Report

Chapter 1: Chapter 1, Problem definition of the proposed machine learning model is given along with background history, motivation for developing this system. Scope and objective of projected system is discussed.
Chapter 2: Chapter 2, will talk about literature referred for getting this idea of project design. Actual working and functionalities of the proposed machine learning model is described here. Section 2.1 will be brief about economical, operational and technical feasibility of the given system. Risk analysis is done in section 2.2 along with planning and effort allocation for development of proposed system in section 2.3 and 2.4.
Chapter 3: Chapter 3, will discuss about requirements of machine learning model. Hardware as well as software requirements are described in section 3.1 and section 3.2. Section 3.3 will give functional and non functional requirements of the the model. And software requirements specification is stated in section 3.4.
Chapter 4: Chapter 4 will give pictorial representation of the system. Section 4.1 will show architecture of the machine learning model using simple block diagram. Section 4.2 will give data flow diagrams (DFDs). UML diagrams of the proposed system are given in section 4.3.
Chapter 5: In chapter 5, conclusion about whatever work had done is discussed in section 5.1. Future work are discussed in section 5.2. At the last, project report embody bibliography too.

1.8 Summary

The given chapter thoroughly describes background and motivation behind selecting this machine learning project. The problem definition, scope and objective are also stated in the given chapter. The organization of report gave the overall structure about the chapters described throughout in this report upfront. The next chapter will discuss about system analysis.

Chapter 2

Project Planning and Management

Requirements analysis is a software engineering task that has bridged the gap between system level requirements engineering and software design requirement engineering activities resulting in the specification of software's operational characteristics (functional, data, and behavioral), that indicates software's interface with other system elements and establishes constraints that software must meet. Gathering of requirements related to the proposed system is done. Analysis is done by comparing these requirements with working and function provided by 'existing forum system'.
The proposed chapter will discuss about feasibility study of the proposed system in section 2.1. The section 2.2 discuss about risk analysis. The project scheduling of system is described in section 2.3. The section 2.4 will discuss about effort allocation. Cost estimation is discussed in section 2.5. Finally the summary in last section.

2.1 Feasibility Study

While considering the feasibility of any system, the system must meet its best performance by three sides i.e. technical, operational and economical feasibility. Feasibility should be considered in any organization since it helps in selection of best system for the job. It is necessary to evaluate the feasibility of the project at the early stages of software development. **If project risk is great, the feasibility of product is reduced.** A well designed feasibility study should provide a description of the product or service, details of operation and working of system. A feasibility study evaluates the project potential for success; The types of feasibility study are:

2.1.1 Technical Feasibility

Minimum hardware requirements: - 1.66 GHz Pentium Processor or Intel compatible processor. 1 GB RAM. Internet Connectivity. 80 MB hard disk space.

2.1.2 Operational Feasibility

The Application will reduce the time consumed to maintain manual records and is not tiresome and cumbersome to maintain the records. Hence operational feasibility is assured.

2.1.3 Economical Feasibility

Once the hardware and software requirements get fulfilled, there is no need for the user of our system to spend for any additional overhead. For the user, the machine learning model will be economically feasible in the following aspects: The model will reduce a lot of labour work. Hence the Efforts will be reduced. Our machine learning model will reduce the time that is wasted in manual processes.

2.2 Risk Analysis

Risk analysis and management are actions that help a software team to understand and manage uncertainty. A risk is a potential problem it might happen, or it may not. Risk is a uncertain event or condition that, if it occurs, has a positive or negative effect on a project's objectives. Risk always involves two characteristics: **uncertainty**- the risk may or may not happen; that is, there are no 100 percent probable risk and **loss**- if the risk becomes a reality ,unwanted consequences or losses will occur. When risks are analyzed, it is important to quantify the level of uncertainty and the degree of loss associated with each risk. The different types of risks are-

- **Project Risk** threaten the project plan. That is, if project risks become real, it is likely that the project schedule will slip and that cost will increase. Project risk identify potential budgetary, schedule, personnel, resource, stakeholder, and requirements problems and their impact on a software etc.
- **Technical Risk** threaten the project plan. That is, if project risks become real, it is likely that the project schedule will slip and that cost will increase. Project risk identify potential budgetary, schedule, personnel, resource, stakeholder, and requirements problems and their impact on a software etc.
- **Business Risk** threaten the viability of software to be built and often jeopardize the project or the product. Candidates for the top five business risks are:
    1. Building an excellent product or system that no one really wants**(market risk)**.
    2. Building a product that no longer fits into overall business strategy for the company**(strategic risk)**.

3. Building a product that the sales force doesn't understand how to sell**(sale risk)**.
4. Losing the support of senior management due to a change in focus or a change in people**(management risk)**.
5. Losing budget or personnel commitment**(budget risks)**.

Risk	Category	Probability	Impact
Technology will not meet requirements	TE	30%	1
End-user resist the system	BU	40%	3

Table 2.1: Risk Analysis

2.3 Project Scheduling

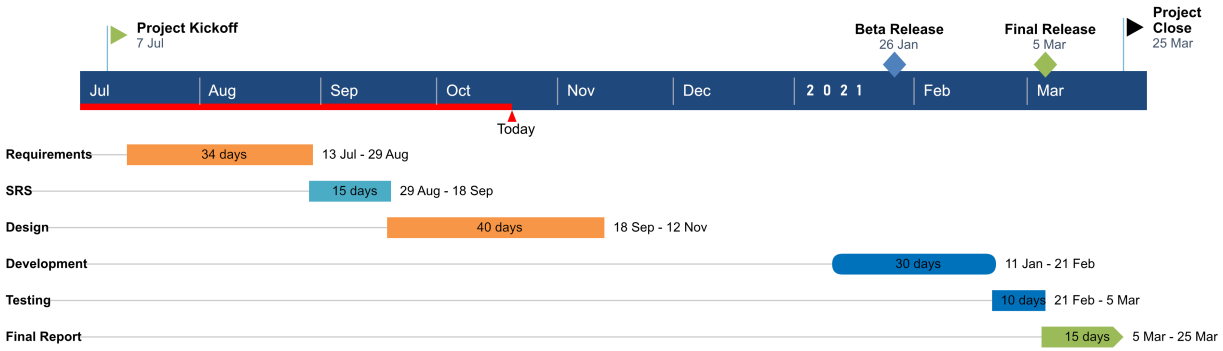

Figure 2.1: Gantt Chart

2.4 Effort Allocation

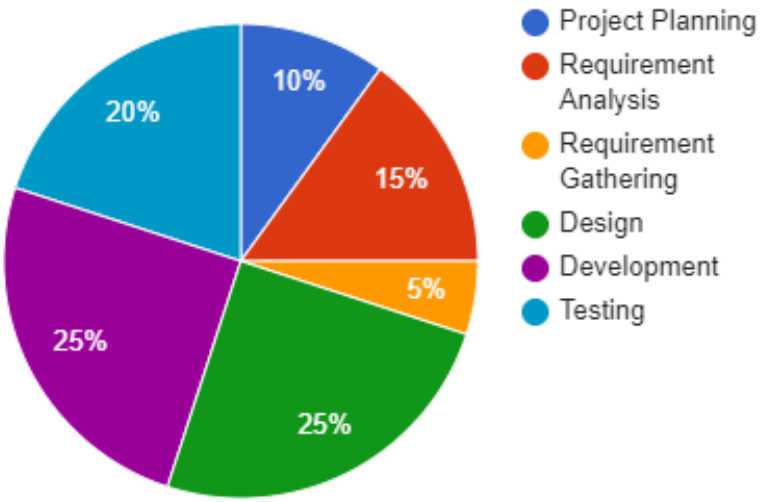

Figure 2.2: Effort Allocation

2.5 Cost Estimation

In order to estimate the cost of the project the COCOMO (Constructive Cost Model) model is used. It is estimated that the project will be have a value of 10 KLOC (Kilo Line of Code). The estimation of cost is explained in the subsection below.

2.5.1 COCOMO Model

The type of the project that is been estimated here is a organic.Where the requires team is adequately small. It is also estimated to have 10 KLOC.
Cost Estimation of the same is done below:

Software Project	a	b	c	d
Organic	2.4	1.05	2.5	0.38

Table 2.2: The constant values a,b,c and d for the organic model.

The estimated KLOC is 10.

$Effort = a(KLOC)^b$

$Effort = 2.4 * (10)^{1.05}$

$Effort = 26.92 Person - Month.$

$Time = c(Effort)^d$

$Time = 2.5 * (26.92)^{0.38}$

$Time \approx 8 Months.$

$PeopleRequired = Effort/Time$

$PeopleRequired = 26.92/8.73$

$PeopleRequired \approx 3.$

2.6 Summary

The given chapter thoroughly describes feasibility study, risk analysis, project scheduling, effort allocation and cost estimation. The next chapter will discuss about analysis.

Chapter 3

Analysis

The chapter contains system requirements of the proposed system. The requirement collection and identification of propose system are described in section 3.1. The section 3.2 will discuss about hardware and software requirements of system. In Section 3.3 functional and non functional requirements of proposed system will be discussed. The software requirement's specification of system will be discussed in section 3.4. Finally the summary in last section.

3.1 Requirement Collection and Identification

Requirement Collection: For this project the requirement collection is done using the project management tools and techniques for requirement gathering. For any project success, project result must meet stakeholder's needs and expectations. Capturing all project management requirements keeps you one step closer to project success. For the success of the project the requirement gathering is done through the following methods of requirement gathering.

- **Expert Judgement**: Experts are the people more knowledgeable in their respective areas. Their knowledge and experience help to gather most of the specific product or project requirements. So in order to replicate the proficiency of their work consulting with the project guide helped a lot.
- **Data Gathering**: Data gathering is an important technique for facilitation and or group creativity. Here, a group of people involves figuring out all project requirements. Ideas evolve through group creativity and help to determine requirements.
    1. Questionnaires and Surveys: In order to address the requirements of the project a set of questionnaires was prepared and answered which gave the requirement specification.

2. Brain Storming: A load amount of brain storming sessions were done with the project team and a session with the project guide too.
3. Bench-marking: In this technique, the comparison was made between existing practices and best practices. This way most of the organizations try to explore best in class practices. And hence set their requirements to improve their current position.

- **Data Analysis**: This technique is also known as Document analysis. As the name suggests, we analyzed existing documents to elicit project requirements.
- **Interpersonal and team skills**:

1. Observation: Also known as a job – shadowing. Here, a potential user or group of users is observed for identifying requirements. E-commerce sites use this technique to identify the patterns of customer's needs.
2. Nominal Group Technique: This technique is generally used to prioritize the requirements. When the requirements were clearly states they were prioritize according to their functionality.

Requirement Identification: Once, the requirements are collected then it is mandatory to identify them on three basis: Data, Functional and Behavioural requirements.

- **Data Requirement**: As far as the machine learning model is concerned the data requirement for the same is very minimal. In order to build a machine learning model you need a enough large dataset of the required genre and that is all about it.
- **Functional Requirement**: The functional requirement of the machine learning model is to predict the correct house prices once the model is trained and tested without error and with the most precision.
- **Behavioural Requirement**: The behaviour of the machine learning model is intended to be precise and accurate.

3.2 Software Requirement Specification

3.2.1 Product Features

- The interface of the application should be user friendly.

- Accurately predict the house prices, which failing to predict can lead to the project failure.
- The project should be easy-to-use for the end user.

3.2.2 Operating Environment

- **Operating System :** Windows/Linux/MacOS
- **Software :** Internet Browser
- **Hardware :** PC or Laptop, Server

3.2.3 Assumption

-

3.2.4 Functional Requirements

The functional requirements can also be stated as the Functional Requirement Specification which are the intended expectations from the software. The expected functional requirements from the proposed machine learning model are:

1. As machine learning is not intended to be understood by the naive user hence, the interface of the application should be user friendly.
2. The most important functional requirement of this project is to accurately predict the house prices, which failing to predict can lead to the project failure.
3. The project should be technically feasible to the end user.
4. As security is the most important concern to any software project, it is a functional requirement to this project too.
5. Machine learning incorporate several complicated calculations which makes the response time slower. To maintain a proper response time will be another functional requirement to the project.
6. Once the project will be built it should be maintainable further.

3.2.5 Non-Functional Requirements

Non-Functional requirements define desired qualities of the system to be developed and often influence the system architecture more than functional requirements do. A non-functional requirement corresponds to a set of restrictions imposed on the system to be developed, establishing, for instance, how attractive, useful, fast, or reliable it is. The type and description of each subcategory is explained in table:

Type	**Description**
Availability (A)	Describe how likely the system is accessible to a user at given point of time.
Look and Feel (LF)	Describes the style of the product's appearance.
Operability (O)	Degree to which a product or system has attributes to make it easy to operate and control.
Performance (PE)	Performance relative to the amount of resource used under stated condition.
Security (SE)	The degree to which a product or system protects information and data, so that persons or other products or system have the degree of data access appropriate to their types or levels of authorization.
Maintainability (MN)	Degree of effectiveness and efficiency with which a product or system can be modified by the intended maintainers.
Scalability (SC)	The degree to which product or system can effectively and efficiently be adapted for different or evolving hardware, software or other operational or usage environments.

Table 3.1: Descriptions of each type of Non-Functional Requirement (NFR) .

3.2.6 External Interfaces (User, Hardware, Software, Communication)

- **User**: Client.
- **Hardware Specification**:

Processor	Intel core i5
SSD	256 GB
RAM	8 GB

Table 3.2: Hardware specifications of the system.

- **Software Specification**:

Internet Browser	Google Chrome
Programming Language	Python 3

Table 3.3: Software specifications of the system.

3.3 Summary

In this chapter all the system requirements specification is described. System design will be discussed in next chapter.

Chapter 4

System Design

Design is a meaningful engineering representation of something that is to be built. It can be traced to system requirements and at the same time assessed for quality against a set of predefined criteria for "good" design. Software design sits at the technical kernel of software engineering and is applied regardless of software process model that is used. Software design is the first of three technical activities-design, code generation and test that are required to build and verify the software. The respective chapter will discuss about the proposed system design. Section 4.1 will discuss about the system architecture. Data Flow Diagram's are discussed in section 4.2. Section 4.3 contains the UML diagrams. Finally the summary in last section.

4.1 System Architecture

Figure 4.1 Architecture of Proposed System shows actual working flow of proposed system. The architecture is quite easy to understand for a new user. Authentication and house price prediction are the two main modules of the architecture. Authentication uses API's to authenticate the user and house price prediction uses the machine learning algorithm to predict prices.

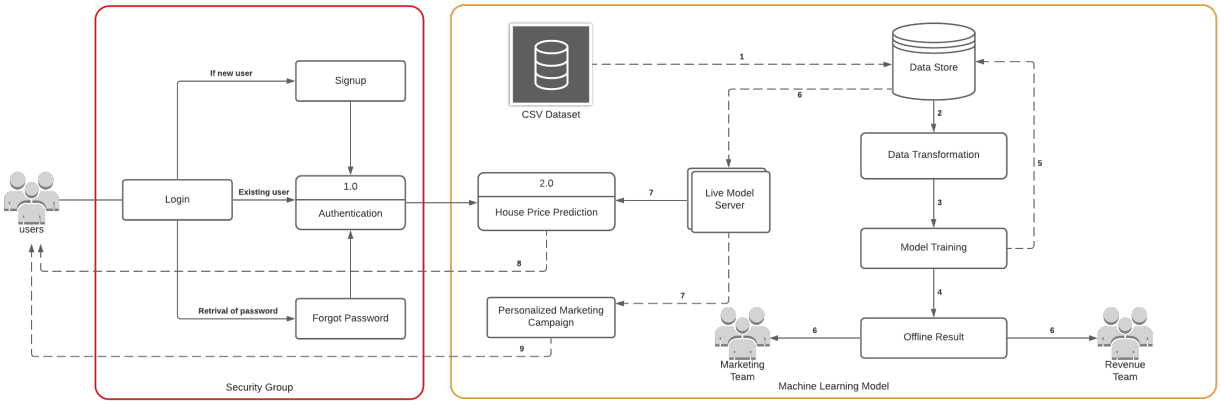

Figure 4.1: Architecture of House Price Prediction Model.

4.2 Data Flow Diagrams

DFD i.e. Data Flow Diagram will show actual flow of data in the system in different levels. Level 0 DFD present view of the data flow in it's simplest form. While Level 1 and Level 2 DFD shows magnified view of data flow in every part of the system.

4.2.1 Level 0 DFD

Figure 4.2 shows the level 0 DFD of house price predictor using machine learning, where a basic flow of data in system is illustrated.

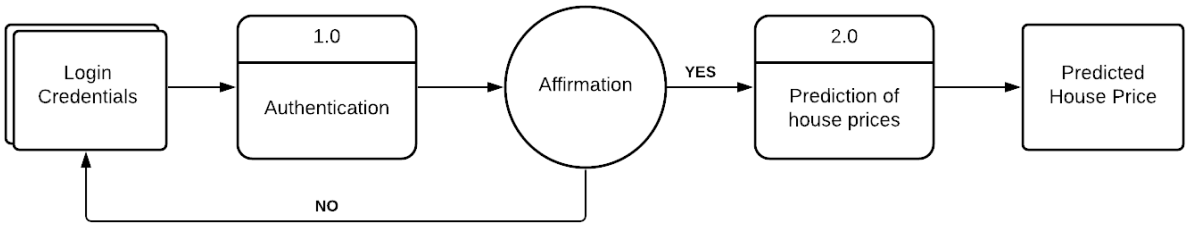

Figure 4.2: Level 0 DFD for house price prediction.

4.2.2 Level 1 DFD

Figure 4.3 presents the level 1 DFD of house price prediction machine learing model, where a slightly magnified view of the system is displayed.

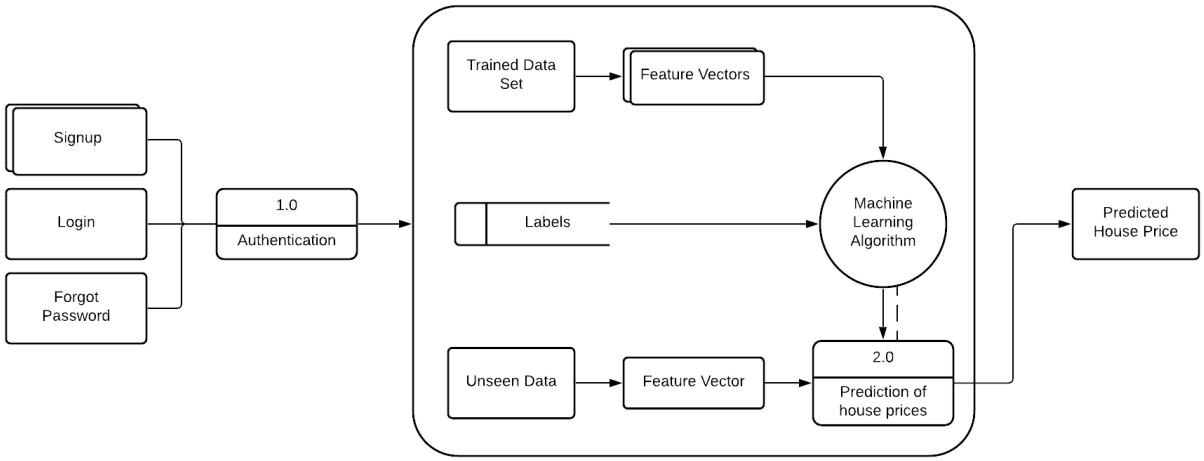

Figure 4.3: Level 1 DFD for house price prediction.

4.2.3 Level 2 DFD

Figure 4.4 presents the level 2 DFD which shows a detailed view of machine learning algorithm used in the system. Below enlisted are the steps involved in machine learning:

1. Gathering Data.
2. Preparing the data.
3. Choosing a Model.
4. Training.
5. Evaluation.
6. Hyper-parameter Tuning.
7. Prediction.

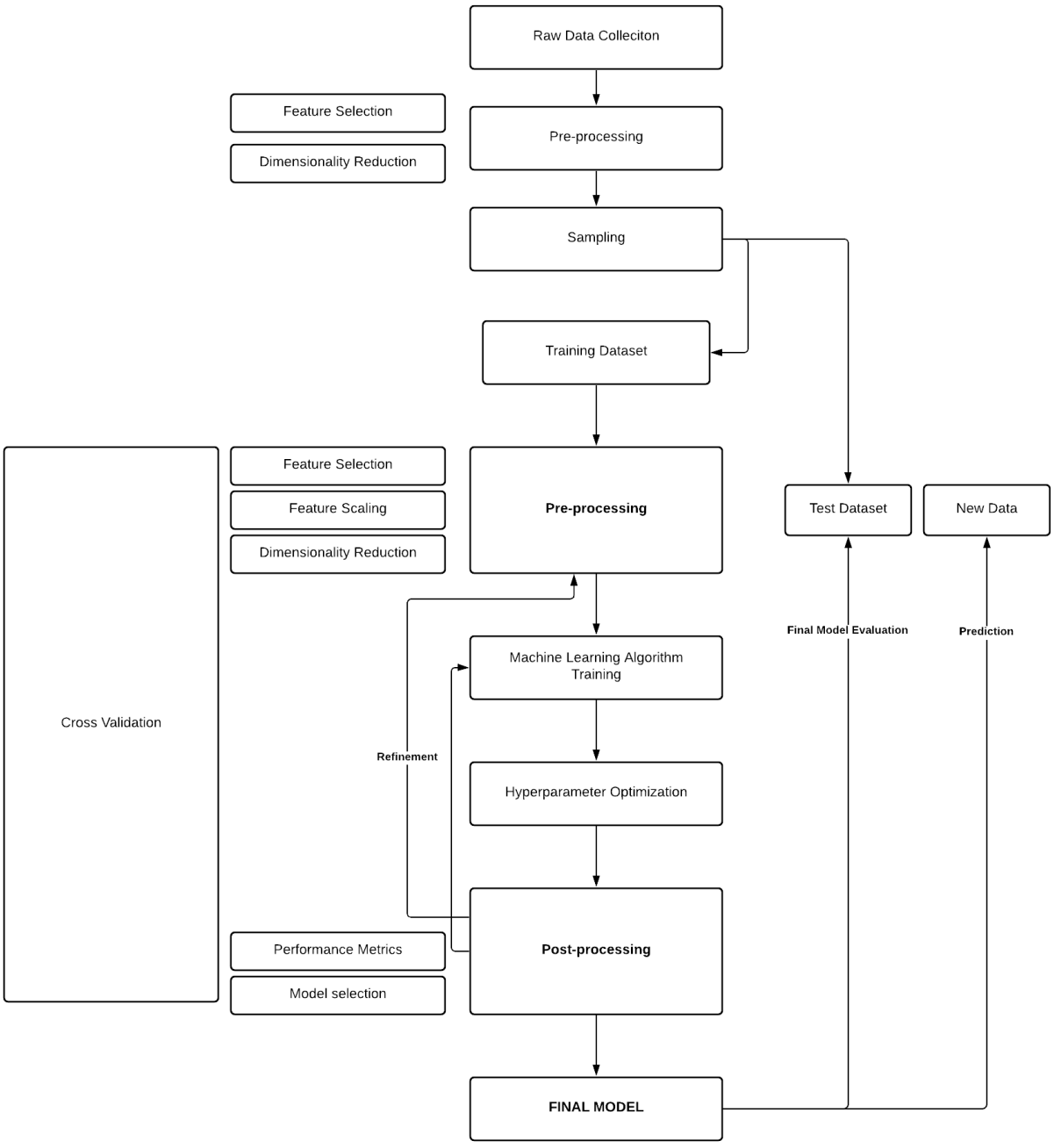

Figure 4.4: Level 2 DFD for house price prediction.

4.3 UML Diagrams

4.3.1 Use Case Diagram

A use case diagram is a dynamic or user's view in UML. Use case diagrams model the functionality of a system using actors and use cases. Below use case diagram exhibit's the same.

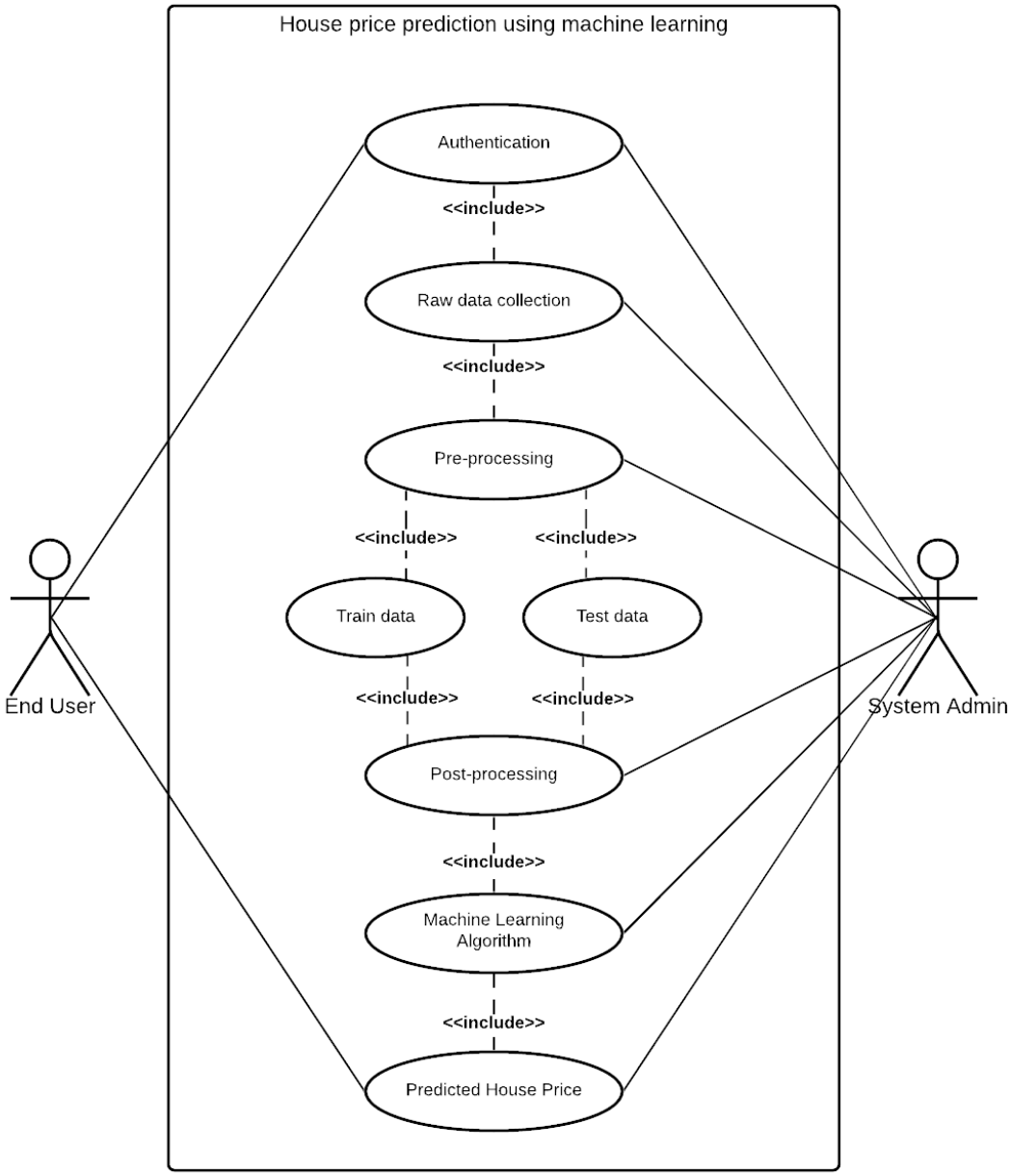

Figure 4.5: Use case diagram for house price prediction.

4.3.2 Sequence Diagram

A sequence diagram shows the object interaction arranged in a time sequence. It depicts the objects involved in the scenario and the sequence of message exchanged between these objects, which are needed to carry out the functionality of the system. In this subsection, two sequence diagrams are drawn with respect to the use cases present in the use case diagram:

- Figure 4.6 presents the sequential working of the system from users point of view.

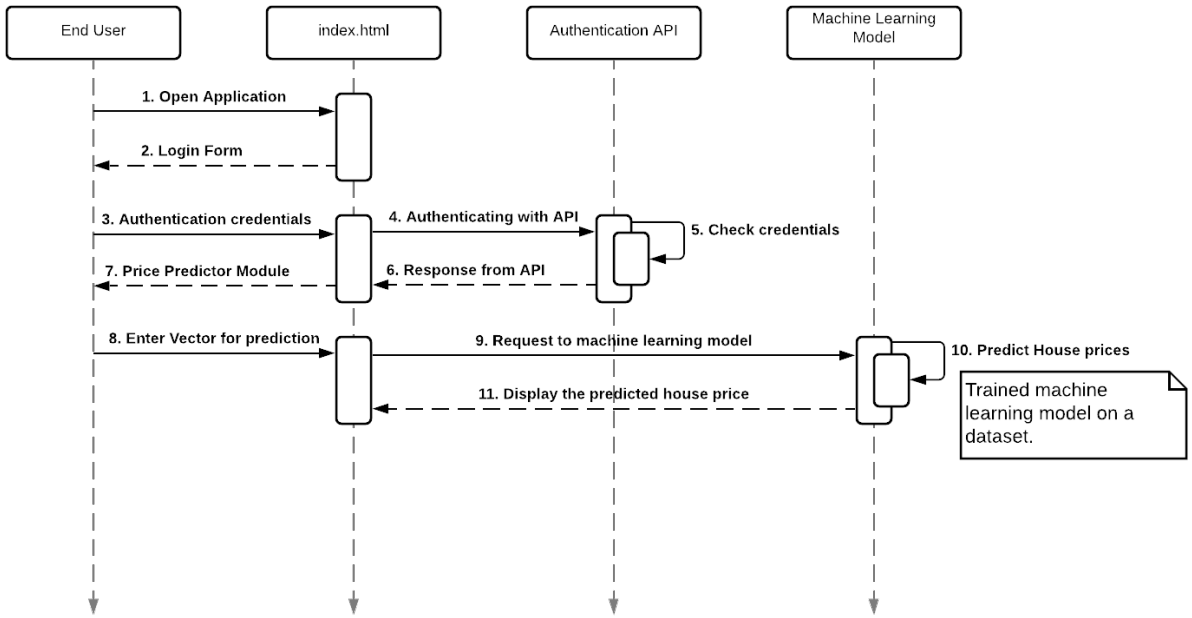

Figure 4.6: Sequence Diagram 1 for house price prediction.

- Whereas, Figure 4.7 presents the sequential working of the system from the system admin point of view.

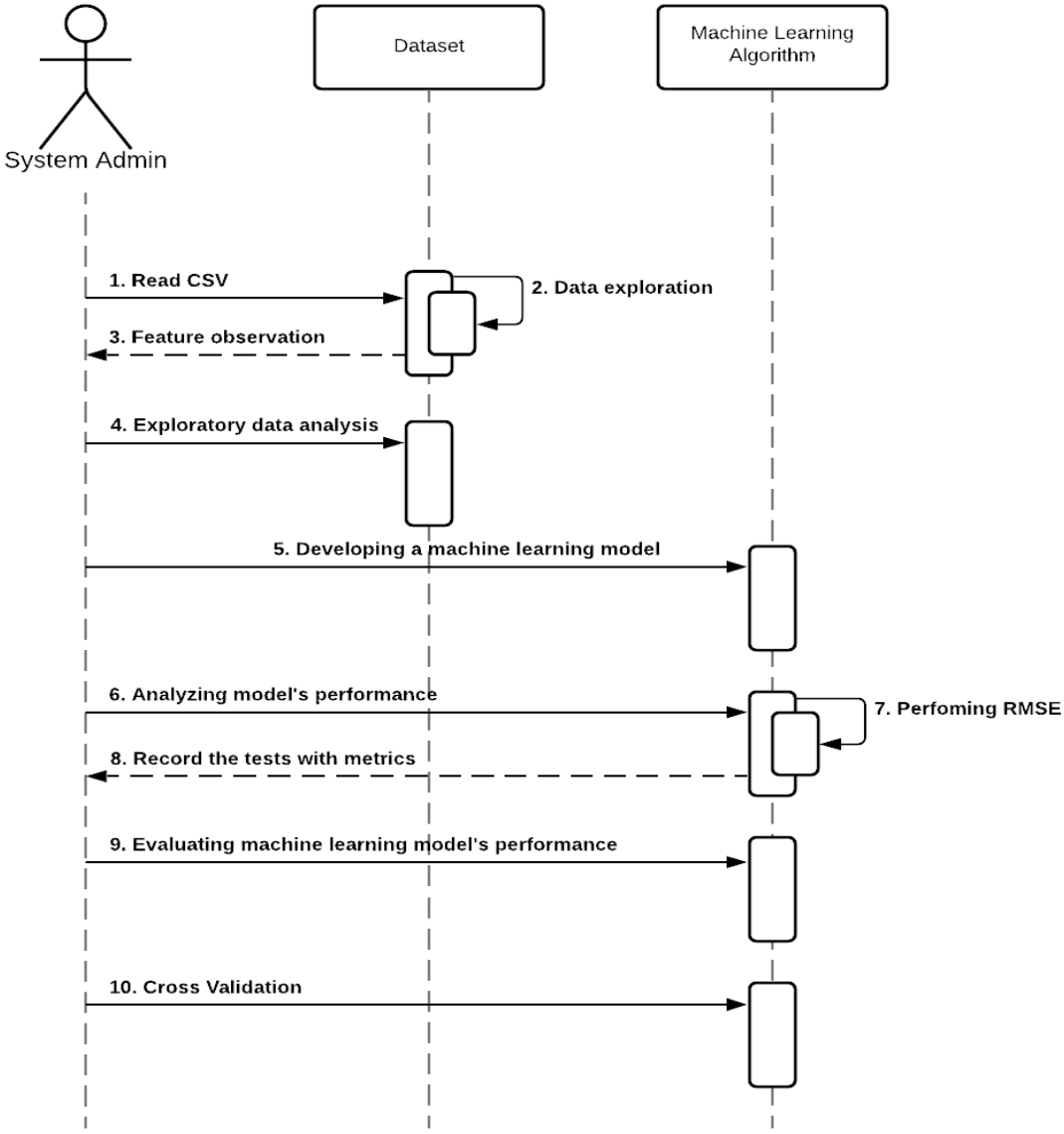

Figure 4.7: Sequence Diagram 2 for house price prediction.

4.3.3 Collaboration Diagram

A collaboration diagram, also known as a communication diagram, is an illustration of the relationships and interactions among software objects in the Unified Modeling Language (UML). These diagrams can be used to portray the dynamic behavior of a particular use case and define the role of each object. Fig 4.8 represents the collaboration between the end-user, index.html, authentication API and the machine learning algorithm. Fig 4.9 exhibits the collaboration of system admin, dataset and machine learning algorithm. These diagrams can be used to portray the dynamic behavior of a particular use case and define the role of each object.

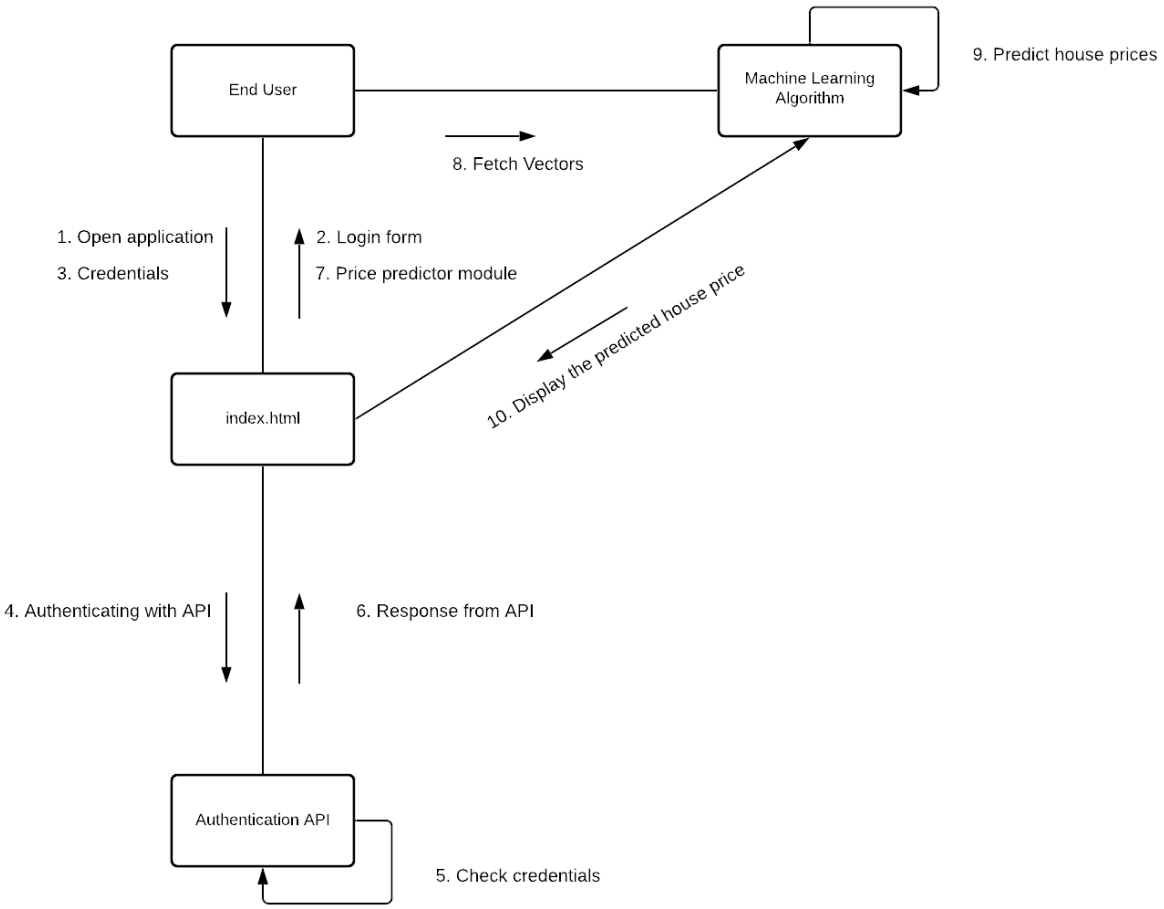

Figure 4.8: Collaboration Diagram 1 for house price prediction.

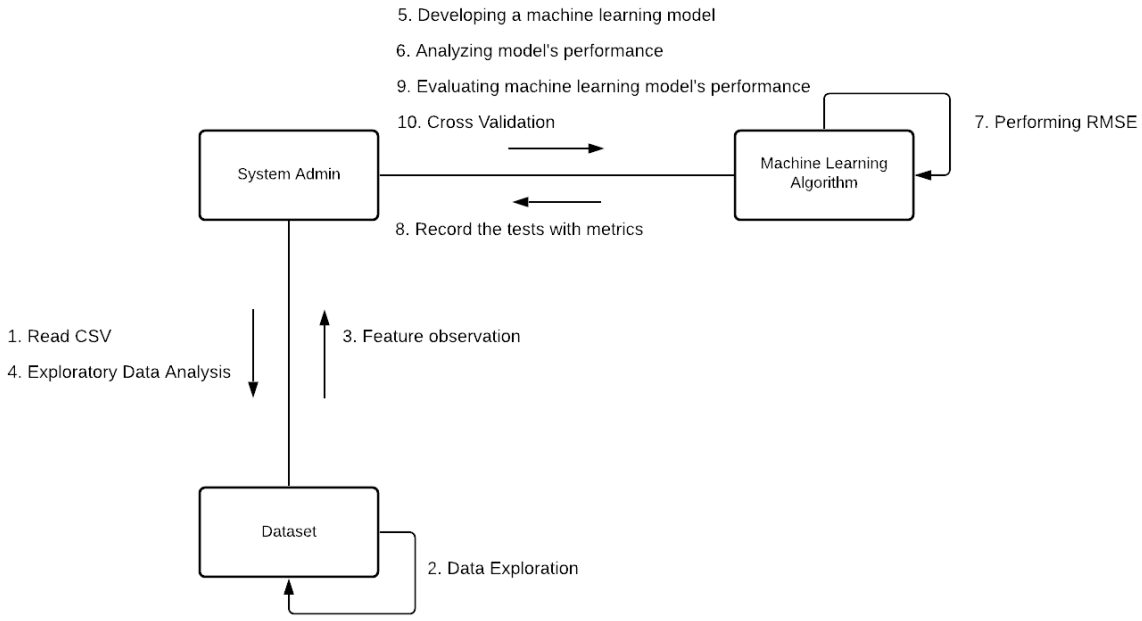

Figure 4.9: Collaboration Diagram 2 for house price prediction.

4.3.4 State Chart Diagram

State chart diagram can be defined as a machine which defines different states of an object and these states are controlled by external or internal events. Fig 4.10 depicts the state chart diagram of the house price predictor machine learning model.

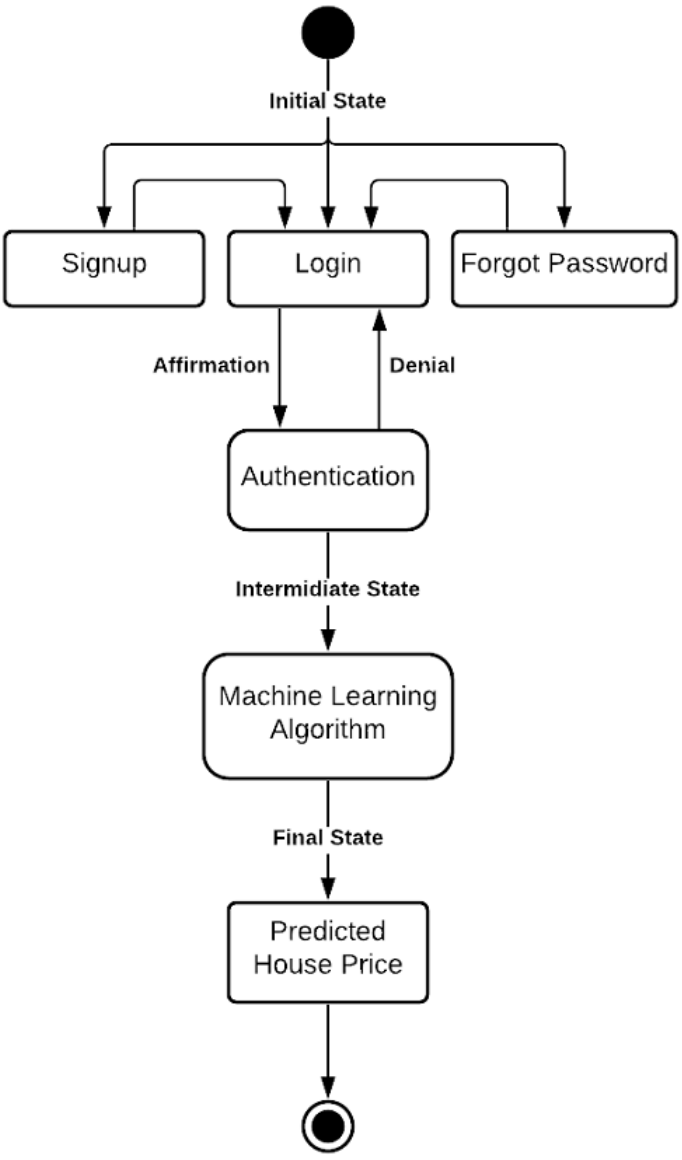

Figure 4.10: State Chart Diagram for house price prediction.

4.3.5 Activity Diagram

An activity diagram is a behavioral diagram i.e. it depicts the behavior of a system. An activity diagram portrays the control flow from a start point to a finish point showing the various decision paths that exist while the activity is being executed.

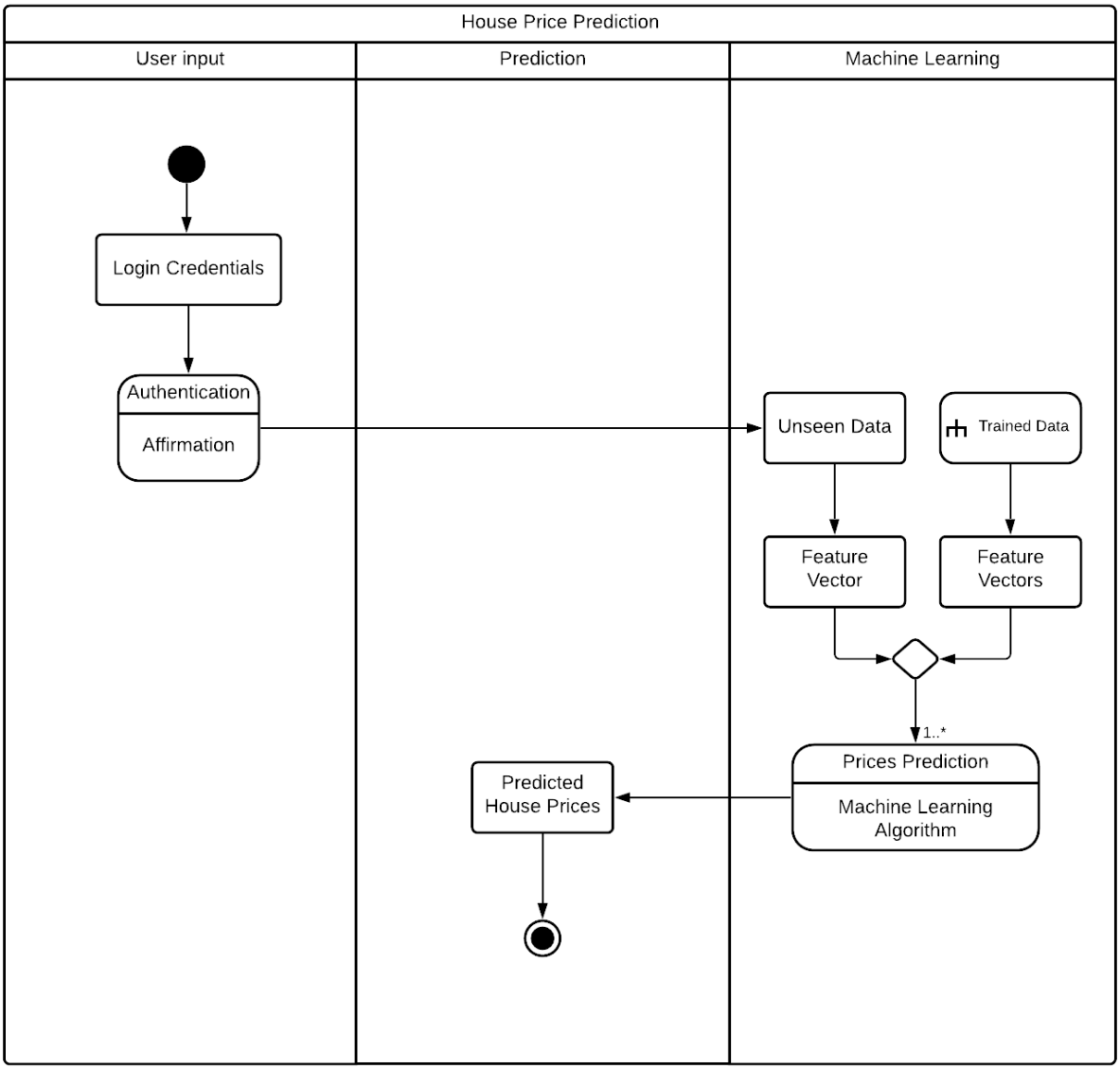

Figure 4.11: Activity Diagram for house price prediction.

4.3.6 Component Diagram

A component diagram, also known as a UML component diagram, describes the organization and wiring of the physical components in a system. Component diagrams are often drawn to help model implementation details and double-check that every aspect of the system's required functions is covered by planned development.

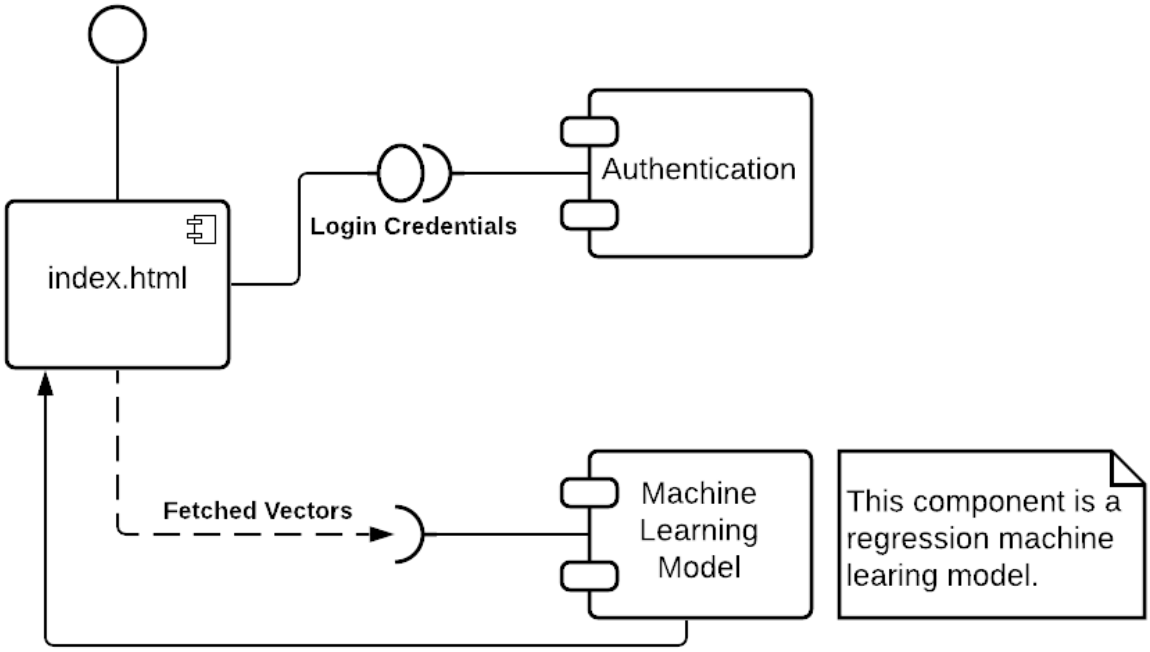

Figure 4.12: Component Diagram for house price prediction.

4.3.7 Deployment Diagram

A deployment diagram is a UML diagram type that shows the execution architecture of a system, including nodes such as hardware or software execution environments, and the middleware connecting them. Deployment diagrams are typically used to visualize the physical hardware and software of a system.
The purpose of deployment diagram shown below is:

- Visualize the hardware topology of a system.
- Describe the hardware components used to deploy software components.
- Describe the runtime processing nodes.

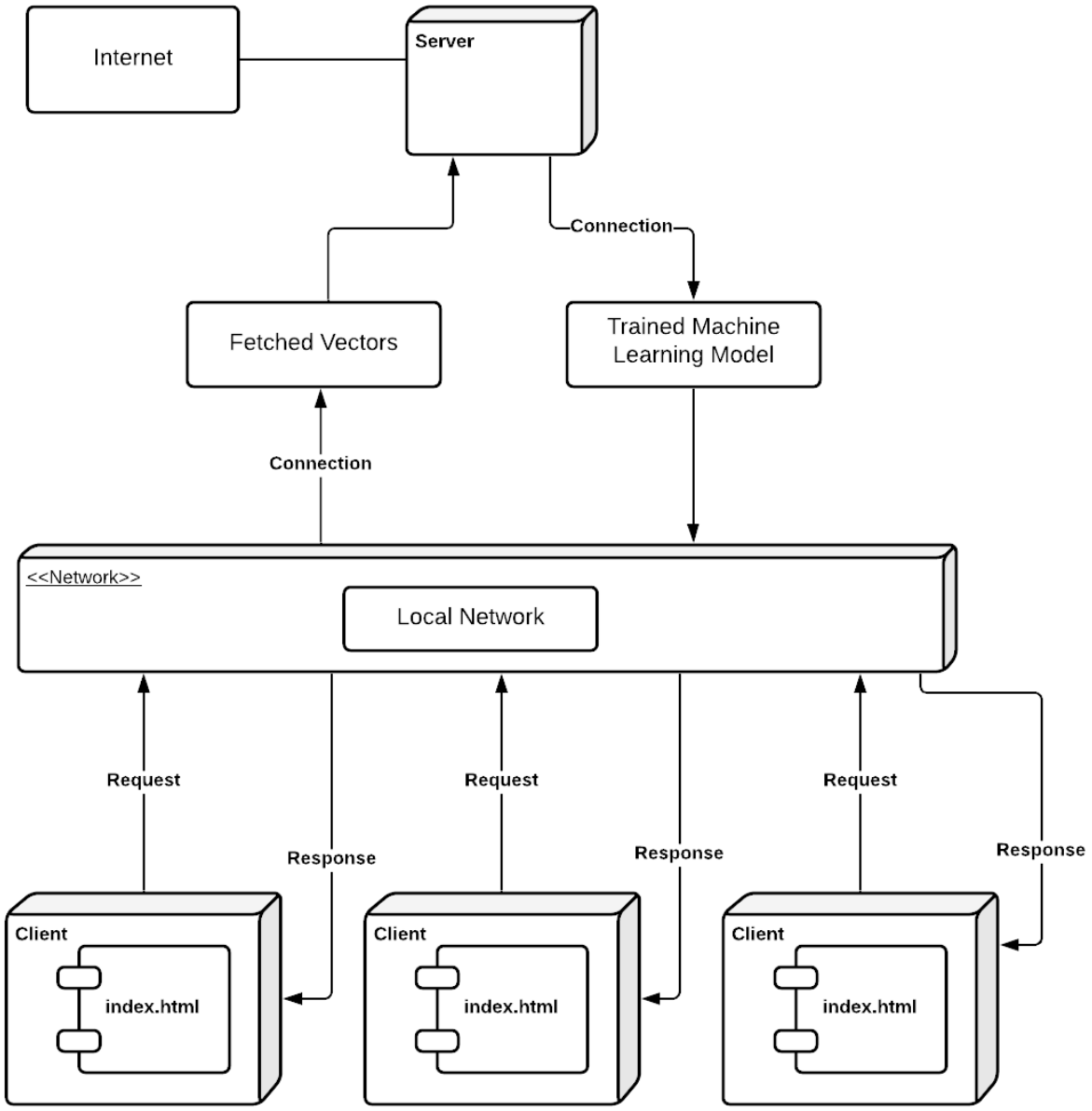

Figure 4.13: Deployment Diagram for house price prediction.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

5.2 Future Work

Behind the increasing interest on machine learning, there are issues on difficulty to use the methods by non-experts, since a typical data analytic project involves a series of complicated tasks. As to resolve the problem, an automated machine learning pipelines has been introduced, which can automate the processes of model selection, hyper-parameter tuning and feature engineering. Designing an effective machine learning model for prediction or classification problem is a tedious endeavor. Significant time and expertise are needed to customize the model for a specific problem. A significant way to reduce the complicated design is by using Automated Machine Learning (AML) that can intelligently optimize the best pipeline suitable for a problem or dataset.