**TRIBHUVAN UNIVERSITY**

**LALITPUR ENGINEERING COLLEGE**

**DEPARTMENT OF COMPUTER ENGINEERING**



PROJECT PROPOSAL ON

REAL ESTATE PRICE PREDICTION

PRABHAT KUMAR YADAV (074|BCT|017)

SANOJ RAJ SHRESTHA (074|BCT|023)

SHIVANI KARNA (074|BCT|024)

SHREEDHAR NEUPANE (074|BCT|025)

JUNE, 30

# ACKNOWLEDGEMENT

The success of this project required a lot of guidance and assistance from many people and we are extremely fortunate to have got this all along the completion of our proposal. Whatever we have done is only due to such guidance and assistance and we would not forget to thank them.

Firstly, we would like to thank Institute of Engineering for including the final year project as a part of our curriculum. Special thanks goes to Lalitpur Engineering College, Department of Computer Engineering for initiating and facilitating this data science project to further enhance our knowledge of machine learning and research oriented work.

We respect and thank our Supervisor Er Ishwor Dhungana for providing us all support and guidance during the working phase.

We are thankful to and fortunate enough to get constant encouragement, support and guidance from all teaching staffs namely Er Praches Acharya , Er Binod Sapkota who helped us in successfully completing our proposal work.

Acknowledgement would be incomplete without mentioning our family members and friends who have been constant source of inspiration during the preparation of the proposal.

# ABSTRACT

A lot of time and research is necessary before buying a property like a permanent home. It is a huge investment and people want the best product for the best price. This project aims to provide an objective price of a house provided parameters like the area of the land, number of rooms, room size, address of the house, etc. The prediction model will be developed using machine learning using linear regression algorithm and trained on the data of thousands of houses on sale in Nepal using the previously mentioned parameters and it will provide an objective value of a house with the given parameters based on the trained data.

It can also be used by real-estate buyers to compare the price of the house to the one given by the model. The model will also recommend similar houses with similar price ranges helping the investment decision further. It will be able to provide some form of assurance to the buyer about the price of the house s/he is about to purchase by providing a predicted price derived from analysis of market data of similar houses.

**Keywords: price-prediction model, linear regression algorithm**

# LIST OF FIGURES

[Figure 1: Use Case Diagram 13](#_Toc80983758)

[Figure 2: Activity Diagram 14](#_Toc80983759)

[Figure 3 : System-Flow Diagram 15](#_Toc80983760)

[Figure 4: Sequence Diagram 16](#_Toc80983761)

[Figure 5: Iterative Model Diagram 17](#_Toc80983762)

[Figure 6:Data Exploring Diagram 19](#_Toc80983763)

[Figure 7: Stages of Data Modeling 21](#_Toc80983764)

[Figure 8: linear Regression 23](#_Toc80983765)

[Figure 9: Gantt-Chart 25](#_Toc80983766)

[Figure 10: Budget Diagram 25](#_Toc80983767)

**TABLE OF CONTENTS**

[ACKNOWLEDGEMENT 1](#_Toc81297180)

[ABSTRACT 2](#_Toc81297181)

[LIST OF FIGURES 3](#_Toc81297182)

[CHAPTER 1: INTRODUCTION 5](#_Toc81297183)

[**1.1 Background** 5](#_Toc81297184)

[**1.2MOTIVATION** 6](#_Toc81297185)

[**1.3 Problem Statement & Related Work** 7](#_Toc81297186)

[**1.4 Goals and objectives** 8](#_Toc81297187)

[**1.5 Scope and application** 9](#_Toc81297188)

[CHAPTER 2: LITERATURE REVIEW 10](#_Toc81297189)

[Chapter 3: REQUIREMENT ANALYSIS 12](#_Toc81297190)

[**3.1Project requirements** 12](#_Toc81297191)

[**3.2 Feasibility Report** 14](#_Toc81297192)

[CHAPTER 4: SYSTEM DESIGN 15](#_Toc81297193)

[**4.1 Use Case Diagram** 15](#_Toc81297194)

[**4.2 Activity Diagram** 17](#_Toc81297195)

[**4.3 System Block Diagram** 18](#_Toc81297196)

[**4.4 Sequence diagram** 19](#_Toc81297197)

[CHAPTER 5: METHODOLOGY AND ANALYSIS 20](#_Toc81297198)

[**5.1 Software Development Approach** 20](#_Toc81297199)

[**5.2 Development Roadmap** 21](#_Toc81297200)

[CHAPTER 6: GANTT-CHART 28](#_Toc81297201)

[CHAPTER 7: EXPECTED OUTCOME 30](#_Toc81297202)

[REFERENCES 31](#_Toc81297203)

# CHAPTER 1: INTRODUCTION

## **1.1 Background**

In the contemporary digital era, useful information of the society can be retrieved from a wide variety of sources and stored in the form of structured, unstructured and semi-structured formats. In the analysis of economic phenomena or social observations, advancement of innovative technology makes it possible to systematically extract the relevant information, transform them into complex data formats and structures, and then perform suitable analyses. Because of these new circumstances, traditional data processing and analytical tools may not be able to capture, process and analyze highly complex information in the social and economic worlds. New techniques have been developed in response to the treatment of the colossal amount of available data.

Previous work on predicting house prices has been based on regression analysis and machine learning techniques. Local linear models and random forest models, fuzzy reasoning, Backpropagation neural networks, and Elman neural networks can be used to forecast real estate prices. Out of these, it is found that the Elman neural network can forecast more accurately and constrict faster than other approaches. Nguyen and Cripps compared the predictive performance of artificial neural networks (ANNs) and multiple regression analysis for single-family housing sales. They found that when enough data points were available for training, ANNs could perform better than multiple linear regressions. Additionally, a latent manifold model with two trainable components can also be used to evaluate house prices where a parametric component is used for predicting the “intrinsic” price of a house, and a nonparametric component can calculate the desirability of the neighborhood.

## **1.2MOTIVATION**

The housing market of Nepal has been a very interesting topic for us to analyze since the industry is booming after the devastating effects of the earthquake. In order to get a better understanding of the most influencing factors of the price of a residential house in Nepal, we are using real data from various real estate websites in Nepal to analyze the Nepalese housing market. After the analysis we will build a residential house price prediction model which could be used by buyers to compare the predicted price (derived from data of the Nepalese housing market) with the actual price to make better investment decisions. We are motivated by the value it brings to real estate buyers in Nepal and the learning opportunity it provides us to be aspiring data scientists.

## **1.3 Problem Statement & Related Work**

In Nepal, there are multiple real estate classified websites where properties are listed for sell/buy purposes such as 99aana, HamroGhar, Gharjagga, and more. However, in each of these websites, we can see a lot of inconsistencies in terms of pricing of a house and there are some cases when similar houses are priced differently. Sometimes the consumers may feel the pricing is not justified for a particular listed house but there is 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 Nepal the transaction prices are quite high, and addressing this issue will help the customers and the real estate industry in the long run. We propose to use machine learning to develop an algorithm that can predict housing prices based on certain features of the house and provide a price as an output by training the model trained on thousands of house data of Nepal.

The business application of this algorithm is that our websites can directly use this algorithm to predict prices of properties that are going to be listed by taking some input variables and predicting the price according to the available market data.

## **1.4 Goals and objectives**

* Develop a model where users can get the predicted price of a house after entering suitable parameters.
* Develop an analysis model of housing market in Nepal.

## **1.5 Scope and application**

Future Scope can be indicated by:

* The models accuracy will be improved by using better data and innovative algorithm which is in trend.
* More parameters will be added so users and clients can get more options to choose from.
* The model will be deployed, in cloud technology like Azure and AWS so continuous integration and development is possible.

# CHAPTER 2: LITERATURE REVIEW

People are careful when they are trying to buy a new house with their budgets and market strategies. The objective of the project is to forecast the coherent house prices for non-house holders based on their financial provisions and their aspirations. By analyzing the foregoing merchandise, fare ranges, and also forewarned developments, speculated prices will be estimated. It is not without reason, considering that housing is one of the core sectors of the economy, and possessing a house, is considered as one of the most valued immovable asset that, apart from providing the core need of shelter and a definite identity in the society, is an asset that appreciates in value over time. The phenomenon of the falling or rising of the house prices has attracted interest from the researcher as well as many other interested parties.

There have been many previous research works that used various regression techniques to address the question of the changes in house prices. The work done by  *D. Banerjee* and *S. Dutta* [1]considers the issue of changing house prices as a classification problem and applies machine learning techniques to predict whether house prices will rise or fall. The work considers two discrete values 0 and 1 as respective classes. If the value of the class is 0 then we consider that the price of the house has decreased and if the value of the class is 1 then we consider that the price of the house has increased. It is essential to remember that the output from following research (*D. Banerjee* and *S. Dutta*) is only useful for certain group of people who are concerned about knowing fall or rise of real state property. Real estate is the least transparent industry in the ecosystem according to another research paper (A. Varma, A. Sharma, S. Doshi, and R. Nair) [2]. Housing prices keep changing day in and day out and sometimes are hyped rather than being based on valuation. Predicting housing prices with real factors is the main crux of the author’s research project. Evaluations were based on every basic parameter that is considered while determining the price. Authors used various regression techniques in this pathway, results were not sole determination of one technique rather it was the weighted mean of various techniques to give the most accurate results.

After reviewing all these papers it can be understood that there can be improvement in real estate price prediction system like people may be more interested in knowing prices rather than just knowing price fluctuation of house. In addition to that the

paper suggested by A. Varma, A. Sharma, S. Doshi, and R. Nair considers their surrounding houses with their surrounding local parameter units like size of house

expressed in sqft and acre, however the size units of houses of other countries obviously differs like in Nepal common units used to express size of a lot is dhur, anna, ropani etc. Scholars like C. R. Madhuri, G. Anuradha and M. V. Pujitha [3] have actually made a remarkable progress by preparing a regression model that can predict price for those who are in need of house but their project accuracy is highly doubtful because of use of redundant regression algorithm which is used to train the model. So due to lack of cutting edge technology, algorithm, local parameters and units being used in all these projects it is quintessential to establish a modern real estate price prediction project which overcomes all the following shortcomings.

# 

# Chapter 3: REQUIREMENT ANALYSIS

## **3.1Project requirements**

**3.1.1Software Requirement**

**Python:**

Python is an interpreted, object-oriented programming language similar to PERL, that has gained popularity because of its clear syntax and readability. Python has a significant number of users. A notable feature of Python is its indenting of source statements to make the code easier to read.

**NumPy:**

NumPy (short for Numerical Python) offers comprehensive mathematical functions, random number generators, linear algebra routines, Fourier transforms, and more. NumPy arrays are like Python’s built-in list type, but NumPy arrays provide much more efficient storage and data operations as the arrays grow to arise.

**Pandas:**

Pandas is a software library written for the Python programming language for data manipulation and analysis. Pandas allow importing data from various file formats such as comma-separated-values, JSON, SQL, and Microsoft Excel.

**Matplotlib:**

Matplotlib is a plotting library for the Python Programming language and its numerical and its numerical mathematics extension NumPy. It provides an object-

oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK.

**Seaborn:**

Seaborn is a library that uses Matplotlib underneath to plot graphs. It will be to visualize random distribution. Seaborn is created for enhanced data visualization.

**Scraping Tools:**

**Beautiful Soup:**

Beautiful Soup is a Python library for getting data out of HTML, XML, and other markup languages. Say you’ve found some web pages that display data relevant to your research, such as data or address information, but that do not provide any way of downloading the data directly. Beautiful Soup helps you pull particular content from a webpage, remove the HTML markup and save the information. It is a tool for web scraping that helps you clean up and parse the documents you have pulled down from the web.

**Selenium:**

Selenium is an open-source and portable automated software testing tool for testing web applications. It has capabilities to operate across different browsers and operating systems. Selenium is not just a single tool but a set of tools that helps testers to automate web-based applications more efficiently.

**Framework:**

**Flask:**

Flask is a micro web framework written in Python. It is classified as a micro framework 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

**3.1.2 Functional Requirements**

* Users should be able to enter the parameters like area of land, the number of rooms, location, etc., and get fairly accurate price predictions.
* Users must get only one price for a particular house/prediction**.**

**3.1.3Non-functional requirements**

The result of the predicted price should be at least 80% to 120% of the mentioned price of the house.

## 

## **3.2 Feasibility Report**

**3.2.1 Economic Feasibility**

The time and effort required to create the project may look a lot but is actually aceptable when compared to other projects. The cost to hire a data scientist is NPR.710 currently in Nepal which is reasonable but for this project it can be negotiated to NPR.400 if there is already available clean data. Hence the project is economically viable.

**3.2.2 Technical Feasibility**

Open source language, libraries and frameworks like python, flask, pandas will be used and learned, since this project does not require powerful GPU or TPU google Colab is not required, however technical expertise in Jupyter notebook and programming is very essential. Hence technical feasibility of the project seems to be fulfilled.

**3.2.3 Operational Feasibility**

Operational feasibility is mainly related to human organizational and political aspects. It refers to evaluation which analyzes how well a system operates. The project requires a laptop or android device to run the app.

# 

# CHAPTER 4: SYSTEM DESIGN

## **4.1 Use Case Diagram**

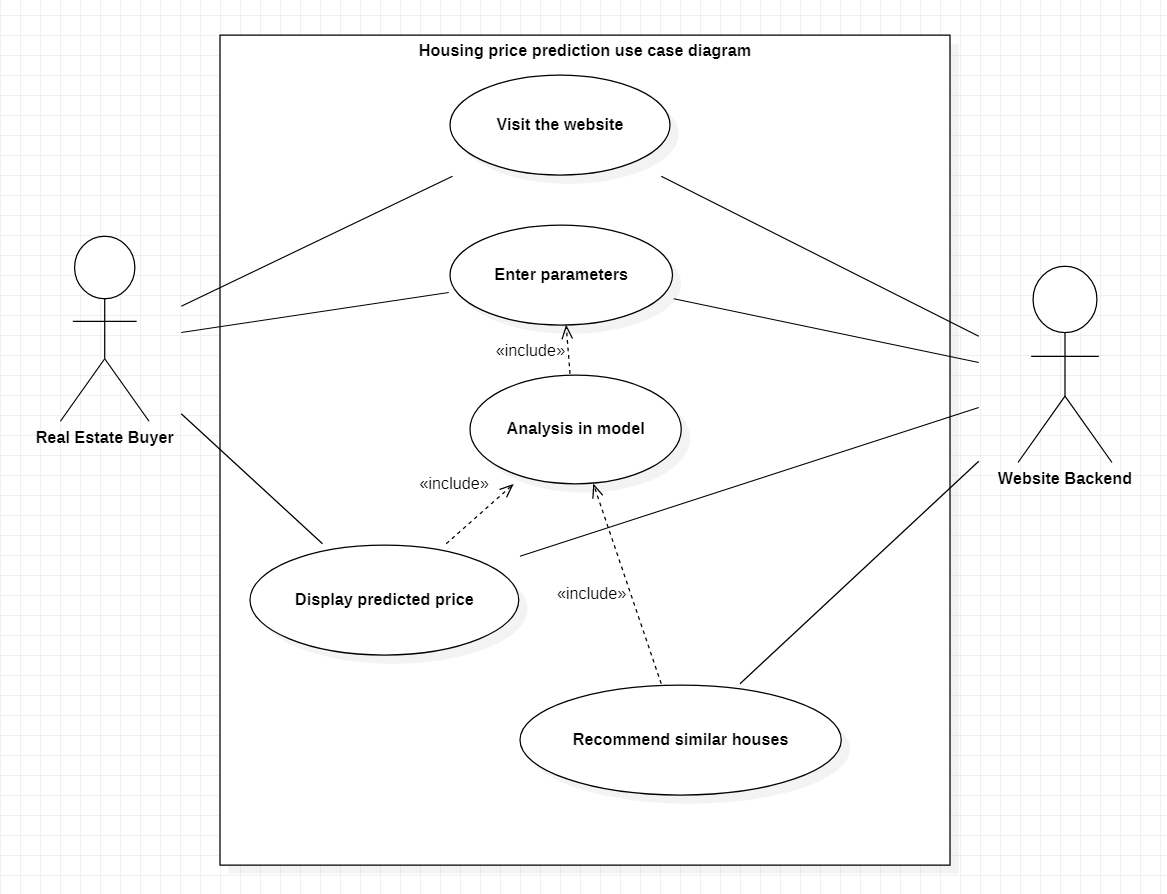
****

Figure 1: Use Case Diagram

## 

## **4.2 Activity Diagram**

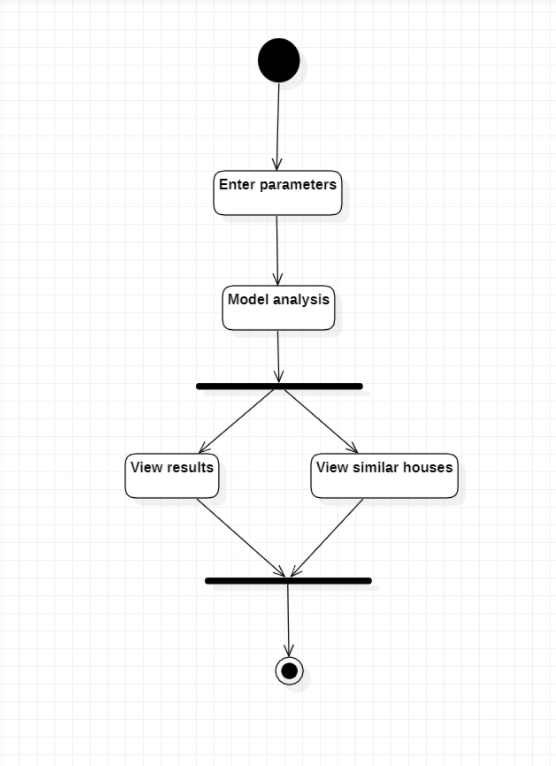
****

Figure 2: Activity Diagram

## 

## **4.3 System Block Diagram**

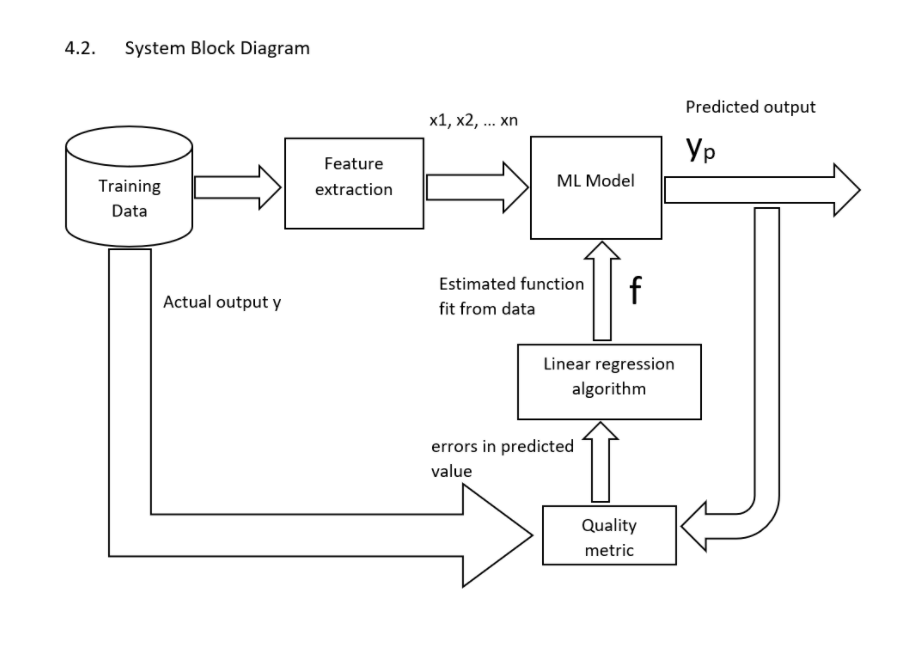
****

Figure 3 : System-Flow Diagram

## **4.4 Sequence diagram**

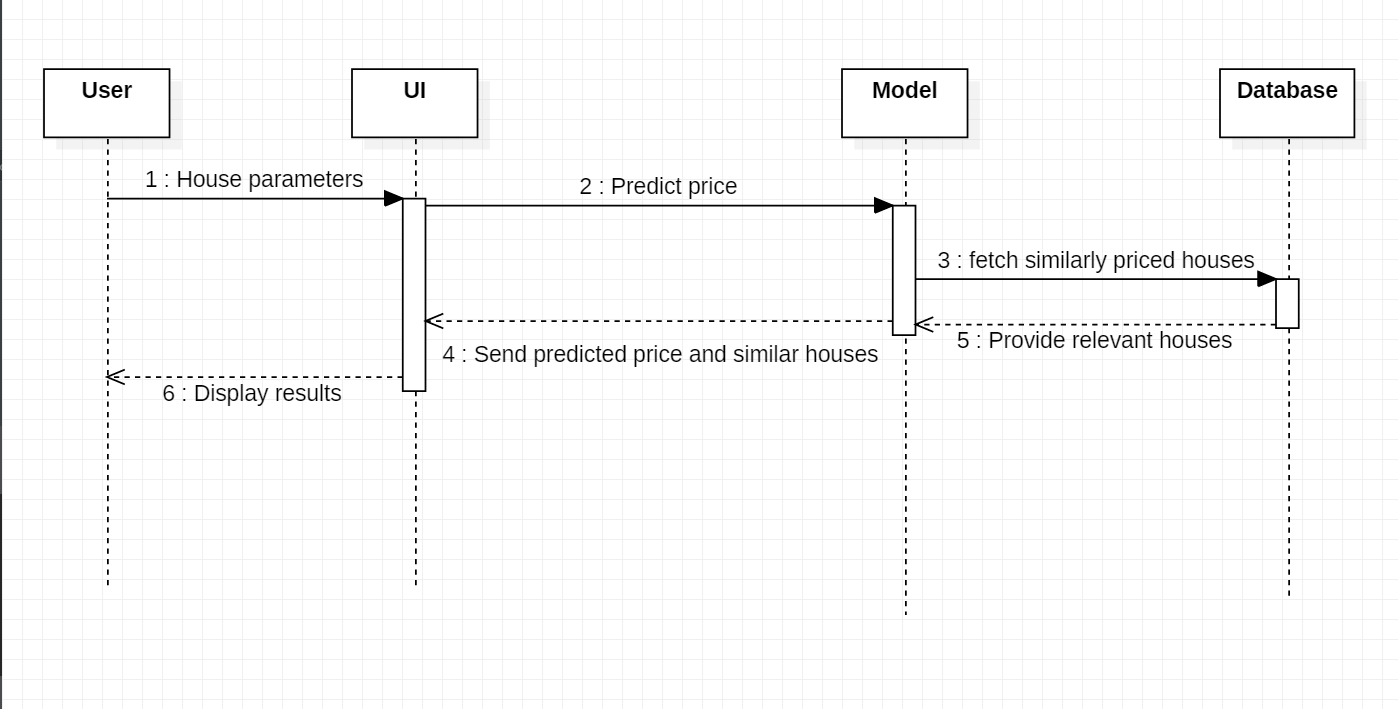


Figure 4: Sequence Diagram

# CHAPTER 5: METHODOLOGY AND ANALYSIS

## **5.1 Software Development Approach**

**5.1.1 Iterative Model**

For our project development we have chosen to follow the iterative Model. In this Model, you can start with some of the software specifications and develop the first version of the software. After the first version if there is a need to change the software, then a new version of the software is created with a new iteration. Every release of the Iterative Model finishes in an exact and fixed period that is called iteration.

The Iterative Model allows accessing earlier phases, in which the variations are made respectively. The final output of the project was renewed at the end of the Software Development Life Cycle (SDLC) process. The iterative model is aparticular implementation of a software development life cycle (SDLC) that focuses on an initial, simplified implementation**,** which then progressively gains more complexity and a broader feature set until the final system is complete. When discussing the iterative method, the concept of incremental development will also often be used liberally and interchangeably, which describes the incremental alterations made during the design and implementation of each new iteration. The basic idea behind this **me**thod is to develop a system through repeated cycles (iterative) and in smaller portions at a time (incremental), allowing software developers to take advantage of what was learned during the development of earlier parts or versions of the system.

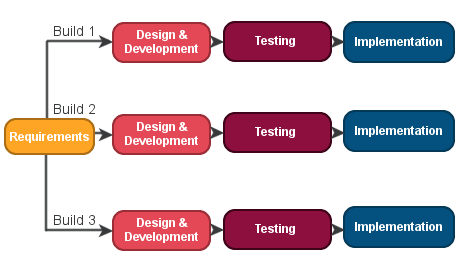


Figure 5: Iterative Model Diagram

## **5.2 Development Roadmap**

The very first step of a data science project is straightforward. We obtain the data that we need from available data sources. To obtain data we used the web scraping method with tools like beautiful soup and selenium. Beautiful Soup provides a few simple methods and Pythonic idioms for navigating, searching, and modifying a parse tree: a toolkit for dissecting a document and extracting what we need.

**Data cleaning**

After obtaining data, the next immediate thing to do is scrubbing data. This process is for us to “clean” and to filter the data. In this process, we convert the data from one format to another and consolidate everything into one standardized format across all data. For example, if data is stored in multiple CSV files, then we will consolidate these CSV data into a single repository so that we can process and analyze them. On top of that, scrubbing data also includes the task of extracting and replacing values. If there are missing data sets or they appear to be non-values, this is the time to replace them accordingly. Lastly, we will also need to split, merge and extract columns. This process focuses on organizing and tidying up the data, removing what is no longer needed, replacing what is missing, and standardizing the format across all the data collected.

We will achieve the process of data cleaning by following steps:

* Dropping irrelevant columns.
* Renaming column names to meaningful names.
* Making data values consistent.
* Imputing input values

We will use pandas and NumPy to clean data.

**Explore Data**

Different data types like numerical data, categorical data, ordinal and nominal data, etc. require different treatments. Data exploration is the initial step in data analysis, where users explore a large data set in an unstructured way to uncover initial patterns, characteristics, and points of interest. This process isn’t meant to reveal every bit of information a dataset holds, but rather to help create a broad picture of important trends and major points to study in greater detail. This process makes deeper analysis easier because it can help target future searches and begin the process of excluding irrelevant data points and search paths that may turn up no results. More importantly, it helps build a familiarity with the existing information that makes finding better answers much simpler.

Many times, data exploration uses visualization because it creates a more straightforward view of data sets than simply examining thousands of individual numbers or names.

In any data exploration, the manual and automated aspects also look at different sides of the same coin. In any situation where we have a massive set of information, data exploration can help cut it down to a manageable size and focus efforts to optimize project analysis.

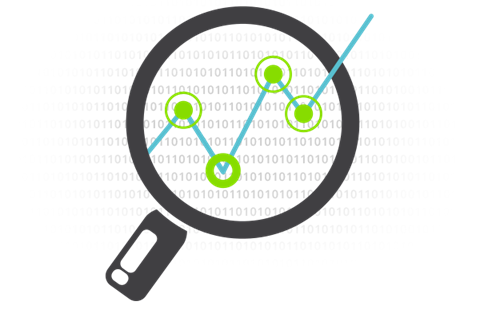


Figure 6:Data Exploring Diagram

**Model data**

We will forecast values using linear regression algorithm. We can also use modeling to group data to understand the logic behind those clusters. In short, we use regression and predictions for forecasting future values, and classification to identify, and cluster to group values. A data model describes information in a systematic way that allows it to be stored and retrieved efficiently in a Relational Database Management System (RDBMS), such as SQL Server, MySQL, or Oracle. The model can be thought of to translate the logic of accurately describing things in the real world and the relationships between them into rules that can be followed and enforced by computer code. Data modeling is sometimes as much art as science. Following the rules of normalization can be straightforward, but knowing when to break them, and what data to optimize for later access, takes perception beyond simply applying rules.

**The Stages of Data Modeling**

There are three stages or types of the data model (called schemas):

* Conceptual – This is the first step in the modeling process, which imposes a theoretical order on data as it exists in relationship to the entities being described, often real-world artifacts or concepts.
* Logical – Taking the semantic structure built at the conceptual stage, the logical modeling process attempts to impose order by establishing discrete entities, key values, and relationships in a logical structure that is brought into at least 4th normal form (4NF).
* Physical – Actually not physical at all, but it would be confusing to use “logical” twice, this step breaks the data down into the actual tables, clusters, and indexes required for the data store.

The outputs of prediction and feature engineering are a set of label times, historical examples of what we want to predict, and features, predictor variables used to train a model to predict the label. The process of modeling means training a machine-

learning algorithm to predict the labels from the features, tuning it for the business need, and validating it on holdout data.

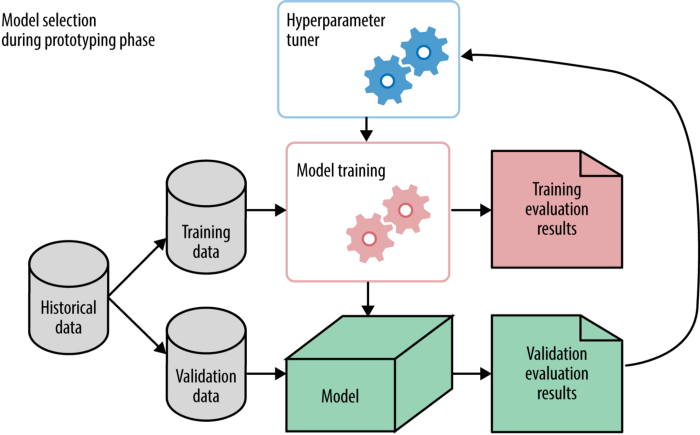


Figure 7: Stages of Data Modeling

**Interpreting Data**

The predictive power of a model lies in its ability to generalize. Interpreting data refers to the presentation of your data to a non-technical layman. We deliver the results to answer the business questions we asked when we first started the project, together with the actionable insights that we found through the data science process.

Actionable insight is a key outcome that we show how data science can bring about predictive analytics and later on prescriptive analytics. In which, we learn how to repeat a positive result, or prevent a negative outcome. We build models to construct rules that we can incorporate into our mental models of processes. A marketing firm for example can build a model that correlates marketing campaign data to financial data to determine what constitutes an effective marketing campaign. A common quote on model interpretability is that with an increase in model complexity, model interpretability goes down at least as fast. To interpret our findings we visualize our output so that layman can understand and all the stakeholders can take decisions accordingly.

**Linear regression**

Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It’s used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them (e.g. cat, dog).

**Simple regression**

Simple linear regression uses the traditional slope intercept form, concerning where mm and bb are the variables our algorithm will try to “learn” to produce the most accurate predictions’ represents our input data and y represents our prediction.

y=mx+b.

**Multivariable regression**

A more complex, multi-variable linear equation might look like this, where w represents the coefficients or weights, our model will try to learn.

F(x,y,z)=w1x+w2y+w3z

The variables x,y,z represent the attributes or distinct pieces of information, we have about each observation. For sales predictions, these attributes might include a company’s advertising spend on radio, TV, and newspapers.

Sales=w1 Radio+w2TV+w3NewsSales=w1 Radio+w2TV+w3News

[**Simple regression**](https://ml-cheatsheet.readthedocs.io/en/latest/linear_regression.html#id13)

Let’s say we are given a dataset with the following columns (features): how much a company spends on Radio advertising each year and its annual Sales in terms of units sold. We are trying to develop an equation that will let us predict units sold based on how much a company spends on radio advertising. The rows (observations) represent companies.

[**Making predictions**](https://ml-cheatsheet.readthedocs.io/en/latest/linear_regression.html#id14)

Our prediction function outputs an estimate of sales given a company’s radio advertising spend and our current values for Weight and Bias.

Sales=Weight.Radio + BiasSales=WeightRatio+Bias

**Weight**

The coefficient for the Radio independent variable. In machine learning, we call coefficients weights.

**Radio**

The independent variable. In machine learning, we call these variables features.

**Bias**

The intercept where our line intercepts the y-axis. In machine learning, we call intercepts.

Our algorithm will try to learn the correct values for Weight and Bias. By the end of our training, our equation will approximate the line of best fit.

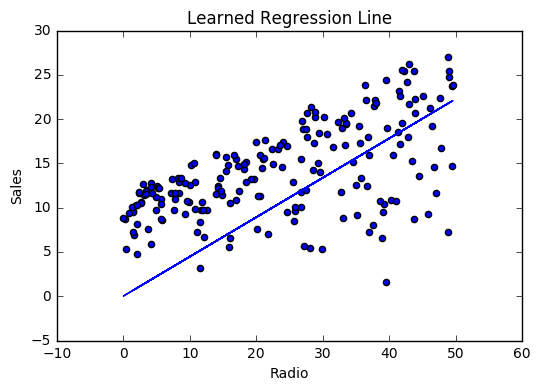


Figure 8: linear Regression

**Supervised learning**

Supervised machine learning algorithms are designed to learn by example. The name “supervised” learning originates from the idea that training this type of algorithm is like having a teacher supervise the whole process.

When training a supervised learning algorithm, the training data will consist of inputs paired with the correct outputs. During training, the algorithm will search for

patterns in the data that correlate with the desired outputs. After training, a supervised learning algorithm will take in new unseen inputs and determine which label the new inputs will classify based on prior training data. The objective of a supervised learning model is to predict the correct label for newly presented input data. At its most basic form, a supervised learning algorithm can be written simply as:

https://miro.medium.com/max/158/0*jBtf04r_NW3UKkj6

Where Y is the predicted output that is determined by a mapping function that assigns a class to an input value x. The function used to connect input features to a predicted output is created by the machine learning model during training.

# 

# CHAPTER 6: GANTT-CHART

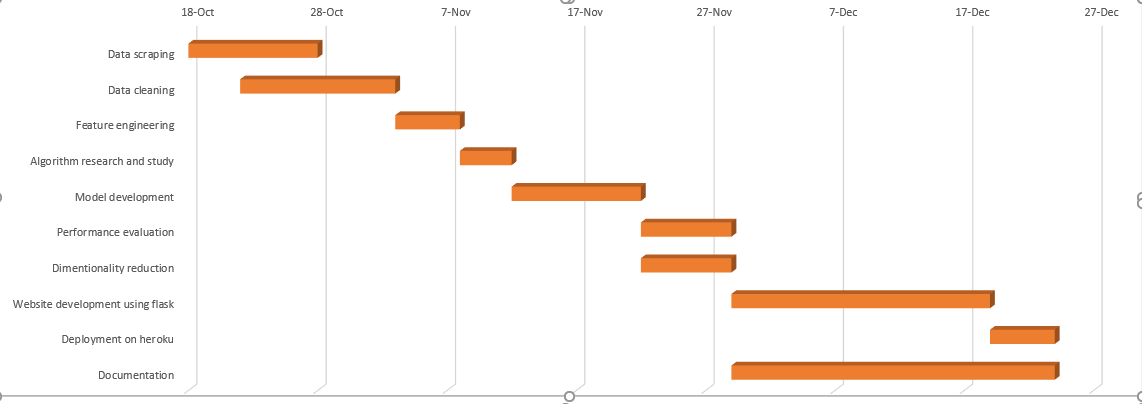


Figure 9: Gantt-Chart

**Budget:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Resources and manpower | Cost per month  in NPR | Number  Of people | Number of months | Calculation |
| Developers | 40000 | 3 | 3 | 360000 |
| Researchers | 50000 | 2 | 1 | 100000 |
| Documentation | 20000 | 4 | 3 | 240000 |
| Electricity | 900 | 4 | 3 | 10800 |
| Grand Total | | | | 710800 |

Figure 10: Budget Diagram

# CHAPTER 7: EXPECTED OUTCOME

By the end of this project we expect to have a website built using python and flask on which a user can enter the parameters of a house and the website will provide the predicted price of the house using our model and also provide some similarly priced houses from the database in the site.

# REFERENCES

1. C. R. Madhuri, G. Anuradha and M. V. Pujitha, "House Price Prediction Using Regression Techniques: A Comparative Study," 2019 International Conference on Smart Structures and Systems (ICSSS), 2019, pp. 1-5, doi: 10.1109/ICSSS.2019.8882834.
2. D. Banerjee and S. Dutta, "Predicting the housing price direction using machine learning techniques," 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), 2017, pp. 2998-3000, doi: 10.1109/ICPCSI.2017.8392275.
3. A. Varma, A. Sharma, S. Doshi and R. Nair, "House Price Prediction Using Machine Learning and Neural Networks," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, pp. 1936-1939, doi: 10.1109/ICICCT.2018.8473231.