

# **KANTIPUR ENGINEERING COLLEGE**

**(Affiliated to Tribhuvan University)**

**Dhapakhel, Lalitpur**



**[Subject Code: CT755]**

## **A MAJOR PROJECT PROPOSAL ON STOCK PRICE PREDICTION**

**Submitted by:**

**Aman Devkota [28808]**

**Ankur Karmacharya [28811]**

**Prashad Adhikary [28852]**

**A MAJOR PROJECT SUBMITTED IN PARTIAL  
FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE  
OF BACHELOR IN COMPUTER ENGINEERING**

**Submitted to:**

**Department of Computer and Electronics Engineering**

**June, 2023**

# **STOCK PRICE PREDICTION**

**Submitted by:**

**Aman Devkota [28808]**

**Ankur Karmacharya [28811]**

**Prashad Adhikary [28852]**

**A MAJOR PROJECT SUBMITTED IN PARTIAL  
FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE  
OF BACHELOR IN COMPUTER ENGINEERING**

**Submitted to:**

**Department of Computer and Electronics Engineering**

**Kantipur Engineering College**

**Dhapakhel, Lalitpur**

**June, 2023**

## ABSTRACT

Stock market prediction is when people try to figure out what the value of a stock will be in the future. They do this to make money by buying and selling stocks at the right time. Deep learning models are used to help predict stock prices. Recurrent neural networks are a type of deep learning model that is often used. There are different types of deep learning models that can be used depending on the situation. Predicting stock prices is difficult because there are many factors that can affect them. These factors can include things like politics, global economic conditions, and a company's financial performance. This project performs a comparative analysis of three deep learning models-the Long Short-term Memory (LSTM), Gated Recurrent Unit (GRU), and Vanilla RNN (VRNN)- in predicting the next day's closing price of the Nepal Stock Exchange (NEPSE) index. The performances of employed models are compared using the standard assessment metrics-Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and Correlation Coefficient (R).

**Keywords**— *LSTM, GRU, VRNN, NEPSE, Deep learning*

## ACKNOWLEDGMENT

We would like to express sincere gratitude to Department head Er. Rabindra Khati, Project Co-ordinator Er. Bishal Thapa and all the faculty members of Kantipur Engineering College for the continuous support during this project for their patience, motivation, enthusiasm, and immense knowledge. Their guidance helped us in all time of research, development and implementation of this project.

Aman Devkota [28808]

Ankur Karmacharya [28811]

Prashad Adhikary [28852]

# TABLE OF CONTENTS

<b>Abstract</b>	<b>i</b>
<b>Acknowledgment</b>	<b>ii</b>
<b>List of Figures</b>	<b>v</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background . . . . .	1
1.2 Problem Statement . . . . .	1
1.3 Objectives . . . . .	2
1.4 Project Features . . . . .	2
1.5 Application Scope . . . . .	2
1.6 System Requirement . . . . .	3
1.6.1 Development Requirements . . . . .	3
1.6.1.1 Software Requirements . . . . .	3
1.6.1.2 Hardware Requirements . . . . .	3
1.6.2 Deployment Requirements . . . . .	3
1.6.2.1 Software Requirements . . . . .	3
1.6.2.2 Hardware Requirements . . . . .	3
1.7 Project Feasibility . . . . .	3
1.7.1 Technical Feasibility . . . . .	3
1.7.2 Operational Feasibility . . . . .	4
1.7.3 Economic Feasibility . . . . .	4
1.7.4 Schedule Feasibility . . . . .	4
<b>2 Literature Review</b>	<b>5</b>
2.1 Related Projects . . . . .	5
2.1.1 Zillow . . . . .	5
2.1.2 Redfin's Home Value Estimator . . . . .	5
2.1.3 House Canary . . . . .	5
2.1.4 Propmix.io . . . . .	5
2.2 Related Works . . . . .	5
<b>3 Methodology</b>	<b>7</b>
3.1 Working Mechanism . . . . .	7
3.1.1 Data set . . . . .	8

3.1.2	Add dummy variables . . . . .	10
3.1.3	Filter the data set . . . . .	10
3.1.4	Split the data set into train set and test set . . . . .	10
3.1.5	Statistical Calculations . . . . .	10
3.1.6	Representing the normal equation of multiple linear regression in matrix form . . . . .	11
3.1.7	Calculation of intercept and coefficients using Gauss Elimination	11
3.1.8	Calculation and Evaluation of predicted price . . . . .	12
3.2	System Diagram . . . . .	13
3.2.1	Use case diagram . . . . .	13
3.2.2	DFD level diagram . . . . .	14
3.2.3	Software Development Model . . . . .	16
<b>4</b>	<b>Results and Discussion</b>	<b>17</b>
4.1	Result . . . . .	17
4.2	Discussion . . . . .	18
4.3	Limitations . . . . .	19
<b>5</b>	<b>Conclusion and Future Enhancements</b>	<b>20</b>
5.1	Conclusion . . . . .	20
5.2	Future Enhancements . . . . .	20
	<b>References</b>	<b>20</b>

## LIST OF FIGURES

3.1	Working mechanism of Stock Price Prediction System . . . . .	7
3.2	ATR . . . . .	9
3.3	RSI . . . . .	9
3.4	MFI . . . . .	9
3.5	Use case Diagram of House Price Prediction System . . . . .	14
3.6	DFD Level 0 diagram of House Price Prediction System . . . . .	15
3.7	DFD Level 1 diagram of House Price Prediction System . . . . .	15
3.8	Incremental Model . . . . .	16
4.1	Home Page . . . . .	17
4.2	Heat map for correlation . . . . .	17
4.3	Scatter plot . . . . .	18
4.4	Evaluation metrics . . . . .	18

# CHAPTER 1

## INTRODUCTION

### 1.1 Background

The stock price is determined by the highest price a buyer is willing to pay or the lowest price a seller is willing to accept. Supply and demand are key factors that can affect stock prices. High demand can lead to an increase in stock price, while high supply can lead to a decrease. However, it is difficult to determine the exact factors that contribute to changes in demand and supply. Stock market prediction involves forecasting the future value of a stock. Financial analysts use two main schools of thought for analyzing and predicting stock markets: technical analysis and fundamental analysis. [?]

Machine learning models can learn a function by analyzing data without explicit programming. The performance of these algorithms depends on the representation of the data. However, stock market time-series data is difficult to map and is best described as a random walk, making feature engineering and prediction challenging. Therefore, deep learning models are the best available tool for stock market prediction.[?]

There are two schools of thought in developing predictive models for estimating stock prices. Classical thinking uses historical facts and indicators to predict future stock prices, often through variations of Autoregressive Integrated Moving Average (ARIMA) models. However, these statistical models may not efficiently capture the noisy and non-linear behavior of stock market data. Modern theory assumes that historical data cannot reflect the exact upcoming structure due to inherent non-linearity in stock market data. With the rapid advancement of artificial intelligence and machine learning techniques, availability of large-scale data, and increased computational capabilities, robust machine learning models can capture nonlinear behavior and predict stock prices.[?]

### 1.2 Problem Statement

The stock market is a frequent topic in the media, with news outlets reporting on its daily fluctuations. Whenever the market experiences a new peak or dip, it garners significant attention. Developing an effective algorithm to forecast short-term stock prices could



potentially boost investment rates and create more business prospects in the market.

### **1.3 Objectives**

The primary objectives of this projects are as follows:

- i. Gathering the multifaceted information of NEPSE index, putting them together into a common framework, and constructing a reliable model for accurate predictions.
- ii. Conducting extensive data driven experimentation using customized parameters of LSTM, GRU, and VRNN models.
- iii. Performing comparative study of deep learning models (LSTM, GRU, and VRNN) for the best fit and forecasting under the identical conditions.
- iv. Conducting statistical experiment to validate and verify the reliability and robustness of the model.

### **1.4 Project Features**

The project will be able to accomplish following:

- Reliable
- Accurate
- User friendly
- Efficient

### **1.5 Application Scope**

Stock price prediction has various applications such as investment decision-making, risk management, algorithmic trading, portfolio optimization, and economic forecasting. Predicting stock prices can help investors make informed decisions, manage risk, optimize portfolios, and provide insights into the economy. It is a valuable tool for investors, financial institutions, and economists seeking to make informed decisions and manage risk.

## **1.6 System Requirement**

### **1.6.1 Development Requirements**

#### **1.6.1.1 Software Requirements**

- Windows/Linux/Mac
- HTML/CSS/JS
- Jupyter Notebook
- Python IDE

#### **1.6.1.2 Hardware Requirements**

- PC with at least 4-8 GB RAM
- Higher graphics of at least 2 GB

### **1.6.2 Deployment Requirements**

#### **1.6.2.1 Software Requirements**

- Web browser
- Visual studio code
- Pycharm

#### **1.6.2.2 Hardware Requirements**

- More than 1.5 GHz clock speed
- Minimum 4 GB RAM

## **1.7 Project Feasibility**

### **1.7.1 Technical Feasibility**

The technical feasibility assessment is focused on gaining in understanding of the present technical resources required by the system and their applicability to the expected needs of the proposed system. Regarding the proposed system, the technical requirement includes a PC.

### **1.7.2 Operational Feasibility**

The user will not need any formal knowledge about programming so our project is operationally feasible.

### **1.7.3 Economic Feasibility**

The purpose of the economic feasibility assessment is to determine the positive economic benefits to the user that the proposed system will provide. Most of the software used for the development is free. Thus, the project is economically feasible.

### **1.7.4 Schedule Feasibility**

## **CHAPTER 2**

### **LITERATURE REVIEW**

#### **2.1 Related Projects**

##### **2.1.1 Zillow**

Zillow is a popular online real estate marketplace that provides an automated valuation model called the Zestimate. This model uses a combination of data from public records, user-submitted data, and machine learning algorithms to predict the value of a home.

##### **2.1.2 Redfin's Home Value Estimator**

Redfin is another online real estate marketplace that provides a home value estimator tool that uses machine learning algorithms to predict the value of a home based on its location, features, and recent sales in the area.

##### **2.1.3 House Canary**

HouseCanary is a real estate analytics company that provides a range of services, including home value estimates and forecasts, neighborhood insights, and market analytics. Their home value estimates are based on a proprietary machine learning model that analyzes millions of data points.

##### **2.1.4 Propmix.io**

PropMix.io is a real estate data and analytics platform that offers a range of tools for real estate professionals, including a home value estimator that uses machine learning algorithms to predict the value of a home based on its features and location.

#### **2.2 Related Works**

Anirudh Kaushal and Achyut Shankar researched in detail about house price prediction using multiple linear regression method. In the paper “House Price Prediction Using Multiple Linear Regression” published on April 25, 2021 there is explanation about filtering of data set, data processing, training and evaluating multiple linear regression model. [?] Manasa, J., Gupta, R., & Narahari, N. S. studied and compared the algorithms for estimation of price of houses in city of Bengaluru in the paper “Machine learning based predicting house prices using regression techniques”. [?] M. Thamarai, S P. Malarvizhi have compared the decision tree and regression algorithms in the paper

“House Price Prediction Modeling Using Machine Learning”. The basis of comparison was accuracy, MAE, MSE and RMSE. [?] Rana, V. S., Mondal, J., Sharma, A., & Kashyap, I. have studied in detail about the machine learning algorithms and compared the results obtained by each to learn about the algorithm most suitable to use in the house price prediction system in their paper “House Price Prediction Using Optimal Regression Techniques”. [?] Madhuri, C. R., Anuradha, G., & Pujitha, M. V. studied on different regression algorithms. In the paper “House price prediction using regression techniques: a comparative study” published on March, 2019, there is explanation about different types of regression methods and their accuracy to predict the values. [?]

## CHAPTER 3 METHODOLOGY

### 3.1 Working Mechanism

The development of stock price prediction system involves major steps which is depicted in the diagram given below:

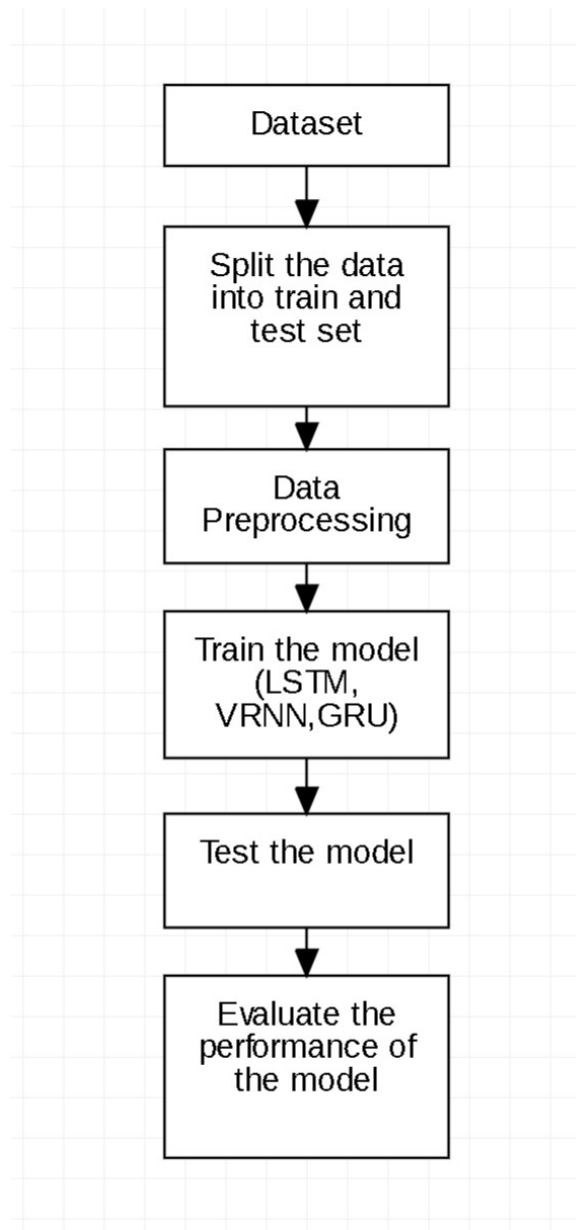


Figure 3.1: Working mechanism of Stock Price Prediction System

### 3.1.1 Data set

The data set contains financial data, macroeconomic data and technical indicators. Fundamental data is significant or historical data that provides important information needed for stock trading. It has an opening price, closing price, high price, low price and volume. The opening price is the first trading price at the opening of the trading day and the closing price is the last trading price of the stock that day. Likewise, the high and low prices are the highest and lowest prices for that day. Volume refers to the number of products traded in a trading day and indicates traders' interest in the market. High volume means more satisfaction and repetition. All business history data is daily data entry from Share Sansar portal, is one of the popular products providing detailed information about Nepal stock market.

Macroeconomic data is a macroeconomic variable that affects stock market performance. Under the umbrella of macroeconomic factors, the representative features that affect stock price forecasts are remittances (RMT), inflation rate (IR), commercial bank interest rates (CBIR), treasury bills (TRB), consumer price index (CPI) and exchange rate (ER). Remittance is the money Nepalese workers send from abroad. Inflation is the rate of increase in prices over a given period of time. The commercial bank interest rate is the bank rate at which Nepal Rastra Bank lends money to domestic financial institutions. The treasury bill is typically a promissory maturity note issued by a government as a primary instrument for regulating money supply and raising funds via open market operations. The Consumer Price Index is a measure of the average change overtime in the prices paid by urban consumers for a market basket of consumer goods and services. Exchange rate is the value of one country's currency against another country's currency.

Technical indicator includes Moving Average Convergence Divergence (MACD), Average True Range (ATR), Relative Strength Index (RSI), and Money Flow Index (MFI). MACD is calculated by subtracting the 26-day exponential moving average (EMA) from the 12-day EMA. ATR measures the market volatility and is defined as follows:

$$\text{True Range (TR)} = \max\{(H - L), |H - C_p|, |L - C_p|\}$$

$$\text{Average True Range (TR)} = \frac{1}{n} \sum_i^n TR_i$$

Figure 3.2: ATR

where,  $H, L, C_p$  represent current high, current low, previous close prices and  $TR_i, n$  represent a particular true range, the time period respectively.

The RSI is a momentum indicator that signals whether a security is overbought or oversold with current price levels, which is computed as follows:

$$RSI = 100 - \frac{100}{1 + \frac{\text{Average gain}}{\text{Average loss}}}$$

Figure 3.3: RSI

where average gain and average loss are the average percentage gain and loss calculated over the certain look-back period.

The MFI communicates a possible reversal in time and provides a signal for further investment. Its calculation starts by finding the typical price (TP), the average of high, low, and close prices for each trading day. If the current TP is higher than the previous one, then positive money flow is calculated by multiplying the current TP with its volume. Similarly, a negative money flow is obtained if the current typical price is lower than the previous one. If the TP does not change, both positive and negative money flow will be zero. Summing all positive money flow indexes leads to positive money flow for a particular period. Negative money flow is calculated similarly. Mathematically, MFI is defined as:

$$MFI = 100 - \frac{100}{1 + \text{Money Ratio}}$$

Figure 3.4: MFI



### 3.1.2 Add dummy variables

The parameters in the data set like addresses are string values but the multiple linear regression can only accept numerical values so they are need to be converted into numerical values. In order to convert it we need to give them the value 0 and 1 on the basis that they are available but there may arise dummy trap problem (when two independent parameters affect each other which can cause conflict in multiple linear regression). So to solve this problem the dummy variables should be one less than the n number of string valued parameters.

### 3.1.3 Filter the data set

All the parameters in the raw data set are not needed for multiple linear regression as some of the parameters have little or no significance in changing the price of the house. So to filter out the useless parameters, we find the single correlation between price of the house and the parameters. Correlation coefficients are used to measure the strength of the linear relationship between two variables. A correlation coefficient greater than zero indicates a positive relationship while a value less than zero signifies a negative relationship. If the correlation is very low or near to zero, they can be neglected.

Mathematical calculation for correlation

$$(r_{xy}) = \frac{(n \sum xy - \sum x \sum y)}{(\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}} \quad (3.1)$$

### 3.1.4 Split the data set into train set and test set

The data set is divided in train set and test set so that the train set can be used for training the model.

### 3.1.5 Statistical Calculations

For the multiple linear regression, we need to find the values for  $\sum x$ ,  $\sum xy$ , ...  $\sum x_n$  etc. which are calculated in this stage.

### 3.1.6 Representing the normal equation of multiple linear regression in matrix form

The equation of plane is

$$x = w_0 + w_1x_1 + w_2x_2 + \dots w_Nx_N \quad (3.2)$$

Here,  $x_1, x_2, \dots, x_N$  are independent parameter variables and  $x$  is dependent parameter variable. The intercept is  $w_0$  and coefficients are  $w_1, w_2, \dots, w_N$ .

Now to find the values of  $w_0, w_1, w_2, \dots, w_N$ , we need the normal equations which are as follows:

$$\sum x = nw_0 + w_1 \sum x_1 + w_2 \sum x_2 + \dots w_N \sum x_N \quad (3.3)$$

$$\sum xx_1 = w_0 \sum x_1 + w_1 \sum x_1^2 + w_2 \sum x_1x_2 + \dots w_N \sum x_1x_N \quad (3.4)$$

$$\sum xx_2 = w_0 \sum x_2 + w_1 \sum x_1x_2 + w_2 \sum x_2^2 + \dots w_N \sum x_2x_N \quad (3.5)$$

:

$$\sum xx_N = w_0 \sum x_N + w_1 \sum x_1x_N + w_2 \sum x_2x_N + \dots w_N \sum x_N^2 \quad (3.6)$$

Python does not recognize equations so we need to represent these equations in matrix form for further calculations.

### 3.1.7 Calculation of intercept and coefficients using Gauss Elimination

To find the values of  $w_0, w_1, w_2, \dots, w_N$ , we need to solve the matrices. For this we use the Gauss Elimination method.

Algorithm:

1. Start
2. Declare the variables and read the order of the matrix N
3. Take the coefficients of the linear equations as:  
Do for k= 1 to n  
Do for j = 1 to n+1  
Read a[k][j]

```

    End for j
  End for k
4. Do for k= 1 to n-1
  Do for i= k+1 to n
    Do for j= k+1 to n+1
       $a[i][j] = a[i][j] - a[i][k]/a[k][k]*a[k][j]$ 
    End for j
  End for i
End for k
5. Compute  $x[n] = a[n][n+1]/a[n][n]$ 
6. Do for k= n-1 to 1
  sum = 0
  Do for j = k+1 to n
    sum +=  $a[k][j]*x[j]$ 
  End for j
   $x[k] = 1/a[k][k] * (a[k][n+1] - \text{sum})$ 
End for k
7. Display the result  $x[k]$ 
8. Stop

```

### 3.1.8 Calculation and Evaluation of predicted price

Now the calculated values of intercept and coefficients are inserted in the equation (3.2) to calculate the predicted price.

We evaluated model's performance using metrics: the coefficient of determination  $R^2$ , Root Mean Squared Error(RMSE).

RMSE: It can be defined as the standard sample deviation between the predicted values and the observed ones. It is to be noted that unit of RMSE is same as dependent variable  $y$ . The lower RMSE values are indicative of a better fit model. If the model's primary objective is prediction then RMSE is a stronger measure.

R-squared: The R-square value provides a measure of how much the model replicates the actual results, based on the ratio of total variation of outcomes as explained in the model. The higher the R-squared, the better the model fits the data given. The R-squared

value ranges from 0 to 1, representing the percentage of a squared correlation between the target variable's expected and real values. But in case of multiple linear regressions, R-squared value may increase with increasing features even though the model is not actually improving. A related, Adjusted R-squared statistic can be used to address this disadvantage. This measures the model's goodness and penalizes the model to use more predictors.

## **3.2 System Diagram**

### **3.2.1 Use case diagram**

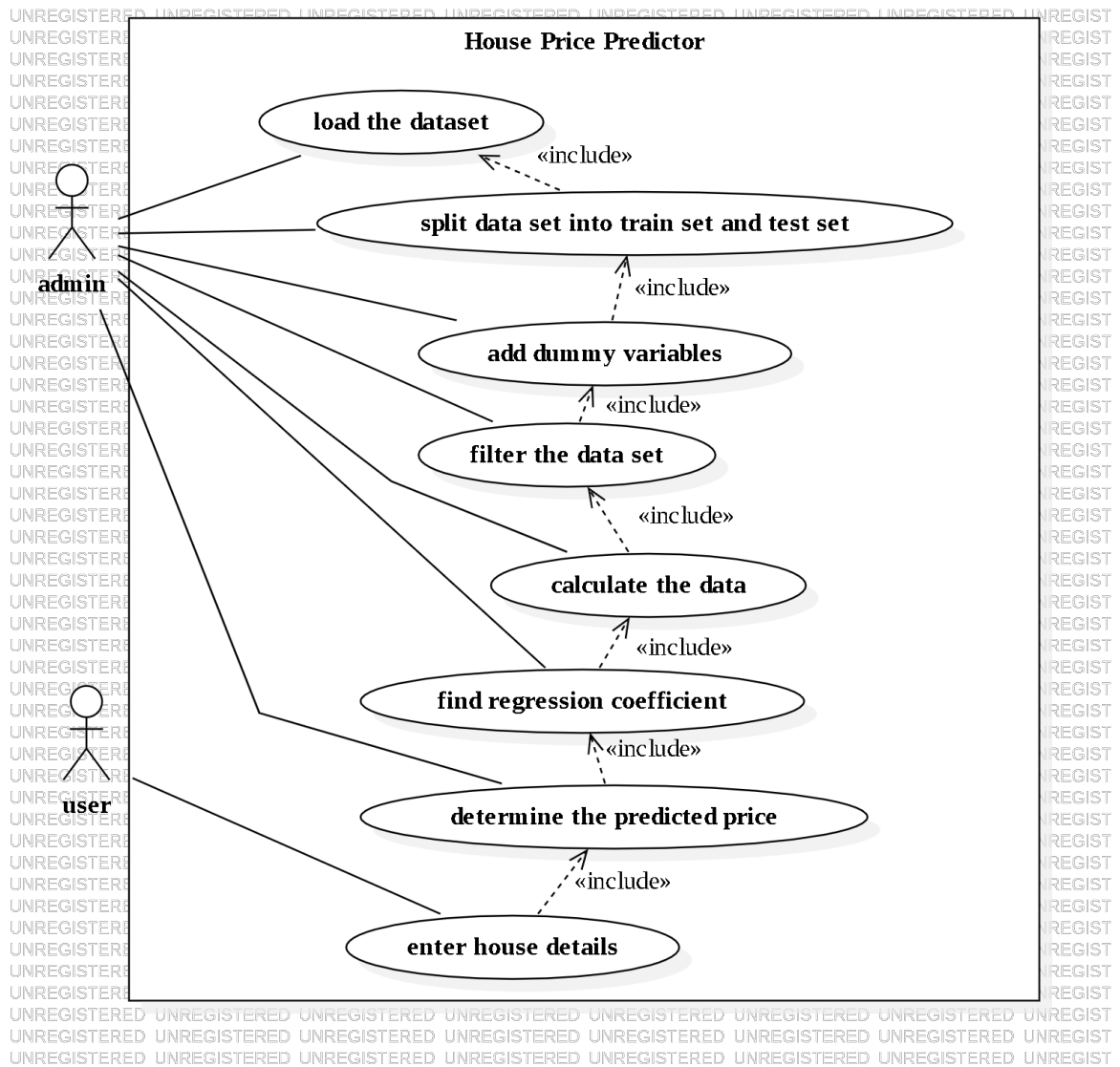


Figure 3.5: Use case Diagram of House Price Prediction System

### 3.2.2 DFD level diagram

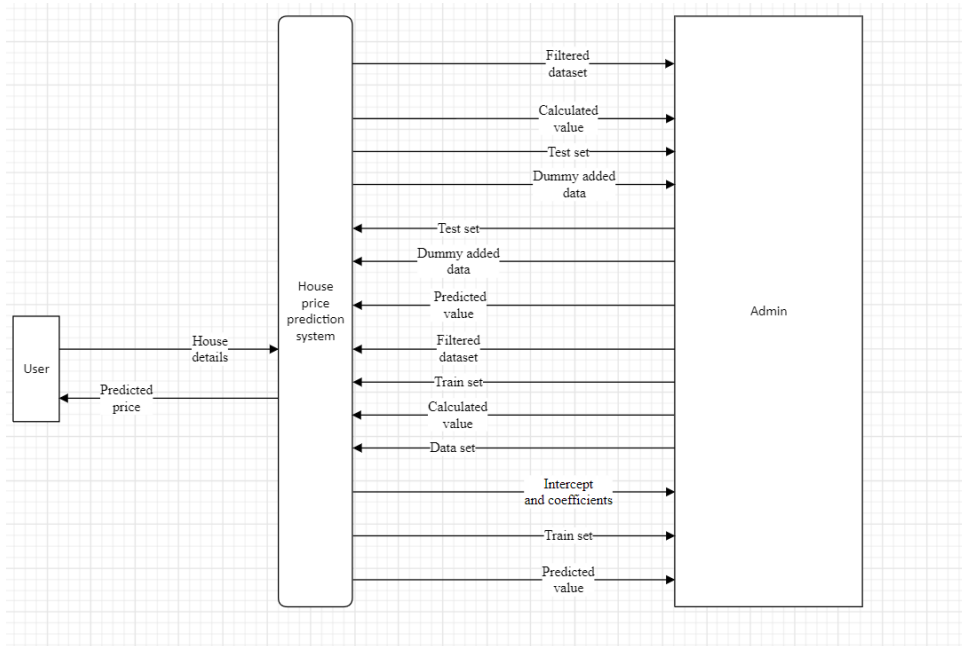


Figure 3.6: DFD Level 0 diagram of House Price Prediction System

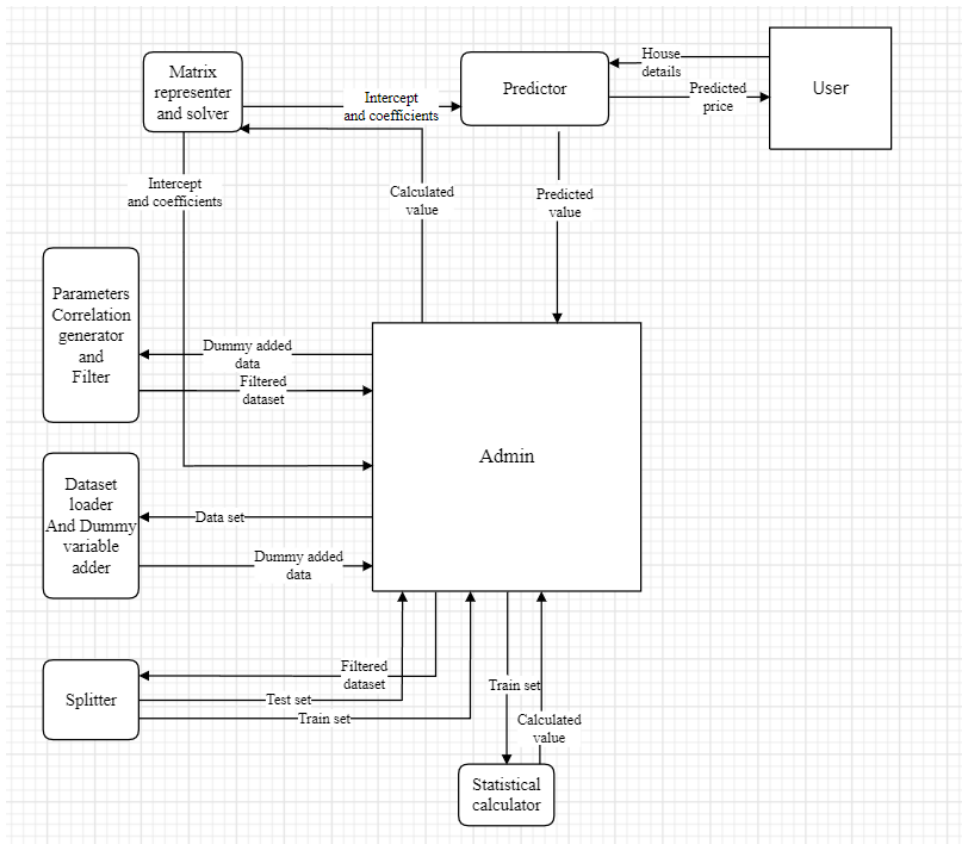


Figure 3.7: DFD Level 1 diagram of House Price Prediction System

### 3.2.3 Software Development Model

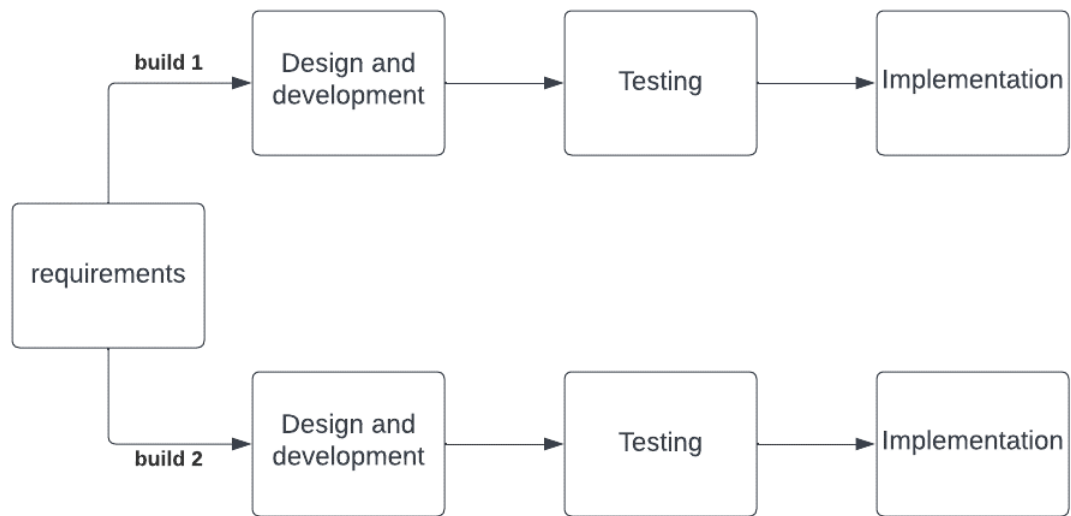


Figure 3.8: Incremental Model

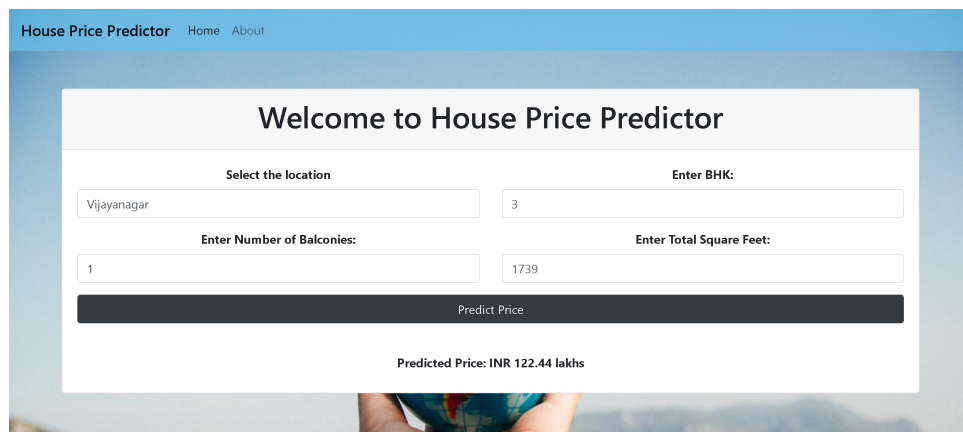
Incremental model is a method of software engineering that combines the elements of waterfall model in iterative manner. It involves both development and maintenance. In this model requirements are broken down into multiple modules. Incremental development is done in steps from analysis design, implementation, testing/verification, maintenance. Each iteration passes through the requirements, design, coding and testing phases. The first increment is often a core product where the necessary requirements are addressed, and the extra features are added in the next increments. The core product is delivered to the client. Once the core product is analyzed by the client, there is plan development for the next increment.

## CHAPTER 4

### RESULTS AND DISCUSSION

#### 4.1 Result

We have completed the design and development of the project along with obtaining the desired output of the project. The project currently takes in location, area(square feet), bhk and balcony and predicts the price of the house through linear regression. The interface of the project is shown below:



The screenshot shows a web application titled "House Price Predictor" with a navigation bar containing "Home" and "About". The main content area has a heading "Welcome to House Price Predictor". Below this, there are four input fields: "Select the location" (with "Vijayanagar" entered), "Enter BHK:" (with "3" entered), "Enter Number of Balconies:" (with "1" entered), and "Enter Total Square Feet:" (with "1739" entered). A "Predict Price" button is located below these fields. At the bottom, the output is displayed as "Predicted Price: INR 122.44 lakhs".

Figure 4.1: Home Page

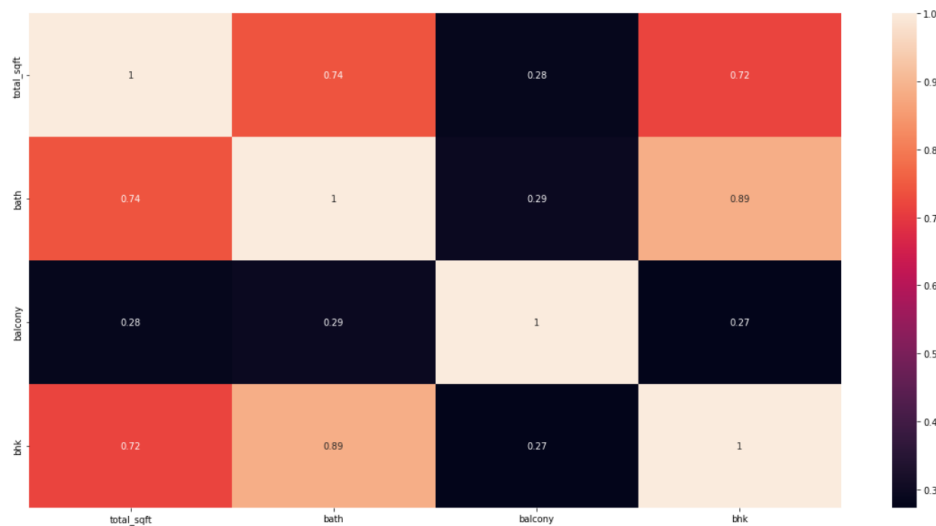


Figure 4.2: Heat map for correlation

From Figure 4.2, we can see that the correlation between bathroom and bhk is very high so we can ignore one of them. In our case, we have ignored bathroom.



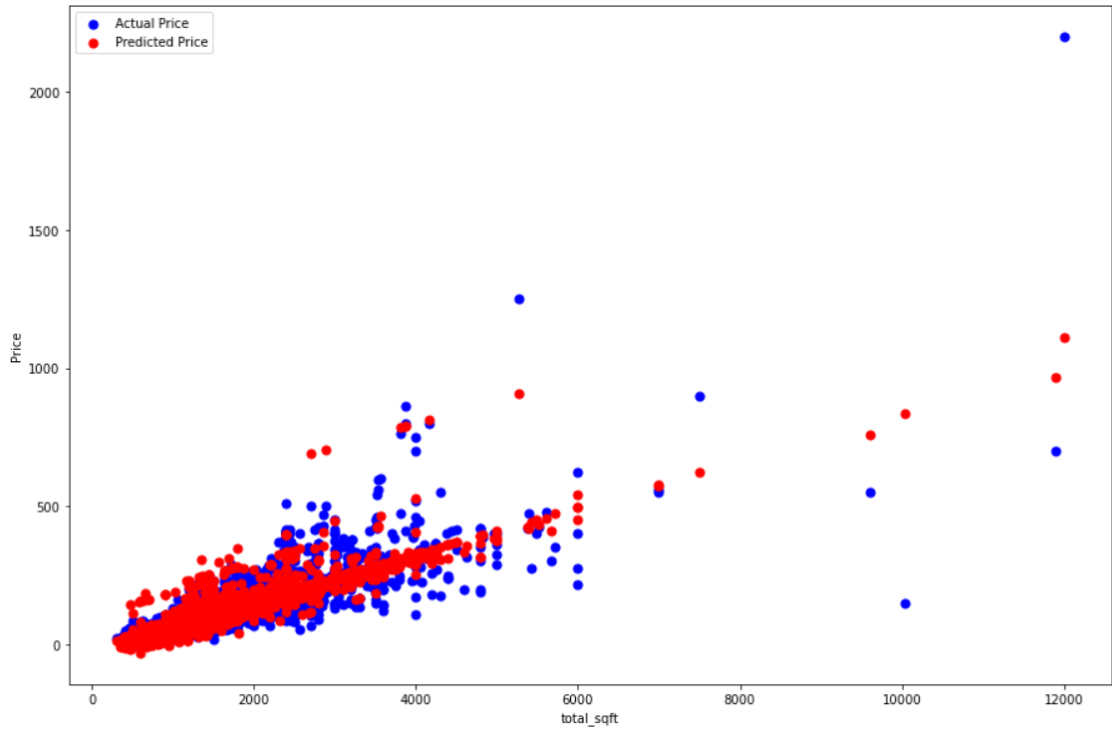


Figure 4.3: Scatter plot

From the above figure, it can be concluded that for some data points, the actual price is very close to the predicted price which means that for some data points the model is highly accurate. However, the figure also shows that for some points the difference between the actual price and the predicted price is large which shows that for some data points the result is less accurate. Overall, we can say that the model has a decent amount of accuracy.

Metric	Train set	Test set
R2	0.84	0.85
MSE	1024.22	181.49
RMSE	32.0035	13.47

Figure 4.4: Evaluation metrics

## 4.2 Discussion

House Price Predictor system is a website that predicts the prices of houses based on location, area, number of balconies and bhk. For front end development of the website, we have used HTML, CSS and JS. We used Pandas and numpy libraries for accessing

dataframe. Matplotlib is used for data visualisation. We have done predictions with the help of Linear Regression, Correlation and Gauss Elimination method. Evaluation of the system was done by R-squared, Mean Squared Error and Root Mean Squared Error. Flask framework was used to connect the front end and back end of our website.

### **4.3 Limitations**

1. Outliers can have huge effects on the regression.
2. Linear Regression also looks at a relationship between the mean of the dependent variables and the independent variables. Just as the mean is not a complete description of a single variable, linear regression is not a complete description of relationships among variables.

## **CHAPTER 5**

### **CONCLUSION AND FUTURE ENHANCEMENTS**

#### **5.1 Conclusion**

In this project, we have used the regression model to predict the price of different houses. It comes under the area of supervised learning which is one of the types of machine learning. All the steps required for the successful completion of the house price prediction system have been completed.

#### **5.2 Future Enhancements**

1. increase and update the data set on a regular basis
2. add locations for Nepal