HOUSE PRICE PREDICTION USING MACHINE LEARNING

INNOVATIVE AND CREATIVE PROJECT

Submitted by

BANUMATHI S (18BCS025)

in partial fulfillment for the award of the degree

of

Bachelor of Engineering

in

Computer Science and Engineering

Dr.Mahalingam College of Engineering and Technology

Pollachi - 642003

An Autonomous Institution

Affiliated to Anna University, Chennai - 600 025

JANUARY 2022

Dr.Mahalingam College of Engineering and Technology

Pollachi - 642003

An Autonomous Institution

Affiliated to Anna University, Chennai -600 025

BONAFIDE CERTIFICATE

Certified that this project report, "HOUSE PRICE PREDICTION USING MACHINE LEARNING"
is the bonafide work of

BANUMATHI S (18BCS025)

who carried out the project work under my supervision.

Dr. G.Anupriya
HEAD OF THE DEPARTMENT
Computer Science and Engineering
Dr. Mahalingam College of Engineering and Technology, Pollachi – 642003

Dr.M.Pandi
SUPERVISOR
Assistant Professor(SG)/CSE
Computer Science and Engineering
Dr. Mahalingam College of Engineering and Technology, Pollachi – 642003

Submitted for the Autonomous End Semester Examination Innovative and Creative ProjectViva-voce held on _____

INTERNAL EXAMINER **EXTERNAL EXAMINER**

HOUSE PRICE PREDICTION USING MACHINE LEARNING

ABSTRACT

House prices increase every year, so there is a need for a system to predict house prices in the future. House price prediction can help the developer determine the selling price of a house and can help the customer to arrange the right time to purchase a house. There are three factors that influence the price of a house which include physical conditions, concept and location.We propose to implement a house price prediction model of Bangalore, India. It's a Machine Learning model which integrates Data Science and Web Development. We have deployed the app on the Heroku Cloud Application Platform. Housing prices fluctuate on a daily basis and are sometimes exaggerated rather than based on worth. The major focus of this project is on predicting home prices using genuine factors. Here, we intend to base an evaluation on every basic criterion that is taken into account when establishing the pricing. The goal of this project is to learn Python and get experience in Data Analytics, Machine Learning, and AI.

ACKNOWLEDGEMENT

First and foremost, I wish to express my deep unfathomable feeling, gratitude to my institution and my department for providing me a chance to fulfill my long cherished dreams of becoming Computer Science Engineers.

I express my sincere thanks to my honorable Secretary **Dr.C.Ramaswamy,** for providing me with required amenities.

I wish to express my hearty thanks to **Dr.A.Rathinavelu,** Principal of my college, for his constant motivation and continual encouragement regarding my+ project work.

I am grateful to **Dr.G.Anupriya,** Head of the Department, Computer Science and Engineering, for her direction delivered at all times required. I also thank her for her tireless and meticulous efforts in bringing out this project to its logical conclusion.

My hearty thanks to my guide **Dr.M.Pandi,** Assistant Professor(Selection Grade) for her constant support and guidance offered to me during the course of my project by being one among me and all the noble hearts that gave me immense encouragement towards the completion of my project.

LIST OF ABBREVIATIONS

HPP: House Price Prediction

ML: Machine Learning

LR: Linear Regression

EDA:Exploratory Data Analysis

LASSO: Least Absolute Shrinkage and Selection Operator

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Machine learning is an relevance of artificial intelligence (AI) that endow with systems the capability to repeatedly learn and improve from experience without being overtly programmed[1][2]. Machine learning do centre of attention on the growth of computer programs that can access data and use it be trained for themselves. Networking Sites using which helps the people to connect with the existing friends, relatives, .group of employees etc.

Investment is a business activity that most people are interested in this globalization era. There are several objects that are often used for investment, for example, gold, stocks and property. In particular, property investment has increased significantly since 2011, both on demand and property selling. One of the increasing of property demand is because of high population in Indonesia. Indonesian Central Bureau of Statistics states that in East Java 50% of the population of East Java classified as a young population who have age approximately at 30 years old. The result of this census indicates that the younger generation will need a house or buy a house in the future. Based on preliminary research conducted, there are two standards of house price which are valid in buying and selling transaction of a house that is house price based on the developer (market selling price) and price based on Value of Selling Tax Object (NJOP). According to Lim, et al the fundamental problem for a developer is to determine the selling price of a house. In determining the price of home, the developer must calculate carefully and determine the appropriate method because property prices always increase continuously and almost never fall in the long term or short.

There are several approaches that can be used to determine the price of the house, one of them is the prediction analysis. Developers are interested to know the future trends for their decision making. In order to accurately estimate property prices and future trends, large amount of data that influences land price is required for analysis, modelling and forecasting. The factors that affect the land price have to be studied and their impact on price has also to be modelled

1.1 Problem Statement

Housing prices are an important reflection of the economy, and housing price range are of great prices.

The goal of this statistical analysis is to help us understand the relationship between house features and how these variables are used to predict house price.

The goal of this project is to create a regression model that are able to accurately estimate the price of the house given the features.

1.2 Objective

The aim is to predict the efficient house pricing for real estate customers with respect to their budgets and priorities.

By analyzing previous market trends and price ranges, and also upcoming developments future pieces will be predicted.

Using two different models in terms of minimizing the difference between predicted and actual rating.

1.3 Scope of the project

This study aims to analyse the accuracy of predicting house prices when using Multiple linear, Lasso, Ridge, Random Forest regression algorithms and Artificial neural network (ANN). In addition, the given datasets should be processed to enhance performance, which is accomplished by identifying the necessary features by applying one of the selection methods 2 to eliminate the unwanted variables since each house has its unique features that help to estimate its price

2. LITERATURE SURVEY

Literature survey is the most important step in any kind of research.The studies approaches to calculate the Price of House are done in previous years.

Changes in the value of the real estate will have an impact on many home investors, bankers, policymakers, and others. Real estate investing appears to be a tempting option for investors. As a result, anticipating the important estate price is an essential economic indicator. According to the 2011 census, the Asian country ranks second in the world in terms of the number of households, with a total of 24.67 crores.

First, this paper is a procedure of finding waves. It is linked to this paper in that it predicts the predicted dangerous waves by be relevant linear regression algorithms shows the result of the linear regression. In most recent two decades forecasting the property worth has turn out to be an important field. Rise in insist intended for property and unpredictable behaviour of financial system induce researchers to come across out a method that forecast the real estate prices devoid of any biases. However the problem is that presentation evaluation is pedestal only on classifiers. Performance comparison of other machine learning algorithms should also be measured. In article, the authors have forecast the stock market prices using linear regression methods.

Machine learning has many application’s out of which one of the applications is prediction of real estate. The real estate market is one of the most competitive in terms of pricing and same tends to be vary significantly based on lots of factor, forecasting property price is an important modules in decision making for both the buyers and investors in supporting budget allocation, finding property finding stratagems and determining suitable policies hence it becomes one of the prime fields to apply the concepts of machine learning to optimize and predict the prices with high accuracy. The study on land price trend is felt important to support the decisions in urban planning. The real estate system is an unstable stochastic process. Investors decisions are based on the market trends to reap maximum returns. Developers are

interested to know the future trends for their decision making. To accurately estimate property prices and future trends, large amount of data that influences land price is required for analysis, modelling and forecasting. The factors that affect the land price have to be studied and their impact on price has also to be modelled.

An analysis of the past data is to be considered. It is inferred that establishing a simple linear mathematical relationship for these time-series data is found not viable for forecasting. Hence it became imperative to establish a non-linear model which can well fit the data characteristic to analyse and forecast future trends. As the real estate is fast developing sector, the analysis and forecast of land prices using mathematical modelling and other scientific techniques is an immediate urgent need for decision making by all those concerned. The increase in population as well as the industrial activity is attributed to various factors, the most prominent being the recent spurt in the knowledge sector viz. Information Technology (IT) and Information technology enabled services. Demand for land started of showing an upward trend and housing and the real estate activity started booming. All barren lands and paddy fields ceased their existence to pave way for multistore and highrise buildings. Investments started pouring in Real estate Industry and there was no uniform pattern in the land price over the years. The need for predicting the trend in land prices was felt by all in the industry viz. the Government, the regulating bodies, lending institutions, the developers and the investors. Therefore, in this paper, we present various important features to use while predicting housing prices with good accuracy. We can use regression models, using various features to have lower Residual Sum of Squares error. While using features in a regression model some feature engineering is required for better prediction. Often a set of features multiple regressions or polynomial regression (applying a various set of powers in the features) is used for making better model fit. For these models are expected to be susceptible towards over fitting ridge regression is used to reduce it. So, it directs to the best application of regression models in addition to other techniques to optimize the result. It is about taking suitable action to maximize reward in a particular situation. It is employed by various software and machines to find the best possible behaviour or path it should take in a specific situation.

3. METHODOLOGY

3.1 Data Collection

The statistics were gathered from Bangalore home prices.The information includes many variables such as area type, availability, location, BHK, society, total square feet, bathrooms, and balconies.

3.2 Linear Regression

Linear regression is a supervised learning technique. It is responsible for predicting the value of a dependent variable (Y) based on a given independent variable (X). It is the connection between the input (X) and the output (Y). It is one of the most machine learning algorithm (m) and c is the intercept

$$(y = mx+c) \qquad (1)$$

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independentvariables.

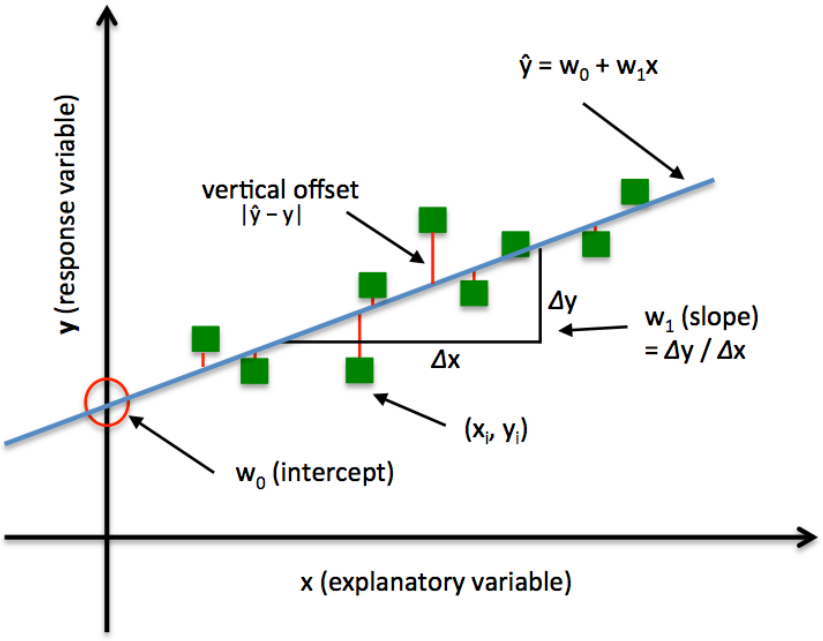

Figure 3.1: Linear Regression illustration

3.3 Random Forest Regression

A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap Aggregation, commonly known as bagging.

Bagging, in the Random Forest method, involves training each decision tree on a different data sample where sampling is done with replacement.

The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

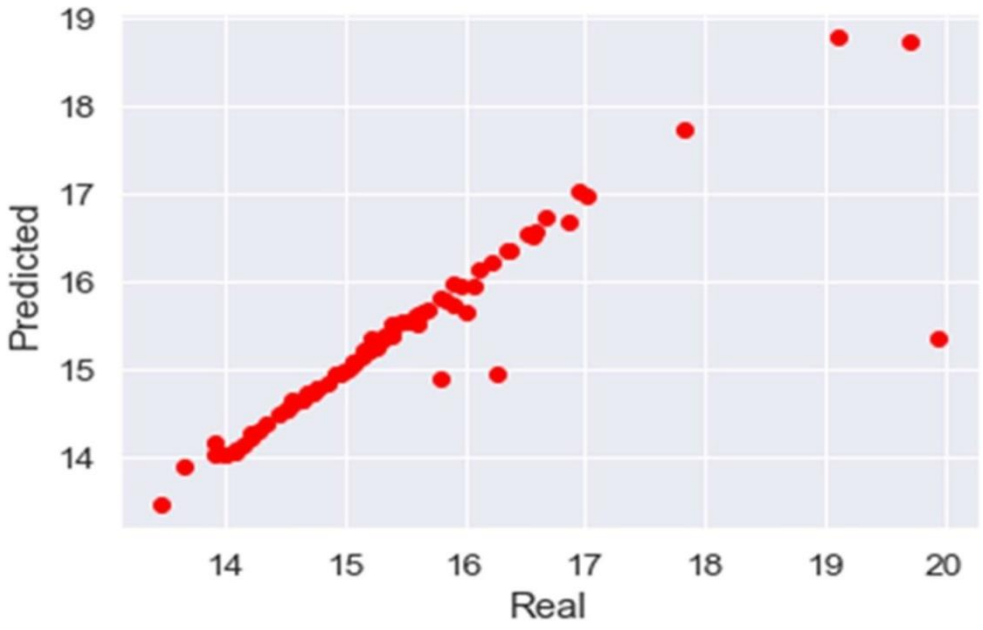

Figure 3.2: Random forest regression

3.4 Regression Tree

It supports both continuous and categorical input variables. Regression trees are regarded as research with various machine algorithms for the regression issue, with the Decision Tree approach providing the lowest loss. The R-Squared value for the Decision Tree is 0.998, indicating that it is an excellent model. A regression tree is built through a process known as binary recursive partitioning, which is an iterative process that splits the data into partitions or branches, and then continues splitting each partition into smaller groups as the method moves up each branch.

3.5 Decision Tree Regression

It is an object that trains a tree-structured model to predict data in the future in order to provide meaningful continuous output. The core principles of decision trees, Maximizing Information Gain, Classification trees, and Regression trees are the processes involved in decision tree regression. The essential notion of decision trees is that they are built via recursive partitioning. Each node can be divided into child nodes, beginning with the root node, which is known as the parent node. These nodes have the potential to become the parent nodes of their resulting offspring nodes. The nodes at the informative features are specified as the maximizing information gain, to establish an objective function that is to optimize the tree learning method.

3.6 Support Vector Regression

Supervised learning is linked with learning algorithms that examine data for classification and regression analysis.

3.7 Algorithms used

Machine Learning offers a wide range of algorithms to choose from. These are usually divided into classification, regression, clustering and association. Classification and regression algorithms come under supervised learning while clustering and association comes under unsupervised learning.

Regression

Regression in machine learning consists of mathematical methods that allow data scientists to predict a continuous outcome (y) based on the value of one or more predictor variables (x).Example: Linear regression

Regression analysis is used to find equations that fit data. Once we have the regression equation, we can use the model to make predictions. One type of regression analysis is linear analysis. When a correlation coefficient shows that data is likely to be able to predict future outcomes and a scatter plot of the data appears to form a straight line, you can use simple linear regression to find a predictive function. If you recall from elementary algebra, the equation for a line is $y = mx + b$. This article shows you how to take data, calculate linear regression, and find the equation $y = a + bx$**.** Note**:** If you're taking AP statistics, you may see the equation written as $b_0 + b_1x$, which is the same thing (you're just using the variables $b_0 + b_1$ instead of $a + b$.

Supervised Learning

Supervised learning, as the name indicates, has the presence of a supervisor as a teacher. Basically supervised learning is when we teach or train the machine using data that is well labeled. Which means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that the supervised learning algorithm analyses the training data(set of training examples) and produces a correct outcome from data.

Unsupervised Learning

Unsupervised learning is the training of a machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here the task of the machine is to group unsorted information according to similarities, patterns, and differences without any prior training of data.

4. PROJECT

4.1 Data

The data is the most important aspect of a machine learning assignment, to which special attention should be paid. Indeed, the data will heavily affect the findings depending on where we found them, how they are presented, if they are consistent, if there is an outlier, and so on. Many questions must be addressed at this stage to ensure that the learning algorithm is efficient and correct.

4.2 Data sets

Table 1: Data sets

Location	Size	Bath	Total_sqft	Price
1st Block Jayanagar	4 BHK	4	2850	428.0
2nd Phase Judicial Layout	2 BHK	3	1450	50.75
5th Block Hbr Layout	2 BHK	4	1670	370.0
Whitefield	3 BHK	2	1530	150.0

4.3 Data Exploration

Data exploration is the first step in data analysis and typically involves summarizing the main characteristics of a data set, including its size, accuracy, initial patterns in the data and other attributes. It is commonly conducted by data analysts using visual analytics tools, but it can also be done in more advanced statistical software, Python. Before it can conduct analysis on data collected by multiple data sources and stored in data warehouses, an organization must know how many cases are in a data set, what variables are included, how many missing values there are and what general hypotheses the data is likely to support. An initial exploration of the data set can help answer these questions by familiarizing analysts with the data with which they are working.

4.4 Data Visualization

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

4.5 Data Selection

Data selection is defined as the process of determining the appropriate data type and source, as well as suitable instruments to collect data. Data selection precedes the actual practice of data collection. This definition distinguishes data selection from selective data reporting (selectively excluding data that is not supportive of a research hypothesis) and interactive/active data selection (using collected data for monitoring activities/events, or conducting secondary data analyses). The process of selecting suitable data for a research project can impact data integrity.

The primary objective of data selection is the determination of appropriate data type, source, and instrument(s) that allow investigators to adequately answer research questions. This determination is often discipline-specific and is primarily driven by the nature of the investigation, existing literature, and accessibility to necessary data sources.

4.6 Data Transformation

The log transformation can be used to make highly skewed distributions less skewed. This can be valuable both for making patterns in the data more interpretable and for helping to meet the assumptions of inferential statistics. It is hard to discern a pattern in the upper panel whereas the strong relationship is shown clearly in the lower panel. The comparison of the means of log-transformed data is actually a comparison of geometric means. This occurs because, as shown below, the anti-log of the arithmetic mean of log-transformed values is the geometric mean.

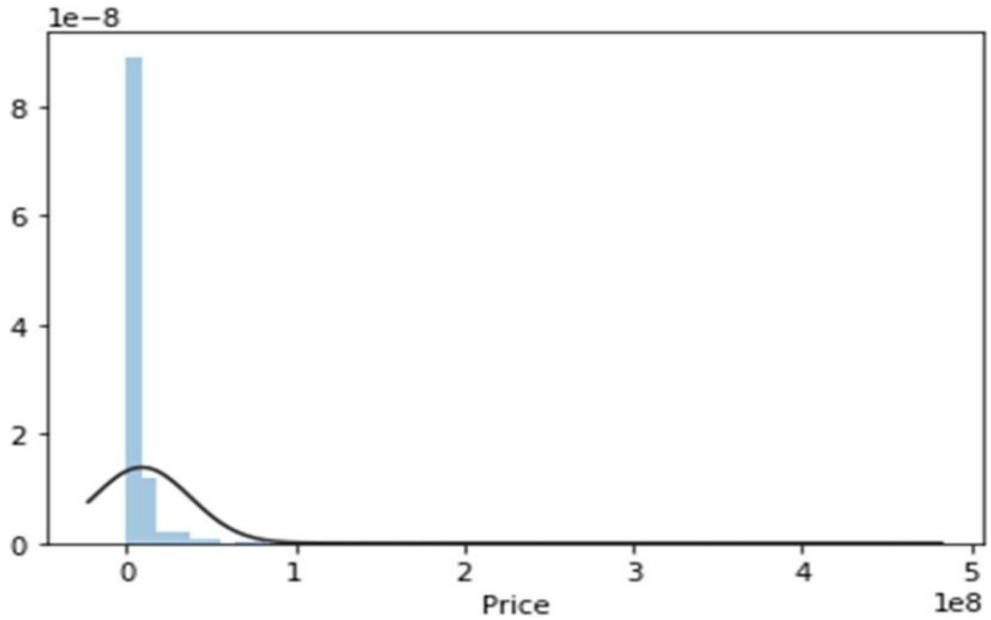

Figure 4.1: Normal price

4.7 Block Diagram

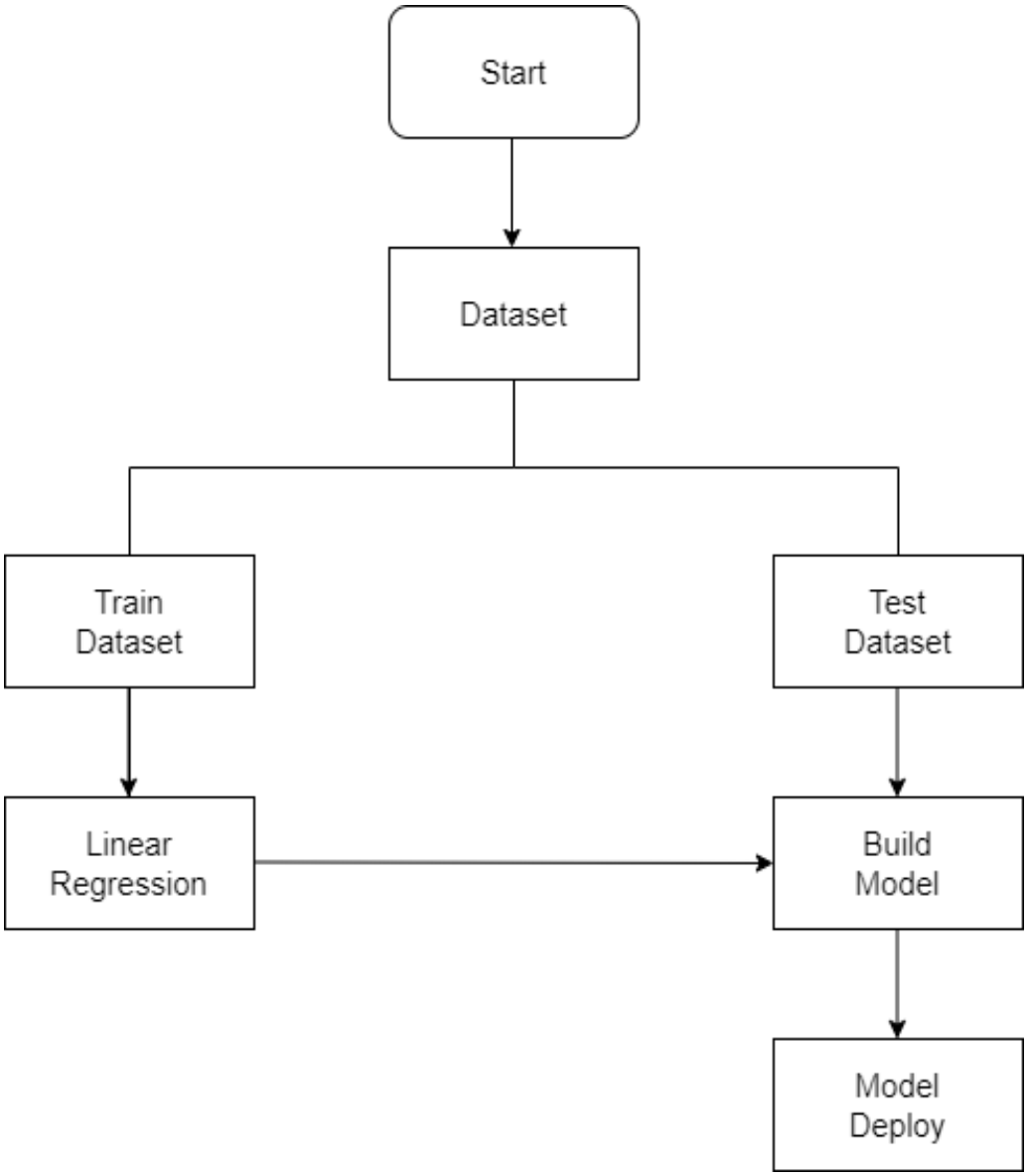

Figure 4.2: Work flow of the system

5. IMPLEMENTATION

5.1 Steps To Create Model

1 . Import libraries

- Pandas
- Numpy
- Scikit Learn
- Maplotlib

2. Load Dataset

3. Exploratory Data Analysis

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns to spot anomalies to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. We performed some analysis on the data to get a better overview of the data and to find outliers in our data-set. Outliers can occur due to some kind of errors while collecting the data and need to be removed so that it affect the performance of our model.

4. Data Cleaning

5. Feature Engineering

6. Outlier Removal using Standard Deviation & Mean

7. Building Model

8. Test the Model for few properties

5.2 Tools used

- Anaconda Prompt
- Jupyter Notebook
- Flask
- Pycharm

5.3 Technologies used

- Python
- HTML
- CSS

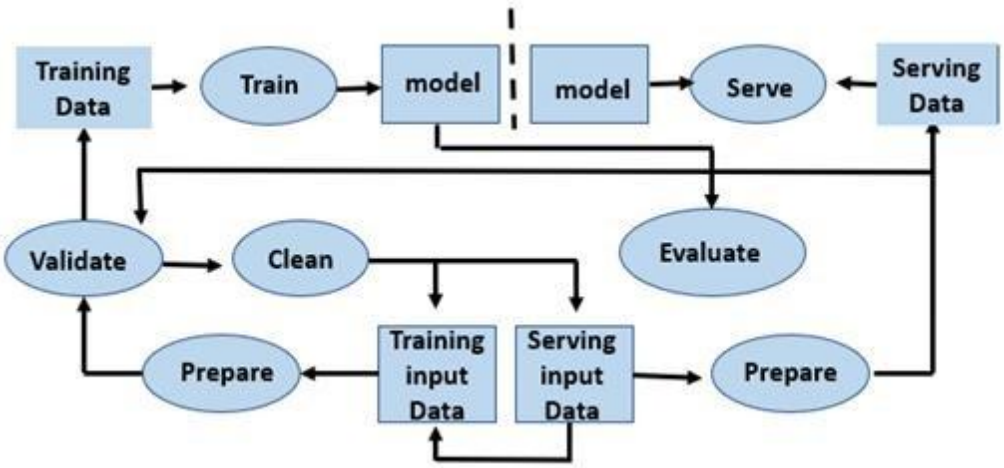

Figure 5.1: Machine Learning Life Cycle

The frontend has been developed using Flask in Python , HTML and CSS for designinga web page and it will open in a local host server in web browser. Jupyter notebook is used to model and analyze the data for prediction.

6. RESULTS

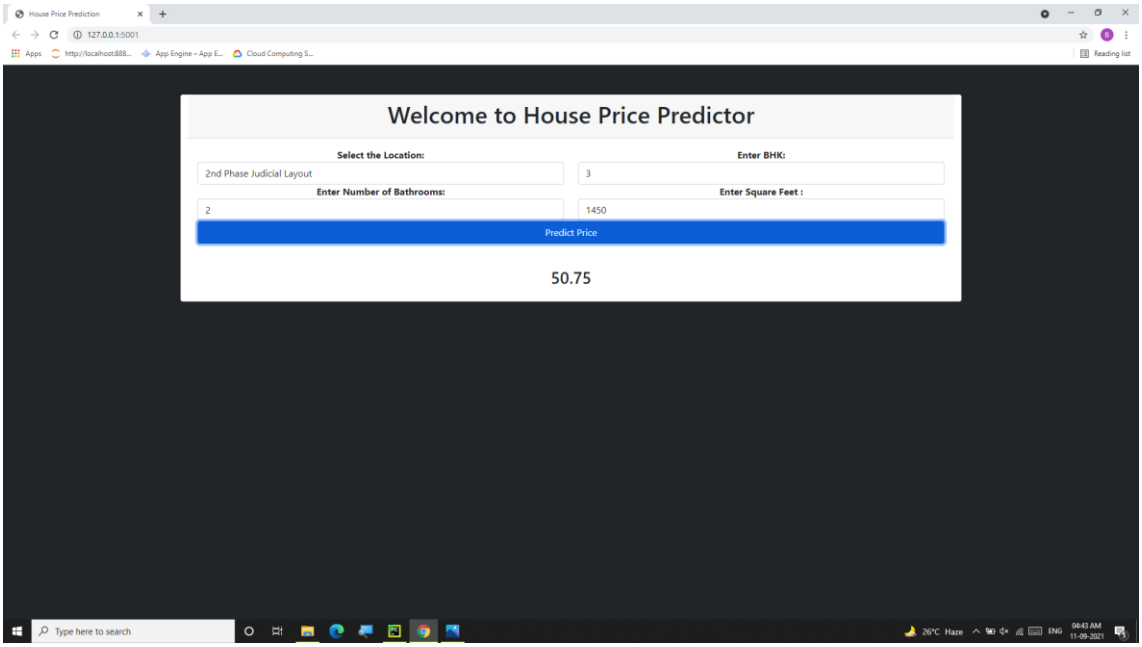

Figure 6.1: House Price Prediction Result

7. CONCLUSION

The goal is to achieve the system which will reduce the human effort to find a house having reasonable price. Proposed system focused on predict the house price according to the area for that image processing and machine learning methods are used. The experimental results showed that this technique that are used while developing system will give accurate prediction of house price.

Buying your own house is what every human wish for. Using this proposed model, we want people to buy houses and real estate at their rightful prices and want to ensure that they don't get tricked by sketchy agents who just are after their money. Additionally, this model will also help Big companies by giving accurate predictions for them to set the pricing and save them from a lot of hassle and save a lot of precious time and money. Correct real estate prices are the essence of the market and we want to ensure that by using this model.

The system is apt enough in training itself and in predicting the prices from the raw data provided to it. After going through several research papers and numerous blogs and articles, a set of algorithms were selected which were suitable in applying on both the datasets of the model. After multiple testing and training sessions, it was determined that the XGBoost Algorithm showed the best result amongst the rest of the algorithms. The system was potent enough for Predicting the prices of different houses with various features and was able to handle large sums of data. The system is quite user-friendly and time-saving.

REFERENCES

[1] Model “BANGALORE HOUSE PRICE PREDICTION MODEL”

[2] Fan C, Cui Z, Zhong X. House Prices Prediction with Machine Learning Algorithms. Proceedings of the 2018 10th International Conference on Machine Learning and Computing ICMLC 2018. doi:10.1145/3195106.3195133.

[3] Supriya Rajankar, Neha sakhrakar and Omprakash rajankar “House price Prediction using machine learning algorithms” International journal of Engineering Research and Technology (IJERT) (June 2019).

[4] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al.* Scikit-learn: Machine Learning in Python The Journal of Machine Learning Research, 12 (2011), pp. 2825-2830

[5] Raschka S, Mirjalili V. Python machine learning: Machine learning and deep learning with Python, scikit-learn, and Tensor Flow (2nd ed.), Packt Publishing, Birmingham (2017).

[6] Varma, A. Sarma, 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.

[7] Furia, Palak, and Anand Khandare. "Real Estate Price Prediction Using Machine Learning Algorithm." eConference on Data Science and Intelligent Computing. 2020.

[8] Musciano, Chuck, and Bill Kennedy. HTML & XHTML: The Definitive Guide: The Definitive Guide. " O'Reilly Media, Inc.", 2002.

[9] Aggarwal, Shalabh. Flask framework cookbook. Packt Publishing Ltd, 2014.

[10] Grinberg, Miguel. Flask web development: developing web applications with python.

APPENDIX A : SAMPLE CODE

```
1. from flask import Flask, render_template, request, jsonify
2. import pandas as pd
3. import pickle
4. import numpy as np
5. app = Flask(__name__)
6. @app.route('/')
7. def index():
8.data =pd.read_csv('C:/Users/HP/PycharmProject/pythonProject/Cleaned_Data.csv')
9. locations = sorted(data['location'])
10. return render_template('index.html', locations=locations)
11. @app.route('/predict', methods=['POST'])
12. def predict():
13. pipe = pickle.load(open("RidgeModel.pkl", "rb")
14. data = request.json
15.  print('Python Coming')
16. location = data['location']
17. print(location)
18. bhk = data['bhk']
19. print(bhk)
20.  bath = data['bath']
```

```
21.  print(bath)
22.  total_sqft = data['total_sqft']
23.  print(total_sqft)
24.  data = pd.read_csv('C:/Users/HP/PycharmProject/pythonProject/Cleaned_Data.csv')
25. columns location, bath, bhk,  total_sqft
26.  response = data.loc['location: location', 'bhk': bhk]
27.  locationData = data[(data['location'] == location) & (data['bhk'] == int(bhk)) &
28. (data['bath'] == int(bath)) & (data['total_sqft'] == float(total_sqft))];
29. print(locationData)
30. locationCount = len(locationData)
31.  if locationCount > 0:
32.  price = locationData['price']
33. print(price);
34. return str(price)
35. else:
36. return 'No Price Matching Found'
37. if __name__=="__main__":
38.  app.run(debug=True, port=5001);
```

APPENDIX B : SCREENSHOTS