HOME CONSTRUCTION COST ESTIMATION USING ML

##### Mrs. Suchetha N V

**Abstract**

Home construction cost estimation is a complex and time-consuming process that has traditionally relied on expert judgment and experience. However, recent advances in machine learning (ML) have made accurate cost prediction possible by analyzing historical data on similar projects. By training ML models on large datasets, accurate cost predictions can be made in a fraction of the time it would take for a human estimator. Additionally, ML algorithms can continuously learn and adapt to new data, making them more accurate over time. ML-based cost estimation can help construction companies optimize their resources and plan their projects more efficiently, leading to improved accuracy, efficiency, and profitability in the construction industry. Additionally, ML can be used to generate home samples and provide location-based insights, allowing for more informed decisions about project planning and execution. By leveraging ML for cost prediction, sample generation, and location-based insights, the construction industry can improve accuracy, efficiency, and profitability.

# Introduction

## Chapter 1

## Project Introduction

The construction industry has long relied on expert judgment and experience to estimate the costs associated with building homes. However, this traditional approach is time- consuming and often inaccurate, leading to costly overruns and delays. As a result, there has been a growing interest in leveraging machine learning (ML) to improve the accuracy and efficiency of cost estimation in the construction industry.

The aim of this project is to develop an ML model that can accurately predict home construction costs. To achieve this, we will leverage historical data on similar projects to identify patterns and insights that can be used to inform our cost predictions. This data will include information on the location of the home, the size of the project, the materials used, and the labor required, among other factors. By analyzing this data, we will train our ML model to make accurate predictions about the costs associated with building a new home.

In addition to cost prediction, our project will also explore the use of ML for generating home samples and providing location-based insights. By generating home samples, we can provide builders and homeowners with a better understanding of what their finished project may look like. Similarly, by providing location-based insights, we can help builders and homeowners make informed decisions about the best location for their project, taking into account factors such as local zoning laws, environmental considerations, and community preferences. Ultimately, our project aims to demonstrate the potential of ML in improving accuracy, efficiency, and profitability in the construction industry.

## Problem Description

The problem is how to develop an accurate and efficient method for home construction cost estimation that overcomes challenges such as reliance on expert judgment and experience, lack of standardized data, and the need for location-based insights and home samples. To address this problem, we will leverage machine learning to analyze historical data on similar projects and identify patterns and insights that can inform cost predictions.

We will also explore the use of ML for generating home samples and providing location- based insights to help builders and homeowners make informed decisions about their projects. Ultimately, our goal is to develop an ML model that improves the accuracy, efficiency, and profitability of home construction cost estimation.

**Literature Review**

## Chapter 2

## Literature Survey

The Decision Tree Regressor and Random Forest Regressor both had the best accuracy, according to the authors [1] Thuraiya Mohd and Suraya Masrom. Ridge and Linear Regression produce a comparable outcome with a very tiny decrease in Lasso. There is no significant difference between any of the feature selection groups, strong or weak, across the board. It is a positive indication that model over-fitting can be reduced when only the purchase prices are used to predict the selling prices. For all feature selections, the Root Square Mean Error (RMSE) shows the same pattern of outcomes.

Sayan Putatunda et.al. [2], employed machine learning techniques to forecast prices by utilizing attribute variables. They applied sophisticated machine learning algorithms, including Random Forest, Gradient Boosting, and Artificial Neural Networks, to a real-world construction dataset and conducted a comparative analysis of these methodologies. Their findings revealed that the Random Forest method outperformed the others in terms of predictive accuracy.

The authors [3] B.Balakumar and P.Ravira used the machine learning algorithms to predict houseprices. These feature sets were then given as an input to four algorithms, and they have used linear regression algorithm. They done analysis on the Bengaluru house dataset and in that dataset they focus on those features on which buyers are interested to find out their fit one.

The authors [4] M Thamaraj and S P Malarvizhi in IEEE in 2020. The study presents an approach to predicting house prices using machine learning techniques, with a focus on the city of Chennai in India. To develop their model, the authors collected data from multiple sources, including property websites, real estate agents, and government websites. They used various feature selection techniques to determine the most relevant features for predicting house prices. The authors proposed a novel machine learning algorithm called Boosted Regression Tree (BRT) for predicting house prices. They compared the performance of BRT with other popular machine learning algorithms, including linear regression, decision tree, and random forest. The results of the study

showed that the BRT algorithm outperformed other algorithms in terms of accuracy and robustness. The authors also conducted a detailed analysis of the important features affecting house prices and their impact on the prediction.

The authors [5] Ismail Ibrahim presents a study on predicting house prices in Malaysia using two machine learning models, Random Forest and Extreme Gradient Boosting. The author compared the performance of the two models in terms of accuracy and robustness, and the results showed that the XGBoost model outperformed the RF model. The paper concludes that the XGBoost model is a more effective approach for predicting house prices and provides valuable insights for the real estate industry. The findings of the study can be used to develop a pricing strategy for properties in Malaysia.

## Comparative Analysis of the Related Work

The table 2.1 discusses the comparative analysis of the current systems in light of the suggested proposal.

**Table 2.1 Comparative analysis**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sl. No.** | **Author(s)** | **Algorithms/Techniques** | **Performance**  **Measures** |
| **1.** | Machine Learning Housing Price Prediction in Petaling Jaya, Selangor, Malaysia | Support Vector Regression (SVR), Random Forest Regression (RFR), Artificial Neural Network (ANN) | Accuracy |
| **2.** | Using machine learning algorithms for housing price prediction: The case of Fairfax County,  Virginia housing data | Random Forest Regression (RFR), Artificial Neural Network (ANN) | Accuracy |
| **3.** | Vision-based real estate price estimation | Computer vision techniques, convolutional neural network (CNN), regression models | Accuracy |

## Chapter 3

**Problem Formulation**

## Problem Statement

In India, there are several real estate classified websites that list properties for construction purposes, including popular ones like 99acres, No Broker, Housing, Magic Bricks, and many others. However, these platforms often exhibit discrepancies in property pricing, with instances where similar properties are priced differently. This lack of uniformity raises concerns about transparency and accuracy. Customers may sometimes question whether a listed house's value is justified, but they lack a means to verify the accuracy of the data.

Accurate property evaluations and fair pricing can significantly enhance transparency and trust within the real estate sector. This is particularly vital in India, where property transactions often involve substantial sums. Addressing this concern will ultimately benefit both customers and the real estate industry over the long term.

## Objectives of the Present Study

The objectives of the proposed project are as follows:

* + 1. To provide real time construction cost estimation.
    2. To provide home samples based on estimated construction price.
    3. To Provide top view of sight location using global positioning system (GPS).
    4. To Provide chat bot for human computer interaction.

## Summary

The best solution for home construction cost estimation is using machine learning techniques. The regression algorithms are more accurate when compared to traditional regression techniques. Developing an early home construction cost estimation system can be useful for many clients as well as buyers by helping them to know the estimated price of the home for a particular location in Bangalore.

## Chapter 4

**Requirements and Methodology**

## Hardware Requirements

The hardware requirements for the proposed project are depicted in Table 4.1.

**Table 4.1: Hardware requirements**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Hardware/Equipment** | **Specification** |
| 1. | Graphics Card | Intel 621 Graphics card or 2GB |
| 2. | RAM | 4GB or above |

## Software Requirements

The software requirements for the proposed project are depicted in Table 4.2.

**Table 4.2: Software requirements**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Software** | **Specification** |
| 1. | Visual Studio Code | 1.6 GHz or faster processor 1GB of RAM |
| 2. | React | V16.13.00 |
| 3. | db.sqlite3 | 3.41.2 |
| 4. | Python | 3.9.16 |
| 5. | Django | 3.2.16 |

## Methodology Used

The proposed home cost estimation system is implemented using the following steps:

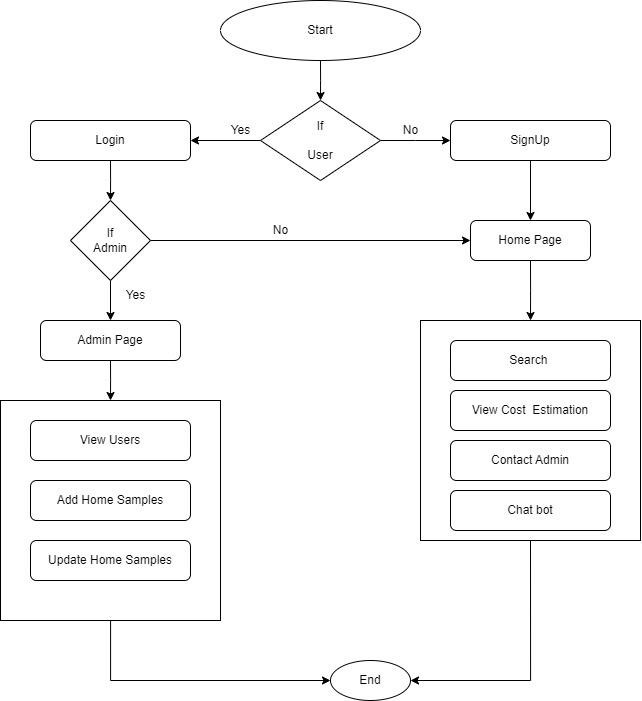
1. **Data Collection:** To collect data from publicly available sources.
2. **Data pre-processing:** This may involve removing duplicates, filling in missing values, and converting data into a format that is compatible with machine learning algorithm.
3. **Model Development:** This may involve selecting an appropriate algorithm and splitting the data into training and testing sets.
4. **Front-End Development:** This may include creating a user interface using ReactJS that allows users to input data.
5. **Back-End Development:** This will involve creating APIs using Django Rest Framework that allow the front-end to communicate with the machine learning model and retrieve predictions based on user input.
6. **Location-Based Search:** This will allow users to input their location and view properties by integrate mapping API such as Google Maps into the front-end.
7. **Chatbot Integration:** This will allow users to ask questions about home construction costs and receive responses from the chatbot.

## Chapter 5

**System Design**

## Architecture of the Proposed System

Figure 5.1 shows the architecture of the proposed system.



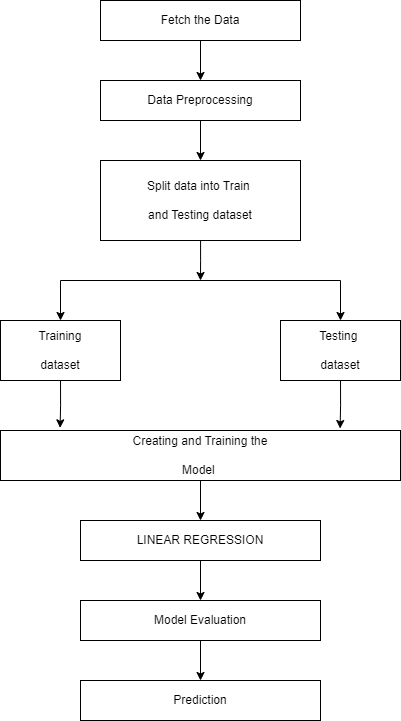
**Figure 5.1: Architecture of the proposed system**

The website follows a standard user flow. Initially, users need to register on the website. Upon successful registration, they are redirected to the login page, where they enter their credentials. If the credentials match, the user is then directed to the home page. The home page features a chat bot for queries and a cost estimation page. The chat bot assists in predicting the price of a home by requesting inputs related to the property. Once the user clicks the price estimation button, they can view the estimated price based on the given inputs, such as the location on a map and displayed home samples. On the admin page, the admin can access user and price details for homes.

The website is built using ReactJS for the frontend and Django Rest Framework for the backend. The backend integrates a machine learning model, which is embedded in Django. Through a REST API, the predicted cost is sent to the frontend, allowing users to view the estimated price of the house.

## System Flowchart

A system flowchart is a way of depicting how data flows in a system and how decisions are made to control events. Figure 5.2 depicts the system flowchart.



**Figure 5.2: System flowchart**

The system flow of our Machine Learning model, depicting the sequential progression of data throughout the training and prediction processes. Within the realm of Machine Learning, data flow pertains to the systematic passage of data through a model. The data flow commences with data pre-processing, encompassing activities such as data cleansing, normalization, and transformation, aiming to prepare the raw data for compatibility with the machine learning model. Subsequently, the preprocessed data is partitioned into distinct datasets: training, validation, and testing. These datasets are leveraged to train the model and assess its performance. To ensure the accuracy and efficacy of a predictive model, it is a common practice to partition the available data into two distinct sets: a training set and a test set. The training set is utilized to train the model, where input variables are coupled with their corresponding variable, which target the model endeavors to predict. In the case of regression training for the model on models, two widely the training set adopted algorithms are the XGBoost regressor and linear regression.

## Chapter 6

**System Testing, Results and Discussion**

## System Testing

**Table 6.1: Unit test cases**

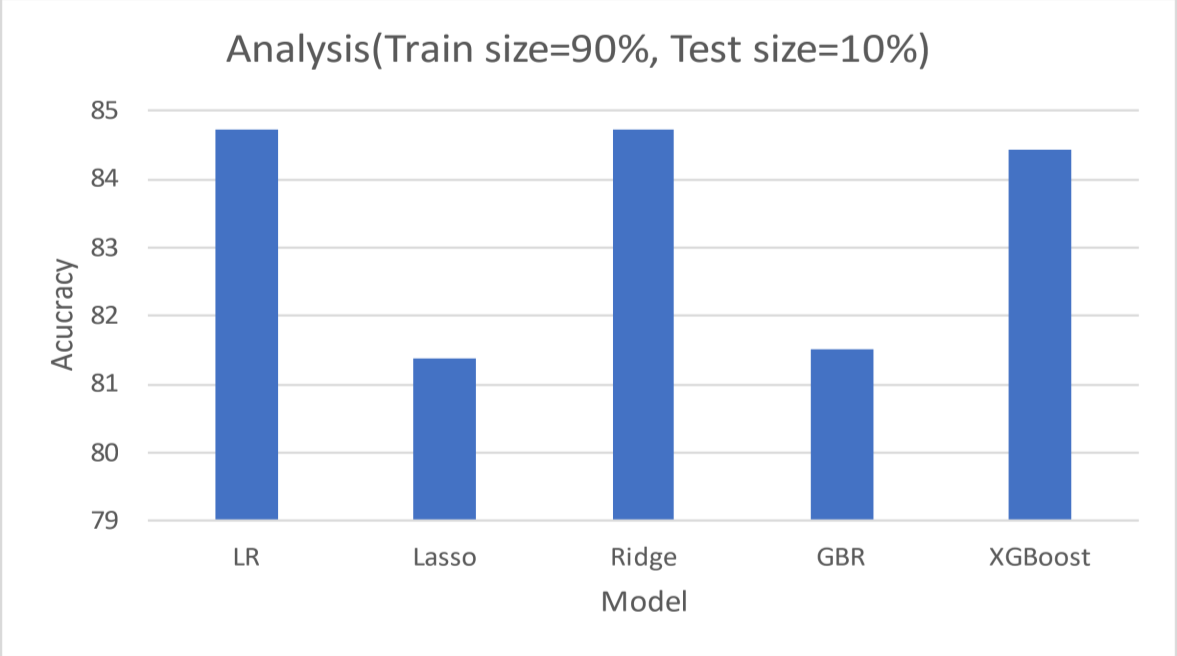
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Test case number** | **Input** | **Stage** | **Expected behavior** | **Observed behavior** | **Status P=Pass**  **F=Fail** |
| 1 | Signup credentials | Signup page | If details valid, the signup is  successful | As expected | P |
| 2 | Login credentials | Login page | If details matched, the input page  appear | As expected | P |
| 3 | Enter input values from the test set | Input page | The result should appear as Estimated  price | As expected | P |

## Result Analysis

The main aim of the project was to predict the home cost estimation using machine learning algorithms. Table 6.2 shows the analysis that was performed on the five algorithms they are Linear Regression(LR), Lasso Regression(Lasso), Ridge Regression(Ridge), Gradient Boosting Regression(GBR), XGBoost Regression(XGBoost) with different training and testing sizes. It was found that Linear Regression(LR) was the most accurate in all the cases.

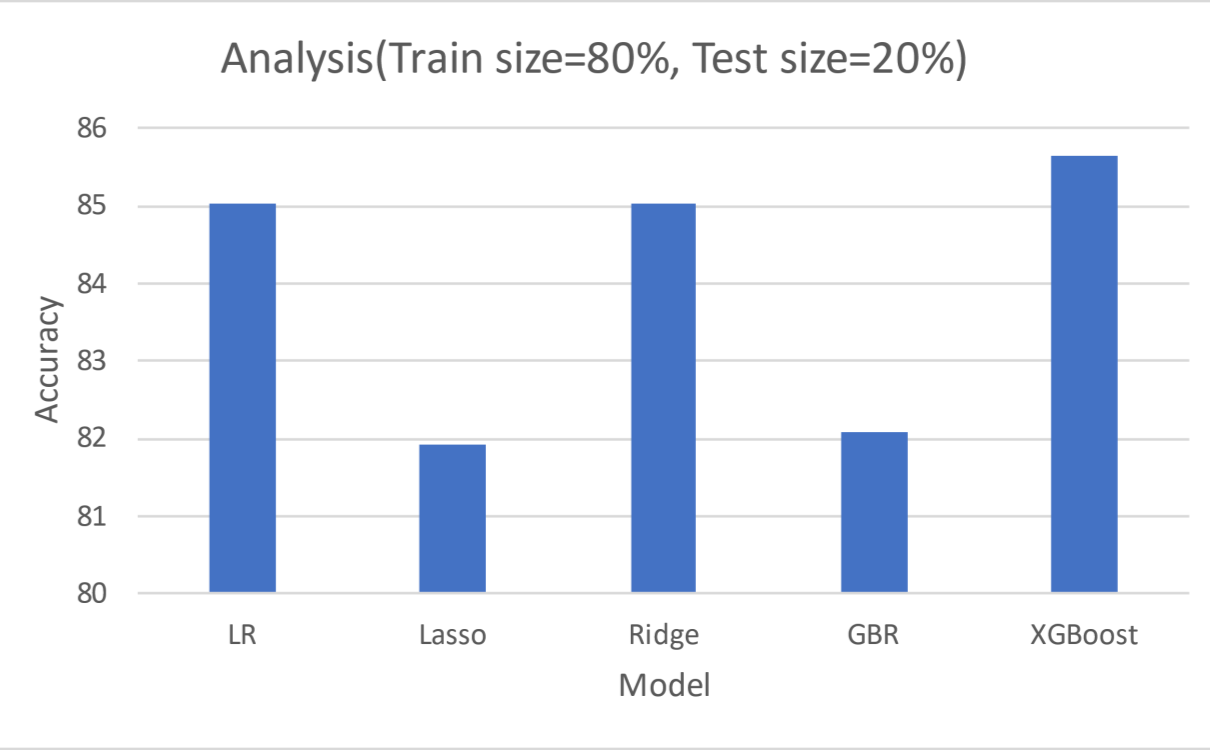
**Table 6.2: Analysis of the five algorithms**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Test Case** | **Training Size** | **Testing Size** | **Accuracy (%)** | | | | |
| **LR** | **Lasso** | **Ridge** | **GBR** | **XGBoost** |
| 1 | 90% | 10% | 84.74 | 81.38 | 84.74 | 81.51 | 84.42 |
| 2 | 80% | 20% | 85.04 | 81.93 | 85.04 | 82.10 | 85.65 |
| 3 | 70% | 30% | 82.97 | 80.29 | 82.97 | 76.65 | 78.77 |
| 4 | 60% | 40% | 85.10 | 81.99 | 85.15 | 83.36 | 84.76 |

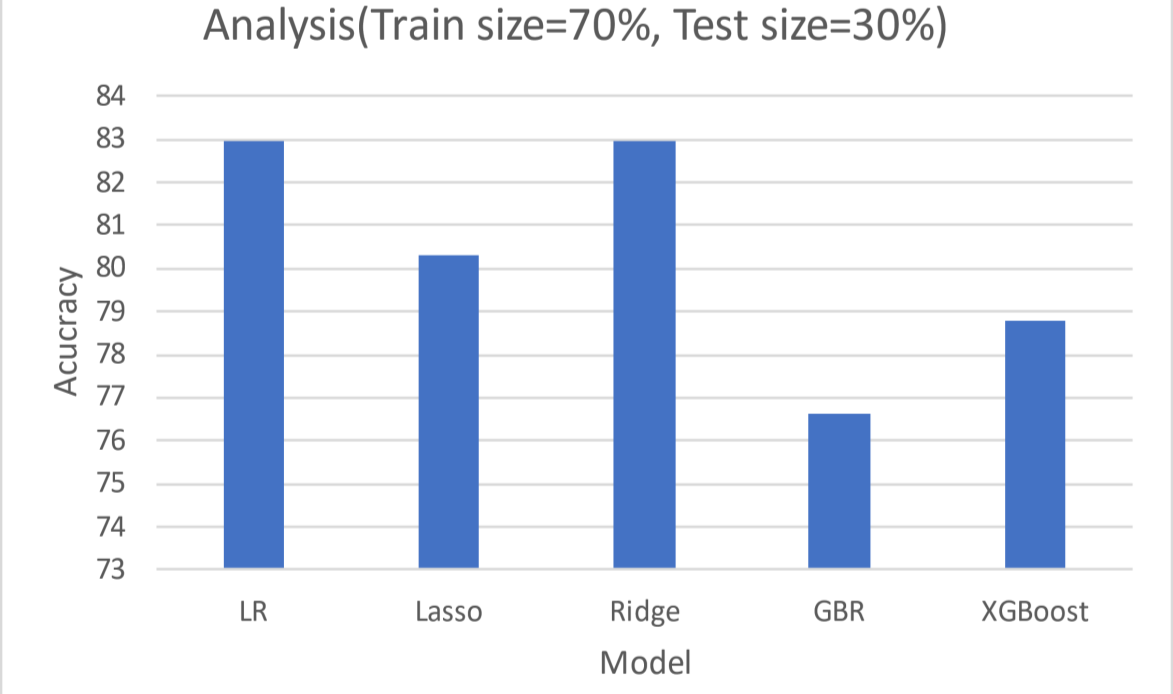
Figure 6.1 shows the bar graph for the accuracy of the five algorithms where the train set size was 90% and the test set size was 10%.

**Figure 6.1: Graph analysis of the test case 1**

Figure 6.2 shows the bar graph for the accuracy of the five algorithms where the train set size was 80% and the test set size was 20%.

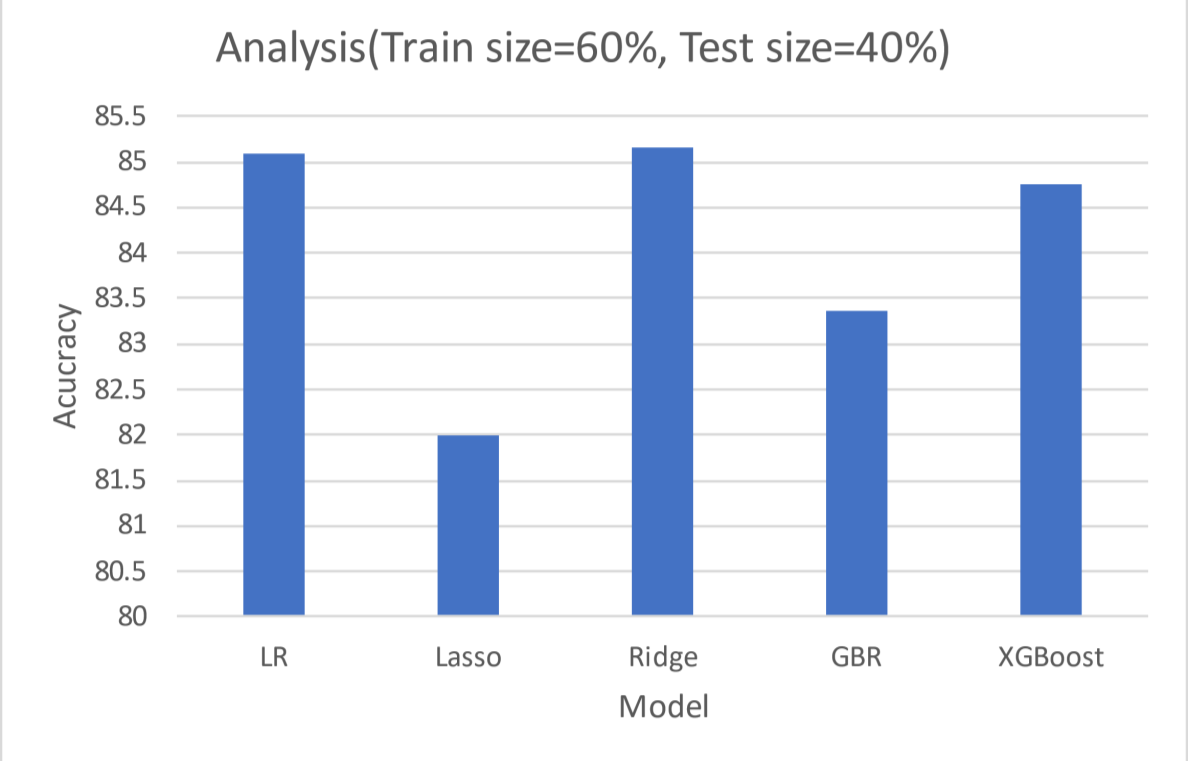


**Figure 6.2: Graph analysis of the test case 2**

Figure 6.3 shows the bar graph for the accuracy of the five algorithms where the train set size was 70% and the test set size was 30%.

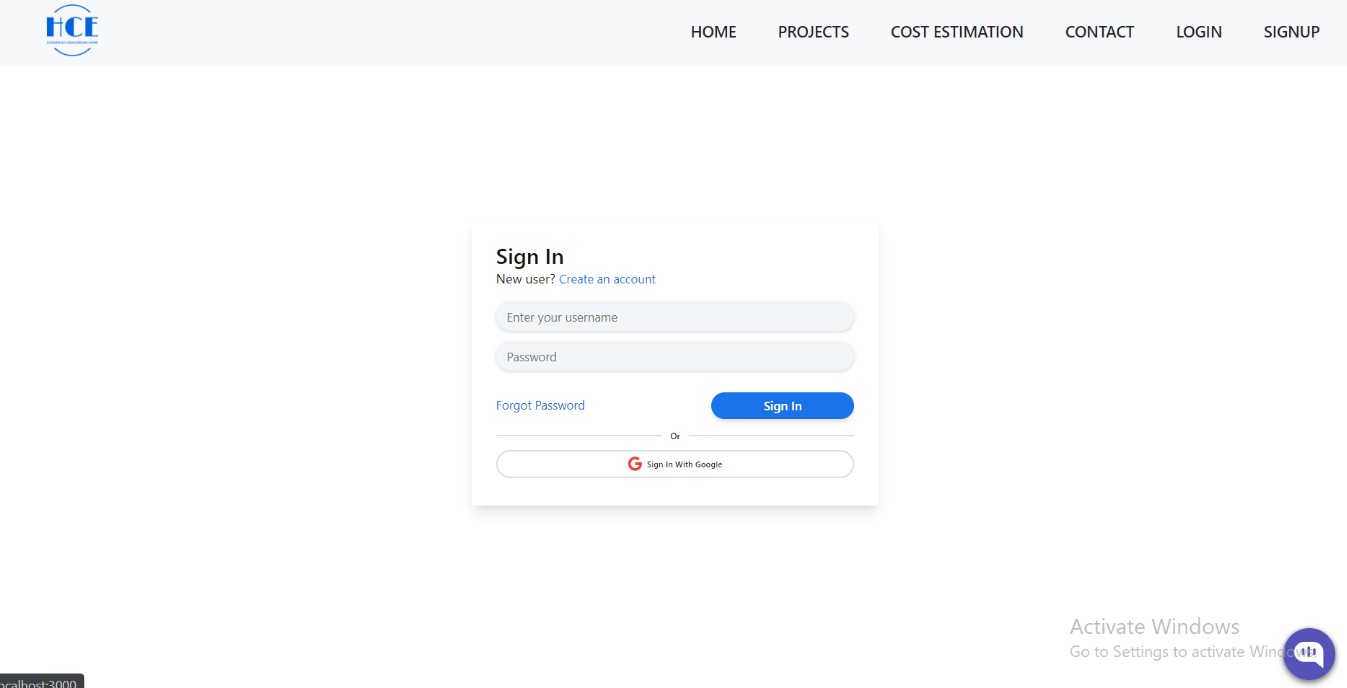
**Figure 6.3: Graph analysis of the test case 3**

Figure 6.4 shows the bar graph for the accuracy of the five algorithms where the train set size was 60% and the test set size was 40%.



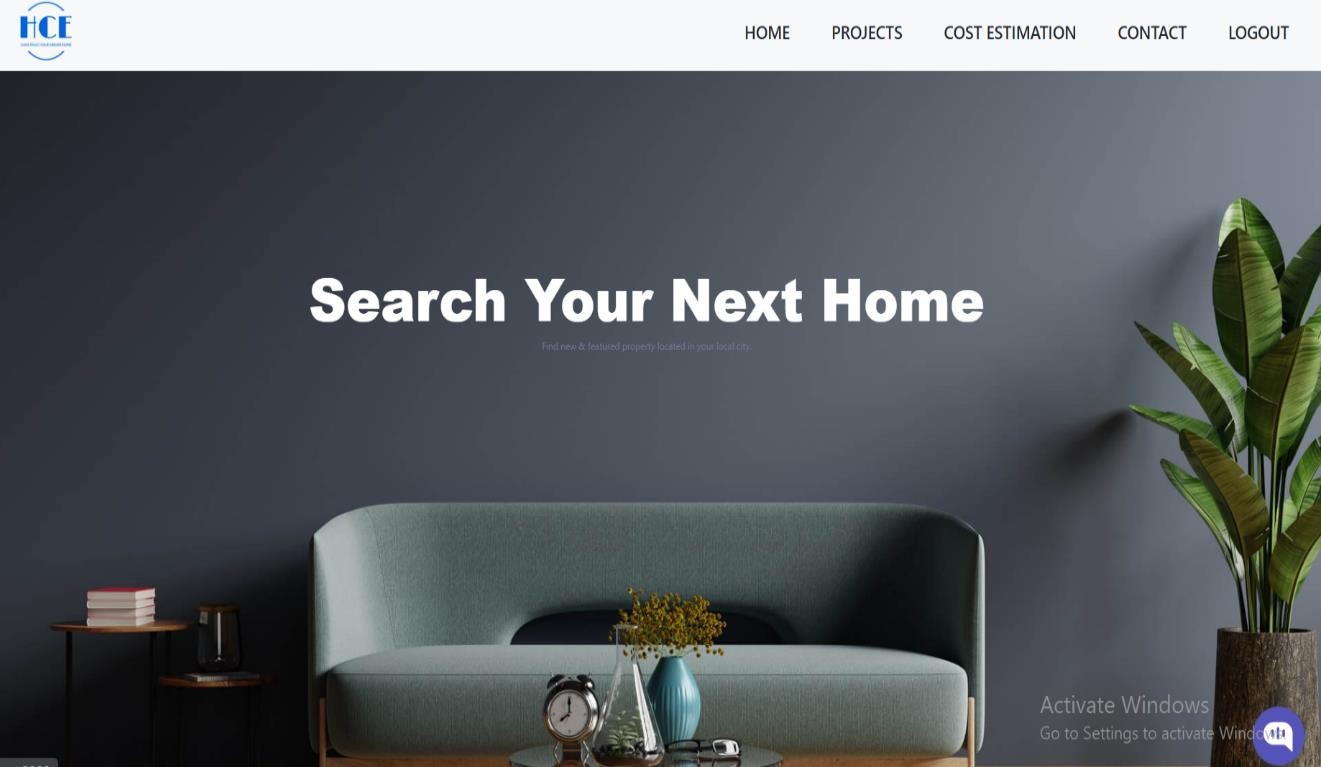
**Figure 6.4: Graph analysis of the test case 4**

Figure 6.5 is the login page for the users who use this application.



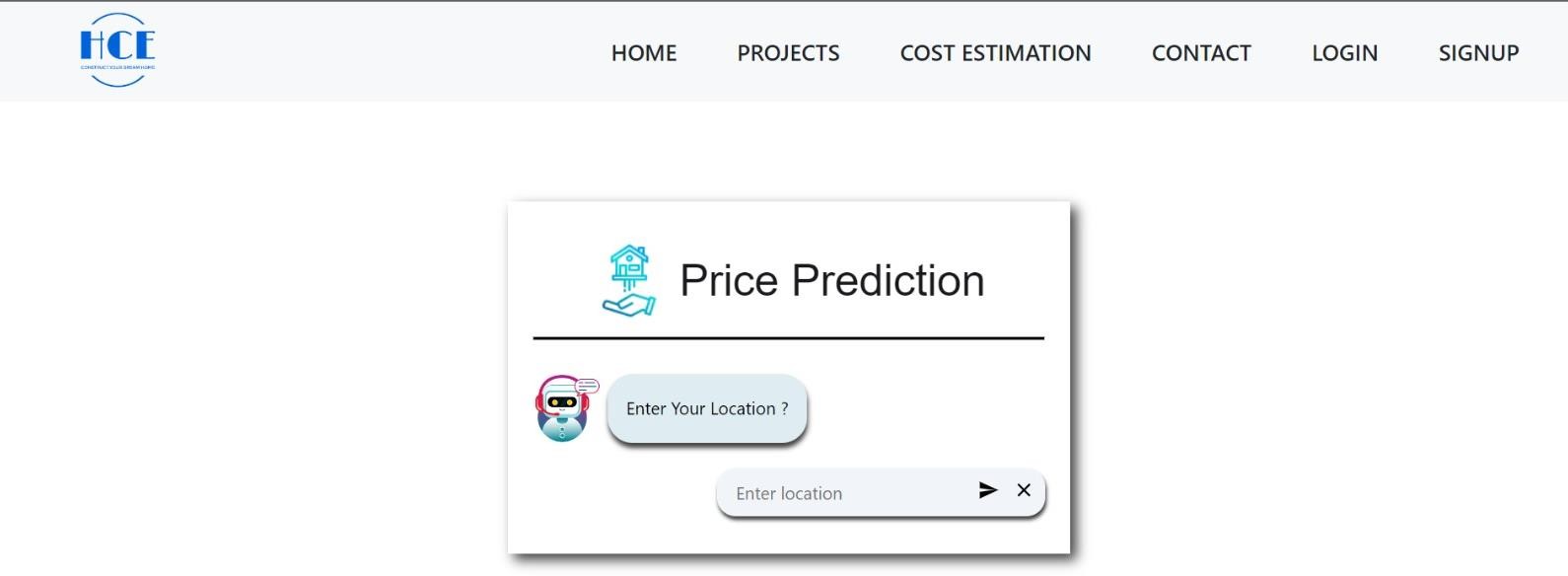
**Figure 6.5: Login page**

Figure 6.6 is the home page Of Home construction cost estimation using ML.



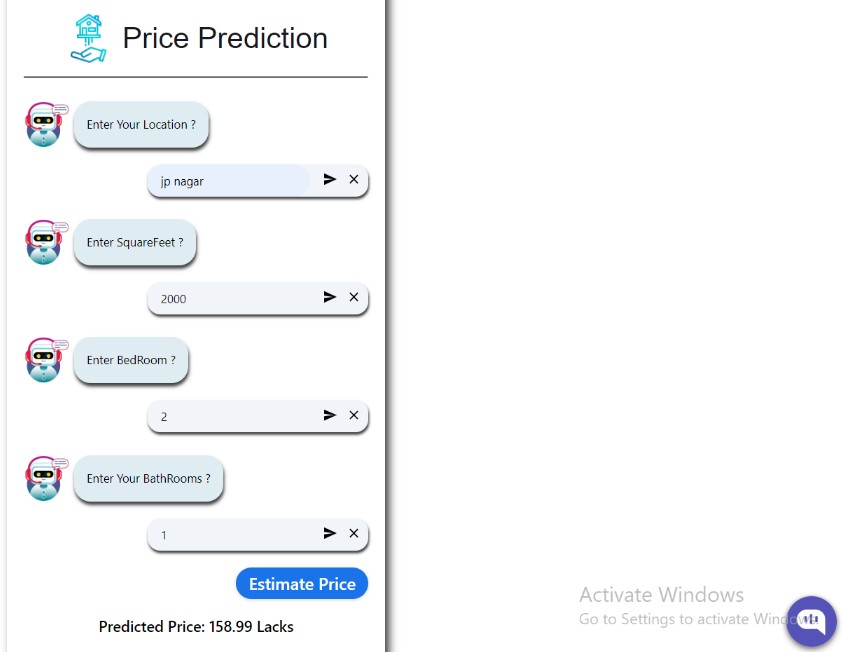
**Figure 6.6: Home page**

Figure 6.7 is the input page. Here, the user will enter the parameters. Four features are present. These features are the ones that are responsible for the result.



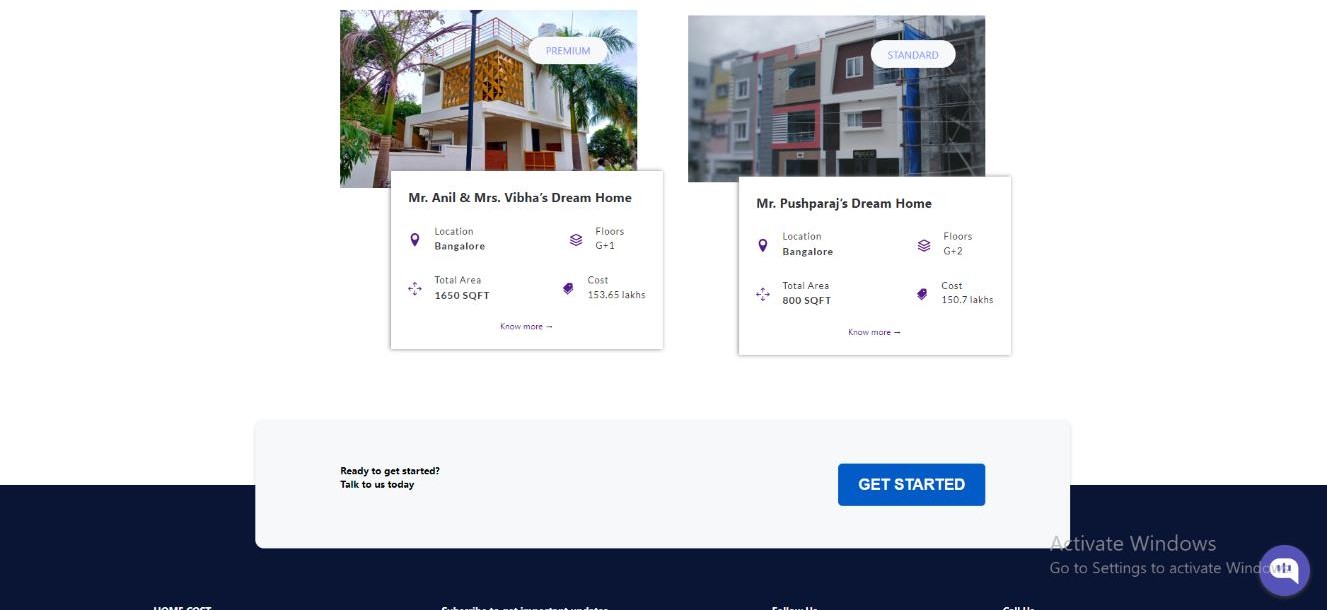
**Figure 6.7: Input page**

Figure 6.8 is the page where the result is obtained as Estimated price.



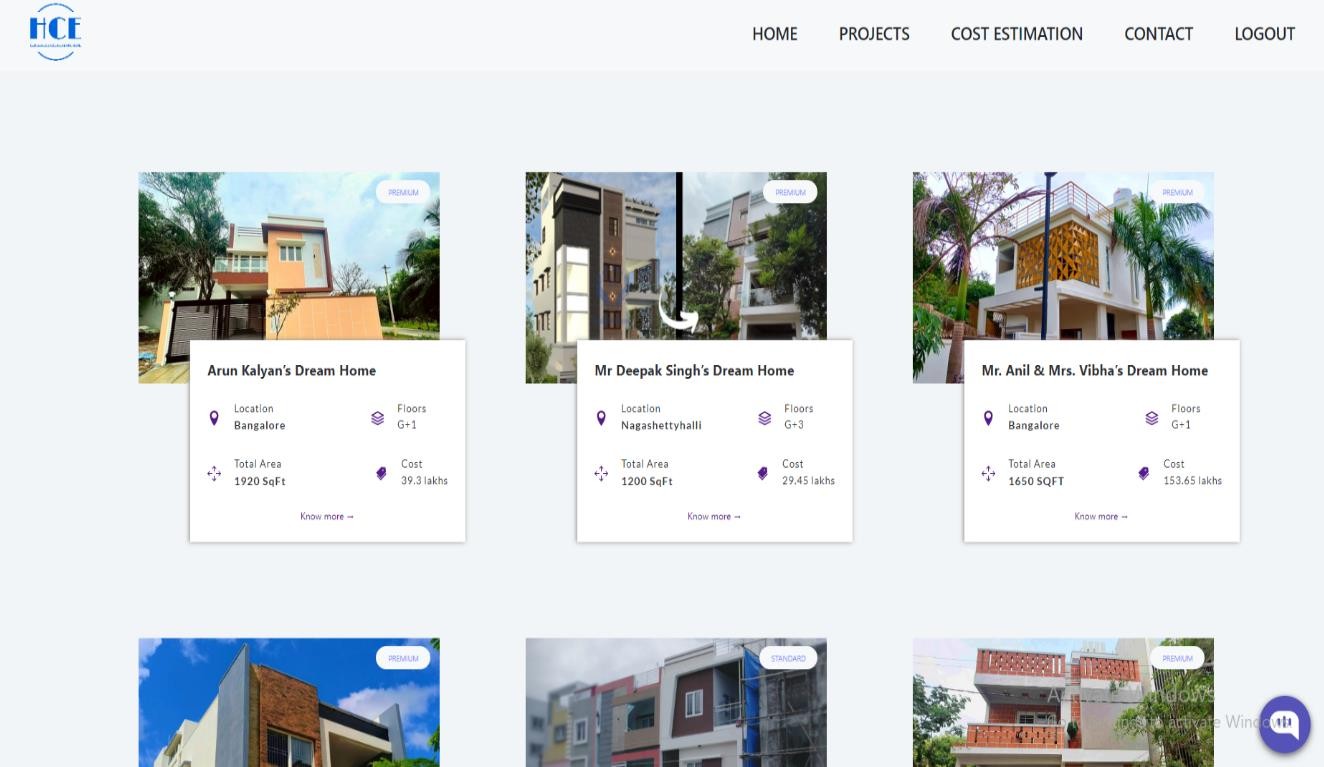
**Figure 6.8: Result value as estimated price**

Figure 6.9 is the result page where users can see home samples.



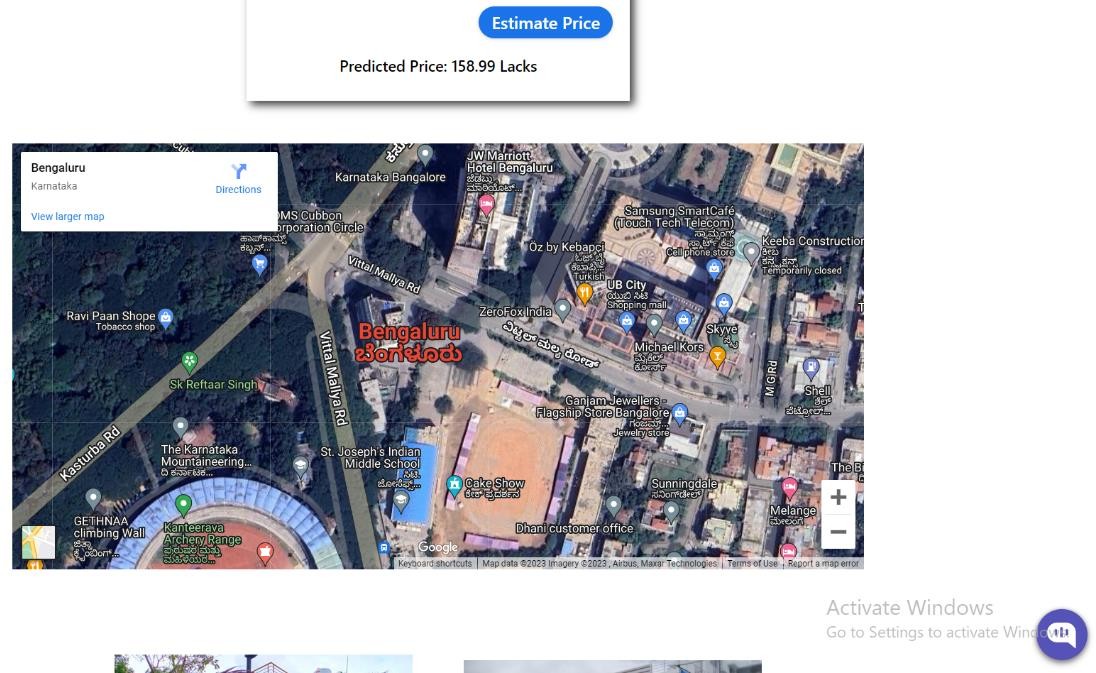
**Figure 6.9: Result home samples**

Figure 6.10 shows the projects page where the recently done projects displaced.



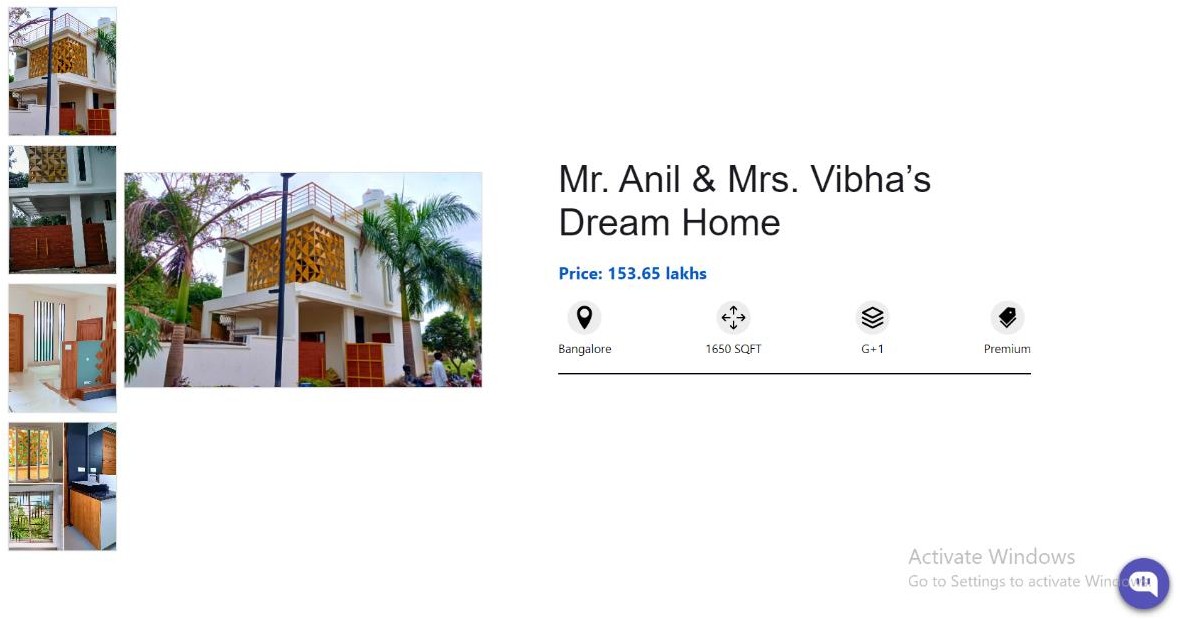
**Figure 6.10: Projects page**

Figure 6.11 shows the Location based input page where user can see the location based on entered location in input page.

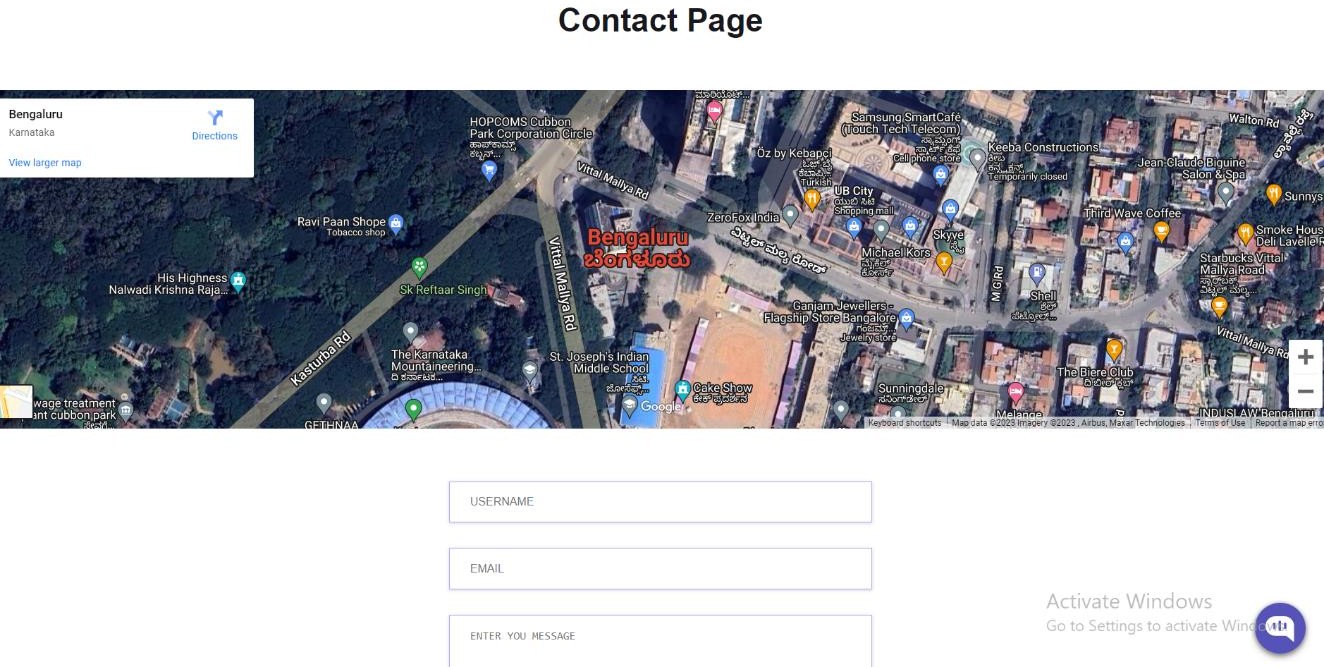


**Figure 6.11: Location based on input page**

Figure 6.12 shows the description of the home samples shown in the projects.



**Figure 6.12: Description about home samples**

Figure 6.13 shows the contact page where user can write opinions about the web page or user can ask query.

**Figure 6.13: Contact page**

## Summary

The application was developed using the Django framework. The programming languages that were used were Python, React, and Sqlite3. The figures in the previous section showed the snapshots of various pages of the application. Since Linear Regression was found to be the most accurate among the five algorithms, the prediction model was created using it.

## Chapter 7

**Conclusion and Scope for Future Work**

## 7.1 Conclusion

In recent years, machine learning (ML) techniques have shown great potential in home cost estimation. By taking into account numerous factors such as location, size, age, and features of a property, ML models can generate more accurate predictions than traditional methods. Linear regression, decision trees, and neural networks are some of the algorithms that have been explored for home cost estimation using ML. The choice of algorithm should depend on the specific requirements of the problem, as each algorithm has its strengths and weaknesses. Data quality and availability are crucial factors that can impact the performance of ML models. Effective data pre-processing and feature engineering techniques can help to extract essential information from raw data and enhance the model's accuracy. Utilizing ML for home cost estimation has the potential to revolutionize the real estate industry by providing more reliable and accurate estimates. However, further research is necessary to address some of the challenges such as data quality and model interpretability. In conclusion, the use of ML for home cost estimation holds great promise and has the potential to significantly improve the accuracy of cost estimates in the real estate industry.

## Scope for Future Work

In the field of home cost estimation using machine learning, there are several potential areas for future research and development. These areas include integrating additional data sources such as economic indicators and demographic data, developing more interpretable models to increase trust and confidence in machine learning-based models, exploring different machine learning techniques such as ensemble methods and deep learning, incorporating temporal data to capture trends and changes in the market, and evaluating the impact of machine learning on the real estate industry. By addressing these areas, researchers and real estate professionals can improve the accuracy of home cost estimation models and gain a better understanding of how machine learning will impact the industry in the future. These advancements have the potential to revolutionize the way home cost estimation is performed and provide more reliable and accurate cost estimates, ultimately benefiting the real estate industry as a whole.

**References**

1. Thuraiya Mohd, Suraya Masrom, Noraini Johari, "Machine Learning Housing Price Prediction in Petaling Jaya, Selangor, Malaysia ", International Journal of Recent Technology and Engineering (IJRTE), Volume-8, Issue-2S11, 2019.
2. Sayan Putatunda, "PropTech for Proactive Pricing of Houses in Classified Advertisements in the Indian Real Estate Market".
3. B. Balakumar, P.Raviraj, S.Essakkiammal , "Predicting Housing Prices using Machine Learning Techniques".
4. M Thamarai, S P Malarvizhi, " House Price Prediction Modeling Using Machine Learning", International Journal of Information Engineering and Electronic Business (DJIEEB), VoL12, No.2, pp.15- 20, 2020. DOI: 10.5815/ijieeb.2020.02.03.
5. Zulkifley, Nor & Rahman, Shuzlina & Nor Hasbiah, Ubaidullah & Ibrahim, Ismail. (2020). House Price Prediction using a Machine Learning Model: A Survey of Literature. International Journal of Modern Education and Computer Science. 12. 46- 54. 10.5815/ijmecs.2020.06.04.
6. Kuvalekar, Alisha and Manchewar, Shivani and Mahadik, Sidhika and Jawale, Shila, House Price Forecasting Using Machine Learning (April 8, 2020). Proceedings of the 3rd International Conference on Advances in Science & Technology(ICAST)2020.
7. Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh - “A hybrid regression technique for house prices prediction” 2017, IEEE.
8. Akshay Babu, Dr. Anjana S Chandran, "Literature Review on Real Estate Value Prediction Using Machine Learning”, International Journal of Computer Science and Mobile Applications, Vol: 7 Issue: 3, 2019
9. M. Venkataraman, V. Panchapagesan, and E. Jalan, ‘‘Does Internet search intensity predict house prices in emerging markets? A case of India,’’ Property Manage., vol. 36, no. 1, pp. 103–118, Feb. 2018, doi: 10.1108/PM-01- 2017-0003.

[10] C. R. Madhuri, G. Anuradha, and M.V. Pujitha, ‘‘House price prediction using regression techniques: A comparative study,’’ in Proc. Int. Conf. Smart Struct. Syst. (ICSSS), Mar. 2019, pp. 1–5.