

Premier University, Chattagram



Project Report

On

“House Price Prediction Using Machine Learning”

Submitted By

Shafayet Ullah Ramim ID: 2104010202219

Shihabul Alam Sakib ID: 2104010202221

Khalid Ahamed Rahi ID: 2104010202202

In Partial Fulfillment of the Requirements for the Degree of

Bachelor of Science in Computer Science & Engineering

Under the Supervision of

Ms. Adiba Ibnat Hossain

Lecturer

September 2024

Contents

Abstract	3
1 Introduction	4
1.1 Conceptual Study of the Project	4
1.2 Objectives of the Project	4
1.3 Difficulties	5
1.4 Related Work	5
2 Literature Review	6
3 Methodology	7
3.1 Diagram for House Price Prediction	7
3.2 Data Collection	7
3.3 Data Exploration	7
3.4 Data Description	8
3.5 Data Visualization	8
3.6 Data Selection	9
3.7 Data Transformation	9
4 Implementation	11
4.1 Language	11
4.1.1 Python	11
4.2 Models	11
4.2.1 Linear Regression Model	11
4.2.2 Random Forest Regression Model	12
4.2.3 XGBoost Regression Model	12
4.3 Source Code	12
5 Result Analysis	14
5.1 Output	14
5.2 Output Analysis	15
6 Future Work	17
6.1 Model Enhancement	17
6.2 Data Expansion	17
6.2.1 Model Validation and Evaluation	17
6.2.2 User Interface and Experience	17
6.2.3 Integration and Deployment	18
6.2.4 Ethical and Social Considerations	18

7 Conclusion	19
8 Contribution	20

Abstract

This research looks at how different factors affect residential property prices over time. It shows how important it is to prepare and use data properly for accurate predictions. The main factors we considered include the average income, average house age, average number of rooms, average number of bedrooms, and average population in an area. These factors help us understand how property prices change based on different neighborhood features and housing conditions. We used the USA.csv dataset [5] for our analysis and tested three types of prediction models: Linear Regression, Random Forest Regression, and XGBoost Regression. Our results showed that Linear Regression gave the most accurate predictions compared to the other models. The project was built using Django, which helped us create a functional and easy-to-use application. We also used HTML, CSS, and Python to support different parts of the project. This research shows that Linear Regression works well for predicting property prices and highlights the importance of choosing the right model for accurate predictions.

Chapter 1

Introduction

House price prediction is complex, involving physical characteristics, location, neighborhood, and public perception. Determining an objective price is challenging due to variability in market willingness to pay. Automated Valuation Models (AVMs) could enhance accuracy for buyers, sellers, and banks. Research generally supports the hedonic approach, which uses physical and locational variables to predict prices [3]. Location effects like spatial dependence and heterogeneity are important [1] [2]. Challenges include spatial correlations between nearby properties and variations across space. Methods range from using proxies to advanced models like spatial econometrics and machine learning. This literature review synthesizes methods and data types for house price prediction, with a focus on geospatial components. It identifies trends and gaps, noting the prevalence of conventional models and emerging advanced techniques.

1.1 Conceptual Study of the Project

Shelter is a fundamental human need, covering various types of residences like houses, villas, or flats. When deciding to purchase a home, understanding its pricing is crucial. This report addresses these concerns by examining factors that influence house prices and providing a comprehensive case study on regression problems in data science. The study focuses on variables such as average area income, average house age, average number of rooms, average number of bedrooms, and average population density to understand how they affect property prices in different neighborhoods. By applying various regression models, we demonstrate how effective data analysis and model selection can lead to precise price predictions. This analysis underscores the importance of choosing the right model and using data effectively to achieve reliable results in house price prediction.

1.2 Objectives of the Project

- To analyze various parameters such as average income and area to predict house prices.
- To aid customers in making informed real estate investments without relying on agents.
- To offer a reliable and efficient method for determining house prices.
- To provide a transparent pricing model that prevents users from being misled.

1.3 Difficulties

- **Data Collection and Preparation:** Gathering and preprocessing data on property prices and relevant features like location, size, and neighborhood characteristics.
- **Feature Analysis:** Identifying and analyzing factors affecting property prices, such as average income, house age, and number of rooms.
- **Model Selection and Development:** Applying and comparing regression models (e.g., Linear Regression, Random Forest, XGBoost) to predict house prices.
- **Model Evaluation:** Evaluating model performance using metrics like MAE, MSE, and R-squared, and fine-tuning for accuracy.
- **Application and Implementation:** Creating a tool or application to provide real-time price estimates and integrating models into real estate platforms.
- **Insights and Recommendations:** Offering insights into pricing trends and recommendations for real estate decisions.
- **Challenges and Considerations:** Addressing issues like data quality and market fluctuations to ensure reliable predictions.

1.4 Related Work

- **Hedonic Pricing Models:** These models estimate house prices based on property attributes like size, age, and location, assuming prices reflect the value of individual features.
- **Spatial Econometrics:** This research examines how geographic location impacts property values using methods like spatial lag and error models to account for spatial dependence and heterogeneity.
- **Machine Learning Approaches:** Techniques such as Random Forests, Gradient Boosting, and Neural Networks are used to capture complex relationships between features and prices, improving prediction accuracy.
- **Temporal Analysis:** Studies focus on how historical price trends influence current values, using time-series and dynamic regression models to address temporal dependencies.
- **Geospatial Data Integration:** Incorporating geospatial data (e.g., GIS, location coordinates) into models helps understand the impact of spatial factors on property values.
- **Hybrid Models:** Combining hedonic pricing with machine learning methods aims to enhance prediction accuracy and provide detailed insights into price determinants [6].
- **Advanced Techniques:** Recent research explores deep learning methods, including CNNs and RNNs, to model complex patterns and interactions in property data.

Chapter 2

Literature Review

The recent global financial crisis has rekindled interest in housing prices and their impact on economic cycles. Lamer (2007) notes that the housing market often predicts economic recessions, highlighting its role in the business cycle. Vargas and Silva (2008) argue that changes in housing prices are crucial for determining the phase of the economic cycle, with booms driving prices up and downturns leading to slower declines. The delay in falling nominal house prices during downturns reflects homeowners' reluctance to reduce prices. Several studies have explored the influence of housing prices on economic indicators. Forni et al. (2003) [4] and others have found that housing market fluctuations can signal GDP growth and inflation trends (Gupta Kabundi, 2010). Rapach and Strauss (2007) used ARDL models to predict housing price movements, showing their effectiveness compared to benchmark models. Gogas and Pragidis (2011) used risk premiums to forecast future house prices, suggesting that current data can help predict trends. Gupta and Das (2010) employed Spatial Bayesian VARs to forecast declines in real house prices, finding that while BVAR models are useful, they may under-predict downturns due to a lack of fundamental information.

Chapter 3

Methodology

3.1 Diagram for House Price Prediction

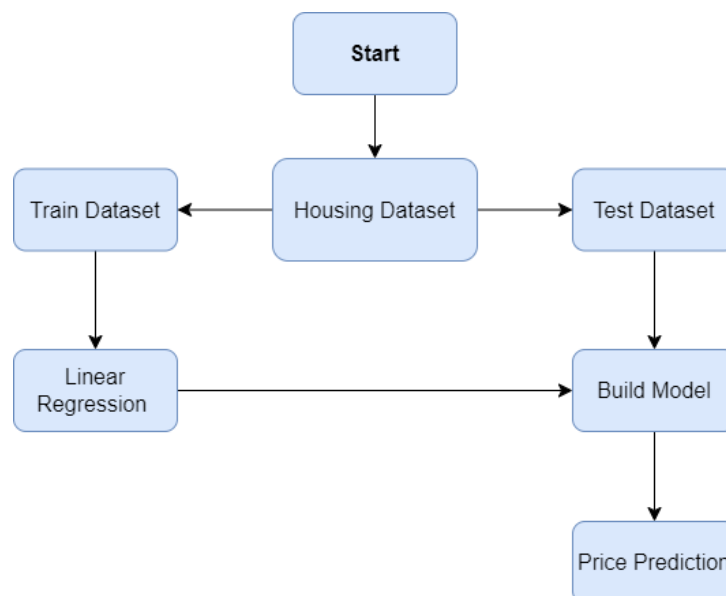


Figure 3.1: Flow Chart

3.2 Data Collection

Here we have web scrapped the data from <https://www.kaggle.com/datasets/aariyan101/usahousingcsv> which is one of the real estate website. Our data contains USA housing only.

3.3 Data Exploration

Data exploration is the initial phase of data analysis and involves summarizing key characteristics of a dataset, such as its size, accuracy, emerging patterns, and other attributes. This process is often carried out by data analysts using visual analytics tools, but can also be performed using advanced statistical software like Python. Before analyzing data collected from various sources and stored in data warehouses, organizations need to understand the number of cases in the dataset, the included variables, the extent of missing values, and the general hypotheses the

data may support. An initial exploration of the dataset helps analysts familiarize themselves with the data and answer these critical questions

3.4 Data Description

```
Data types of all columns:
Avg. Area Income      float64
Avg. Area House Age   float64
Avg. Area Number of Rooms float64
Avg. Area Number of Bedrooms float64
Area Population        float64
Price                 float64
Address               object
dtype: object
```

Figure 3.2: Data Description

3.5 Data Visualization

Data visualization involves representing information and data in graphical formats. By utilizing visual elements such as charts, graphs, and maps, it offers an accessible means to identify trends, outliers, and patterns within the data. In the context of Big Data, visualization tools and technologies are crucial for analyzing large volumes of information and facilitating data-driven decision-making.

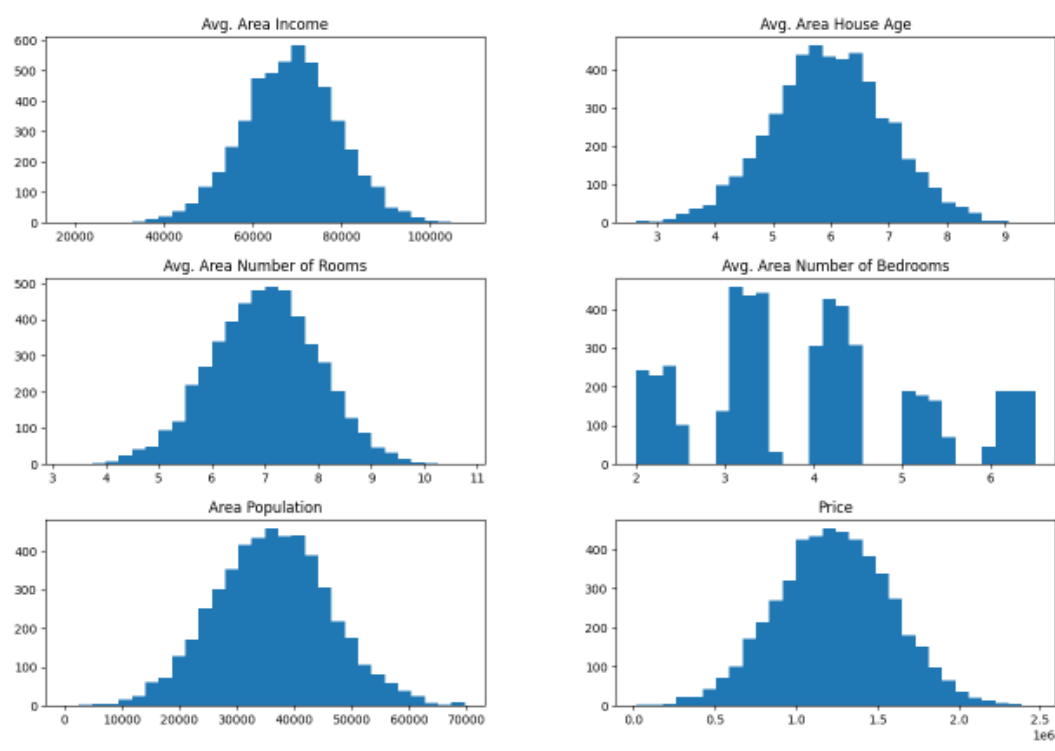


Figure 3.3: Histogram for numerical features

3.6 Data Selection

Data selection refers to the process of identifying the appropriate data types and sources, as well as suitable tools for data collection. This step occurs before the actual data collection takes place. It is important to differentiate data selection from selective data reporting (which involves excluding data that does not support a research hypothesis) and interactive or active data selection (where collected data is used for monitoring activities or secondary analyses). Choosing suitable data for a research project is crucial for maintaining data integrity. The main goal of data selection is to determine the right data type, source, and instruments that enable researchers to effectively address their research questions. This determination is often specific to the discipline and influenced by the nature of the investigation, existing literature, and access to relevant data sources.

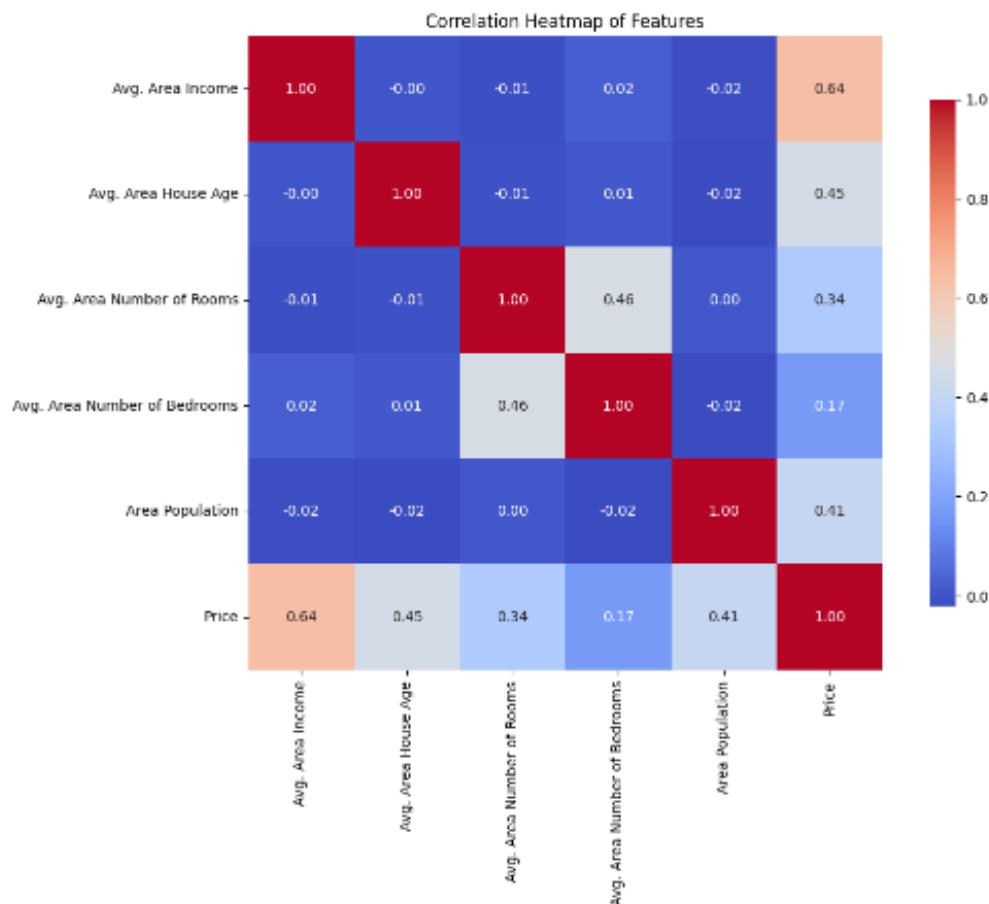


Figure 3.4: Heatmaps

3.7 Data Transformation

Log transformation is useful for reducing the skewness of highly skewed distributions. This can enhance the interpretability of data patterns and help meet the assumptions required for inferential statistics. In the upper panel, it's difficult to identify a clear pattern, while the lower panel clearly displays a strong relationship. When comparing the means of log-transformed

data, it effectively compares geometric means. This is because the anti-log of the arithmetic mean of log-transformed values yields the geometric mean.

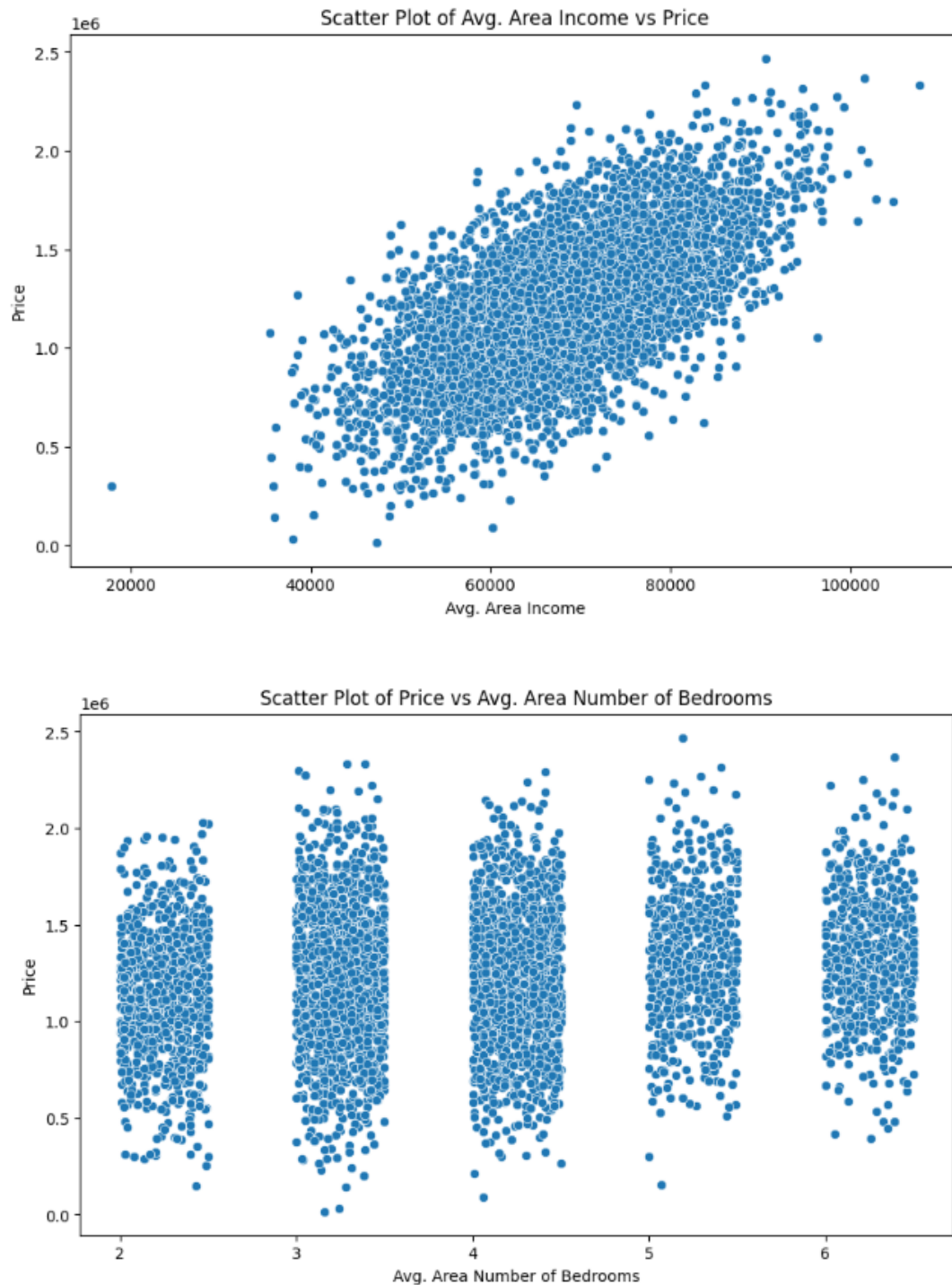


Figure 3.5: Scater Plot

Chapter 4

Implementation

4.1 Language

4.1.1 Python

Python is extensively utilized in scientific and numerical computing:

- **SciPy:** A suite of packages designed for mathematics, science, and engineering.
- **Pandas:** A library focused on data analysis and modeling.
- **IPython:** A robust interactive shell that allows for easy editing and recording of work sessions, and it supports visualizations and parallel computing.
- **The Software Carpentry Course:** Offers foundational skills for scientific computing, conducts bootcamps, and provides open-access teaching materials.

Libraries Used for This Project:

- Pandas
- NumPy
- Matplotlib
- Seaborn
- Scikit-learn
- XGBoost

4.2 Models

4.2.1 Linear Regression Model

Linear regression is a machine learning algorithm that operates under supervised learning. It is used for regression tasks, modeling a target prediction value based on independent variables. This technique is primarily employed to explore relationships between variables and make forecasts

4.2.2 Random Forest Regression Model

Random Forest is an ensemble technique that performs both regression and classification tasks using multiple decision trees and Bootstrap Aggregation (bagging). In this method, each tree is trained on different data samples selected with replacement, allowing for a combination of outputs to improve accuracy over individual decision trees.

4.2.3 XGBoost Regression Model

XGBoost stands for eXtreme Gradient Boosting and implements the gradient boosting decision tree algorithm. It uses boosting, an ensemble technique where new models are added sequentially to correct errors made by existing models, continuing until no further improvements can be achieved.

4.3 Source Code

```
views.py

1  from django.shortcuts import render
2  import pandas as pd
3  import numpy as np
4  from sklearn.model_selection import train_test_split
5  from sklearn.linear_model import LinearRegression
6
7  def home(request):
8      return render(request, 'home.html')
9
10 def predict(request):
11     # Your prediction logic here
12     return render(request, 'predict.html')
13
14 def result(request):
15     data = pd.read_csv(r"E:\ai_project\HousePricePrediction\USA_Housing.csv")
16     data = data.drop(['Address'], axis=1)
17
18     X = data.drop('Price', axis=1)
19     Y = data['Price']
20     X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.30)
21     model = LinearRegression()
22     model.fit(X_train, Y_train)
23
24     var1 = float(request.GET['n1'])
25     var2 = float(request.GET['n2'])
26     var3 = float(request.GET['n3'])
27     var4 = float(request.GET['n4'])
28     var5 = float(request.GET['n5'])
29
30     pred = model.predict(np.array([[var1, var2, var3, var4, var5]]))
31     pred = round(pred[0])
32
33     price = "The Predicted Price is $" + str(pred)
34
35     return render(request, "predict.html", {"result2": price})
36
```

Snipped

Figure 4.1: Source Code

Chapter 5

Result Analysis

5.1 Output

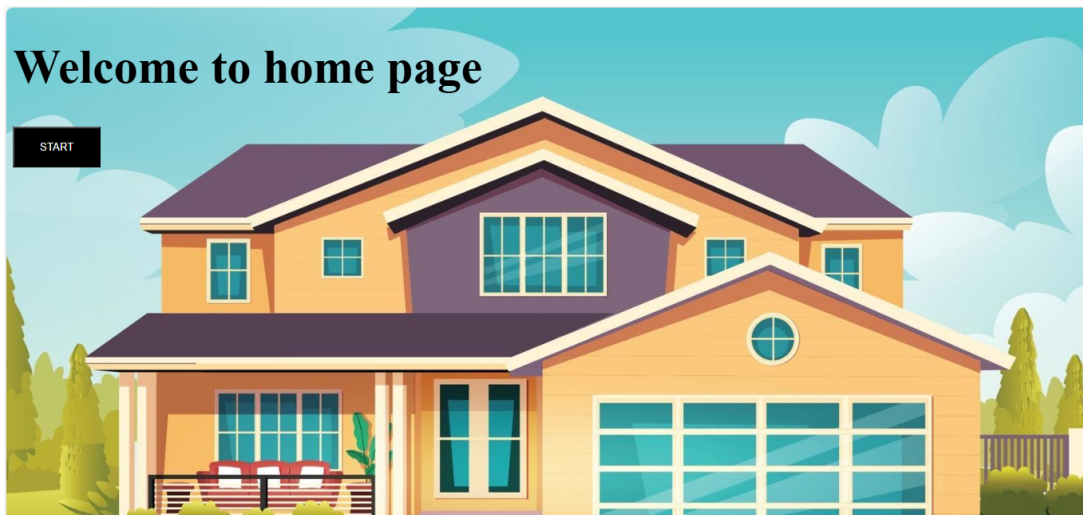


Figure 5.1: Home Page

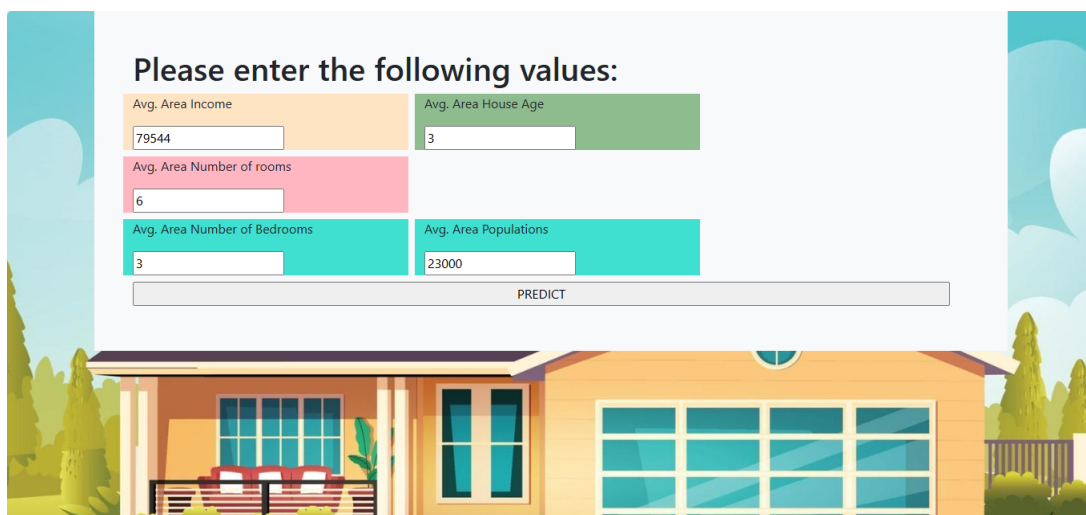
A form titled "Please enter the following values:" with five input fields. The fields are arranged in a grid: "Avg. Area Income" (orange background, value 79544), "Avg. Area House Age" (green background, value 3), "Avg. Area Number of rooms" (pink background, value 6), "Avg. Area Number of Bedrooms" (teal background, value 3), and "Avg. Area Populations" (teal background, value 23000). A "PREDICT" button is located at the bottom right of the form. The background of the form is a light blue sky with white clouds, and the bottom of the form features a colorful illustration of a house.

Figure 5.2: Predict Form

Please enter the following values:

Avg. Area Income	Avg. Area House Age
<input type="text"/>	<input type="text"/>
Avg. Area Number of rooms	
<input type="text"/>	
Avg. Area Number of Bedrooms	Avg. Area Populations
<input type="text"/>	<input type="text"/>

PREDICT

The Predicted Price is \$656268

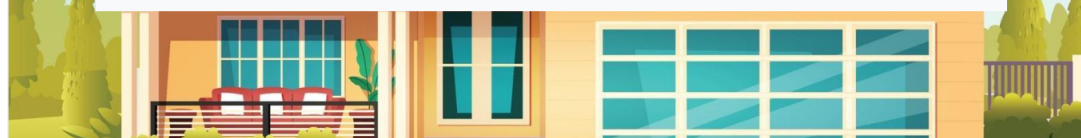


Figure 5.3: Result

5.2 Output Analysis

	Model	Mean Squared Error	R ² Score	Accuracy Percentage
0	Linear Regression	3.220663e+10	0.744719	74.471925
1	Random Forest	3.620154e+10	0.713054	71.305419
2	XGBoost	4.055519e+10	0.678546	67.854571

Figure 5.4: difference between models

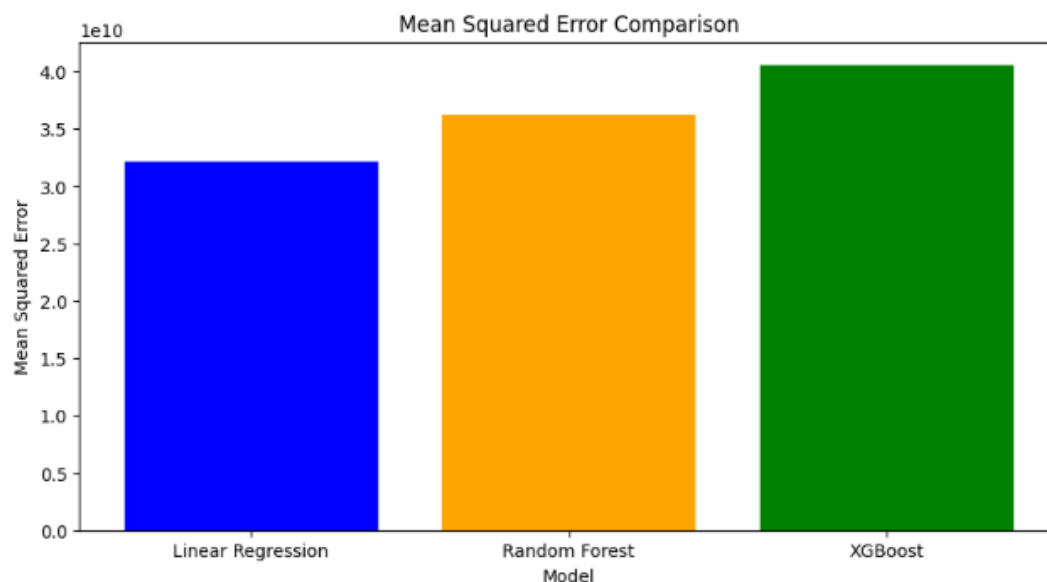


Figure 5.5: Mean squared error comparison

- **Mean Squared Error (MSE):**

- Linear Regression has the lowest MSE (3.22×10^{10}), indicating that its predictions are, on average, closer to the actual values compared to the other models.

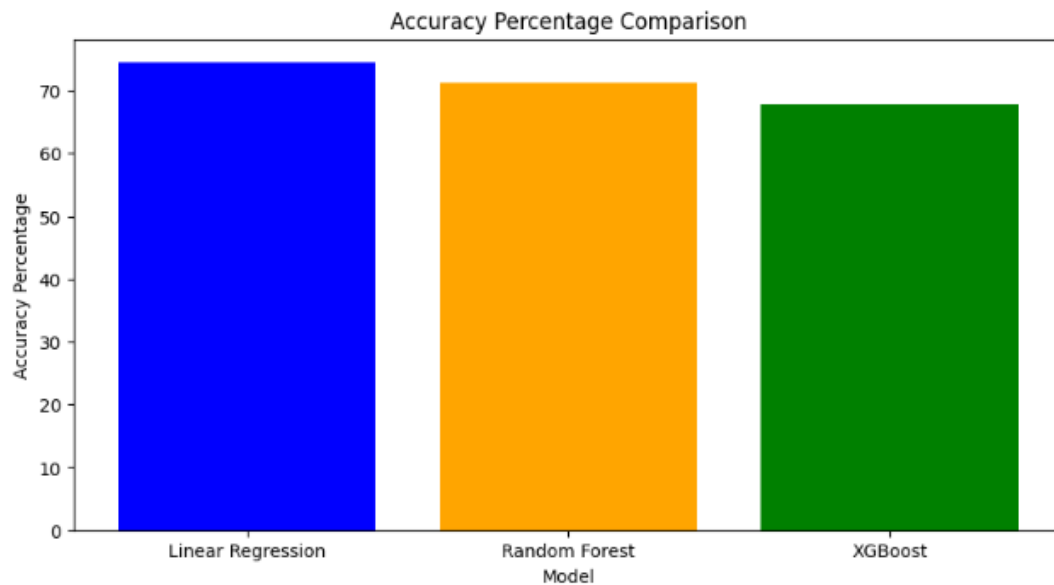


Figure 5.6: accuracy percentage comparison

- Random Forest and XGBoost have higher MSEs, suggesting less accurate predictions.
- **R² Score:**
 - Linear Regression again shows the highest R² score (0.74), meaning it explains approximately 74% of the variance in house prices. This indicates a strong relationship between the independent variables and the target variable.
 - Random Forest and XGBoost have lower R² scores (0.71 and 0.68), indicating that they explain less of the variance in house prices.

We can see that Linear Regression is the best model among the three for this dataset based on both the Mean Squared Error and R² Score. It performs the best in terms of accuracy and model fit.

Chapter 6

Future Work

The house price prediction project has successfully demonstrated the application of linear regression techniques to estimate property values based on various features. However, there are several areas where the project could be expanded and improved in future iterations:

6.1 Model Enhancement

- **Advanced Algorithms:** While linear regression provides a solid baseline, exploring more advanced machine learning models such as Decision Trees, Random Forests, Gradient Boosting, or Neural Networks could improve prediction accuracy.
- **Feature Engineering:** Further investigation into feature engineering could uncover additional relevant features or interactions between existing features, enhancing model performance.

6.2 Data Expansion

- **Inclusion of Additional Data:** Integrating more diverse datasets, such as historical price trends, neighborhood demographics, and economic indicators, could provide a more comprehensive view and improve prediction accuracy.
- **Temporal Analysis:** Incorporating time-series analysis to account for market trends and seasonal variations could lead to more dynamic and responsive predictions.

6.2.1 Model Validation and Evaluation

- **Cross-Validation:** Implementing cross-validation techniques to assess model performance and generalizability across different subsets of the data.
- **Evaluation Metrics:** Utilizing additional evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared to gain deeper insights into model performance.

6.2.2 User Interface and Experience

- **Interactive Visualization:** Developing an interactive web interface with visualizations to allow users to explore predictions and model insights more intuitively.

- **Customization Options:** Providing options for users to input custom feature values and see predictions in real-time.

6.2.3 Integration and Deployment

- **API Development:** Creating an API for the prediction model to facilitate integration with other applications or platforms.
- **Scalability:** Ensuring the system is scalable to handle large datasets and high user traffic efficiently.

6.2.4 Ethical and Social Considerations

- **Bias and Fairness:** Addressing potential biases in the model to ensure fair and equitable predictions across different demographic groups.
- **Data Privacy:** Implementing measures to safeguard user data and ensure compliance with data protection regulations.

Chapter 7

Conclusion

In this report , several tests have been performed using linear regression algorithm to perform house price prediction. This algorithm is to predict prices of new properties that are going to be listed by taking some input variables and predicting the correct and justified price. It was a great learning experience building this predictive Sale Price model. In Future Using different methods that match the time-series data will be used in the research to obtain smaller error prediction values and using more data to get the better result.

Chapter 8

Contribution

Name	ID	Contribution
Shafayet Ullah Ramim	2104010202219	60
Shihabul Alam Sakib	2104010202221	20
Khalid Ahamed Rahi	2104010202202	20

Table 8.1: Team Members and Their Contributions

Bibliography

- [1] Ayse Can. “Specification and estimation of hedonic housing price models”. In: *Regional science and urban economics* 22.3 (1992), pp. 453–474.
- [2] Yuhao Kang et al. “Understanding house price appreciation using multi-source big geo-data and machine learning”. In: *Land use policy* 111 (2021), p. 104919.
- [3] Sherwin Rosen. “Hedonic prices and implicit markets: product differentiation in pure competition”. In: *Journal of political economy* 82.1 (1974), pp. 34–55.
- [4] G Naga Satish et al. “House price prediction using machine learning”. In: *Journal of Innovative Technology and Exploring Engineering* 8.9 (2019), pp. 717–722.
- [5] $USA_{Housing}$. “USA Housing”. In: Kaggle. Nov. 2017.