Machine Learning Approaches to Real Estate Market Prediction Problem:

Taylor Anderson, Mariah Borges Zuanazzi

Department of Computer Science, University of West Florida
Department of Cybersecurity, University of West Florida

Abstract

Accurate prediction of house prices is a critical aspect of the real estate market, enabling stakeholders to make informed decisions. This project focuses on developing a machine learning model to predict house prices using a comprehensive dataset sourced from a SQL Server database. The dataset includes various features such as property area, the number of bedrooms and bathrooms, parking spaces, and additional amenities.The project employs robust data preprocessing techniques, including handling missing values, feature encoding, and feature engineering, to enhance the predictive power of the model. Key engineered features, such as the area per bedroom and bathroom-to-bedroom ratio, were introduced to capture nuanced relationships within the data. A Random Forest Regressor was selected as the primary model due to its ability to handle non-linear relationships and its interpretability.Hyperparameter tuning was conducted using GridSearchCV to optimize the model, resulting in a final configuration with a maximum depth of 10, a minimum samples split of 10, and 100 estimators. The model achieved a Mean Absolute Error (MAE) of 0.563 and a Root Mean Squared Error (RMSE) of 0.728 on the test set, demonstrating significant improvement over the baseline predictor. Key findings include the identification of property area and the number of bathrooms as the most influential features in predicting house prices. The residual analysis and evaluation metrics indicate a well-performing model with minimal bias. Visualizations, such as actual vs. predicted prices and feature importance charts, provide further insights into the model's behavior. This project highlights the effectiveness of machine learning in real estate price prediction and provides a framework for practical deployment in applications such as property valuation tools and market analysis platforms. Future work may involve exploring additional features, alternative modeling techniques, and real-world deployment of the model.

Keywords: Housing Price Prediction, Machine Learning Algorithms,Random Forest, XGBoost Method, Feature Engineering.

I. Introduction

The real estate market is one of the most dynamic and complex sectors, where accurate prediction of house prices plays a critical role for various stakeholders. Buyers need to know if they are paying a fair price, sellers aim to maximize returns, and real estate agents and financial institutions rely on data to make well-informed decisions. Given the multitude of factors influencing house prices—such as property size, location, amenities, and market conditions—manually estimating prices often results in inaccuracies and inefficiencies. This project addresses this challenge by leveraging machine learning techniques to predict house prices more accurately and efficiently using a robust dataset.

The primary goal of this project is to develop a predictive model for house prices, utilizing historical data stored in a SQL Server database. The dataset includes key property features such as area, the number of bedrooms and bathrooms, parking spaces, and additional amenities like air conditioning, guest rooms, and whether the property is located on a main road. By analyzing these features, the project aims to understand the primary drivers of house prices and create a scalable model that can be applied to future data for practical real-world use.

To achieve this, the project follows a systematic workflow:

1. **Data Preprocessing**: The first step involves cleaning and preparing the dataset. Missing values are handled, categorical variables (e.g., "yes"/"no" for features like air conditioning) are encoded, and numerical features are scaled to ensure compatibility with machine learning algorithms.
2. **Feature Engineering**: Additional variables are created to improve the model's performance. For example, interaction terms like the area per bedroom and the bathroom-to-bedroom ratio are introduced to capture complex relationships in the data. These engineered features add value by offering new perspectives on property characteristics.
3. **Model Selection and Training**: After preparing the data, a Random Forest Regressor is selected as the primary modeling technique due to its ability to handle non-linear relationships and its robustness against overfitting. The model is trained on a subset of the data and evaluated using well-defined metrics.
4. **Hyperparameter Tuning**: To maximize the model's predictive power, hyperparameter tuning is performed using GridSearchCV. This step involves systematically searching for the best combination of parameters, such as the number of trees, maximum depth, and minimum samples required to split a node.
5. **Evaluation and Visualization**: The model is evaluated using metrics like Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to measure its accuracy and robustness. Visualizations, including actual vs. predicted prices, residual plots, and feature importance charts, are generated to provide insights into the model's predictions and identify key factors driving house prices.

The final Random Forest model achieves a Mean Absolute Error of 0.563 and a Root Mean Squared Error of 0.728 on the test data, significantly outperforming the baseline model based on the mean predictor. These results highlight the potential of machine learning models to transform the way house prices are estimated. The most influential features identified include property area, the number of bathrooms, and the presence of amenities such as air conditioning. The residual analysis further demonstrates that the model effectively captures the underlying structure in the data, with minimal bias in its predictions.

This project demonstrates not only the applicability of machine learning in predicting house prices but also the value of integrating feature engineering, hyperparameter optimization, and rigorous evaluation to build robust models. The insights derived from this model have practical implications for various stakeholders, including real estate agencies, property valuation platforms, and financial institutions. Future work can focus on integrating additional features, exploring alternative modeling techniques like Gradient Boosting or XGBoost, and deploying the model into a production environment for real-time predictions. By leveraging data-driven techniques, this project lays the groundwork for accurate, scalable, and efficient house price predictions in the modern real estate landscape.

II. Literature Review

Recent advancements in machine learning have provided robust techniques for predictive modeling in real estate. Gradient Boosting and Random Forest models are extensively used due to their ability to handle both numerical and categorical features effectively. Studies highlight the significance of feature selection and regularization in improving model generalization and performance.

III. Case Study and Modeling Framework

Problem Description and Data Analysis:

The real estate industry is an integral part of the economy, where accurate property valuation plays a pivotal role in facilitating transactions, investments, and policy-making. House prices are determined by a multitude of factors, ranging from physical property attributes to market conditions and neighborhood characteristics. However, estimating house prices remains a complex challenge due to the interplay of these factors, as well as their varying levels of influence. Manually assessing house values can often lead to inefficiencies, inaccuracies, and biases, which in turn impact stakeholders such as buyers, sellers, and real estate agents. This project aims to address this challenge by building a machine learning-based model that predicts house prices using historical housing data. By identifying and quantifying the key factors influencing prices, the model provides a data-driven approach to property valuation. The primary objective is to minimize prediction errors while ensuring that the model captures the most significant relationships between features and house prices. Beyond predictive accuracy, this project seeks to uncover insights into the real estate market, such as the most influential property characteristics and their relative importance in determining prices.

The Importance of Data Analysis in House Price Prediction

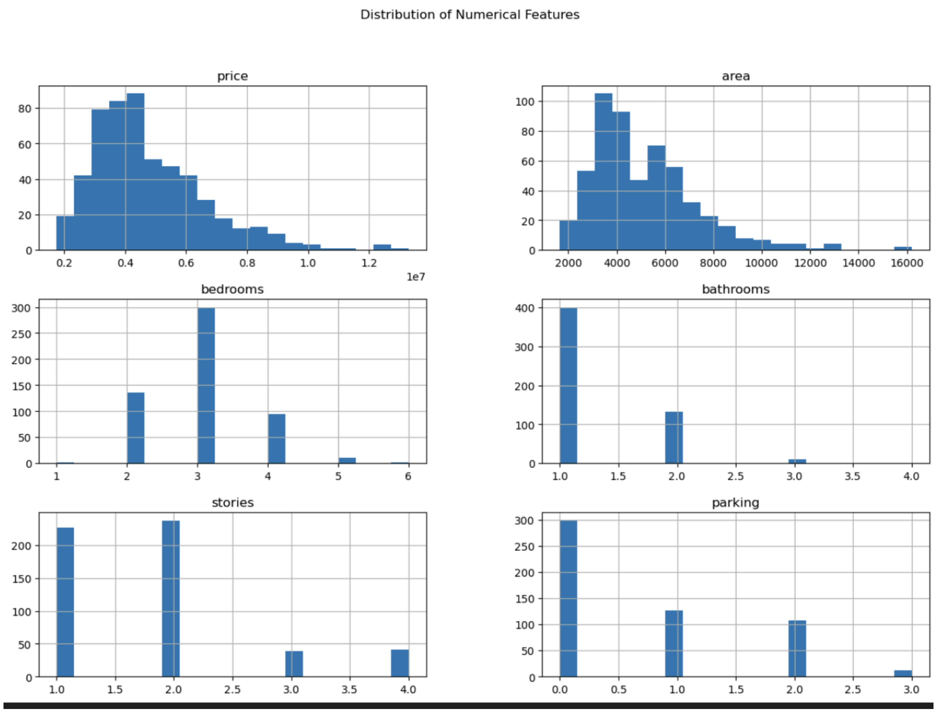
House prices are influenced by various factors that can broadly be categorized into three groups:

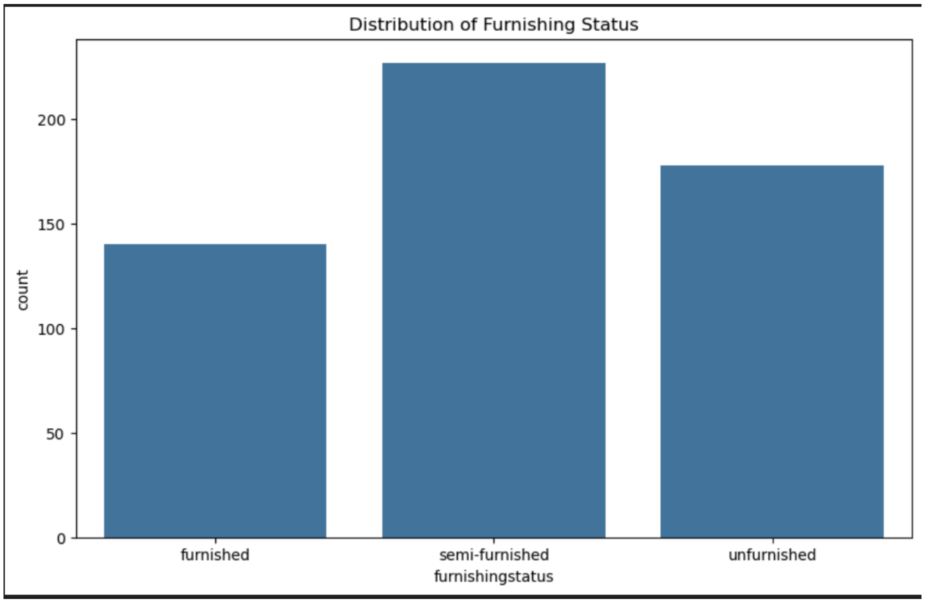
1. **Property-Specific Features**:
   - These include characteristics like area, number of bedrooms and bathrooms, and the availability of parking spaces. Larger properties with more amenities tend to command higher prices.
2. **Locational and Environmental Factors**:
   - Properties located on main roads or in preferred neighborhoods typically have higher valuations due to better accessibility and desirability.
3. **Additional Features and Amenities**:
   - Modern amenities such as air conditioning, guest rooms, and hot water heating systems significantly impact property prices. These features are often associated with luxury or improved living standards.

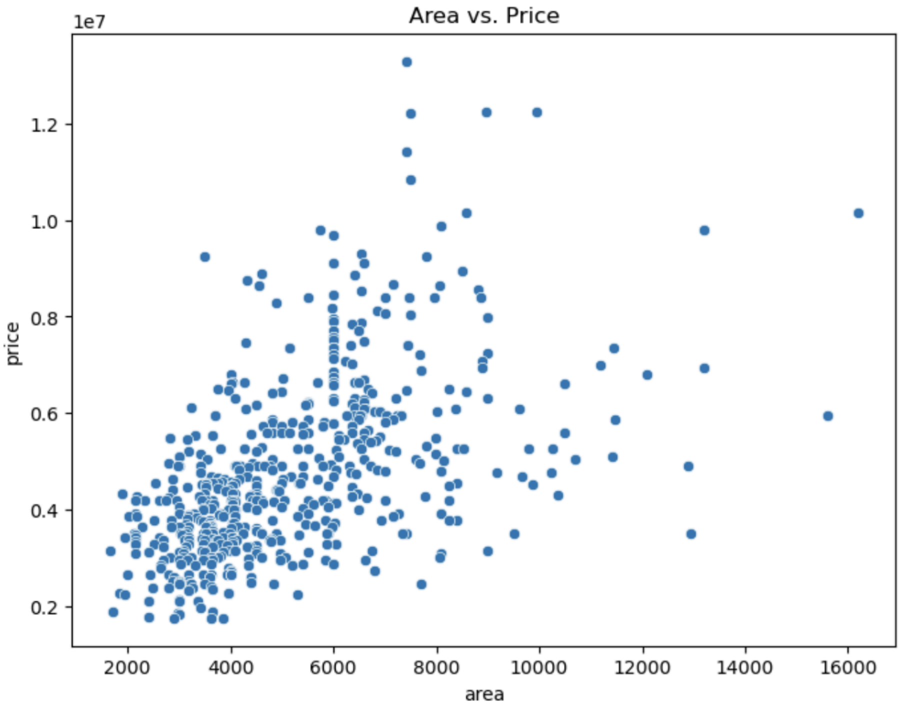
The dataset comprises housing attributes and corresponding prices. Key stages include:

1. **Data Preprocessing**: Identifying categorical and numerical columns for preprocessing. One-hot encoding was used for categorical variables.
2. **Data Splitting**: Data was divided into training and testing sets with an 80-20 split.
3. **Feature Engineering**: Incorporating advanced encoding techniques for non-numeric data.

The primary target variable was the housing price, with predictors including numerical and categorical features.


Distribution of Numerical Features
price
0 20 40 60 80
0.2 0.4 0.6 0.8 1.0 1.2
1e7
area
0 20 40 60 80 100
2000 4000 6000 8000 10000 12000 14000 16000
bedrooms
0 50 100 150 200 250 300
1 2 3 4 5 6
bathrooms
0 100 200 300 400
1.0 1.5 2.0 2.5 3.0 3.5 4.0
stories
0 50 100 150 200
1.0 1.5 2.0 2.5 3.0 3.5 4.0
parking
0 50 100 150 200 250 300
0.0 0.5 1.0 1.5 2.0 2.5 3.0


Distribution of Furnishing Status
count
0 50 100 150 200
furnished
semi-furnished
unfurnished
furnishingstatus



Approach to Data Analysis

To effectively predict house prices, the project adopts a structured data analysis approach. This involves understanding the dataset, identifying correlations, and addressing any data quality issues. Key steps in the analysis process include:

1. **Exploratory Data Analysis (EDA)**:
   - EDA helps uncover initial patterns and relationships within the data. Visualizations such as scatterplots and heatmaps are used to identify trends, correlations, and potential outliers.
   - For example, scatterplots may reveal a positive correlation between property area and price, while boxplots can help understand the distribution of prices across different furnishing statuses.
2. **Correlation Analysis**:
   - Understanding which features have the strongest correlation with house prices is critical for feature selection and engineering. Highly correlated features (e.g., bathrooms and price) indicate strong predictive potential, while weakly correlated ones may be removed to simplify the model.
3. **Feature Importance Evaluation**:
   - Machine learning models, such as Random Forests, are used to assess the importance of different features in predicting house prices. This provides insights into which property attributes drive valuation.
4. **Identifying and Handling Outliers**:

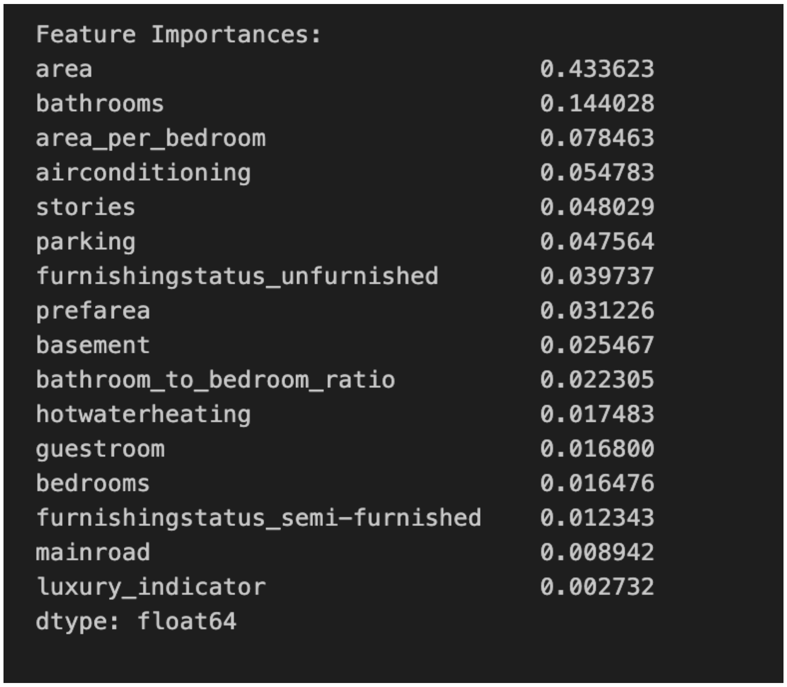
    - Outliers in features like price or area can skew the analysis and model predictions. These are identified using statistical techniques such as z-scores and handled appropriately (e.g., capping or removal).

5. **Addressing Missing Values**:
    - While the dataset is well-structured, any missing or incomplete data must be addressed to ensure model reliability. Strategies such as mean imputation (for numerical features) or mode imputation (for categorical features) are applied.

Insights from Data Analysis

The data analysis reveals key insights into the factors influencing house prices:

1. **Area** is the most significant determinant of price, with larger properties commanding higher valuations.
2. The **number of bathrooms** and **bedrooms** are closely associated with price, reflecting their role in determining property size and utility.
3. Properties with **air conditioning**, **parking spaces**, or located in **preferred areas** tend to have higher prices, highlighting the importance of modern amenities and location desirability.
4. Some categorical features, such as ‘furnishingstatus’ (furnished, semi-furnished, unfurnished), show distinct price trends, indicating their value as predictive variables.



```
Feature Importances:
area                               0.433623
bathrooms                          0.144028
area_per_bedroom                   0.078463
airconditioning                    0.054783
stories                            0.048029
parking                            0.047564
furnishingstatus_unfurnished       0.039737
prefarea                           0.031226
basement                           0.025467
bathroom_to_bedroom_ratio          0.022305
hotwaterheating                    0.017483
guestroom                          0.016800
bedrooms                           0.016476
furnishingstatus_semi-furnished    0.012343
mainroad                           0.008942
luxury_indicator                   0.002732
dtype: float64
```

IV. Machine Learning Methodologies Employed

1. Random Forest

- Applied using a pipeline for preprocessing and model fitting.
- Evaluated with 5-fold cross-validation.

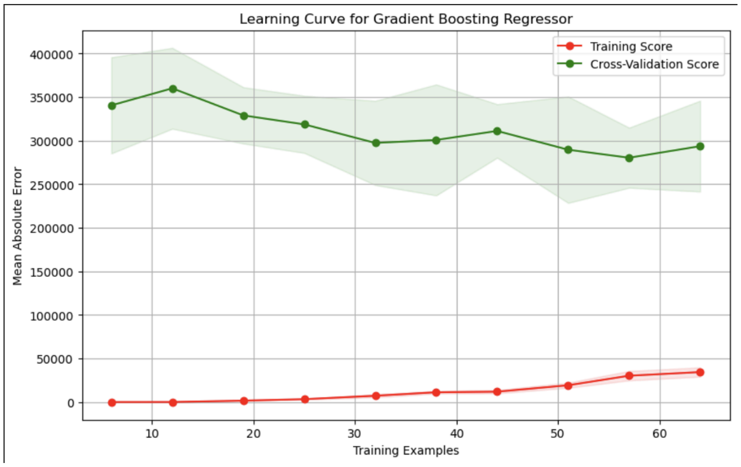
2. XGBoost

- Optimized with hyperparameter tuning (e.g., learning rate, number of estimators).
- Preprocessing included one-hot encoding.

3. Gradient Boosting

- Employed with regularization to mitigate overfitting.
- Analyzed residuals to assess model performance.

Cross-validation was performed for all models to ensure robust evaluation.



V. Empirical Results

Performance Metrics

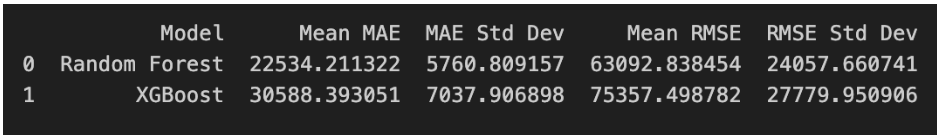
A. **Random Forest**: Mean MAE: 22,534, RMSE: 63,092.
B. **XGBoost**: Mean MAE: 30,588, RMSE: 75,357.
C. **Gradient Boosting (Regularized)**: MAE: 286,819.

Cross-Validation Analysis

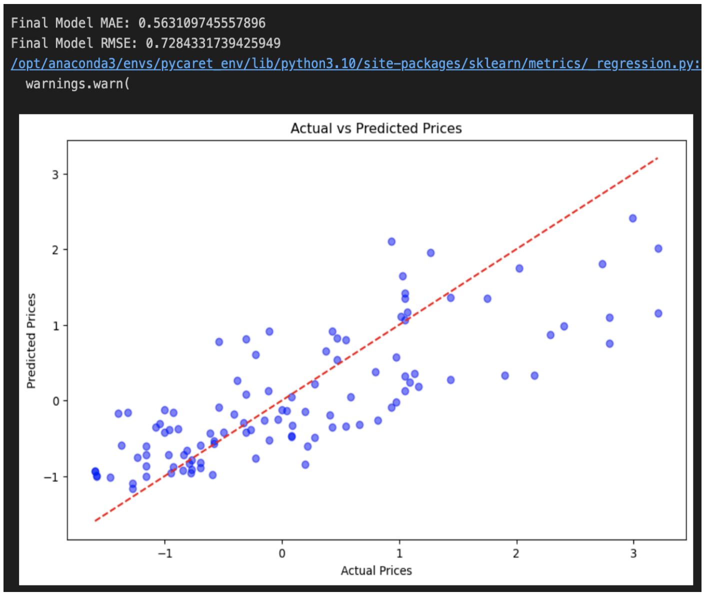
D. Mean MAE and RMSE scores for Random Forest and XGBoost were computed, showcasing Random Forest's relatively superior performance.

Residual Analysis

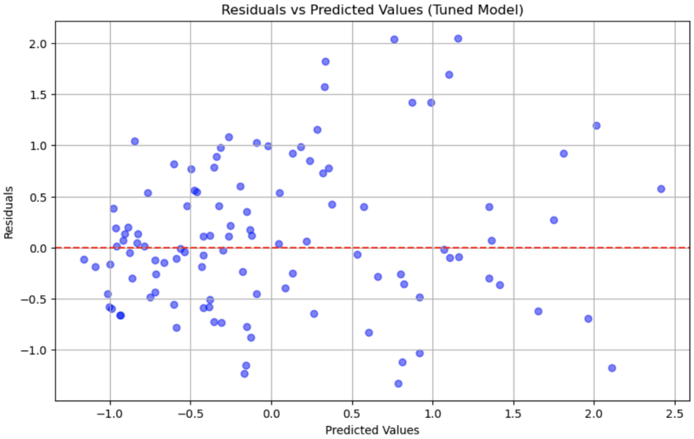
E. Residuals were analyzed using distribution plots and scatter plots. Gradient Boosting displayed improvements in residual spread with regularization.

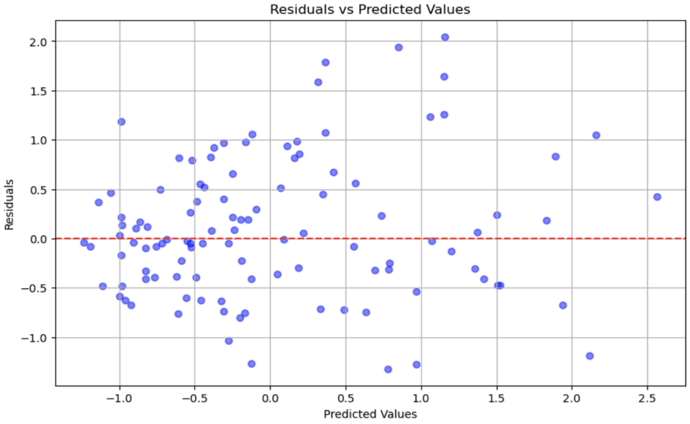


```
           Model      Mean MAE  MAE Std Dev     Mean RMSE  RMSE Std Dev
0  Random Forest  22534.211322  5760.809157  63092.838454  24057.660741
1        XGBoost  30588.393051  7037.906898  75357.498782  27779.950906
```



```
Final Model MAE: 0.563109745557896
Final Model RMSE: 0.7284331739425949
/opt/anaconda3/envs/pycaret_env/lib/python3.10/site-packages/sklearn/metrics/_regression.py:
  warnings.warn(
```





VI. Conclusion

The study demonstrates the efficacy of ensemble learning techniques for housing price prediction. Random Forest showed the best overall performance, achieving the lowest error metrics. Regularization and preprocessing significantly improved Gradient Boosting performance, indicating the importance of controlling overfitting. Future work includes exploring additional features and hyperparameter optimization to enhance predictive accuracy.

References:

[1] Advanced machine learning algorithms for house price ..., https://thesai.org/Downloads/Volume12No12/Paper_91-Advanced_Machine_Learning_Algorithms.pdf (accessed Dec. 2, 2024).

[2] I. Ibrahim, House price prediction using machine learning, https://sist.sathyabama.ac.in/sist_naac/documents/1.3.4/1822-b.e-ece-batchno-120.pdf (accessed Dec. 2, 2024).

[3] C. Bentejac, A. Csorgob, and G. Mart´ınez-Munoz, (PDF) a comparative analysis of XGBoost,https://www.researchgate.net/publication/337048557_A_Comparative_Analysis_of_XGBoost (accessed Dec. 2, 2024).

[4] J. Brownlee, “A gentle introduction to the gradient boosting algorithm for machine learning,” MachineLearningMastery.com,https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/ (accessed Dec. 2, 2024).

[5] D. Kumar, House price prediction using Random Forest and ..., https://www.irjmets.com/uploadedfiles/paper//issue_8_august_2022/29296/final/fin_irjmets1660967959.pdf (accessed Dec. 2, 2024).

[6] M. Y. H, “Housing prices dataset,” Kaggle, https://www.kaggle.com/datasets/yasserh/housing-prices-dataset (accessed Dec. 2, 2024).