Feature Importance Analysis and Predictive Modeling for Boston Prices Using Regression Algorithms

Josh Kenn A. Viray
University of Santo Tomas, joshkenn.viray.cics@ust.edu.ph

Abstract - This study investigates the relationship between various housing features and their impact on house prices in the Boston housing dataset. Using a Random Forest Regressor, feature importance scores were computed to evaluate the predictive power of individual features. Key findings reveal that the average number of rooms per dwelling (RM) and the percentage of lower-status population (LSTAT) are the most significant predictors of house prices. The study also identifies features with minimal importance, such as the proportion of residential land zoned (ZN) and accessibility to radial highways (RAD)—these insights aid in simplifying predictive models without compromising accuracy in applying features to various regression models.

Index Terms–Boston housing prices, feature importance, regression analysis, predictive modeling.

I. INTRODUCTION

House prices are influenced by multiple factors, ranging from environmental conditions to structural attributes. Understanding these relationships is essential for building predictive models and making informed decisions in real estate. This study utilizes the Boston housing dataset to examine feature importance and identify key predictors of house prices. The Gradient Boosting Regressor, known for its robust feature importance estimation, was employed to compute the significance of individual features.

II. METHODOLOGY

A. Dataset (Boston Housing Dataset)

The Boston housing dataset comprises 13 input features and one target variable (MEDV), representing the median house prices in $1,000s. Features include:

- **CRIM**: Per capita crime rate by town.
- **ZN**: Proportion of residential land zoned for lots over 25,000 sq. ft.
- **INDUS**: Proportion of non-retail business acres per town.
- **CHAS**: Charles River dummy variable.
- **NOX**: Nitric oxide concentration.
- **RM**: Average number of rooms per dwelling.
- **AGE**: Proportion of owner-occupied units built prior to 1940.
- **DIS**: Weighted distances to five Boston employment centers.
- **RAD**: Index of accessibility to radial highways.
- **TAX**: Full-value property tax rate per $10,000.
- **PTRATIO**: Pupil-teacher ratio by town.
- **B**: Proportion of Black residents by town.
- **LSTAT**: Percentage of lower-status population

B. Exploratory Data Analysis

For the exploratory data analysis, a heatmap was used with pair plots to have an initial understanding of the data and how they correlate.

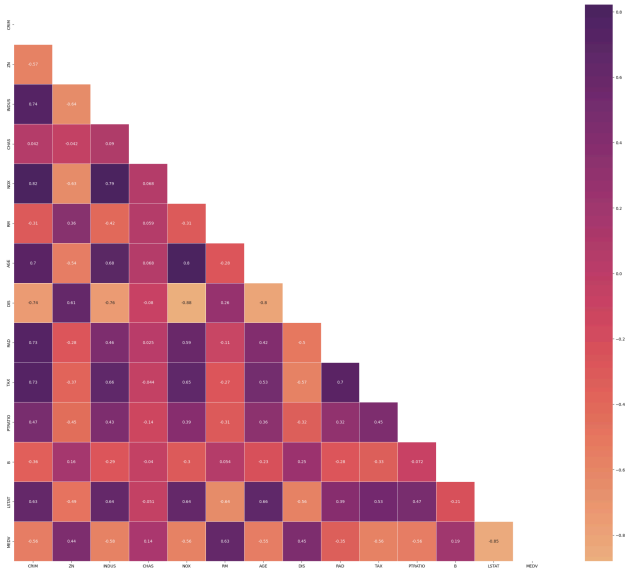

Fig. 1.
Heatmap of all features from the (base) dataset

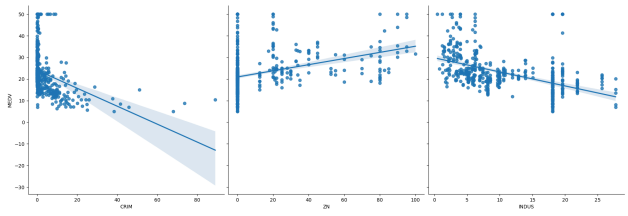

Fig. 2.
Pairplot of CRIM, ZN, and INDUS against MEDV

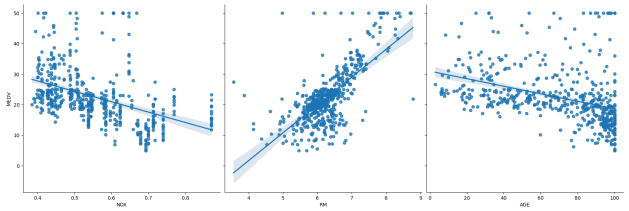

Fig. 3.
Pairplot of NOX, RM, and AGE against MEDV

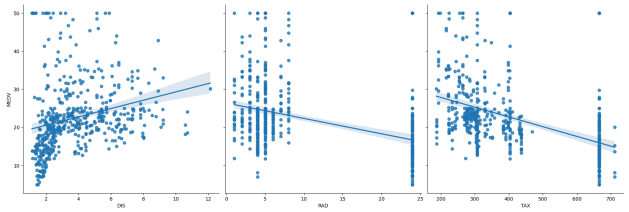

Fig. 4.
Pairplot of DIS, RAD, and TAX against MEDV

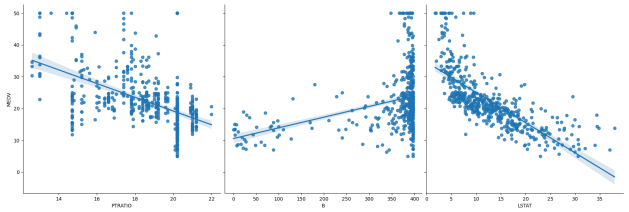

Fig. 5.
Pairplot of PTRATIO, B, and LSTAT against MEDV

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000
mean	3.613524	11.363636	11.136779	0.069170	0.554695	6.284634	68.574901	3.795043	9.549407	408.237154	18.455534	356.674032	12.653063	22.532806
std	8.601545	23.322453	6.860353	0.253994	0.115878	0.702617	28.148861	2.105710	8.707259	168.537116	2.164946	91.294864	7.141062	9.197104
min	0.006320	0.000000	0.460000	0.000000	0.385000	3.561000	2.900000	1.129600	1.000000	187.000000	12.600000	0.320000	1.730000	5.000000
25%	0.082045	0.000000	5.190000	0.000000	0.449000	5.885500	45.025000	2.100175	4.000000	279.000000	17.400000	375.377500	6.950000	17.025000
50%	0.256510	0.000000	9.690000	0.000000	0.538000	6.208500	77.500000	3.207450	5.000000	330.000000	19.050000	391.440000	11.360000	21.200000
75%	3.677083	12.500000	18.100000	0.000000	0.624000	6.623500	94.075000	5.188425	24.000000	666.000000	20.200000	396.225000	16.955000	25.000000
max	88.976200	100.000000	27.740000	1.000000	0.871000	8.780000	100.000000	12.126500	24.000000	711.000000	22.000000	396.900000	37.970000	50.000000

Table 1.
Data description of all columns

The analysis of feature correlations with housing prices in Boston reveals distinct relationships between various factors and the target variable (MEDV). Weakly correlated features include CRIM (per capita crime rate), which shows a negative correlation with housing prices, as lower crime rates generally correspond to higher property values. Similarly, ZN (proportion of residential land zoned for large lots) has a positive correlation, with higher zoning proportions linked to higher prices. INDUS (non-retail business acres) negatively correlates with housing prices, indicating that an increase in industrial areas tends to reduce property values.

On the other hand, several features exhibit moderate to strong correlations. NOX (nitric oxide concentration) has a negative correlation, where higher pollution levels correspond to lower prices. RM (average number of rooms per dwelling) stands out as one of the strongest predictors, showing a positive correlation, as houses with more rooms generally have higher values. AGE (older homes) negatively correlates with prices, while DIS (distance to employment centers) positively influences housing costs. Accessibility features such as RAD (highway accessibility) and TAX (property tax rate) have negative correlations, as proximity to highways and higher tax rates tend to lower home values. Additionally, PTRATIO (pupil-teacher ratio) negatively correlates, suggesting that areas with better education systems command higher prices.

While B (proportion of Black residents) has a weak positive correlation, LSTAT (percentage of lower-status population) strongly negatively correlates, with higher percentages of lower-income residents associated with lower housing prices.

These findings highlight that most features exhibit linear relationships with housing prices, though their predictive significance varies. As the next step, we will implement Linear Regression as a baseline model and apply feature selection techniques to retain the most influential variables, ensuring optimal predictive performance. Given the continuous nature of the target variable, regression techniques will be employed to develop an effective housing price prediction model.

C. Data Cleaning and Preprocessing

In the initial analysis of the Boston housing dataset, it was observed that several features exhibited a high number of outliers. These extreme values can distort statistical analyses and negatively impact machine learning models by introducing bias and reducing generalization performance.

To mitigate the effect of outliers, trimming the tails of the data distribution was applied. This technique involves removing extreme values from both ends of the distribution while retaining the central portion of the data.

In this study, percentile-based trimming was applied to remove extreme values from both ends of the distribution. Specifically:

- The 1st percentile (lower 1%) and the 99th percentile (upper 1%) were selected as cut-off points.
- Any values below the 1st percentile or above the 99th percentile were removed.

The data was also made sure that it would not contain any blanks or null values in the features presented in the dataset.

Feature scaling is a crucial preprocessing step in machine learning, particularly for models like Linear Regression, Ridge, Lasso, and Elastic Net, which are sensitive to feature magnitudes. Without proper scaling, features with larger numerical ranges can dominate the model, leading to biased predictions. In the Boston housing dataset, numerical features such as CRIM (crime rate) and ZN (proportion of residential land zoned) have significantly different ranges, which can hinder model performance.

To address this, StandardScaler was applied, ensuring that all numerical features have a mean of 0 and standard deviation of 1, thereby bringing them onto a similar scale. This standardization improves model convergence, stability, and interpretability. The formula used for standardization is

$$X_{scaled} = \frac{X-\mu}{\sigma} \quad (1)$$

Where X represents the original value, μ is the mean, and σ is the standard deviation.

Additionally, handling categorical variables is another essential preprocessing step. CHAS (a binary variable indicating whether a property is near the Charles River) was the only categorical feature in this dataset. Given that CHAS is a dummy variable (taking values of 0 or 1), it was already in a suitable format for the model and did not require additional encoding. Since it is a binary indicator, models can naturally interpret it without transformation. Proper feature scaling and encoding contribute to a well-prepared dataset, allowing the model to learn effectively and make accurate predictions.

D. Model

Multiple regression models were used in order to best find an algorithm that would fit the data the best with having mean standard error (MSE) and r-squared as quantifiable data to provide an understanding of the model's perceived understanding of the dataset, namely:

1. Linear Regression
2. Ridge Regression
3. Lasso Regression
4. Elastic Net Regression
5. Decision Tree Regressor
6. Random Forest Regressor
7. Gradient Boosting Regressor.

The objectives of this study and the given algorithms were to accomplish the following:

1. Train on the dataset using all parameters.
2. Compute feature importance scores using its inherent mechanism based on impurity reduction.
3. Select high-value features based on the feature importance function.

III. RESULTS AND DISCUSSIONS

A. Baseline Model

A Linear Regression model was implemented as a starting point for predicting housing prices. Linear Regression is a simple yet powerful statistical method that assumes a linear relationship between the independent variables (features) and the dependent variable (target). In this case, the features include variables such as CRIM, RM, and LSTAT, while the target variable is MEDV, representing the median house price.

The performance of the Linear Regression model was evaluated using two key metrics: **Mean Squared Error (MSE)** and **R-squared (R^2)**. The model achieved an MSE of **25.06**, which reflects the average squared difference between the predicted and actual housing prices. While this value suggests some degree of error in the predictions, it provides a baseline for comparison with more sophisticated models.

Additionally, the R-squared value was **0.655**, indicating that approximately 65.5% of the variance in housing prices could be explained by the features in the dataset. While this is a reasonable starting point, the relatively low R^2 score suggests that there is room for improvement, either by using more advanced regression techniques or by incorporating feature engineering.

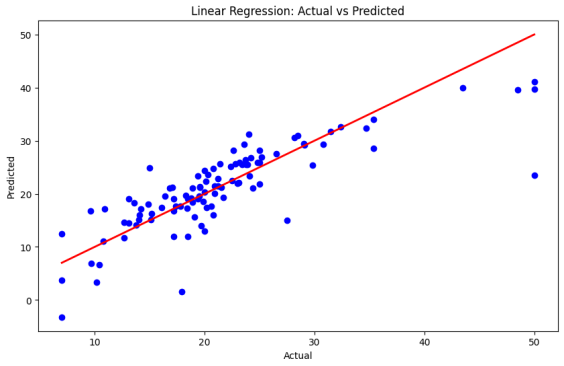

Fig. 6.
Actual vs. Predicted plot for the Linear Regression baseline model.

The **Actual vs Predicted plot** further illustrates the model's performance. The red line represents perfect predictions (where predicted values equal actual values), while the blue points are the actual predictions made by the model. Although the points generally follow the trend of the red line, some points deviate significantly, indicating that the model struggles with certain data points. These deviations highlight the potential influence of outliers, non-linear relationships, or interactions between features that a simple linear model cannot capture.

In summary, Linear Regression provides a strong foundation for understanding the dataset and establishing a baseline. However, the results indicate that more complex models may be necessary to achieve better predictive performance and account for the intricacies of the housing market, given a relatively rich feature set.

B. Advanced Models

In this section, we delve deeper into the performance of advanced regression algorithms applied to the Boston housing dataset. The **Model Evaluation Summary** below compares the algorithms based on two key metrics: **Mean Squared Error (MSE)** and **R-squared (R^2)**. These metrics assess the predictive accuracy and explanatory power of the models, respectively.

1. Linear and Regularized Regression Models
   Linear Regression serves as a baseline model for comparison. It assumes a linear relationship between the features and the target variable (MEDV) and achieved an MSE of 25.06 with an R^2 of 0.655, meaning it explains 65.5% of the variance in house prices. While this provides a good starting point, the model's assumptions may limit its ability to capture non-linear relationships in the data. To address potential overfitting and multicollinearity issues, Ridge Regression and Lasso Regression were applied:
   - **Ridge Regression** incorporates L2 regularization, penalizing large coefficients and stabilizing predictions. It achieved a similar MSE (**25.33**) and R^2 (**0.652**) as Linear Regression, indicating that regularization did not significantly improve performance in this case.
   - **Lasso Regression**, which uses L1 regularization, performed slightly worse with an MSE of **25.76** and an R^2 of **0.646**. This may be due to Lasso's tendency to shrink coefficients to zero, reducing the influence of certain features in the dataset.
   - **Elastic Net Regression**, a hybrid of Ridge and Lasso, yielded an MSE of **26.05** and R^2 of **0.642**, showing no substantial improvement over individual regularization techniques. These results suggest that linear and regularized models may not be sufficient to capture the complexity of the Boston housing dataset.

2. Tree-Based Models
   Tree-based models, which excel at capturing non-linear relationships, were then employed.
   - Decision Tree Regression achieved an MSE of 20.52 and an R^2 of 0.718, a notable improvement over linear models. Decision Trees partition the feature space into smaller regions, allowing for better adaptability to non-linear patterns. However, Decision Trees are prone to overfitting, especially with deeper trees.
   - Random Forest Regression, an ensemble method of multiple decision trees, significantly outperformed individual trees with an MSE of 8.48 and R^2 of 0.883. By averaging predictions from multiple trees, Random Forest reduces overfitting and provides robust predictions. Its high R^2 value suggests that it captures a substantial amount of variance in the target variable.

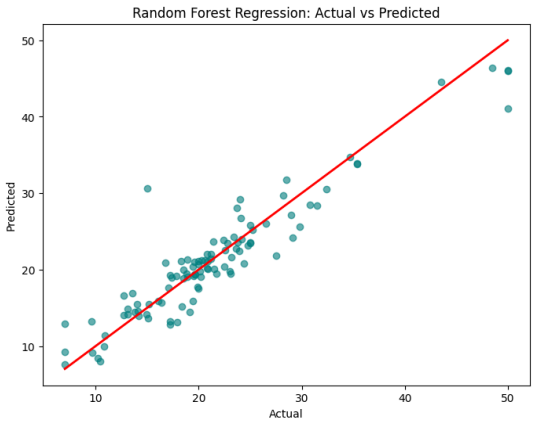

Fig 6.
Actual vs. Predicted plot random forest regression

3. Gradient Boosting (Ensemble)
The Gradient Boosting Regressor emerged as the best-performing model in this study, with an MSE of 5.67 and an R^2 of 0.922, indicating that it explains over 92% of the variance in housing prices. Gradient Boosting builds trees sequentially, where each subsequent tree corrects the errors of the previous ones. This iterative approach makes Gradient Boosting highly effective at capturing complex relationships in the data, albeit at the cost of increased computational time.

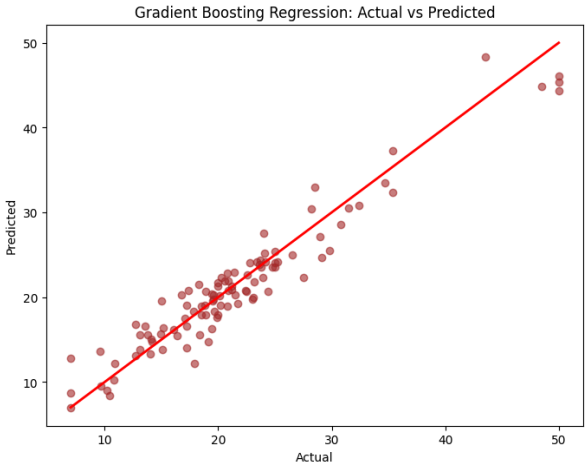

Fig 8.
Actual vs. Predicted plot gradient boosting regression

C. Hyperparameter Tuning
In this study, hyperparameter tuning was employed to optimize the performance of a Random Forest Regressor and identify the most important features for predicting house prices. Using a Grid Search with cross-validation, the optimal hyperparameters for the model, including the number of estimators (n_estimators) and the maximum tree depth (max_depth), were determined. The best model obtained through Grid Search was then used to compute feature importance using permutation importance, which evaluates the impact of each feature on the model's predictions.

The computed feature importance scores were sorted in descending order, and a threshold of **0.4** was applied to select the most influential features. This threshold ensures that only features with significant contributions to the model's predictive power are retained for further analysis. For example, features such as the average number of rooms (RM) and the lower-status population (LSTAT) percentage were identified as highly significant predictors. In contrast, less important features like the proportion of residential land zoned (ZN) were excluded. This process streamlined the dataset by focusing on the most impactful features, ultimately improving the interpretability and efficiency of the predictive model with the following coefficient values.

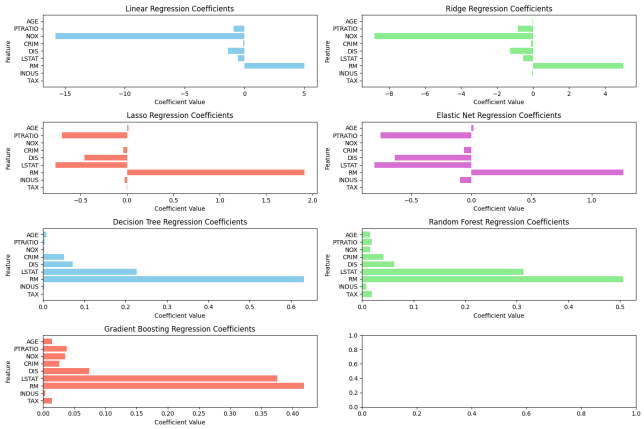

Fig 9.
Coefficient values of regression algorithms

After applying the 0.4 threshold, we are given Figure 4. In the figure above, there is not much value in the INDUS feature, which is consistent with all of the seven algorithms, which is why it would benefit us later if we dropped it from our feature set. However, due to the nature of the purpose of the column, it would not be that helpful to remove it as we continue on to building our initial model.

D. Model Evaluation

The model evaluation process was carried out by analyzing the performance of various regression algorithms applied to the Boston housing dataset. Three key metrics were used to assess the models: **Mean Squared Error (MSE)**, **R-squared (R^2)**, and **Root Mean Squared Error (RMSE)**. Each metric provides unique insights into the predictions' accuracy, variance, and interpretability. Below is a detailed analysis of the models and their performance.

R^2 measures the proportion of variance in the target variable (MEDV) explained by the features. Higher R^2 values indicate better explanatory power.

- Linear models explain approximately **64–65%** of the variance, as indicated by their R^2 scores of **0.64–0.65**.
- The **Decision Tree Regression** improves this to **0.718**, showing its capacity to capture more variability in the target.
- Ensemble models outperform others, with **Random Forest Regression achieving R^2 = 0.883** and **Gradient Boosting Regression achieving R^2 = 0.922**, indicating that Gradient Boosting explains over **92%** of the variance in house prices.

RMSE, the square root of MSE, provides a scale-consistent measure of prediction error. Lower RMSE values signify better accuracy.

- Linear models show higher RMSE values, reaffirming their limitations in predictive accuracy.
- Ensemble methods significantly reduce RMSE, with **Gradient Boosting Regression** achieving the lowest value, indicating the smallest average error in predicting house prices.

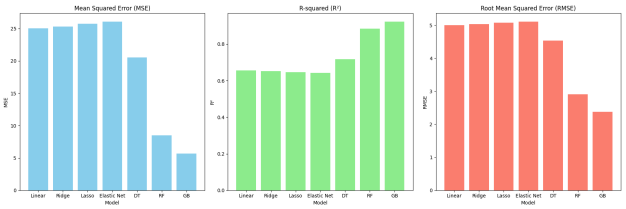

Fig 10.
Benchmarked MSE, R-squared, and RMSE

E. Analysis and Interpretation

Linear models, such as Linear Regression, Ridge Regression, and Lasso Regression, serve as strong baselines but fall short of capturing non-linear interactions. Regularization techniques, including Ridge and Lasso, do not provide significant improvements over simple Linear Regression, suggesting that the dataset does not exhibit severe multicollinearity.

The non-linear nature of Decision Tree Regression enables it to outperform linear models, as evidenced by lower MSE and higher R^2 values. However, Decision Trees are prone to overfitting, which can limit their generalizability.

Ensemble methods outperform all other models, including Random Forest Regression and Gradient Boosting Regression. Random Forest leverages bagging to reduce overfitting, achieving an MSE of 8.48 and an R^2 of 0.883. Gradient Boosting Regression, the best-performing model, refines predictions iteratively by correcting residual errors from prior iterations. This method achieves an MSE of 5.67 and an R^2 of 0.922, demonstrating its superior ability to model complex relationships.

After further testing, by removing the INDUS feature that has been stated, III.C. provided the model with better overall performance for all algorithms, with the highest being from Gradient Boosting with an MSE of 5.26 and 0.9276 R^2.

Model	**MSE**	**R^2**
Linear Regression	25.11	0.6548
Ridge Regression	25.51	0.5491
Lasso Regression	25.86	0.6444
Elastic Net Regression	26.40	0.637
Decision Tree Regression	8.45	0.8838
Random Forrest Regression	8.07	0.889
Gradient Boosting	5.26	0.9276

Table 2.
MSE and R^2 values of the updated feature set

The refined feature set and advanced models demonstrate the importance of feature selection in improving predictive performance. Gradient Boosting Regression was identified as the optimal model, offering the best balance of accuracy and complexity.

IV. CONCLUSIONS

This study examined the application of various regression techniques for predicting Boston housing prices, emphasizing the significance of feature selection and model evaluation in improving predictive accuracy. The findings indicate that certain variables, such as the average number of rooms per dwelling (RM) and the percentage of the lower-status population (LSTAT), are among the most influential factors in determining housing prices. In contrast, features like the proportion of non-retail business acres (INDUS) and residential land zoning (ZN) contributed minimally to model performance. Implementing a data-driven feature selection process based on feature importance thresholds streamlined the dataset, enhancing both model interpretability and computational efficiency.

Among the regression models evaluated, linear approaches such as Ridge, Lasso, and Elastic Net served as effective baseline methods but struggled to capture the dataset's inherent non-linear relationships. In contrast, tree-based models, including Decision Tree and Random Forest Regression, significantly improved predictive accuracy due to their ability to model complex feature interactions. The Gradient Boosting Regressor emerged as the most effective approach, achieving the lowest Mean Squared Error (MSE) and the highest R-squared (R^2) value. Its iterative refinement of predictions allowed it to capture intricate patterns in the data, ultimately reaching an R^2 of 0.9276 with the refined feature set.

Furthermore, this study highlighted the critical role of data preprocessing techniques in optimizing model performance, such as outlier removal, feature scaling, and the appropriate handling of categorical variables. The exclusion of weakly contributing or redundant features resulted in improved accuracy across all regression models, underscoring the importance of robust feature engineering in predictive modeling.

These findings demonstrate the effectiveness of ensemble-based methods, particularly Gradient Boosting, in handling complex, structured datasets for predictive modeling tasks. The success of this approach suggests that machine learning can serve as a powerful tool for real estate price prediction, offering valuable insights for data-driven decision-making. Future research may further optimize model hyperparameters, integrate additional data sources to enhance prediction robustness, and explore alternative algorithms to improve generalization and scalability.

In conclusion, this study underscores the importance of integrating advanced preprocessing techniques, feature selection strategies, and machine learning models to develop accurate and interpretable predictive frameworks. These insights contribute to the broader field of real estate analytics and reinforce the potential of data-driven approaches in addressing complex, real-world problems.

V. CITATIONS

GeeksforGeeks. (n.d.). *Hyperparameter tuning*. Retrieved from https://www.geeksforgeeks.org/hyperparameter-tuning/

GeeksforGeeks. (n.d.). *ML | Gradient Boosting*. Retrieved from https://www.geeksforgeeks.org/ml-gradient-boosting/

Kwak, S. K., & Kim, J. H. (2017). Statistical data preparation: Management of missing values and outliers. *Korean Journal of Anesthesiology, 70*(4), 407–411. https://doi.org/10.4097/kjae.2017.70.4.407

Scikit-learn contributors. (n.d.). Supervised learning–Regression (documentation). Retrieved from https://scikit-learn.org/stable/supervised_learning.html