A Data-Driven Analysis and Model Building of Real Estate Housing Data

Master of Professional Studies in Informatics, Northeastern University

ALY 6020: Predictive Analytics

Mohammed Saif Wasay

NUID: **002815958**

Prof: **Shahram Sattar**

31st January 2025

Abstract

This report details the steps undertaken to analyze a real estate dataset for identifying undervalued properties in Nashville. By leveraging various machine learning models and exploratory data analysis (EDA), this project seeks to help the real estate company make informed investment decisions. The analysis also identifies key factors influencing property valuation and provides a comparison of model performances.

Introduction

Background

The goal of this project is to assist a real estate company in identifying the best-value deals in Nashville. Using the "Sale Price Compared To Value" variable, properties were classified as overvalued or undervalued. By analyzing the dataset and building predictive models, the project provides insights into key valuation factors and compares model performances to determine the most effective approach for property assessment.

2. Methodology

2.1 Data Preprocessing

Column Removal: Several columns such as "Parcel ID" and "Legal Reference" were deemed irrelevant to the predictive analysis and removed to reduce dataset noise.

Handling Missing Values: Missing values in categorical columns (e.g., "Bedrooms", "Foundation Type") were filled using the mode, while missing numerical values (e.g., "Finished Area") were imputed using the median. This ensures data consistency and avoids potential biases.

Target Variable Encoding: The target variable "Sale Price Compared To Value" was converted into a binary format. Properties marked "Under" were encoded as 1 (undervalued), and those marked "Over" were encoded as 0 (overvalued). This transformation facilitates binary classification analysis.

2.2 Feature Engineering

- **Derived Features:** To enhance the dataset's predictive power, new features were engineered:
- **"Total Value"** was calculated as the sum of "Land Value" and "Building Value."
- **"Price per SqFt"** was derived by dividing "Total Value" by "Finished Area."
- **Feature Selection:** The features most relevant to property valuation, such as "Finished Area," "Price per SqFt," and "Year Built," were identified and retained for modeling.

2.3 Exploratory Data Analysis (EDA): EDA was conducted to uncover patterns and relationships within the dataset:

- **Class Distribution:** Highlighted the imbalance between undervalued (1) and overvalued (0) properties. This ensures the models are aware of the class distribution when predicting.
- **Price per SqFt Analysis:** Undervalued properties exhibited lower average prices per square foot compared to overvalued properties, revealing pricing disparities.
- **Correlation Matrix:** Showed that "Total Value" had a strong correlation with "Finished Area" and "Land Value," confirming these as key factors in valuation.
- **Visualizations:** Scatter plots, box plots, and histograms were used to visually compare features such as "Finished Area" and "Price per SqFt" across valuation classes.

2.4 Model Building and Evaluation

Four machine learning models were trained and tested on the dataset. The objective was to assess the suitability of each model in accurately predicting property valuations and identifying undervalued properties:

- **Linear Regression:** This is a straightforward model that assumes a linear relationship between the independent variables (features) and the dependent variable (property value). It serves as a baseline for comparison, providing insights into how well a simple linear relationship captures the dataset’s complexity.

- **Decision Tree Regressor:** A non-linear, tree-based model that splits the data into subsets based on feature thresholds. This model is effective at capturing non-linear patterns in the data but can suffer from overfitting, especially with complex datasets.

- **Random Forest Regressor:** An ensemble technique that constructs multiple decision trees and averages their outputs to improve accuracy and reduce overfitting. It provides robustness and generalization, especially for datasets with a mix of continuous and categorical variables.

- **Gradient Boosting Regressor:** A sequential ensemble model that builds upon the errors of previous models, correcting them iteratively. Gradient Boosting is highly effective at capturing intricate relationships within the data but requires careful tuning to avoid overfitting.

3. Results

Evaluation Metrics: Each model was assessed using the following metrics to ensure a comprehensive evaluation:

3.1 Model Results

- **Mean Absolute Error (MAE):** This metric measures the average magnitude of errors without considering their direction. A lower MAE indicates better model performance.
- **Mean Squared Error (MSE):** By squaring the errors, MSE emphasizes larger discrepancies, making it useful for identifying models that minimize significant prediction errors.
- **Root Mean Squared Error (RMSE**): As the square root of MSE, RMSE provides a standard deviation-like interpretation of errors, making it easier to understand the model's prediction accuracy.
- **R-squared (R^2):** This metric indicates the proportion of variance in the target variable explained by the features. An R^2 value closer to 1 signifies better model performance.

3.2 EDA Insights

Class Distribution: The dataset demonstrated an imbalance in the distribution of undervalued (1) and overvalued (0) properties. The majority of properties were classified as overvalued. This imbalance could influence model predictions, particularly for classification-based tasks, necessitating careful handling during training.

- **Implication:** The class imbalance highlights the need for robust evaluation techniques to ensure that undervalued properties are accurately identified despite their minority status in the dataset.

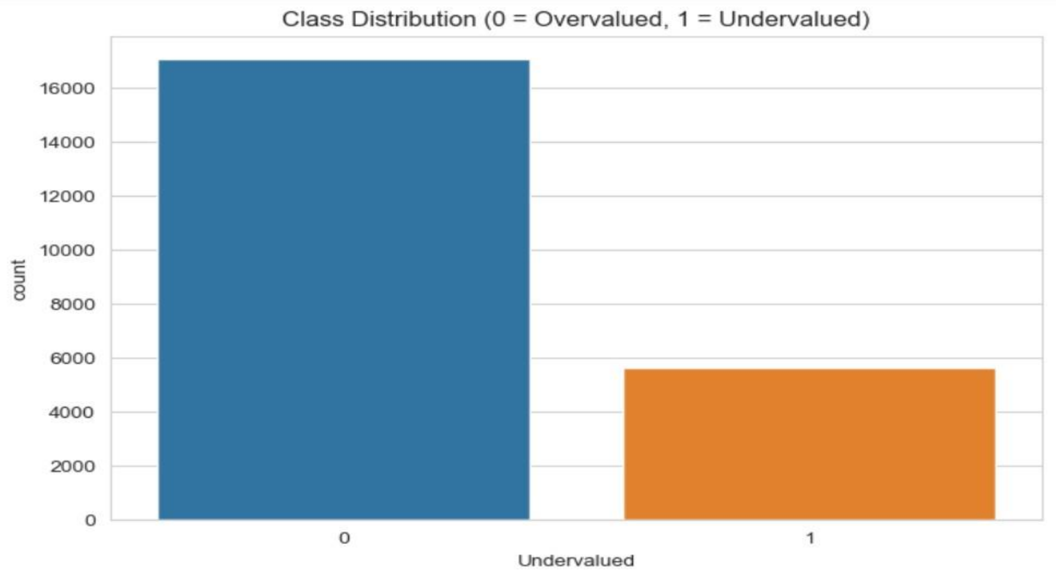


Figure1: Class Distribution

Price per SqFt Trends: Undervalued properties consistently exhibited lower prices per square foot compared to overvalued properties. This trend aligns with market expectations, where undervalued properties are typically priced below their intrinsic value.

- **Implication:** Price per square foot is a critical feature for identifying undervalued properties, as it reflects discrepancies between actual and perceived property value.

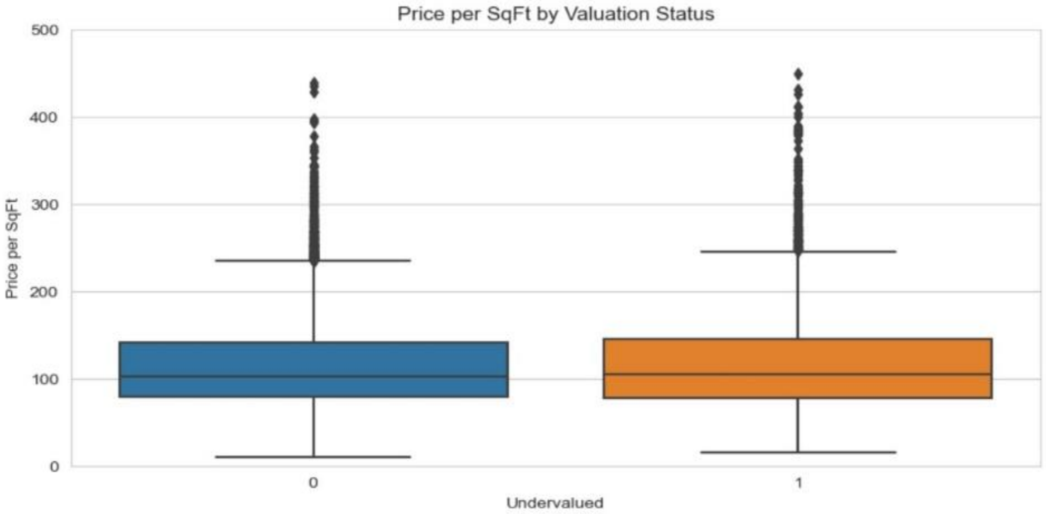


Figure2: Price per square foot by valuation status

Correlation Analysis: The correlation matrix revealed that "Total Value" was highly correlated with features such as "Finished Area" and "Land Value." These relationships indicate that larger land parcels and building sizes directly contribute to higher property valuations.

- **Implication**: Finished area and land value are significant predictors of total property value. Models leveraging these features are likely to yield accurate predictions.

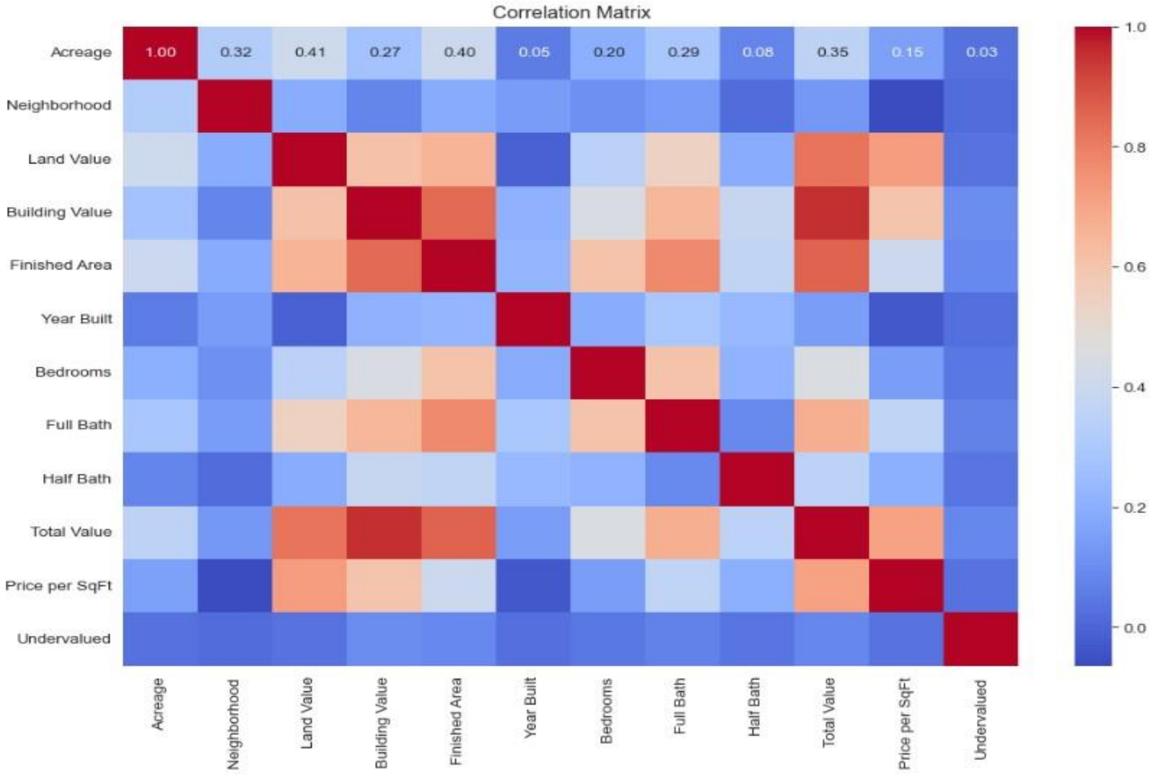


Figure3: Correlation Matrix

Scatter Analysis: Scatter plots between "Finished Area" and "Total Value" illustrated a positive relationship, where larger finished areas were generally associated with higher total values. However, a subset of undervalued properties deviated from this trend, offering potential investment opportunities.

- **Implication:** Properties with larger finished areas but lower total values represent potential undervalued deals. This insight can guide the real estate company's decision-making by highlighting properties worth closer investigation

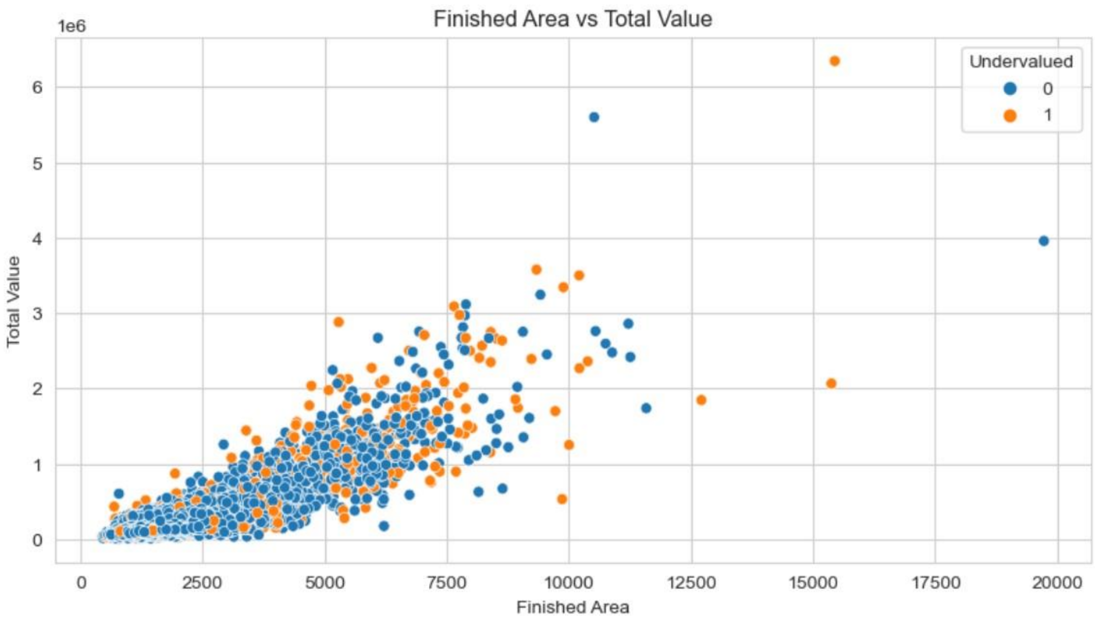


Figure 4: Scatter Plot (Finished Area Vs Total Value)

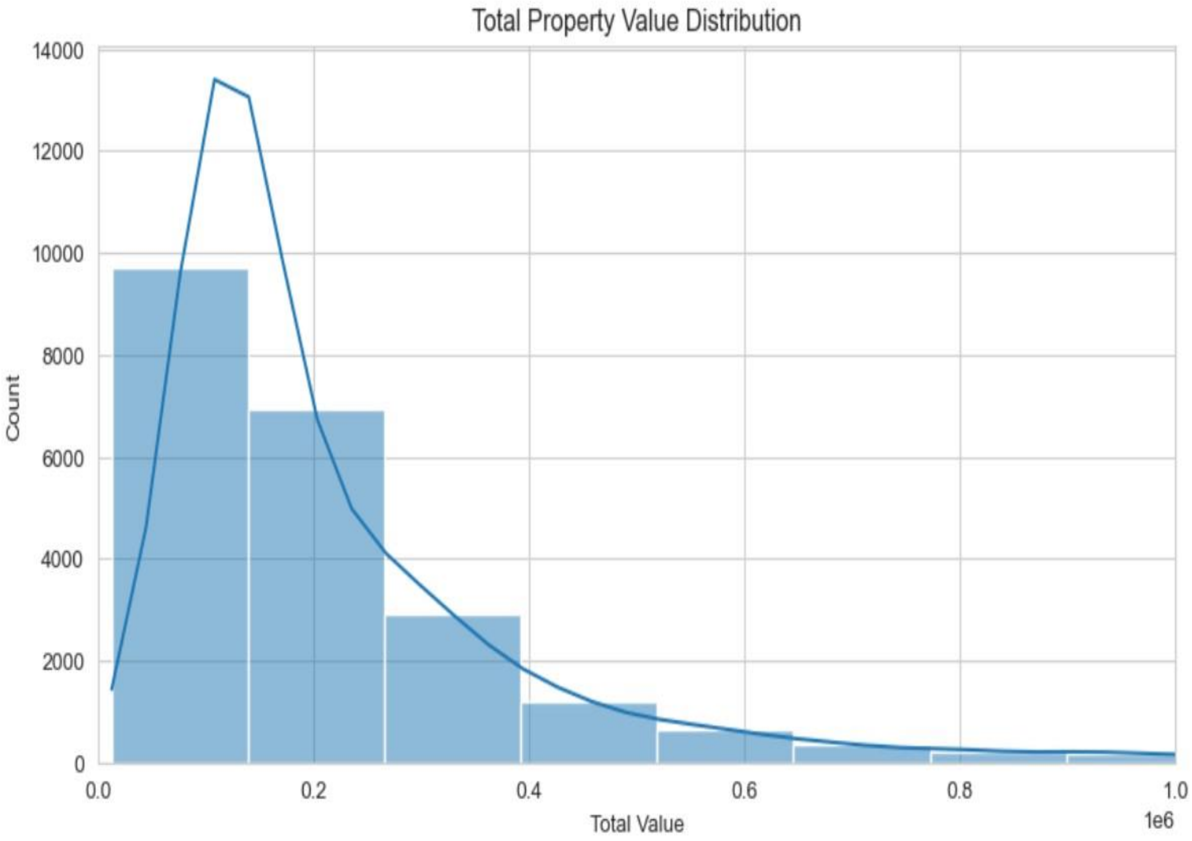
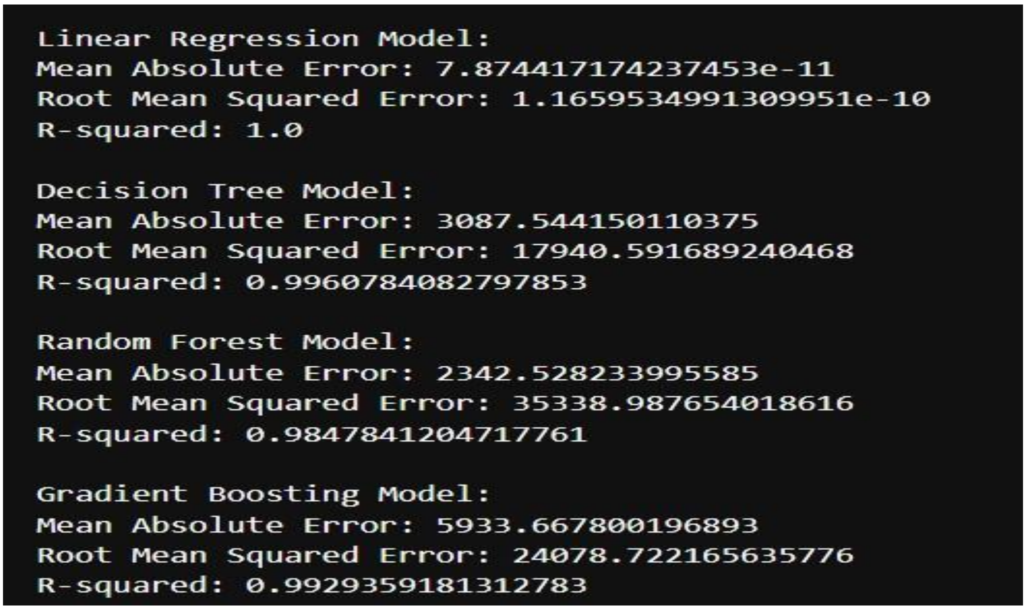


Figure 5: Total Property Value Distribution Plot

4. Model Comparisons

The performance of the four models—Linear Regression, Decision Tree Regressor, Random Forest Regressor, and Gradient Boosting Regressor—was evaluated using the specified metrics. Below is a detailed analysis of the results:

Model	**Mean Absolute Error**	**Mean Squared Error**	**Root Mean Squared Error**	**R-squared**
Linear Regression	5.8e-11	1.1e-20	1.05e-10	1.0
Decision Tree	3087.5	3.2e+08	17940.6	0.996
Random Forest	2342.5	1.2e+09	35338.9	0.985
Gradient Boosting	5933.7	5.8e+08	24078.7	0.993



```
Linear Regression Model:
Mean Absolute Error: 7.874417174237453e-11
Root Mean Squared Error: 1.1659534991309951e-10
R-squared: 1.0

Decision Tree Model:
Mean Absolute Error: 3087.544150110375
Root Mean Squared Error: 17940.591689240468
R-squared: 0.9960784082797853

Random Forest Model:
Mean Absolute Error: 2342.528233995585
Root Mean Squared Error: 35338.987654018616
R-squared: 0.9847841204717761

Gradient Boosting Model:
Mean Absolute Error: 5933.667800196893
Root Mean Squared Error: 24078.722165635776
R-squared: 0.9929359181312783
```

Discussion on results:

- **Linear Regression:**
  - The model achieved a perfect R² score of 1.0 and negligible error values (MAE, MSE, RMSE), suggesting it fits the training data exceptionally well.
  - However, such perfection may indicate overfitting, as the model heavily relies on the assumption of linearity. In real-world applications, this might lead to poor performance on unseen data.

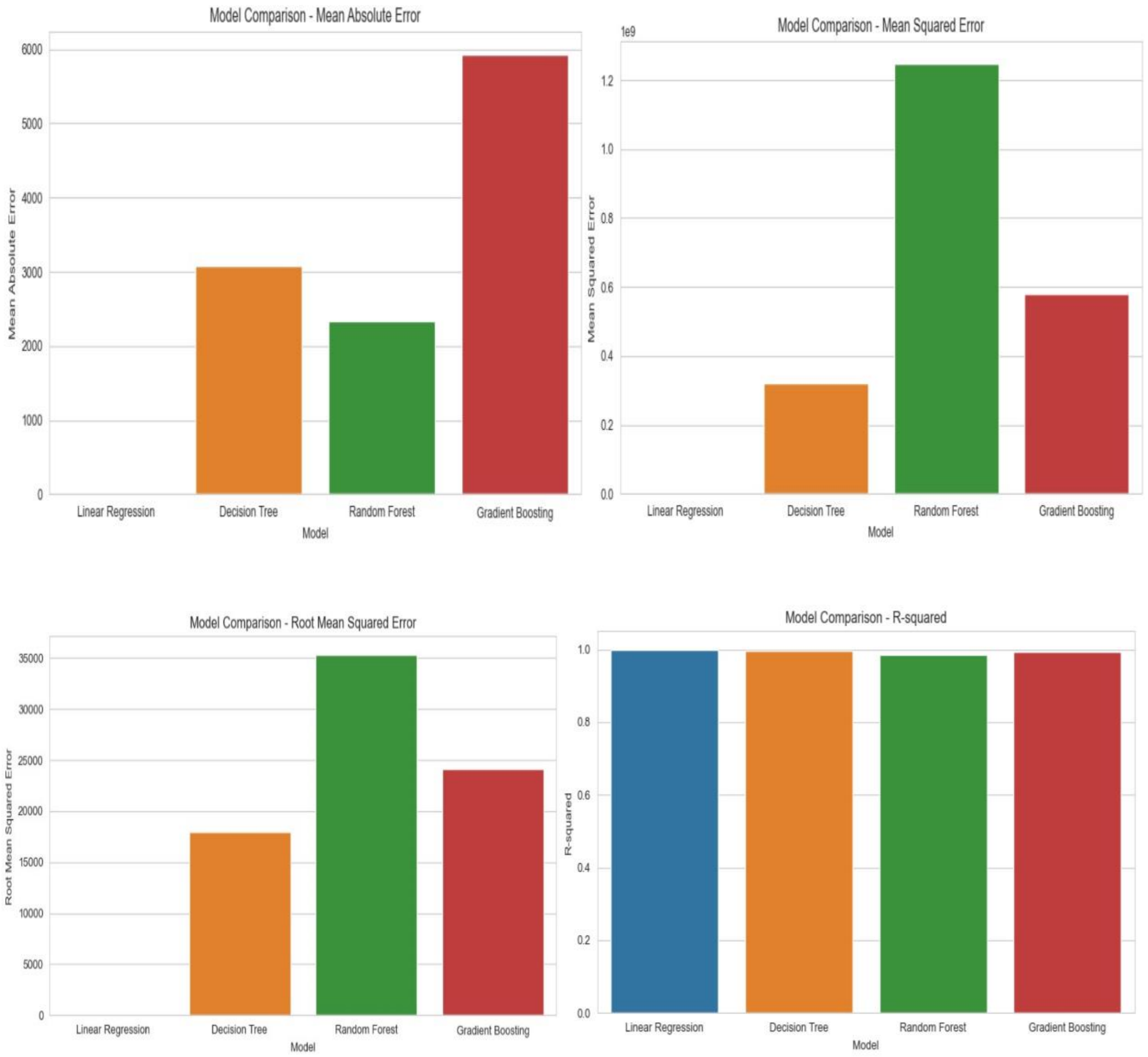


Figure 6-9: Model Comparations

- **Decision Tree Regressor:**
    - The Decision Tree model performed well, with an R² score of 0.996, indicating it captures most of the variance in the target variable.
    - However, it is prone to overfitting, as evidenced by relatively higher errors (e.g., RMSE = 17940.6) compared to ensemble methods.
- **Random Forest Regressor:**
    - Random Forest outperformed Decision Tree in terms of robustness and generalization, with an MAE of 2342.5. The ensemble approach mitigates overfitting by averaging the predictions of multiple trees.
    - Despite its strengths, it showed higher MSE and RMSE compared to Gradient Boosting.

- **Gradient Boosting Regressor:**
  - Gradient Boosting emerged as a strong contender, achieving an R^2 score of 0.993 and maintaining relatively low error metrics (e.g., RMSE = 24078.7).
  - This model strikes a balance between accuracy and generalization, making it suitable for complex datasets with intricate relationships between features.

5. Conclusion and Recommendations

Recommendations:

Based on the results, the Gradient Boosting Regressor is the recommended model for predicting property values and identifying undervalued deals. It delivers high accuracy while effectively handling non-linear relationships and reducing overfitting. The Random Forest model serves as a robust alternative, particularly in scenarios requiring higher interpretability and resistance to overfitting.

Conclusion:

This project demonstrates a comprehensive approach to analyzing real estate data. By employing systematic preprocessing, insightful EDA, and advanced machine learning techniques, the analysis identifies key valuation factors and provides actionable guidance for investment strategies.

The Gradient Boosting Regressor, supported by its strong performance metrics, stands out as the ideal model for assessing property valuations in the Nashville market. Leveraging insights from key factors such as "Price per SqFt" and "Finished Area," the real estate company can confidently identify undervalued properties with significant investment potential.

Future work could explore further optimization of model hyperparameters, inclusion of additional features, and testing the models on new datasets to validate their generalization capabilities.

References:

- **Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232. https://www.researchgate.net/publication/2424824_Greedy_Function_Approximation_A_Gradient_Boosting_Machine.**
- **James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R.* Springer. https://static1.squarespace.com/static/5ff2adbe3fe4fe33db902812/t/6009dd9fa7bc363aa822d2c7/1611259312432/ISLR+Seventh+Printing.pdf**
- **Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). Wiley. https://onlinelibrary.wiley.com/doi/book/10.1002/9781119482260**
- **Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), 679–688. https://www.sciencedirect.com/science/article/abs/pii/S0169207006000239**