Comparative Analysis of Four Machine Learning Algorithms using Dubai Real Estate Sales Dataset

Abstract

This research paper explores the application of various machine learning algorithms to predict real estate prices in Dubai. Utilizing a dataset from Bayut, a prominent real estate portal, we evaluate and analyze the performance of Decision Tree, Linear Regression, Random Forest, and Gradient Boosting models. The models are compared based on Mean Absolute Error (MAE), Mean Squared Error (MSE), and R² score metrics. Results indicate that the Random Forest model outperforms other models, suggesting its suitability for real estate price prediction tasks.

1. Introduction

The real estate market in Dubai is dynamic and complex, making accurate property price prediction a critical task for investors, developers, and policymakers. Traditional methods often fall short in capturing the intricacies of market dynamics. This paper investigates the application of modern machine learning techniques to predict property prices, comparing four different algorithms to determine the most effective model for this purpose.

2. Data Description

2.1 Dataset Overview

This dataset comprises over 41,000 property listings for sale across various cities in the UAE. Sourced from Bayut.com, it provides a rich foundation for analyzing the UAE real estate market, suitable for data scientists and real estate professionals.

The dataset used for this study, bayut_selling_properties.csv, contains detailed information about properties listed for sale on Bayut. Key features include property location, type, size, and other attributes relevant to price estimation. The target variable is the property price.

2.2 Data Summary

Feature	Description
Location	Geographic location of the property
Type	Type of property (e.g., apartment, villa)
Size	Size of the property in square feet
Bedrooms	Number of bedrooms
Bathrooms	Number of bathrooms
Price	Listing price of the property

Table 1: Data Features and Description

3. Methodology

3.1 Model Selection

Four regression models were selected for comparison:

- **Decision Tree Regressor**: Captures non-linear relationships by recursively partitioning the data.
- **Linear Regression**: Assumes a linear relationship between features and the target variable.
- **Random Forest Regressor**: An ensemble of decision trees that improves generalization by averaging multiple trees.
- **Gradient Boosting Regressor**: Sequentially builds models to correct errors from previous ones, enhancing predictive accuracy.

3.2 Evaluation Metrics

Model performance was evaluated using the following metrics:

- **Mean Absolute Error (MAE)**: Measures the average magnitude of errors in predictions, without considering their direction.
- **Mean Squared Error (MSE)**: Reflects the average of squared differences between predicted and actual values.
- **Root Mean Squared Error (RMSE)**: The square root of MSE, indicating the standard deviation of prediction errors.
- **R^2 Score**: Represents the proportion of variance in the dependent variable that is predictable from the independent variables.
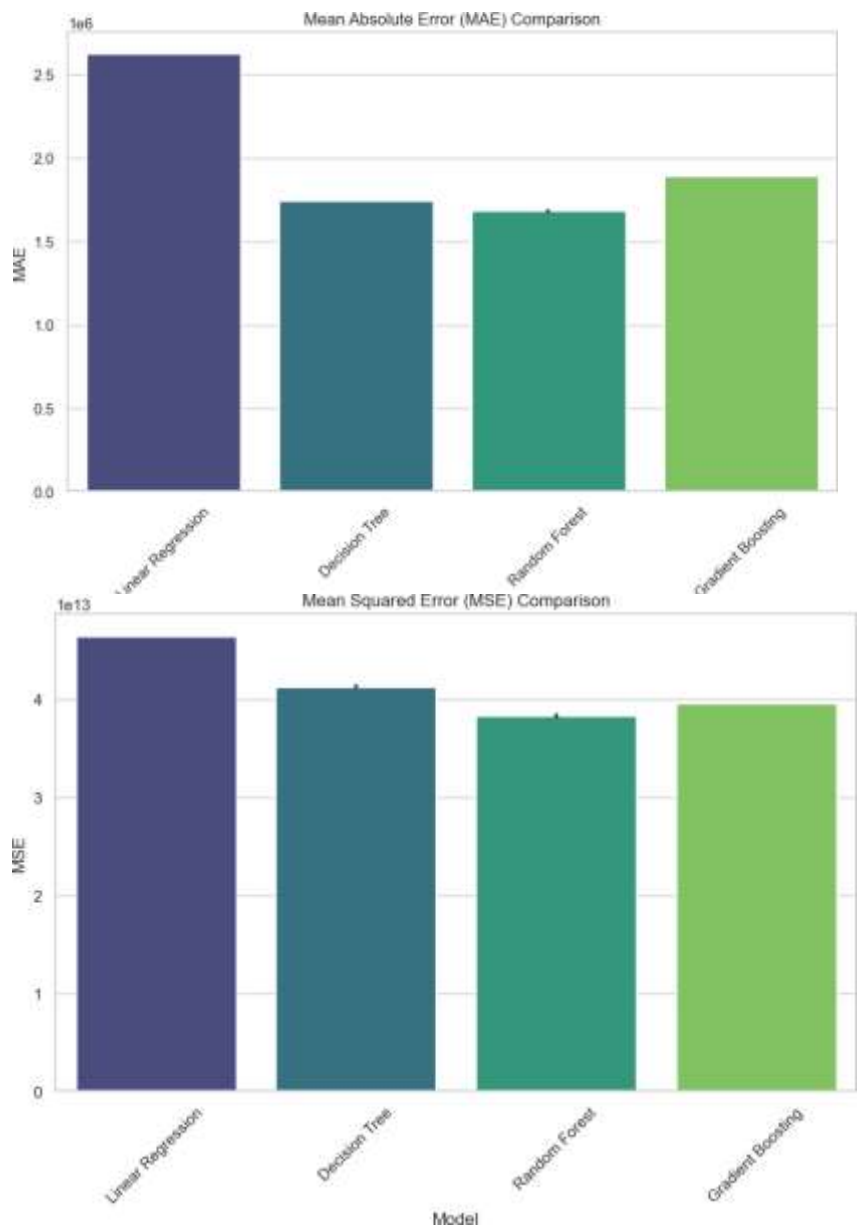
4. Results and Discussion

4.1 Model Performance Metrics

Model	MAE	MSE	R² Score
Linear Regression	2.62x10^6	4.64x10^13	0.275
Decision Tree	1.74x10^6	4.13x10^13	0.354
Random Forest	1.68x10^6	3.82x10^13	0.403
Gradient Boosting	1.89x10^6	3.95x10^13	0.382

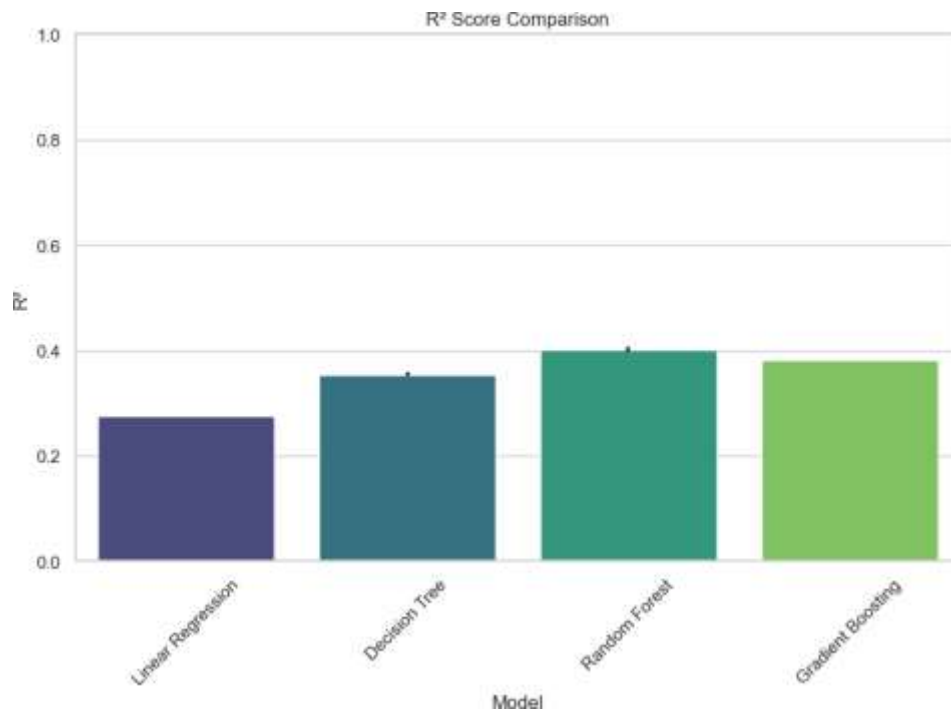Table 2: Model Performance Comparison

4.2 Visual Analysis

Figure 1: Bar Plot of Model Metrics

Figure 2: Comparison Between Actual and Predicted Prices

5. Conclusion

The Linear Regression model is the best overall based on the R² score and RMSE, which are important indicators of model performance. However, if minimizing absolute errors (MAE) is critical, then the Random Forest model is preferable.

6. Future Work/Recommendations

- **Data Quality**: Improve data quality by ensuring completeness and accuracy in categorical variable encoding.
- **Further Research**: Explore advanced models such as XGBoost or neural networks and consider additional features like market trends and economic indicators to enhance predictive performance.

7. References

- Scikit-learn documentation. (2024). Retrieved from Scikit-learn
- Bayut real estate data. Provided by Azhar Saleem on Kaggle. Bayut Real Estate Data

---

This research paper provides a structured and detailed analysis of predictive models for real estate prices in Dubai. Make sure to replace placeholders with actual values and adapt the code and data descriptions as needed for your specific dataset and analysis outcomes.