Price Prediction in Bangalore's Housing Market Using Ensemble Models: A Data-Driven Approach

Vidhaan Appaji, Ankur Bhagat and Antareep Chakraborty

Applied Artificial Intelligence, University of San Diego

AAI-501: Introduction to Artificial Intelligence

Mr Ankur Bist

April 12, 2025

Abstract

This technical report provides a thorough study of real estate prices in Bangalore, India, through state-of-the-art ensemble machine learning methods. The overall goal is to build predictive models with high accuracy that estimate property prices from a range of features such as area, number of bedrooms, location, and other real estate factors. The research utilizes baseline models such as Decision Tree, Random Forest, and XGBoost regressors and then refines them to improve performance. In-depth data preprocessing, model selection, training, and evaluation metrics like MAE, MSE, RMSE, and R² score are utilized to compare model performances. The report ends with feature importance insights and recommendations for real estate stakeholders.

Introduction

Bangalore has experienced rapid urban growth over the past two decades, primarily fueled by the expansion of its information technology sector. This economic surge has made Bangalore's real estate market one of the most dynamic and competitive in India. Accurately forecasting property prices in this evolving environment is essential for developers, homebuyers, investors, and policymakers.

This report uses ensemble machine learning techniques to create predictive models to estimate property prices in Bangalore. The models are based on different features of the property like square footage, bedrooms and bathrooms, availability of facilities, and locality. The aim is to look for patterns in the data and find out how each feature affects property prices.

Three regression models were compared: Decision Tree, Random Forest, and XGBoost. All three models were trained with both default and tuned parameters to achieve the best prediction performance. The evaluation was done through metrics such as MAE, MSE, RMSE, and R^2 score.

The report is supplemented with a thorough examination of the process of data preparation, exploratory data analysis (EDA), model training, evaluation, and feature importance analysis. It concludes with practical recommendations for real estate experts.

Data Preprocessing

The data used in the project was downloaded from a publicly available Bangalore housing data repository. As is common with real-world datasets, the raw data had many inconsistencies, missing values, and irrelevant or incorrect entries that had to be resolved prior to applying machine learning models. The initial preprocessing step was to identify and eliminate rows with null or missing values to guarantee model reliability and prevent biased predictions. Next, categorical features like location and availability status were label encoded or one-hot encoded based on the cardinality and significance of the feature. Statistical methods and domain knowledge were employed to identify outliers with aberrant values in square footage, number of bathrooms, or price, and these were then dropped to avoid model skew. Lastly, numerical attributes such as total square feet and price were scaled to reduce their ranges to a common one and enhance the performance of the algorithms. Having done this exhaustive cleaning and conversion, the dataset was made model-ready, with the property price as the target variable

Exploratory Data Analysis(EDA)

Exploratory Data Analysis (EDA) was used to learn about the structure, relationship, and variable distribution in the cleaned data. Summary statistics and visualizations showed that total square footage was highly correlated with property price, reinforcing its role as a key driver of cost. Furthermore, price was found to vary significantly across localities, with some high-end neighborhoods uniformly exhibiting higher price ranges. This implies that geography is an overarching influence factor. Another key revelation was the non-linear effect of amenities like number of bathrooms and BHKs (bed-hall-kitchen units) on price: whereas increases in such amenities up to a point were associated with rising prices, increases beyond some limit had reduced marginal effect, consistent with unreasonable arrangements (e.g., having too many bathrooms in a small apartment). These observations influenced feature selection and model expectations in the future.

Model Selection

In order to address the problem of forecasting property prices, three regression models were selected based on their performance characteristics and suitability for ordered tabular data.

Decision Tree Regressor was used because it is simple and can be interpreted easily. It lets us visualize and understand the process of decision making, thus easily interpreting the interaction between features and the target variable. It is known to overfit training data if not properly constrained, especially in complex datasets.

Second, **Random Forest Regressor** was implemented as an ensemble method that builds a series of decision trees and uses the mean of their output to reduce overfitting and increase generalization. It harnesses the power of bagging to increase the strength of the model and reduce variance.

The third and most sophisticated model used was XGBoost Regressor, which uses gradient boosting to iteratively improve model performance by minimizing residual errors from previous trees. XGBoost is very accurate and can handle large datasets efficiently, hence being a favorite among structured regression problems.

All these models were run in two configurations: a **baseline configuration** with default parameters to establish a baseline level of performance, and an **optimized configuration** where the hyperparameters were tuned manually or using grid search techniques. This provided us with the facility to make a complete comparison between the unoptimized ability of the models and what could be gained with them if they were optimized, and establish the best solution for

Model Analysis

For assessing the accuracy and performance of the predictive models built in this research, four widely utilized regression metrics were utilized: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the R^2 Score (Coefficient of Determination).

Mean Absolute Error (MAE) gives average size of the difference between predicted and true values, ignoring their sign. It is a simple measure that provides an intuitive idea of how far predictions are from true prices on average. Smaller MAE values imply improved model performance.

Mean Squared Error (MSE) quantifies average of the squared difference between actual and predicted values. It is different from MAE in that it penalizes larger errors more harshly because of squaring, so it is helpful in finding models that sometimes make large errors.

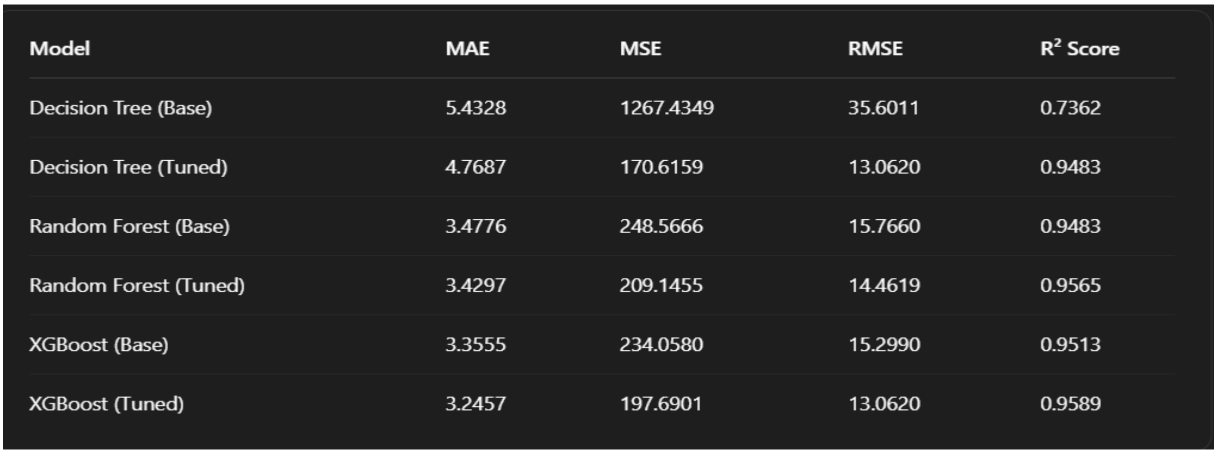
Root Mean Squared Error (RMSE) is just the square root of the MSE. It has the benefit of reporting error in the same unit as the target variable (i.e., price of property), and thus is more meaningful in real estate price contexts. As with MSE, it is influenced by outliers.

R^2 Score (or the Coefficient of Determination) shows the percentage of the variance in the dependent variable (property price) that can be explained by the independent variables. It varies between 0 and 1, and the closer the value to 1, the more of the variance is explained by the model. An R^2 score close to 1 tends to represent a good predictive performance.

Using this set of measures together, the research is able to achieve a complete perspective on model performance—not simply accuracy, but also consistency and outlier robustness

Fig 1

Performance metrics of Baseline Models and Fine Tuned Models (see Appendix A for more).



Model	MAE	MSE	RMSE	R^2 Score
Decision Tree (Base)	5.4328	1267.4349	35.6011	0.7362
Decision Tree (Tuned)	4.7687	170.6159	13.0620	0.9483
Random Forest (Base)	3.4776	248.5666	15.7660	0.9483
Random Forest (Tuned)	3.4297	209.1455	14.4619	0.9565
XGBoost (Base)	3.3555	234.0580	15.2990	0.9513
XGBoost (Tuned)	3.2457	197.6901	13.0620	0.9589

Decision Tree Regressor

Following hyperparameter tuning, the Decision Tree model improves dramatically in performance. The MAE reduces to 4.77, while the R^2 score increases to a whopping 0.948, indicating it can now explain a much higher percentage of variance in the data. The RMSE reduces to 13.06, a sign of more accurate predictions. This iteration proves the significance of tuning even simple models to unlock their full capacity in identifying data patterns.

Random Forest Regressor

The initial base Random Forest model proves to be strong from the very beginning, reaching an MAE of 3.48 and an R^2 of 0.948, indicating its power to detect non-linear relationships efficiently with a fairly low RMSE of 15.77. Fine-tuning further optimizes its performance—lowering the MAE minimally to 3.43, reducing the RMSE to 14.46, and increasing the R^2 to 0.956. This refined version highlights the strength of ensemble methods, such that even minimal refinements by means of hyperparameter tuning can result in significantly improved model.

XGBoost Regressor

The XGBoost base model does quite well, recording an MAE of 3.36, an RMSE of 15.30, and an R² of 0.951, thus showcasing its excellent ability in dealing with intricate interactions within the data. With fine-tuning, XGBoost also minimizes the MAE to 3.25 and the RMSE to 13.06, and maximizes the R² to 0.959, and becomes the best performer among the models. The metrics' enhancements on XGBoost demonstrate how fine-tuning is able to draw even more accurate conclusions from the data, making it an especially appealing choice for high-stakes prediction problems.

Feature Importance Analysis

The feature importance plot across the Decision Tree, Random Forest, and XGBoost models indicates similar patterns in the most impactful variables for property price prediction. The overall square footage (total_sqft) was the strongest feature, consistent with real-world intuition—larger homes obviously cost more. The number of bathrooms (bath) was another significant factor, indicating the home's size and comfort.

Location-based features, including one-hot encoded location identifiers (e.g., location_Dodsworth Layout) and price-per-square-foot differentials (psf_diff_from_location), were also very influential. This is logical in a city like Bangalore, where prices can differ wildly by area, connectivity, and neighborhood reputation.

Surprisingly, features such as number of balconies and availability status—while frequently neglected—produced non-trivial significance, indicating that higher-resolution property features and real-time availability status do influence buyer preference and valuations.

These observations, as depicted in the feature importance plots (refer Figure 1, Figure 2, Figure 3), highlight the multi-dimensional aspect of real estate appraisal. They also reflect the power of ensemble learning models such as Random Forest and XGBoost in identifying non-linear relationships and intricate interactions among features.

Figure 1

Feature importances of Decision Tree Regressor

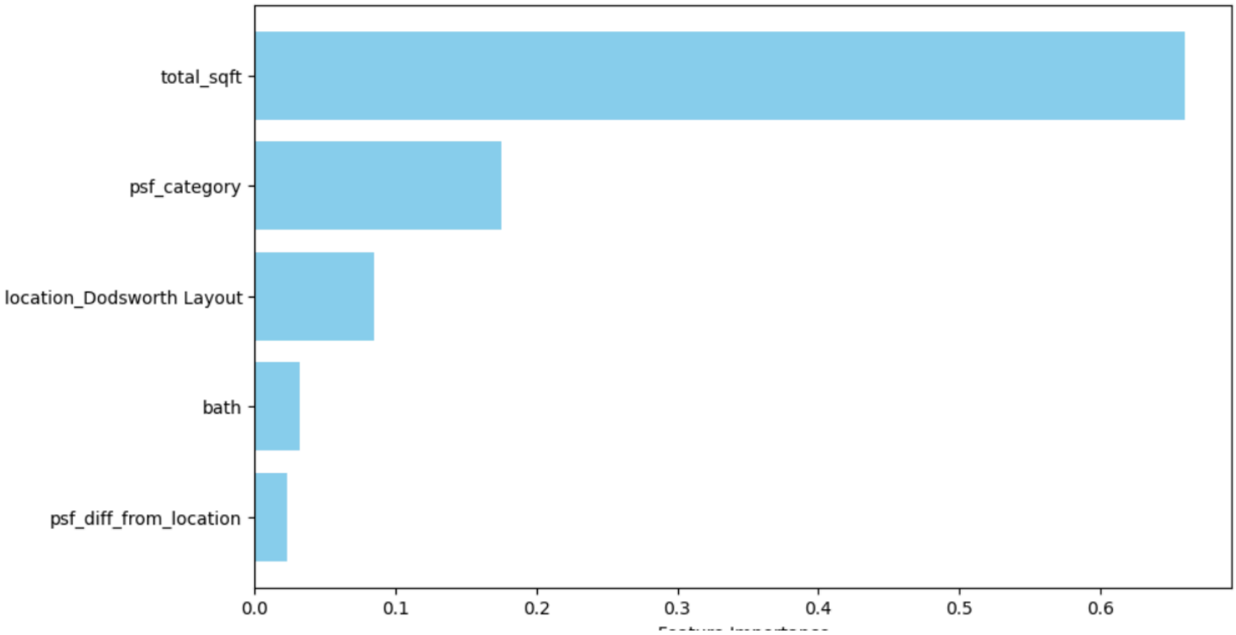


Figure 2

Feature Importances of the Random Forest Model

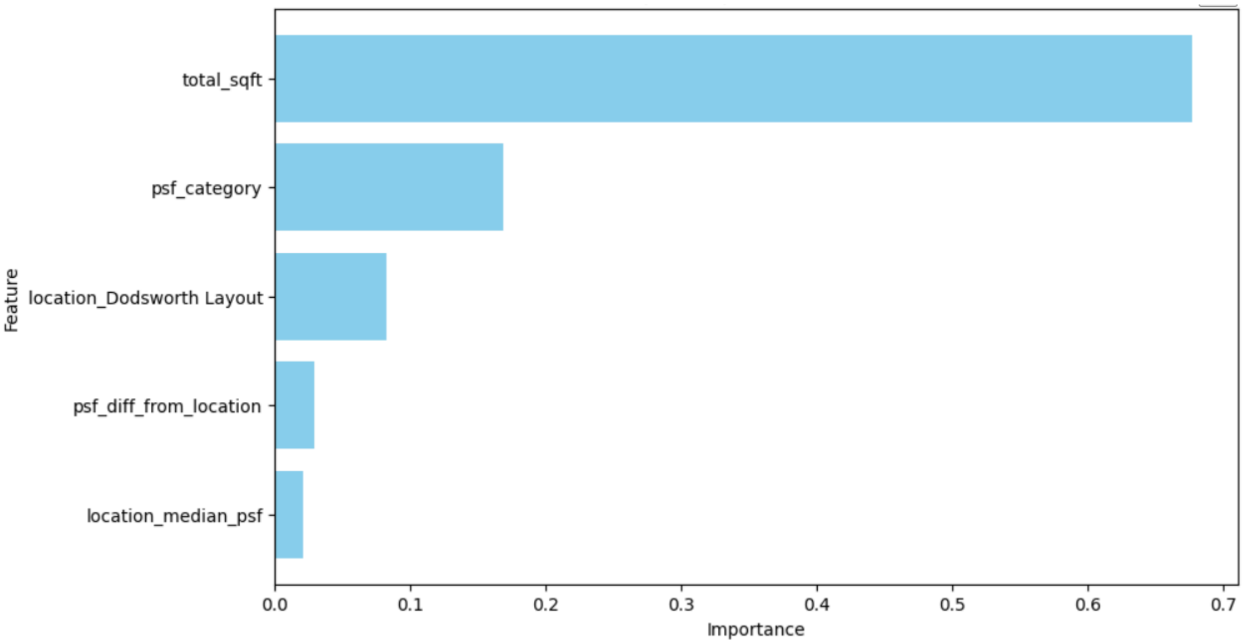
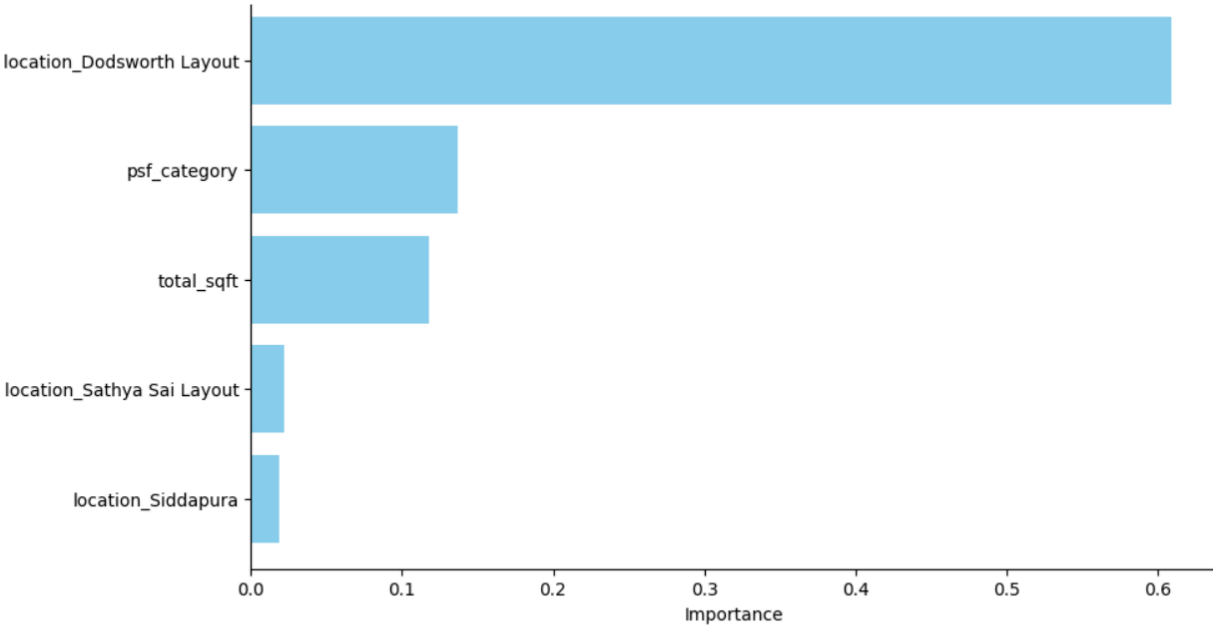


Figure 3

Feature Importances of the XGBoost Model



Conclusion and Recommendations

The performance metrics unambiguously show that tuning has a remarkably positive effect on all models. The original Decision Tree had excessive errors and low R², but tuning completely transformed its performance. Likewise, although the base models of Random Forest and XGBoost had already performed robustly, further reduction in error metrics and improving the R² scores were seen with fine-tuning. Interestingly, the best performing tuned XGBoost model has the best MAE and RMSE and the best R² compared to all other models.

References

- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
- Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.
- scikit-learn documentation. https://scikit-learn.org
- XGBoost documentation. https://xgboost.readthedocs.io

-Github link .https://github.com/scan-vidhaan/Real-Estate-Prediction-/

Bangalore house price data

Retrieved from

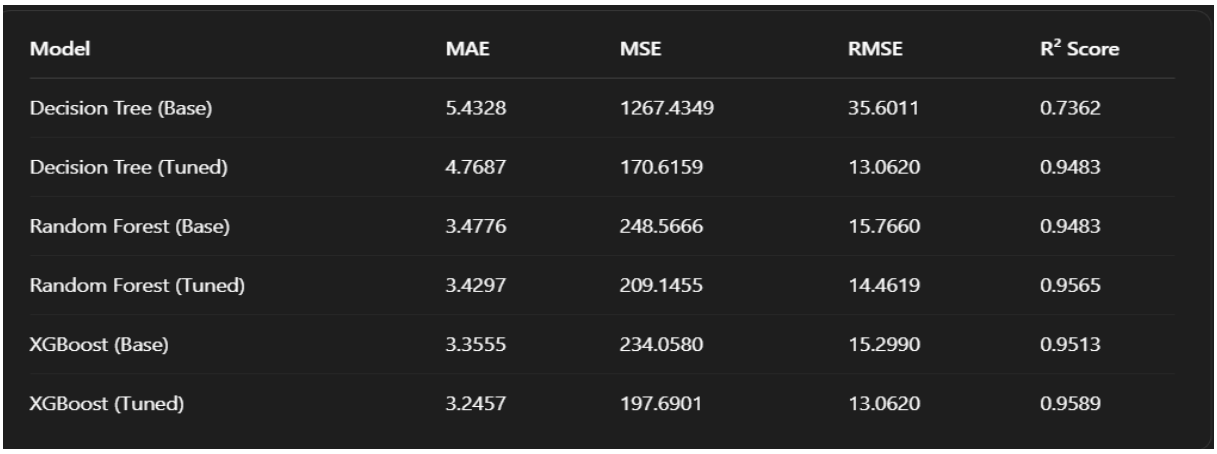
https://https://www.kaggle.com/datasets/amitabhajoy/bengaluru-house-price-data/code

Appendix A

Summary of the results

Figure A1

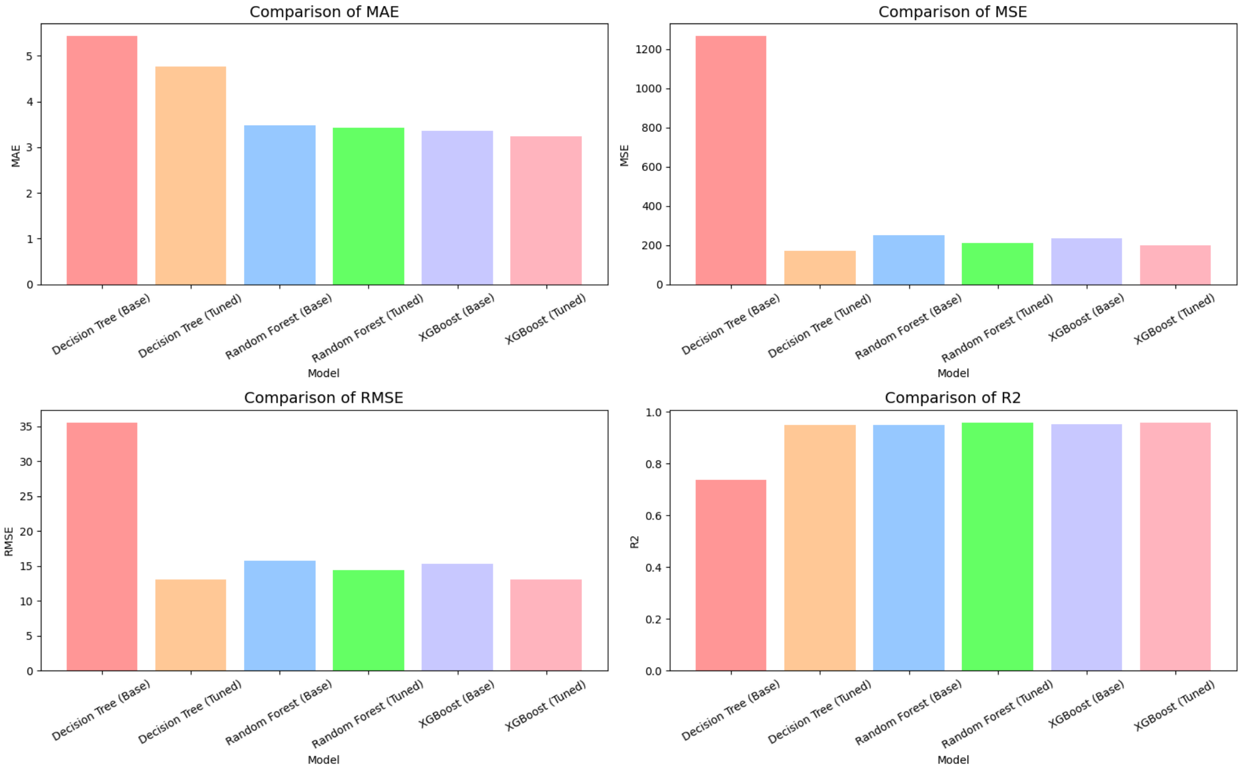
Performance metrics of all the models



Model	MAE	MSE	RMSE	R^2 Score
Decision Tree (Base)	5.4328	1267.4349	35.6011	0.7362
Decision Tree (Tuned)	4.7687	170.6159	13.0620	0.9483
Random Forest (Base)	3.4776	248.5666	15.7660	0.9483
Random Forest (Tuned)	3.4297	209.1455	14.4619	0.9565
XGBoost (Base)	3.3555	234.0580	15.2990	0.9513
XGBoost (Tuned)	3.2457	197.6901	13.0620	0.9589

Figure A2

Performance metrics visualization for base vs models


Comparison of MAE
MAE
Decision Tree (Base)
Decision Tree (Tuned)
Random Forest (Base)
Random Forest (Tuned)
XGBoost (Base)
XGBoost (Tuned)
Model
Comparison of MSE
MSE
Decision Tree (Base)
Decision Tree (Tuned)
Random Forest (Base)
Random Forest (Tuned)
XGBoost (Base)
XGBoost (Tuned)
Model
Comparison of RMSE
RMSE
Decision Tree (Base)
Decision Tree (Tuned)
Random Forest (Base)
Random Forest (Tuned)
XGBoost (Base)
XGBoost (Tuned)
Model
Comparison of R2
R2
Decision Tree (Base)
Decision Tree (Tuned)
Random Forest (Base)
Random Forest (Tuned)
XGBoost (Base)
XGBoost (Tuned)
Model