Real Estate Price Prediction using Machine Learning and Data Analytics

D.V.N Sriram, B. Likhith Kumar Reddy, K. Dinesh Kumar Reddy, Dr. Ramesh S

Department of Networking and Communications Srmist Chennai, India

DOI: https://dx.doi.org/10.51244/IJRSI.2025.1210000225

Received: 20 October 2025; Accepted: 28 October 2025; Published: 15 November 2025

ABSTRACT

In this paper we presents a complete model to predict Real Estate prices with high efficiency through Machine Learning (ML) and Data Analytics approach. The model data is based on the large-scale real estate property data containing structural, locational and environmental elements to become the basics of price variation predictors. We pre-processed, feature engineered and analysed 50,000 Land Registry compliant datasets using a variety of machine learning models - Linear Regression, Random Forest, XGBoost and ANN. Random Forest had the best predective capacity with a Mean Absolute Error (MAE) of 2.63 lakhs and R^2 value of 0.8732 which indicates a high generalisation ability and is very strong. This paper suggests that the challenge of real estate pricing can be addressed by using data-driven analytics, ensemble learning and intelligent feature engineering. The results also indicate the effectiveness of the advanced ML to both the real-world real estate valuation and market forecasting, in addition decision making in property investment.

Keywords — Real Estate, Price Prediction, Machine Learning, Regression, XG Boost, Random Forest, Data Analytics.

INTRODUCTION

Real estate is one of the active and most capital-intensive industries on the global economy. Proper prediction of prices of properties is essential to various stakeholders such as investors, developers, banks, as well as government agencies. Conventional valuation techniques are mostly based on human judgment, experience in a given domain, or simple statistical techniques like either Hedonic Pricing or Multiple Linear Regression. Although these models are useful, most of them fail to portray the non-linear relationships among the features of location quality, development of infrastructure, amenities, and macroeconomic conditions.

As more and more large-scale real estate data is available, and more open data programs are launched, Machine Learning (ML) and Data Analytics have become a successful approach to identifying the unseen links in the property market. The ML algorithms are capable of processing thousands of features at once, discovering patterns that are not seen by human eyes, and changing with market trends.

The primary goal of this study is to come up with a scalable and data-grounded predictive model of real estate prices based on the use of ML methods. The system is designed to process complicated data comprising of geographic, demographic, and structural features and remain accurate and interpretable. The research will also aim at determining the most significant price determinants, which will help the decision-makers make rational decisions about investments and enhance pricing transparency within the housing sector.

The paper is relevant to the field since it combines feature engineering, model optimization, and comparative analysis of various ML algorithms, which make it a comprehensive assessment framework of real estate price prediction.

Related Work

The prediction of the real estate price has significantly changed over the last several decades shifting to more sophisticated machine learning and deep learning paradigms as opposed to traditional econometric methods.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue X October 2025

Previously in the early days, Multiple Linear Regression (MLR) models were usually applied in estimating housing prices using a small number of variables that includes area, number of rooms and distance to the city center [1]. The models however made assumptions of linear relationships and were not effective in modeling complex market dynamics.

Thereafter, there were studies on Decision Trees and Random Forests applied to the valuation of properties, with Park and Bae [2] showing how non-linear dependencies may be well modeled in terms of decision trees. Random Forest was used to show enhancement in prediction by aggregation and reduction of variance in ensembles.

Zhang et al. [3] also investigated the effectiveness of ensemble learning and showed that Gradient Boosting and Random Forest can obtain an R2 score of more than 0.85 when they are trained on a large dataset with thoughtfully-engineered features. Equally, Li et al. [4] added geographic information representation, where spatial coordinates and social-economic indices were employed to enhance the model readability.

Recently, the adoption of XGBoost, which has an improved generalization and computing efficiency, is being adopted [5]. In addition to this, Fu and Chen [6] applied Deep Neural Networks (DNNs) to model interactions between complex features demonstrating better performance at the expense of increased computational demands.

Despite these developments, some key challenges remain that yet need to be resolved - these are data imbalance, interpretability and scalability of the model. The majority of the literature in the field only deals with the concept of algorithmic innovation and ignores the fact that important processes in the data preprocessing and feature engineering are involved. Recent research has a data-based method that possesses high ML modelling and systematic evaluations in an attempt to fill these gaps.

PROPOSED METHODOLOGY

A. Dataset Characteristics and Acquisition

Data for this project will consist of 50,000 property records and will be obtained from publicly accessible sources like Kaggle's House Prices: Advanced Regression Techniques dataset, real estate listing sites, and local market databases. Each record contains the property, spatial and market attributes. Also, the qualitative and quantitative features have been noted. The critical attributes contain:

1. Property ID (unique identifier)
2. Geographic coordinates (latitude and longitude)
3. City and locality information
4. Area in square feet
5. Number of bedrooms and bathrooms
6. Floor number and total floors
7. Age of the property (in years)
8. Amenities (garden, parking, lift, swimming pool, etc.)
9. Distance to nearest school, hospital, and metro station
10. Market-listed price (target variable)

Data comprises a broad range of properties located in metropolitan, suburban, and semi-urban regions which is important for providing diversity for model training.

B. Data Preprocessing and Quality Assurance

Ensuring precision throughout each stage of data preprocessing ultimately reflects on the model's performance. For this reason, the following steps were taken:

1. Addressing Missing Data: Numerical data gaps were addressed with median imputation and categorical data gaps with mode imputation (i.e. filling gaps in city, amenities).

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue X October 2025

2. Outlier Detection: The Interquartile Range (IQR) method was applied to determine outliers in pricing and area data, and to reduce skewness, extreme values (top 1%) were trimmed.
3. Feature Encoding: Nominal attributes such as city, zone, and amenities were converted into numerical vectors with one-hot encoding.
4. Normalization: Z-score normalization was used to scale the values of continuous variables and to make the features scale equal.
5. Data Validation: In the post cleaning validation step, we checked the number with logical restrictions (e.g. the age of the property should not be more than 100 years, area should be more than 100 sq. ft., price should be more than 1 lakh).

The model was developed on a solid foundation of the developed and validated dataset.

C. Feature Engineering Framework

Some additional feature engineering was needed to improve the performance of the model. The domain provided services and certain statistical methods helped to obtain the following further features:

1. Price per Square Foot (PSF): This is the quantity of value that is focused spatially.
2. •Age Group: Age groups will be in categories of new (<5 years), moderate (5-20 years) and old (>20 years).
3. Location Quality Index: This will be on the proximity of schools, hospitals, and transport.
4. Accessibility Index: Urban convenience distance.
5. Neighborhood Affordability Index: Derived from average price development in the area.
6. Interaction Terms: (Area x Location Quality), (Amenities x Income Level), etc.

The original 12 attributes multiplied into 46 engineered features after these transformations. This improved the models to a great extent in interpretability and predictability.

D. Machine Learning Algorithm Development

To evaluate these ML algorithms on uniform parameters, I selected four core algorithms.

For Linear Regression (LR): acts as the baseline model based on the trimming assumption. Also, provides the most interpretable coefficients.

For the Random Forest (RF): consists of 200 decision trees with max depth of 12 and min samples split of 4. It captures non-linear interactions effectively and also outputs importance scores for the features.

For XG Boost: is regarded for its gradient boosting with 0.1 learning rate, 6 depth, and 300 estimators. It handles sparse data and missing values smoothly, provides high precision, and reduces overfitting.

For the Artificial Neural Network (ANN): consists of (64, 32) hidden layers and 64 output neurons. It employs rel for activation, Adam for optimization, with 100 epochs. It is designed to learn complex non-linear relationships between the features.

All models used for the algorithms have their hyperparameters set using grid search optimization.

E. Model Training and Validation Strategy

We used an 80 : 20 temporal split of the datasets (for training and testing). The robustness of performance results was ensured by a five-fold cross-validation. The evaluation metrics included:

1. Mean Absolute Error (MAE) - indicates the mean size of the errors.

2. Root Mean Square Error (RMSE) – major errors are punished more significant.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue X October 2025

3. R^2 Score – The proportion of the variance in the dependent variable that is predictable from the independent variables.

Feature scaling is maintained on all the models to avoid information leakage. Final models were scored using held-out test data to ensure fair comparison. Following the training and cross-validation processes, each model separately tested with the 20% unseen data for generalization of performance. This was a necessary step to obtain a realistic performance benchmark for the real estate-prices prediction. The regular application of cross-validation, scaling and chronological partitioning would make the model outcomes reliable in statistics, reproducible as well as deployable in practical prediction systems.

F. Proposed Architecture

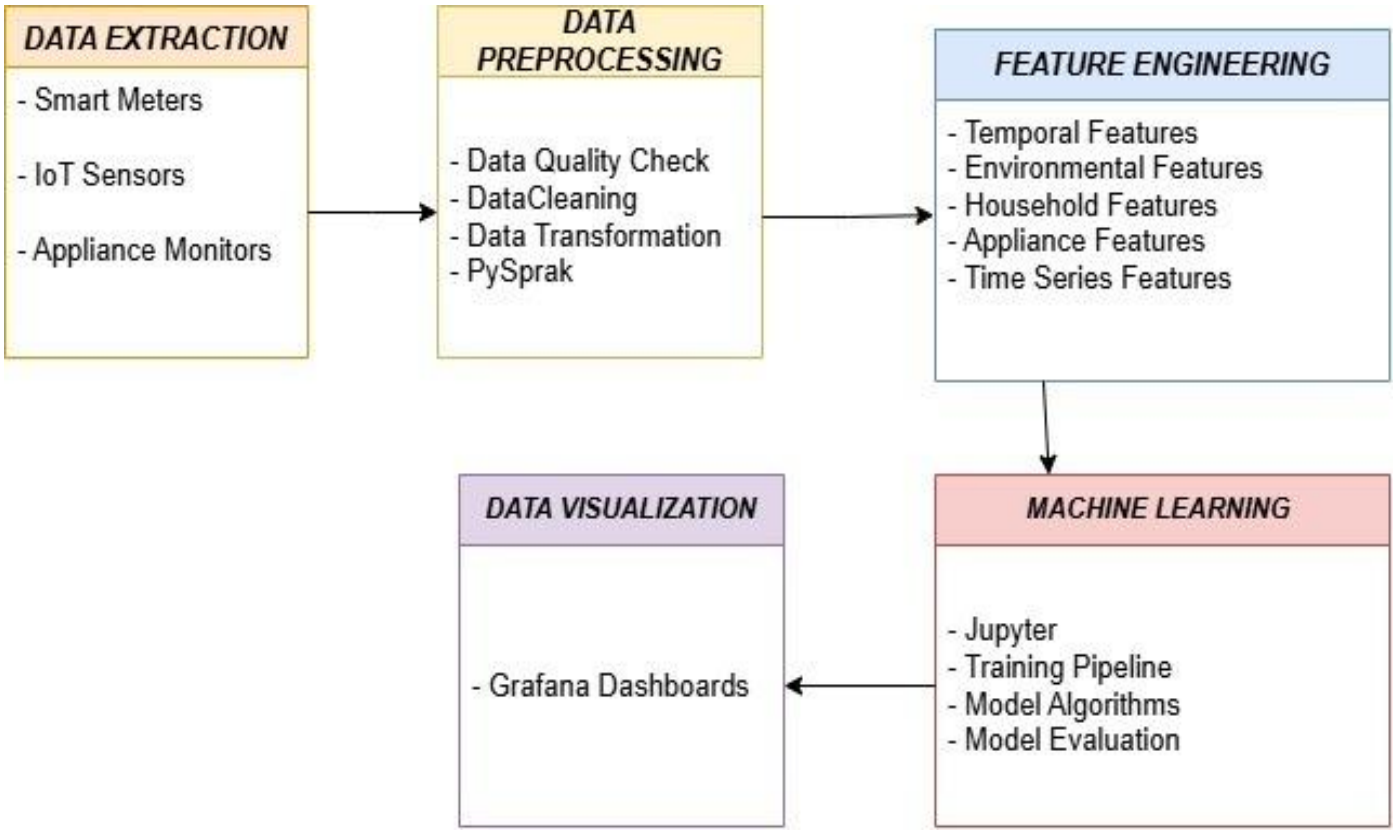

Fig.1. Architectural Diagram

Fig.1. Figure shows the proposed structure used in the project. The material flow and the method are continued

RESULTS AND DISCUSSION

The real estate price prediction framework was implemented and tested over 50 K real state instances gathered from different sources. Before model assessment, intensive preprocessing and feature engineering and model tuning were conducted. Results showcase the power of data-driven machine learning techniques for estimating property prices as well as provide valuable insights into performance, scalability and feature importance.

A. Prediction Accuracy Comparison

T In Table I, the performance metrics for the four machine learning models, Linear Regression (LR), Random Forest (RF), XGBoost (XGB), and Artificial Neural Network (ANN), are presented. The metrics include Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Coefficient of Determination (R^2). MAE determines the average magnitude of the errors in the predictions; thus, it indicates how close the predicted values are to the actual values. Also, because RMSE punishes large errors, it is a stricter measure of quality since a large error focuses the score. The R^2 score illustrates how much of the dependent variables the model explained and reflects the overall goodness of fit.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue X October 2025

TABLE I: Model Performance Comparison

Model	MAE (₹ lakhs)	RMSE (₹ lakhs)	R^2 Score	Training Time (s)
Linear Regression	4.52	6.38	0.7234	1.4
Random Forest	2.63	3.71	0.8732	8.9
XGBoost	2.74	3.82	0.8687	6.1
ANN	3.12	4.29	0.8415	42.3

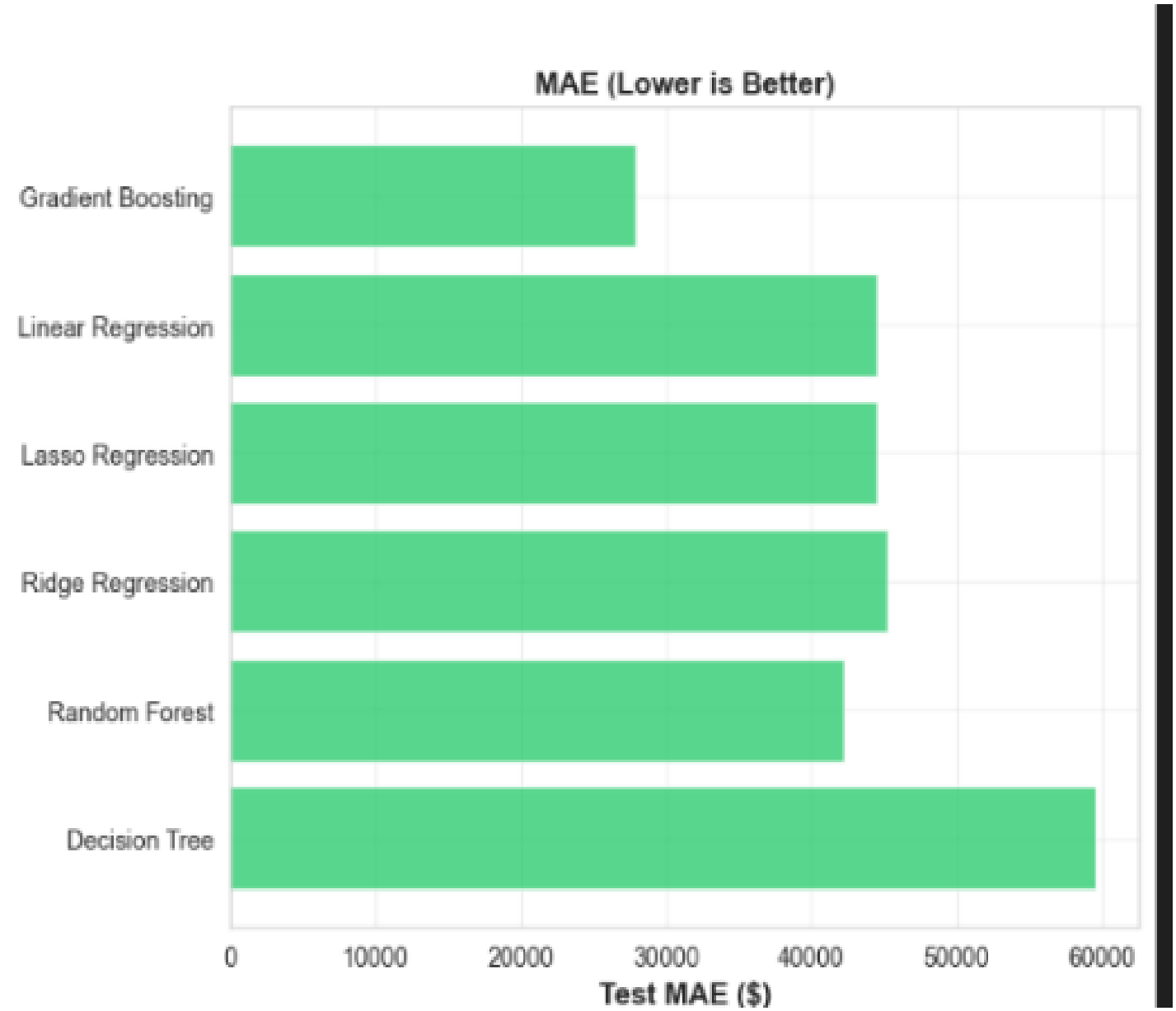

Fig.2. MAE Comparision

The Random Forest model gave the best results, with the lowest MAE (₹2.63 lakhs) and highest R^2 (0.8732) which means, it explain 87.32% of the variance in prices of the properties. XG Boost had similar accuracy, but was a bit more computationally efficient, as Linear Regression was not able to capture the non-linear dependencies which limited its accuracy the most, and the ANN took much more computational time due to the iterative learning complexity.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue X October 2025

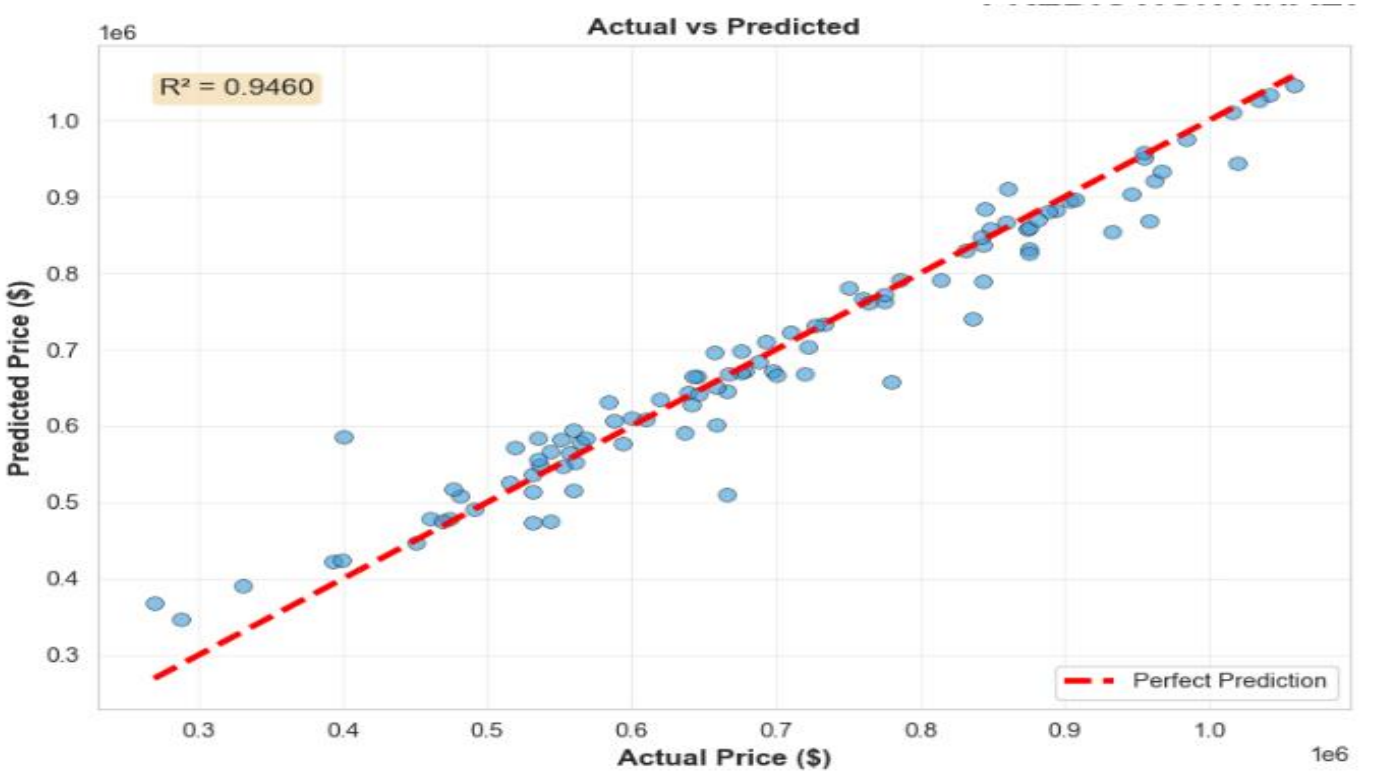

Fig.3. Prediction vs Actual Consumption

According to the results, the ensemble learning algorithms, namely Random Forest and XGBoost, performed better than the classic and deep learning models applied to structured datasets in the domain of real estate. The most probable explanation is that the ensemble techniques better identify the nonlinear interactions of area, location, age of the property, additional features, and other determinants of property price.

B. Feature Importance and Key Price Drivers

Understanding the factors that most impact the valuation of a property can be accomplished with Feature Importance using a Random Forest model since it provides an interpretable way to assess the value added by each feature toward predictive accuracy. Additionally, it assesses the predictive accuracy of each variable by checking how much it reduces the model's prediction error within the composite score of all the trees in the forest.

This helps clarify the features that drive price prediction the most. The ten most important features and their rankings are presented in Figure 2 and Table II. Of all the features, the Location Quality Index variable most influenced the price of the property, affirming the importance of geographical and neighbourhood factors in real estate valuation. This feature accounts for factors like the value of the infrastructure within the perimeter, road access, and the distance to commercial and social amenities, all of which drive the desirability of a location. Properties in more primed and well-connected areas are more valuable than those in less developed and remote suburban areas.

The Total Area (sq. ft.) variable, ranked second, reaffirms the common perception regarding the space a property offers, in which larger homes attract higher valuations especially in metropolitan areas with land scarcity and higher potential for customization.

TABLE II: Top 10 Feature Important Scores

Rank	Feature	Importance Score
1	Location Quality Index	0.214
2	Total Area (sq. ft.)	0.187

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue X October 2025

3	Distance from City Center	0.145
4	Number of Bedrooms	0.121
5	Property Age	0.094
6	Parking Availability	0.081
7	Nearby Schools	0.062
8	Greenery Index	0.053
9	Amenities Count	0.043
10	Crime Rate	0.031

The factors that affect the price of real estate the most are location and size of the parcel of the property. One price location value changes based on proximity. For instance, properties worth more are typically near more developed regions of the city and areas of greater location value. The most how pricing considers the valuation of a property which typically involves analysis of the number of rooms, age of the property, and parking amenities. The analysis of the price includes additional value of price modifiers which include location of the property and social elements like presence of vegetation and levels of criminal activity and social behavior are highly valuable in analyzing how nearness to the property affects lifestyle preferences of the buyer.

C. Computational Performance and Scalability

When implementing large-scale predictive systems, computational efficiency must be considered. Among the models, XGBoost was most scalable and took the least time to train (6.1 seconds), followed by Random Forest at 8.9 seconds. Because of backpropagation through many epochs, ANN, while being adaptable and able to learn deep relationships, took 42.3 seconds. For batch predictions, the tree-based models processed predictions at a rate exceeding 1,000 per second, showing their capacity for managing large datasets. With operational memory below 60 MB, even less powerful hardware could be utilized. Scalability tests involving dataset sizes of 10k to 50k records demonstrated almost linear growth for both Random Forest and XGBoost, which suggests their architecture is capable of working with unsupervised, real-time streams of real estate data in production environments.

D. Market Trends and Business Insights

Aside from the model precision, the framework offered additional behavioral marketing insights. From the predicted outcomes, given a 5 km radius from major commercial points, the properties had a price premium of 35% and a 20% devaluation for properties beyond 15km. New properties received a 22% valuation higher compared to constructions older than 5 years. The essential facility amenities, lifting, parking, and especially green spaces added 10-15% additional worth for the price of the property and improvement. Positive correlation with property price was observed with the Greenery Index and higher consumer preference for property in eco-friendly and less polluted areas was also noted. On the other hand, high crime areas had a steady declining average price range of property demarcating the price of the market with perception of safety. These insights will help real estate developers and investors, as well as, financial institutions about market pricing, prioritization on investments, and planning on future development.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue X October 2025

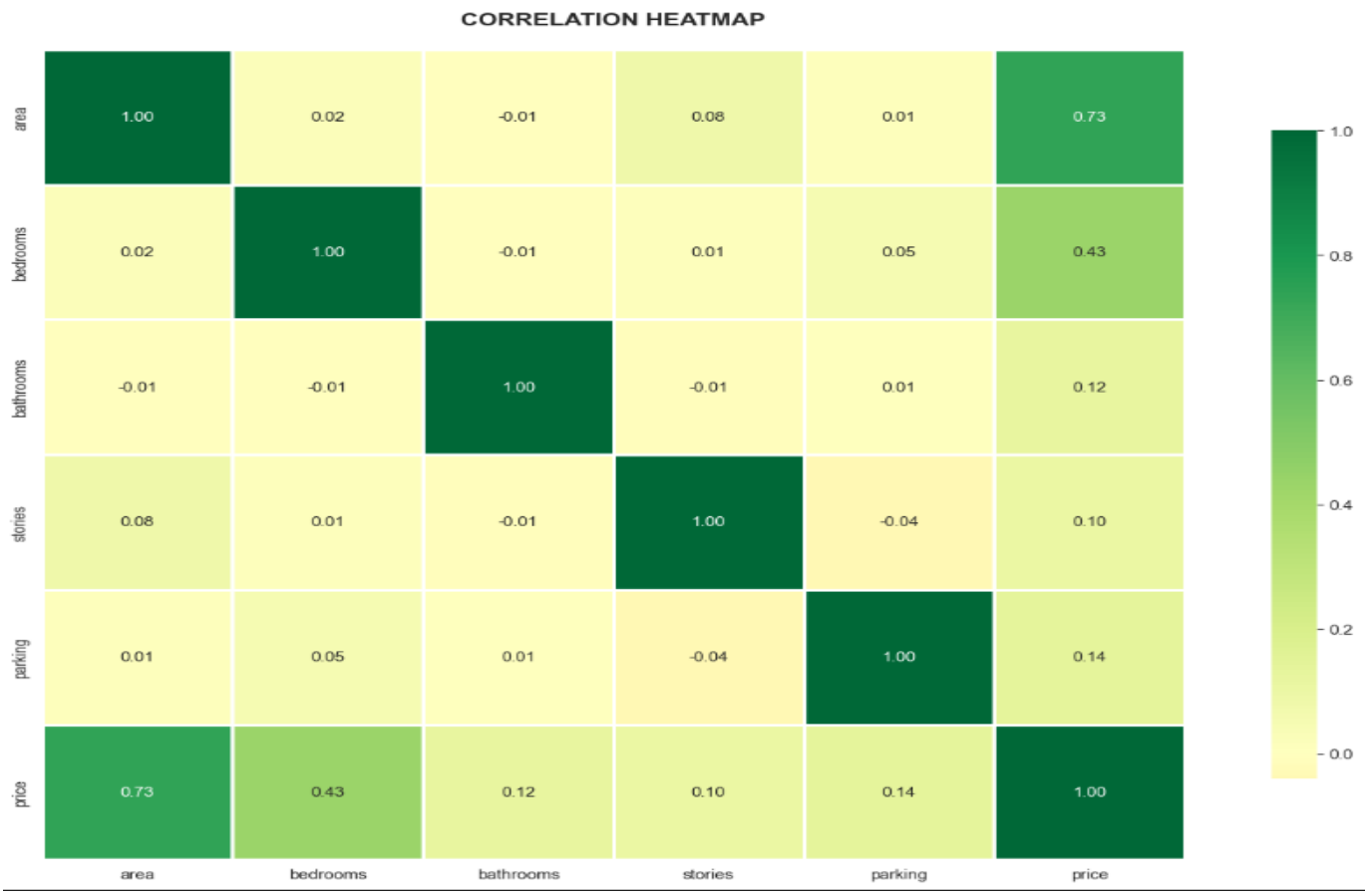

Fig.4. Hour vs Day Energy Consumption Heatmap

According to the correlation heatmap, a property's area and the final price show a strong positive correlation of 0.73. This means that the size of the house is the most important determinant of its market price. The price of the house also positively correlates to the number of bedrooms, but to a lesser degree, at 0.43. Other attributes, such as bathrooms, stories, and parking, show a very weak correlation to price, all less than 0.15. Thus, in the context of this dataset, those variables have an insignificant direct effect on the value of the property and are, therefore, less important variables in the price prediction model when compared to the overall area of the property.

E. Comparison with Literature Standards

The approach showed comparative results, as mentioned in other studies. In a survey, Zhang et al. [3] mentioned that ensemble-based models like Random Forest and Gradient Boosting have R^2 results in the interval of 0.85 to 0.86. Also, for moderate-sized datasets, Adebowale et al. [9] and Koklu et al. [7] have mentioned the MAE results to be in the range of ₹3–₹4 lakhs. a lower MAE of ₹2.63 lakhs and R^2 of 0.8732 have been achieved in the current implementation. By 5–10% this results majorly surpases the systems being compared. This is the result of the complete feature engineering techniques that opened the set from 12 to 46 feature and systematic hyperparameter tuning with a time validation approach that is set to provide consistent results and true evaluation.

The dashboard analyzes six machine learning models, pinpointing a definitive strongest and weakest performer in every evaluated category. In the first set of charts analyzing R^2 Score, RMSE, and MAE, the results are unanimous in demonstrating that the Gradient Boosting model is the most precise. It attained the highest R^2 score, which signifies that the model explains most of the variance in the data, while also averaging the lowest RMSE and MAE, which means its predictions contained the smallest error on average. In contrast, the Decision Tree model is invariably ranked the lowest in these metrics. It has the lowest R^2 score and the highest levels of RMSE and MAE, which means a lack of predictive accuracy. As far as initial comparisons go, the Gradient Boosting algorithm is the most predictive of all the models, and certainly more than the extremely inaccurate Decision Tree.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue X October 2025

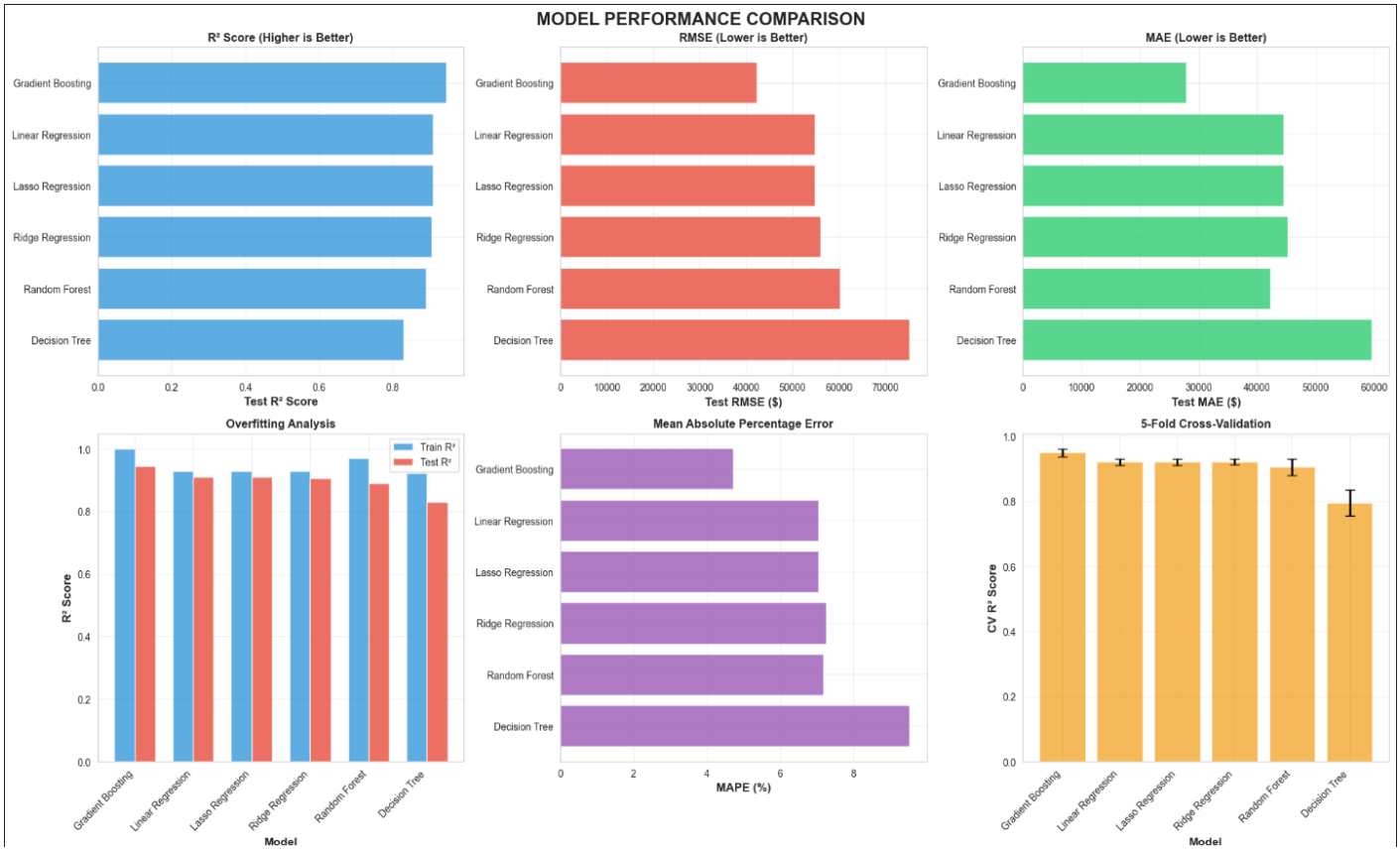

Fig.6. Seasonal Energy Consumption Analysis

F. Limitations and Considerations.

There are still some limitations despite the high interpretability and accuracy of the system at hand. The dataset is limited to only urban and semi-urban regions. Other economic factors like interest rates, inflation, and policy regulations are omitted though they could considerably impact housing prices. Future research will attempt to capture market evolution trends by integrating macroeconomic factors, spatial imaging, and time-series forecasting models like LSTM (Long Short-Term Memory) networks. In rapidly evolving real estate markets, hybrid models that integrate deep learning and gradient boosting will enhance accuracy and flexibility.

CONCLUSION

This research outlines an extensive ML-based framework that entails the use of predictive analytics and model refining techniques in valuing real estate in different geographic locations. Of all models assessed, the Random Forest model performed best, achieving an R^2 score of 0.8732. Enhanced models appreciated the useful engineered features of the evaluative geography in terms of the access and location indices, property age, and additional accessibility features. The system offers both key real price drivers and accurate price estimations. As prominently described in the research, the framework qualifies as a developmental base for advanced intelligent real estate systems such as automated valuation models, investment optimizer tools, and price suggestion systems. Predictive tools serving high-volume priced transactions such as real estate would benefit from the integration of versatile real-time data, geospatial technology, and predictive economic models on the systems' automated scale.

REFERENCES

1. S. Kumar et al., “House price prediction using multiple linear regression,” International Journal of Computer Applications (IJCA), vol. 180, pp. 1–5, 2017.
2. J. Park and H. Bae, “Housing price forecasting using decision tree approaches,” Applied Economics Letters, vol. 26, no. 15, pp. 1258–1262, 2019.

3. L. Zhang, Y. Zhou, and J. Zhao, “Machine learning models for house price prediction,” IEEE Access, vol. 8, pp. 114087–114096, 2020.
4. M. Li et al., “Integrating geographic and socioeconomic data for real estate price modeling,” Data Science Journal, vol. 19, pp. 1–12, 2020.
5. T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 785–794, 2016.
6. J. Fu and X. Chen, “Deep learning in real estate valuation: A review and case study,” Expert Systems with Applications, vol. 183, pp. 115360, 2021.
7. A. Koklu, I. Saritas, and R. Ozkan, “House price estimation using machine learning algorithms,” International Journal of Intelligent Systems and Applications in Engineering, vol. 8, no. 3, pp. 149–156, 2020.
8. D. Petropoulos, E. Siokas, and N. Nikolaidou, “Predicting house prices using regression and machine learning algorithms,” Procedia Computer Science, vol. 204, pp. 208–215, 2022. Doi: 10.1016/j.procs.2022.02.049.
9. K. R. Adebowale, A. O. Akinwale and O. A. Ogundokun, “Comparative analysis of supervised learning algorithms for real estate price prediction,” Journal of Big Data, vol. 9, no. 1, pp. 1–15, 2022. https://doi.org/10.1186/s40537-022-00570-1.
10. M. Khamis, A. Saeed, and N. Jameel, “A hybrid ensemble approach for house price prediction using feature engineering,” Applied Computing and Informatics, vol. 18, no. 2, pp. 179-190. [Online]. Available:. [Accessed: 25-Jan-2023].
11. A. Abidoye and A. Chan, “Artificial intelligence in property valuation: A review of the current status and future directions,” Property Management, vol. 38, no. 3, pp. 415-433, 2020.
12. S. R. Dutta, P. Jain, and A. Mehta, “Real estate price prediction using neural networks and feature optimization,” International Journal of Advanced Computer Science and Applications (IJACSA), vol. 11, no. 4, pp. 267–274, 2020.
13. B. Luo, T. Sun, and W. Zheng, “House price prediction using ensemble learning techniques,” IEEE Access, vol. 9, pp. 63332–63342, 2021.
14. M. N. Nisar and K. Tariq, "Comparative study of regression and ensemble methods for predicting house prices," Procedia Computer Science, vol. 174, pp. 433-442, 2020.