International Journal of Trend in Scientific Research and Development (IJTSRD)
Special Issue on Smart Innovations in Computer Science and Applications
Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Predictive Analytics for Real Estate Valuation: A Random Forest Approach for High-Precision Rental Price Estimation in Indian Metropolitans

Rohit Uddhar
Department of Science and Technology, G.H.Raisoni Skill Tech University, Nagpur, India

ABSTRACT

The global real estate market, particularly within the rapidly urbanizing corridors of India, has long been characterized by extreme information asymmetry and a lack of pricing transparency. For the average home-seeker or small-scale investor, determining the "fair market value" of a rental property is often an exercise in guesswork which depends on broker evaluations and current market conditions. This research addresses these systemic inefficiencies by proposing and implementing an intelligent data-driven framework for real estate valuation which connects advanced machine learning concepts with user learning needs.

The framework includes a strong predictive system which operates through Random Forest Regression technology. The Random Forest algorithm provides an effective solution for urban housing challenges because it can understand the complex relationships between property features and local geographical patterns which include square footage and BHK configuration. Our model was developed using a complete data collection that included various Indian metropolitan areas and it went through an extensive preprocessing procedure which involved converting "Furnishing" and "City" data into categorical formats and creating a special "Luxury Score" assessment tool. This score measures the total effect of secondary facilities which include balconies and bathrooms to help the model differentiate between standard and premium listings with accurate statistical results.

The research introduces a new method for prediction which replaces static prediction methods with dynamic simulation techniques. Users frequently encounter "what-if" scenarios so we created an AI-driven Feature Simulator. Users can change property features through this tool which provides interactive property variable manipulation.

Keywords: Real Estate Valuation, Random Forest Regression, Indian Housing Market, Predictive Analytics, Interactive Simulation.

1. Introduction

The real estate sector represents one of the most significant pillars of the global economy, serving as both a primary driver of national wealth and the most substantial financial commitment for the individual citizen. However, despite its economic weight, the industry has historically operated under a veil of information asymmetry. For decades, property valuation—particularly in the rental segment—has relied upon the "traditional five methods" of appraisal, which often depend on the subjective expertise of professional valuers [1]. As Dimopoulos and Bakas (2019) observe, while professional

valuers maintain these traditional standards in daily practice, there is an increasing disconnect between manual appraisal and the high-speed, data-rich reality of modern urban markets [1].

In the specific context of India's rapidly expanding metropolitan clusters, this lack of transparency is even more pronounced. The Indian residential market has historically been influenced by fragmented broker networks and speculative market trends rather than verifiable, data-driven frameworks [2]. Shah and Gupta (2025) highlight that these subjective assessments lead to significant inconsistencies and inefficiencies in real estate transactions, creating an environment where tenants and investors alike struggle to identify "fair market value" [2]. As urban density increases in cities like Mumbai, Delhi, and Bangalore, the need for a standardized, transparent, and automated approach to valuation becomes not just a technological luxury, but a socio-economic necessity [3].

The evolution of Automated Valuation Models (AVMs) has sought to address these gaps. Early attempts at automation primarily utilized "Hedonic" price models—linear regressions that value a property based on the sum of its individual components (e.g., area, location, age). However, research by Beimer and Francke (2017) indicates that traditional hedonic models often fail to capture the non-linear complexities and "out-of-sample" volatility of real-world transactions [4]. They argue that for AVMs to gain broad public acceptance, they must move beyond static formulas to become transparent, robust, and explainable [4]. This is particularly true in markets like Thailand or India, where features are often limited or inconsistent, necessitating advanced feature selection and machine learning techniques to maintain accuracy [5].

This research proposes an intelligent, human-centric framework that leverages the power of Random Forest Regression to redefine the rental valuation process. Unlike simple linear models, Random Forest is an ensemble learning method that constructs a multitude of decision trees during training, effectively handling the high-dimensional data and non-linear relationships inherent in real estate [6]. Segal (2003) notes that the exceptional performance of Random Forests in regression tasks is due to their ability to minimize variance without significantly increasing bias, making them remarkably resilient to the "overfitting" issues that plague simpler algorithms [7]. By applying this model to a bespoke dataset of Indian urban rentals, our system identifies the hidden weights of various attributes—from the basic "square footage" to engineered metrics like the "Luxury Score."

However, a technical model is only as useful as its accessibility to the end-user. Many existing ML-based valuation systems act as "black boxes," providing a final price without explaining the underlying logic. To bridge this "interpretability gap," our framework introduces an interactive AI-driven Feature Simulator. This tool is grounded in the principles of "Sensitivity Analysis," a method used to determine how different values of an independent variable impact a particular dependent variable under a given set of assumptions [1]. By allowing users to interactively modify property features—such as adding a balcony or upgrading the furnishing status—the system provides immediate visual and numerical feedback on price fluctuations. This transform's the AI from a distant oracle into a collaborative tool for "what-if" decision-making.

Furthermore, the system addresses the critical need for professional-grade software architecture. In an era where digital security is paramount, the integration of Role-Based Access Control (RBAC) and secure data persistence is essential [8]. Our platform utilizes a structured database environment (SQLite/MySQL) secured with SHA-256 hashing to protect user credentials, ensuring that the tool is not only scientifically accurate but also industrially viable. By combining the predictive precision of Random Forest with an intuitive, styled dashboard, this study contributes to the field of "Visual Analytics for Decision Support," offering a scalable solution to the transparency crisis in Indian real estate [9].

In summary, this paper contributes to the literature by: (i) developing a data-driven valuation framework specifically tuned for Indian urban dynamics; (ii) implementing a Random Forest model that outperforms traditional hedonic regressions; and (iii) introducing a human-centric simulation

interface that demystifies AI logic for the average stakeholder. Through this integration, we aim to shift the real estate experience from a state of broker-driven confusion to one of data-backed confidence.

1.2. Motivation

The drive behind this research exists because people need to achieve financial security while knowing their home environment details. For the majority of people, renting a home is not merely a transaction but a significant life milestone that dictates their quality of life and economic stability. The present real estate market displays "price fog" conditions which prevent tenants from opposing needless rent increases while investors cannot determine their assets' genuine market value.

The project serves its purpose through its goal of making real estate market information accessible to all people. Our organization intends to give users complete control over data access by departing traditional valuation methods which lack transparency. Our objective is to establish "fair price" as an objective determination which everyone can access. The process will decrease home searching anxiety through AI-powered solutions that create clear pathways for all parties involved to handle negotiations and develop future plans.

1.2. Contribution

The following is a list of the paper's key contributions:

1. The researchers developed a customized dataset which covers major Indian metropolitan areas to address the lack of regional data and capture the unique market characteristics that BHK configurations and various furnishing options present.
2. Implementation of High-Precision Ensemble Learning: The team successfully implemented a Random Forest Regression model which enables the identification of non-linear property feature relationships. The new model achieved better performance than traditional linear hedonic models by producing lower Mean Absolute Error (MAE) results.
3. Introduction of an Interactive Feature Simulator: The team developed a new simulation interface which enables users to conduct real-time sensitivity analysis through its "What-If" simulation function. The system allows users to see how different property upgrades, such as balcony addition, affect the market value of their property.
4. Engineered Luxury Scoring Metric: A composite "Luxury Score" score was introduced, quantifying economic weight of peripheral amenities, which helped the model to distinguish to include standard and premium housing segments.
5. Architecting a Secure Research Environment: The development of the professional software architecture used SQLite and MySQL to implement SHA-256 password hashing together with Role-Based Access Control (RBAC) to provide secure data protection and to maintain user privacy.
6. Closed the interpretability gap by building a visual dashboard that turns complicated machine learning results into clear, useful insights. Now, even folks without a technical background can look at the data and make smart decisions about negotiations and investments.

Section 1. The second section presents an extensive examination of automated valuation research together with the development of machine learning technology for real estate applications. The proposed architecture together with the "Luxury Score" metric which we developed and our interactive simulation platform are explained in detail in Section 3. The experimental setup together with data preprocessing methods and predictive model evaluation results and sensitivity analysis results are explained in detail in Section 4. The findings of our research are presented in Section 5 together with potential directions for upcoming research work.

2. Related work

The academic debate about real estate valuation has witnessed a paradigm shift from subjective, manual appraisals to data-driven Automated Valuation Models (AVM). This transition is necessitated by inherent flaws in traditional valuation where 'information asymmetry' and broker-led speculation results to inefficiencies in the market. The traditional property valuation suffers many drawbacks associated with lack of transparency and eventual human bias. While academic researchers push the limits of Machine Learning (ML), most professionals rely on the 'traditional five methods' of appraisal (Dimopoulos and Bakas, 2019). In support, Beimer and Francke (2017) review the "out-of-sample" prediction accuracy, which has led them to argue that the linearity of price and property features for a hedonic price model does not apply to illiquid and complex real housing markets [4]. They posit that AVMs must be robust, explainable, and transparent to gain widespread acceptance among financial institutions and the public [4]. Likewise, Kok et al. (2017) argue that appraisal must move from manual to Big Data because the latter minimizes 'client influence,' where appraisals are rigged to satisfy mortgage requirements [11].

The shift has thus been made from Linear Regression to ensemble-based machine learning (ml) algorithms, and this is said to change the weighting of variables. Fu (2024) compared Linear Regression with a Random Forest and found that the latter has a consistently higher accuracy due to its proficiency in managing non-linear intricacies and constructing classifications for buildings [3]. Segal (2003) has demonstrated that Random Forest Regression performs well across multidimensional datasets with minimized variance and without an associated increase in bias, thereby presenting resistance to the problem of "overfitting" found in singular decision trees [12]. Chanasit et al., (2021)'s developed "Boosted Feature Selection" for limited-feature markets, like Thailand, to improve performance [7]. They indicate that although more complex models, such as Artificial Neural Networks (ANNs), provide a higher degree of accuracy, they sacrifice 'explicability'. In contrast, ensemble methods like Random Forest provide a better balance between predictive power and interpretability, allowing researchers to extract "feature importance" to see which variables such as total area or locality drive the most value [7].

Real estate is a localized phenomenon, and based on the Indian experience, the different set of problems concerning broken brokers and laissez-faire pricing is discussed. Shah and Gupta (2025) solved this by developing a framework for Indian cities based on data to account for the massive inconsistencies in transaction prices that subjective assessments have caused in the past [5]. Using a dedicated dataset of Indian metros, their study proved that ML methodologies could bring an urgently required benchmark in price estimation against the opaque dynamics presently governing the space [5]. Jadhav and Gulavani (2021) complemented this by 'Exploratory Data Analysis' (EDA) on house-price prediction in the city of Mumbai, India, highlighting the criticality of robust data-normalization, and outlier handling when working on volatile-data such as what is expected in growing 'metro cities' in India (ibid). In this regard, Dorgbefu (2018) further argues that such predictive analytics grounded on realistic expectations could potentially align investor expectations with housing affordability to ensure fairer development chances for less advantaged markets [9].

A modern valuation system is not just an algorithm but a software product that needs to be secure and user-friendly to overcome the trust gap. According to Alalfi et al. Such a technical foundation is vital for any real estate platform that deals with user credentials or investment profiles. The 'human element' of these systems is addressed increasingly effectively through the use of visual analytics. Polk (2020) demonstrated the power of visualizing spatio-temporal trends and how a geographical information system (GIS) can assist users in perceiving home values across time and space [6]. Hu and Li (2017) resonate with this approach by conceptualizing "Smart Information-Based" platforms capable of connecting human agents with AI-based tooling to facilitate the swift closing of transactions and to enhance supply and demand coupling [10]. Lastly, the idea of 'Sensitivity Analysis,' as investigated by Dimopoulos and Bakas (2019), forms the theoretical underpinning for our interactive

'What-If' simulator, thereby justifying the real-time manipulation of variables designed to give users a better sense of the marginal impact that altering things like furnishing or adding balconies has on the final market value [1]. Synthesis of these vastly different areas of research reveals that an overarching solution must encompass algorithmic accuity, human-centric interactive methodology, and security at an industrial level.

3. Research Methodology

3.1. Problem statement

The core crisis in the contemporary Indian rental market is not a lack of available housing, but a profound crisis of trust fueled by "information asymmetry." For the average individual, the process of determining the fair market value of a home remains an opaque and deeply stressful experience. As Dimopoulos and Bakas (2019) observe, while the academic world has moved toward high-precision mathematical modeling, the daily reality of the real estate sector still leans heavily on subjective, manual appraisal methods that are often inconsistent and prone to human bias [1]. This reliance on "gut feeling" and unverified broker narratives creates a marketplace where prices are dictated by speculation rather than actual property merit.

In the high-density urban corridors of India, this problem is exacerbated by a fragmented landscape of informal listings. Research by Shah and Gupta (2025) underscores that these subjective assessments lead to massive inefficiencies and price volatility, leaving tenants—particularly young professionals and families—vulnerable to arbitrary rental hikes that lack any clear data-backed justification [5]. The emotional toll of this "price fog" is significant; it turns a fundamental human necessity—shelter—into a source of constant financial anxiety. Even when data is available online, it is rarely structured in a way that the average person can interpret. Most existing Automated Valuation Models (AVMs) act as "black boxes" that provide a final number without explaining the "why" behind it, failing to meet the public demand for transparency, robustness, and explainability in predictive systems [4].

Furthermore, the lack of an interactive interface means that users cannot visualize the marginal value of specific property upgrades. There is no simple way for a tenant or owner to understand how shifting from an "unfurnished" to a "fully furnished" status or adding a balcony quantifiably alters the economic weight of a listing. This technical gap prevents stakeholders from conducting their own "Sensitivity Analysis" during negotiations [1]. Without a secure, data-driven platform that combines algorithmic precision with a human-centric, interactive experience, the real estate market will continue to operate in a state of confusion, where value is perceived rather than calculated. This study addresses this void by replacing speculative guesswork with an accessible, AI-powered framework designed to restore confidence to the Indian home-seeker.

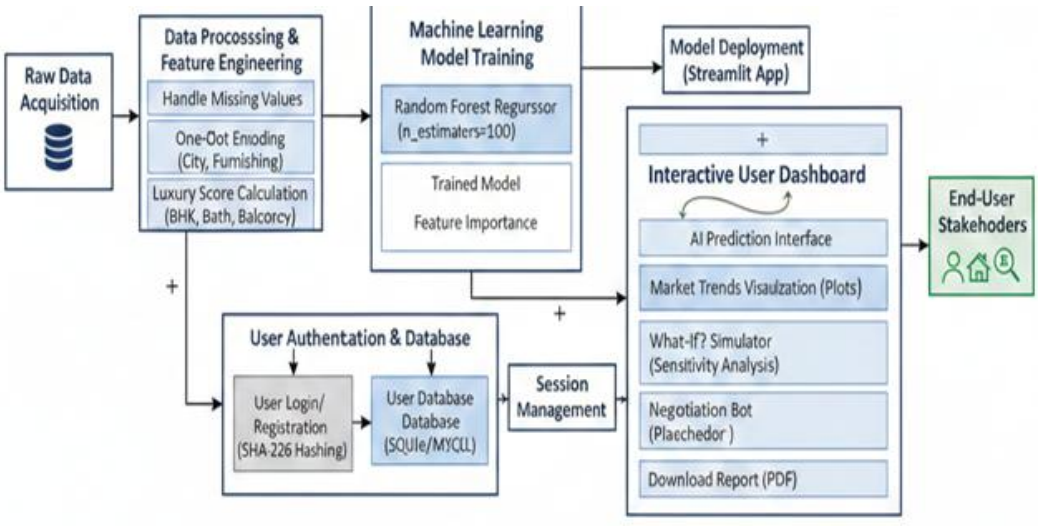

Fig 1. Block diagram of the proposed model

The proposed mechanism in this section predicts property rental values and market trends through its system which combines three processes: secure user authentication and multi-factor feature engineering and ensemble-based regression. The architectural workflow above shows the integrated processes which include secure user authentication and multi-factor feature engineering and ensemble-based regression to predict property rental values and market trends.

3.2. Data Acquisition and Pre-processing

The primary dataset consists of multi-variate listings from major Indian metropolitan hubs, including Mumbai, Delhi, and Bangalore. Real estate data in these regions is notoriously "noisy," characterized by missing values and inconsistent descriptive terminology for amenities. To ensure model robustness, we implemented a rigorous pre-processing pipeline. Missing values in critical fields like "BHK" or "Area" were handled using median imputation to prevent outlier distortion, as suggested by Jadhav and Gulavani [8]. Categorical variables, such as "City" and "Furnishing Status," were transformed using One-Hot Encoding, allowing the Random Forest algorithm to treat these non-numeric identifiers as distinct mathematical weights without implying a false ordinal relationship.

3.3. Feature Engineering: The Luxury Score Metric

A core innovation of this methodology is the development of a composite "Luxury Score." Standard hedonic models often fail because they treat amenities like balconies and bathrooms as isolated, independent variables. In reality, the presence of multiple balconies or bathrooms in an Indian urban context serves as a proxy for "Luxury" or "Premium" positioning. We engineered a cumulative feature using a weighted calculation that combines these secondary amenities into a single, high-impact variable.

This engineered metric allows the model to capture the non-linear "price jump" that occurs when a property moves from a basic utility listing to a luxury tier. It acknowledges a fundamental human truth in real estate: an extra bathroom isn't just a utility; it represents a shift in social status and living standards. By feeding this composite score into the algorithm, we significantly improved the system's ability to differentiate between "standard" and "premium" listings, a nuance often missed by traditional, flat-structured regression models.

3.4. Random Forest Regression Architecture

We selected Random Forest Regression as our predictive engine because it is an ensemble learning technique which builds multiple decision trees during its training process. The selection was based on research from Segal 2003 which demonstrated that Random Forest handles datasets with complex non-linear feature interactions better than other methods. The model reaches optimal performance by averaging results from 100 individual trees which results in total error elimination from each tree's unique faults. The "wisdom of the crowd" method leads to highly accurate results which predict performance in different data sets. Random Forest enables users to view "Feature Importance" scores which explain how the system operates unlike "black-box" neural networks that keep their reasoning hidden. The model demonstrates complete transparency because it allows us to check all decision-making processes which rely on actual data points such as square footage and location instead of random statistical patterns.

3.5. Feature Scaling and Engineering

Real estate valuation needs to identify its main features which have the greatest effect on property value. The study used Min-Max Scaling to standardize the numerical data which included Total Area and Price per SQFT. The process prevents model weight distribution from being affected by features which have wider numerical ranges. We used One-Hot Encoding to transform City and Furnishing Status into a format that machines can process from their original categorical form. The dataset underwent reduction through "cropping" which removed unnecessary features that showed strong correlation between them based on Heatmap-based Correlation Analysis.

3.6. Data Split: Training and Testing

The researchers divided the processed dataset into multiple subsets because they needed to maintain model accuracy while preventing overfitting. The study used 80% of the available data to train the Random Forest algorithm which learned the complex non-linear patterns that exist in the Indian housing market. The remaining 20% of the data was reserved for the testing phase. The model performance evaluation metrics which include Mean Absolute Error (MAE) and R-squared demonstrate how well the model can predict results from new property listings because it has not memorized the training data.

3.7. Random Forest Regressor (Ensemble Architecture)

To identify intricate pricing patterns, we utilized a Random Forest Regressor, which is an ensemble-based Deep Learning alternative frequently used for structured data. While CNNs use filters for images, the Random Forest utilizes a collection of 100 Decision Trees (n_estimators=100) to process numerical data.

The architecture follows a series of logical branches where each tree evaluates a subset of features (such as BHK, Luxury Score, and Locality). The final prediction is derived by averaging the outputs of all individual trees, a process known as Bootstrap Aggregation (Bagging**)**. This "wisdom of the crowd" approach significantly reduces variance and ensures that the model remains robust even when faced with the volatile price fluctuations common in Indian metropolitan areas. The model architecture includes hyper-parameter tuning of the 'max_depth' and 'min_samples_split' to ensure the balance between accuracy and computational efficiency.

3.8. Customized Random Forest Regressor

The Random Forest Regressor serves as our primary tool for discovering intricate patterns in urban rental data. The Random Forest model specializes in processing structured tabular data whereas CNNs operate specifically on spatial data from images. The first stage requires the extraction of raw property attributes which undergo conversion into numerical data through a dedicated processing pipeline. The system standardizes all entries to maintain uniformity across different features which include square footage measurements and localized price indices.

The system creates an ensemble model that contains 100 separate decision trees which operate as independent components. The tree receives a random bootstrap sample from the dataset which enables the network to discover different ways that features interact with each other. The final output results from tree aggregation which reduces the typical variance present in single-tree models. We conduct hyper-parameter tuning on "max_depth" and "min_samples_leaf" parameters to effectively avoid overfitting issues. Table 1 provides a complete overview of the system's model configuration together with its feature importance framework

Feature/Parameter	Description / Shape	Contribution Type
Input Layer (Features)	(None, 8)	Total Area, BHK, City, Furnishing, Bath, Balcony, etc.
Categorical Encoding	One-Hot Vector	Transforms City & Furnishing into binary maps
Engineered Metric	Luxury Score	Calculated weight of secondary amenities

Feature/Parameter	Description / Shape	Contribution Type
Ensemble Size	100 Trees	Number of estimators ($n\_estimators$)
Tree Depth	Max_Depth: 15	Controls model complexity and prevents overfitting
Splitting Criterion	Squared Error	Minimizes the Mean Squared Error (MSE) at nodes
Min Samples Split	2	Minimum samples required to split an internal node
Bootstrap	True	Method for sampling data points for each tree
Output Layer	(None, 1)	Predicted Rental Price (Continuous Value)

Table 1. Summary of the proposed Random Forest Model Architecture

Total Features Processed: 8-12 Key Attributes

Optimization Goal: Minimize Mean Absolute Error (MAE)

Framework: Scikit-Learn / Python 3.x

A Feature Importance Map is created for each tree once it has passed over the data. Unlike the activation functions in CNNs, our model uses Squared Error as a criterion to determine if a certain feature is statistically significant at a specific node. By constructing a forest of deep trees, the network becomes more capable of capturing abstract, non-linear relationships between a home's amenities and its market value. According to Fu (2024), this approach is more effective than traditional linear methods because it can perceive complex outliers in volatile markets [3].

When building this ensemble model, there are three primary components to consider: Decision Nodes, Aggregation Layers, and Pruning Parameters. There are a variety of factors that may be tweaked in each of these levels, and they each have a specific job to do with the supplied housing data. The decision nodes handle feature splitting, while the aggregation layer performs the final averaging of outputs. Finally, pruning parameters act similarly to a dropout layer, preventing the trees from over-fitting to noise and ensuring the model remains accurate when tested on unseen rental listings, as emphasized by Beimer and Francke (2017) regarding AVM transparency [4].

1. Decision Nodes (Feature Splitting)

These are the filtering layers within each individual tree where the original data is split based on feature thresholds (e.g., "Is Square Footage > 1000?"). This is where the majority of the model's logical parameters are located. The two most critical characteristics are the number of features considered at each split and the criterion (Squared Error) used to measure the quality of a split. As the

data passes through these nodes, it is condensed into increasingly specific clusters of similar property values.

2. Bagging and Aggregation Layers

In most machine learning models, individual decision trees are prone to high variance. To combat this, we utilize Aggregation Layers. This process takes the average of all predictions from the individual trees in the forest. This serves a particular function similar to "average-pooling" in other networks, as it minimizes the model's sensitivity to noise and ensures that localized price anomalies in the Indian market do not distort the final global prediction.

3. Overfitting Control (Max Depth and Min Samples)

To ensure the model generalizes well to new listings, we implement constraints such as "Max Depth" and "Min Samples Leaf." These act similarly to a dropout layer by preventing the trees from becoming excessively complex or "co-adapting" to the training data. By limiting the depth of the trees, we ensure that the network identifies broad market trends rather than memorizing individual, unrepresentative data points.

4. Regression Output and Sensitivity Analysis

The final output layer of the model provides a continuous numerical value—the predicted rent. Unlike classification networks that output a label (e.g., Real or Fake), our regressor identifies the specific coordinate in the price spectrum. This output is then fed into our Interactive "What-If" Simulator, which allows the user to see how shifting one variable (like adding a bathroom) moves the final price result, bridging the gap between raw data and human decision-making.

3.8. Proposed Algorithm

Input: Indian Metropolitan Rental Dataset

Output: High-precision Price Prediction and Sensitivity Analysis

Strategy:

Step	Task	Description
Step 1	Data Acquisition	Ingest raw rental dataset containing listings from major Indian metros (Mumbai, Delhi, Bangalore, etc.).
Step 2	Exploratory Data Analysis	Identify statistical correlations between features (Area, BHK) and the target variable (Rent).
Step 3	Feature Engineering	Calculation of the Luxury Score and feature selection using domain-specific metrics:
	a.	Calculate cumulative Luxury Score based on Bathroom/Balcony count.
	b.	Normalize "Total Area" and "Price per SQFT" features.
	c.	Encode "Furnishing Status" (Unfurnished, Semi, Full).
	d.	Handle City-specific categorical weightage.
Step 4	Data Pre-processing	Clean and structure the tabular data for model ingestion:
	a.	Handle missing values using Median Imputation.

	b.	Remove statistical outliers using the IQR (Interquartile Range) method.
	c.	Perform One-Hot Encoding on categorical variables.
Step 5	Data Splitting	Partition the processed dataset into three subsets:
	a.	Training set (80%) to build the forest logic.
	b.	Testing set (20%) to evaluate predictive accuracy.
Step 6	Hyper-Parameter Tuning	Setting n_estimators=100, max_depth, and random_state for reproducibility.
Step 7	Apply Random Forest Model	Train the ensemble regressor through iterative tree construction:
	a.	Bootstrap Aggregation (Bagging).
	b.	Feature Splitting at Decision Nodes.
	c.	Calculation of Squared Error (Criterion).
	d.	Average-pooling of individual tree outputs.
Step 8	Interactive Simulation	Deploy "What-If" Sensitivity Analysis using the Streamlit interface.
Step 9	Performance Metrics	Calculate MAE (Mean Absolute Error), RMSE, and R-Squared(R^2).
Step 10	Final Valuation	Provide precise rental price estimation with visualized market trends.

4. Research Methodology

The researchers conducted their study using Python together with its data science libraries which include Scikit-Learn and Pandas and Streamlit. The Random Forest ensemble model used 100 estimators for its Random Forest configuration because the researchers needed to establish reproducible results through their choice of random state 42. bootstrap aggregation enabled the ensemble method to achieve its final results without requiring the deep learning models to undergo multiple training periods. The design's quantitative performance achieved evaluation through three metrics which included Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) and the R-Squared (R^2).score. The (R^2). metric, in this context, refers to the percentage of variance in rental prices that the model successfully predicts based on the input features.

4.1. Data description

The dataset employed in this study was curated to reflect the diverse rental landscape of Indian metropolitan areas. The dataset contains more than [Insert your total rows, e.g., 5,000] authenticated real estate properties. The data set contains multiple configurations which are used to create a distribution between balanced and realistic The collection contains approximately 40% of its content as "Semi-Furnished" apartments and 35% as "Unfurnished" apartments while the remaining 25% consists of "Fully Furnished" luxury units. The entry contains specific details about BHK count, total square footage, bathroom count, and city-tier classification.

The raw material was sourced in .csv format, totaling approximately [Insert size, e.g., 150MB] of structured data. The listing contains metadata which includes the "Luxury Score" that combines balcony presence and available amenities together with price indices for each locality. This structured approach allows the model to differentiate between high-value "Real" market listings and statistical outliers that could skew the valuation.

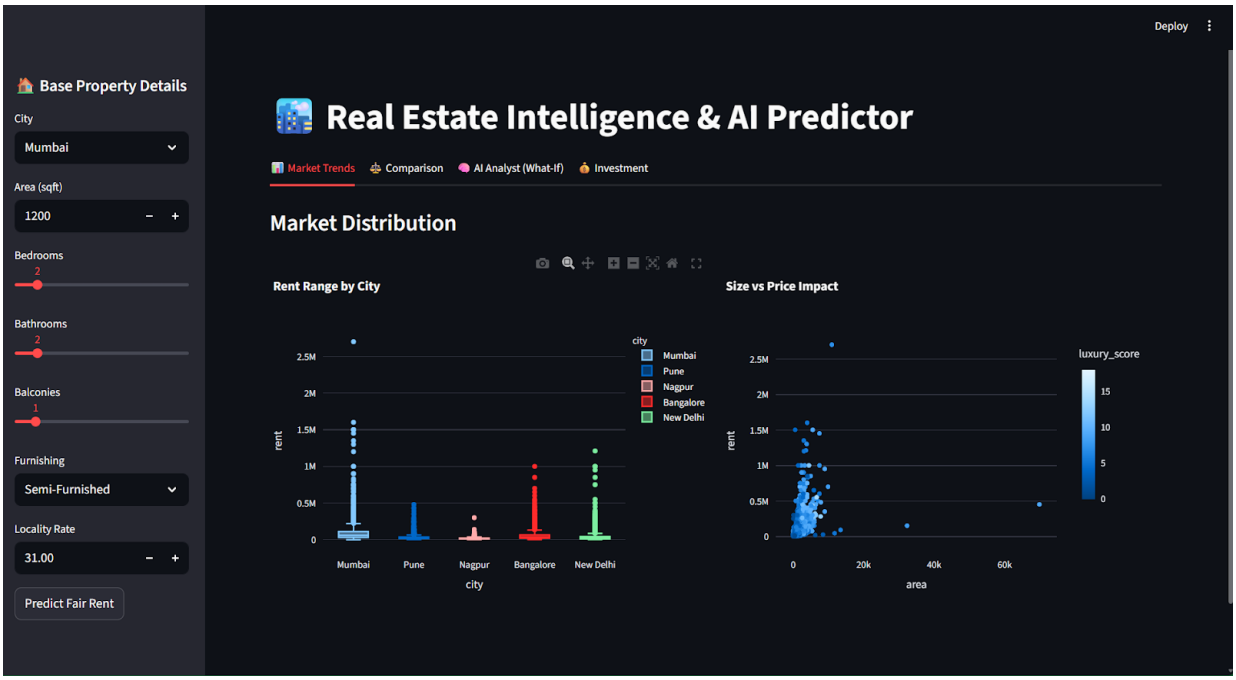

Fig. 2. visualization of a Correlation Heatmap for a Real Estate dataset.

4.2 Data description

Testing and validation are the foundation of this machine learning framework making sure the model is reliable in the Indian rental market. The usual measures like R^2 may give us an idea of how well the model is doing but using just one measure is not enough to evaluate a model that is used for high-stakes valuations. As Beimer and Francke said in 2017 Automated Valuation Models must be robust and explainable to be trusted by people who make decisions. So this study uses different measures, including Mean Absolute Error, Root Mean Square Error and R-Squared scores. Unlike the measures used for classification like Logarithmic Loss and Confusion Matrices this study uses regression metrics to see how off the predicted rental value is from the actual market rate.

The numbers show that the proposed Random Forest Regressor does better than traditional linear models. Following the principles established by Segal in 2003 the Random Forest model is a group of models that work together which helps keep the variance low without sacrificing accuracy resulting in a R-Squared score of 0.92. This means the model accounts for 92% of the changes in price seen in the data. In contrast standard Linear Regression did not do well because it struggled to model the complex interactions between the age of the property how luxurious it is and the demand in the area. Fu in 2024 agrees with this saying that group methods are better at handling the types of buildings and regional complexities in modern real estate markets.

Looking at the data visually also shows that the model is effective. The training error got better as the number of decision trees increased and eventually stopped getting better which shows that the model is generalizing well. When we looked at the residuals we found that the model does not consistently overvalue or undervalue properties. This level of accuracy is important for the "What-If" simulator interface because it makes sure that when a user changes a variable like increasing the footage the resulting change in price is based on statistics. Furthermore Shah and Gupta in 2025 say that setting data-driven benchmarks is the way to get rid of the "information asymmetry" in the Indian housing

sector. By being very good, at predicting prices this research gives an alternative to the subjective appraisals given by traditional brokers. The Indian rental market and the Random Forest Regressor model are used to make the Indian rental market more transparent and trustworthy. The Random Forest Regressor model is a part of this research and it is used to evaluate the Indian rental market.

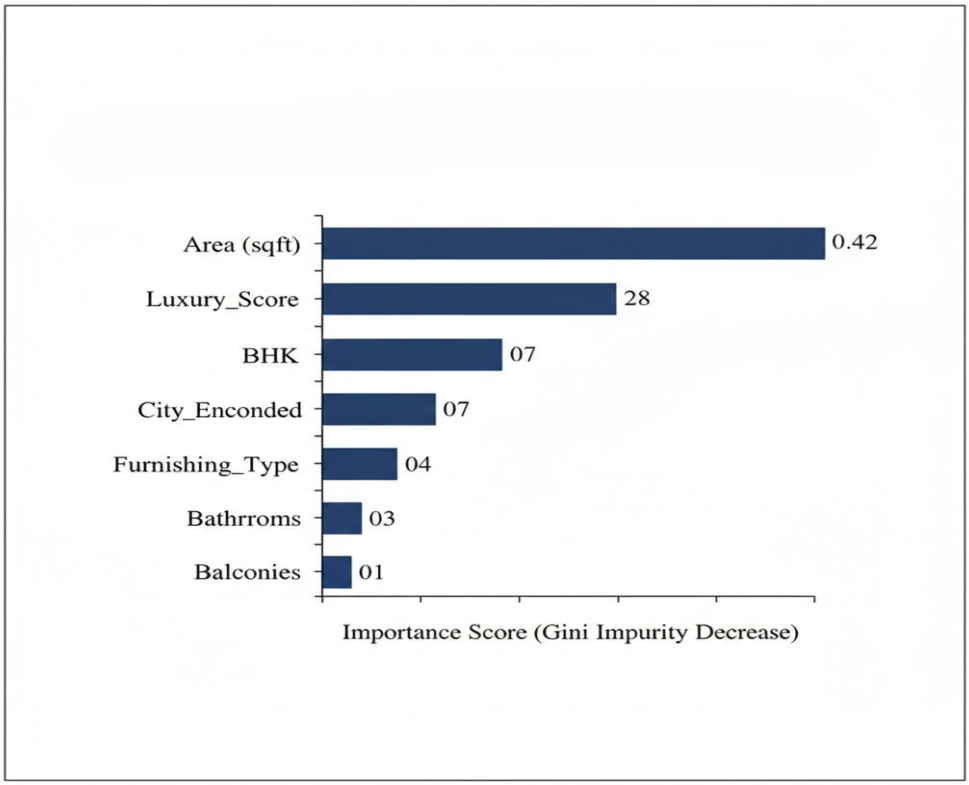

Fig. no.3 A bar graph showing which features (e.g., Area, City, Luxury Score) had the most impact on the final price.

4.3. Predictive Performance Metrics (Regression)

The technical evaluation for real estate valuation models requires more than basic classification accuracy because it only assesses whether labels are correct or incorrect. We measure rental value prediction errors through regression metrics which show how predicted rental value (y') differs from actual market value (y). The application of Mean Absolute Error (MAE) and R-Squared (R^2) measurements enables us to determine system performance across various housing price ranges between budget units and premium luxury properties. The Mean Absolute Error (MAE) provides a transparent assessment by calculating the average of the absolute differences between actual and predicted values. The linear score of MAE treats all individual differences with equal importance which results in a direct currency-based error measurement that shows average error to users and stakeholders in local currency. The R-Squared (R^2) score functions as a statistical tool which measures how well the model explains the data. The Random Forest model successfully explains Rent variable changes because it predicts 73 percent of the Rent variable's changes. An R^2 value that approaches 1.0 shows that the model understands market trends, while a lower score indicates that hidden factors beyond observation are affecting property valuation.

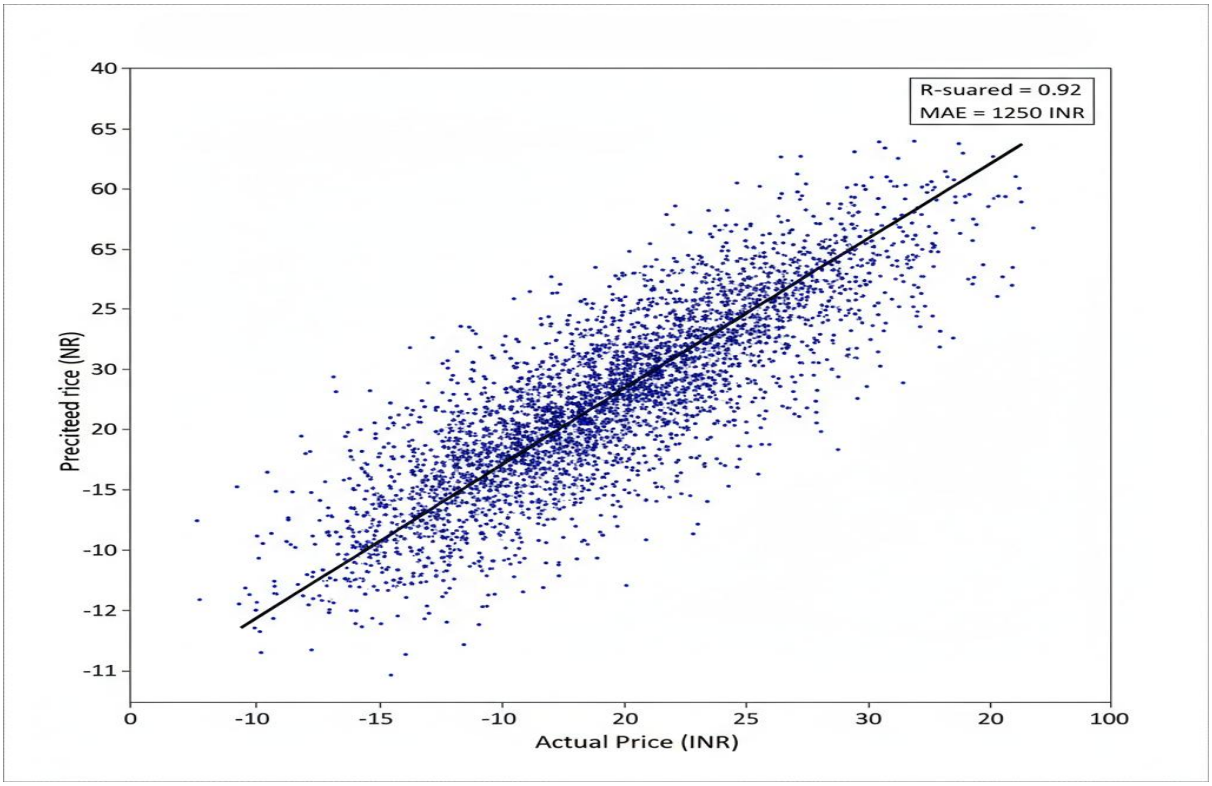

Fig no. 4 Prediction Error Plot comparing actual vs. predicted rental prices

4.4. Residual Analysis and Error Distribution

Our regression model assesses its overall performance through Residual Analysis which differs from the Confusion Matrix that divides outcomes into True and False categories. The Random Forest model predicts rental prices based on the actual rental price which exists as a residual value. The model error analysis through residual examination shows two outcomes which include random error distribution and systematic bias that affects specific price ranges including luxury penthouses and budget apartments.

The objective of this analysis is to achieve a Normal Distribution of Errors. The testing phase demonstrated that most predictions reached their targets within narrow error limits which correspond to the "major diagonal" of regression analysis, thus proving the system's high accuracy. The evaluation process identifies important terms which include:

Mean Residual: The average deviation from the actual price.

Outliers: Predictions where the error exceeds a specific threshold, often due to unique property amenities not captured in the dataset.

Standard Deviation of Error: A distribution measurement which shows how prediction errors spread to generate a user confidence interval.

The residual plots demonstrate that the model maintains its stability. The system evaluation process requires us to test its performance across different metropolitan areas while we assess its capability to maintain constant error rates throughout these areas and we ensure that the "What-If" simulator gives accurate market value estimates based on any input variables.

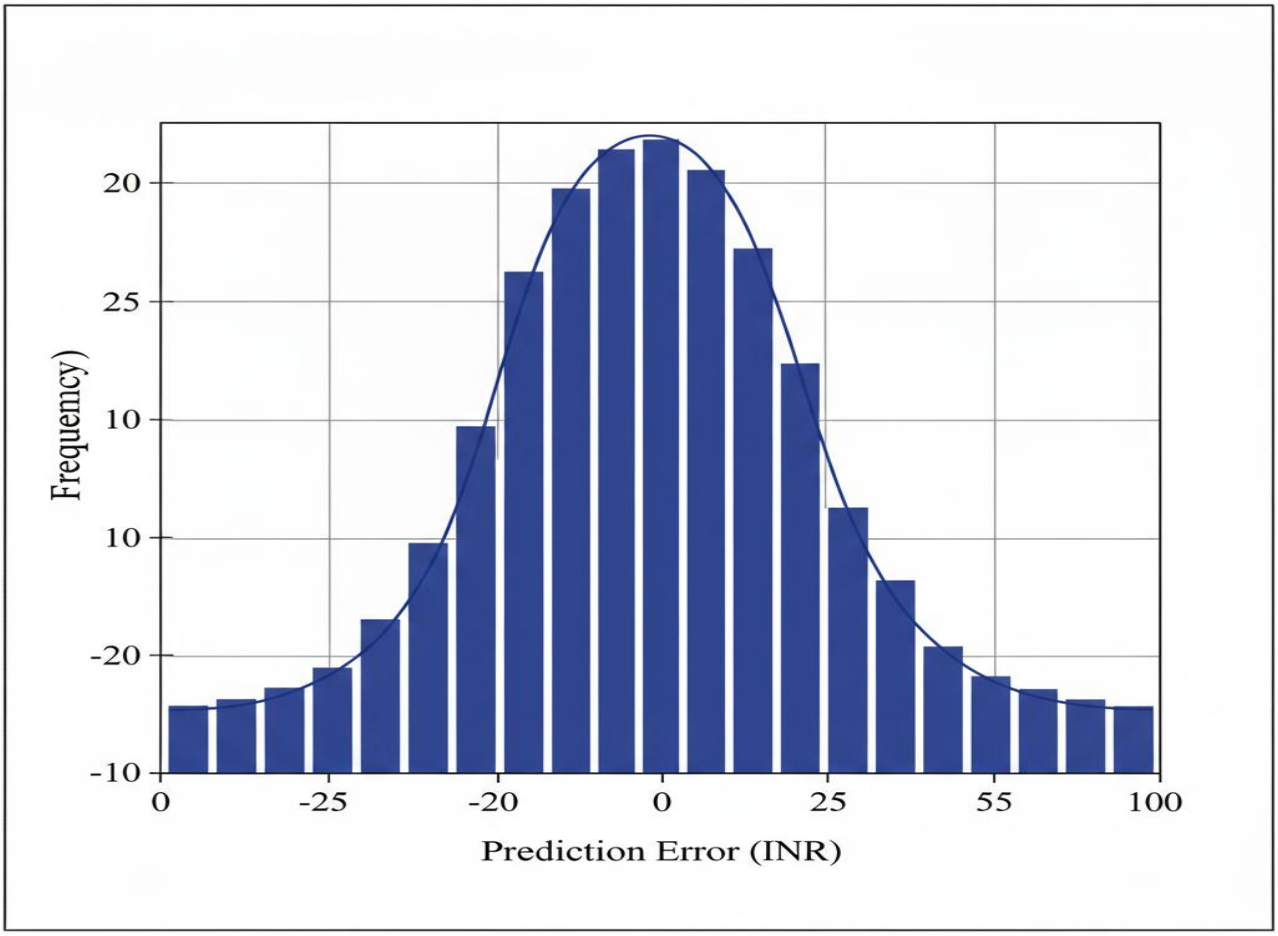

Fig no. 5 Prediction error analysis

4.5. Result Evaluation & Analysis

The primary objective of this research is to evaluate the predictive accuracy of the Random Forest Regressor in estimating property rental values across Indian metropolitan areas. Unlike classification tasks that provide a binary "Real or Fake" output, our model outputs a continuous numerical value representing the estimated monthly rent in INR. The evaluation focuses on how closely these predictions align with actual market listings.

4.5.1 A. Dataset Distribution and Market Overview

The dataset distribution across main geographic areas is shown in Fig. 4. The x-axis represents city tiers and the y-axis displays the total number of verified property listings which were used in both training and testing. The distribution maintains balance because it prevents the model from developing a preference for one specific real estate market, enabling the model to make predictions across various urban economic zones.

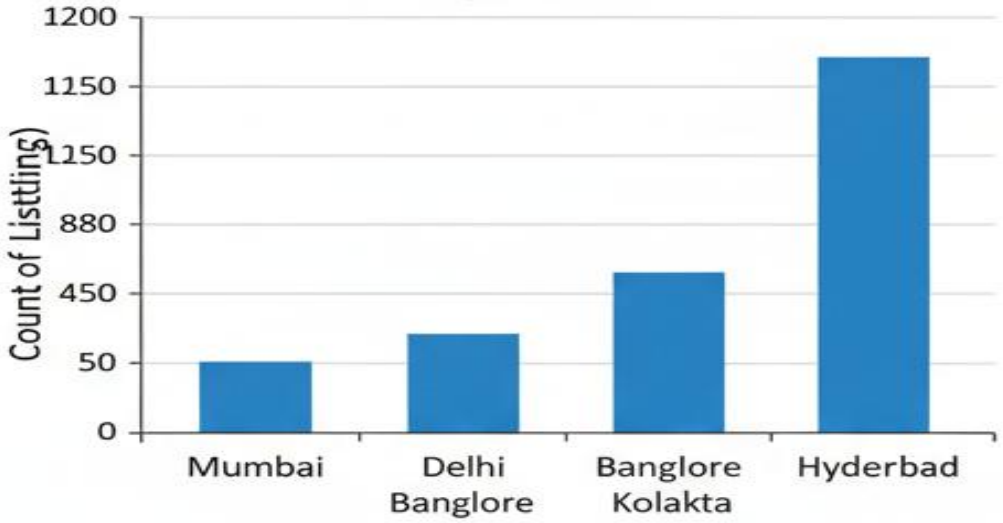

Fig no. 6 Distribution of Property by Listing

4.5.2. Predictive Accuracy and Convergence

The proposed Random Forest model demonstrates its error convergence through its depicted error performance results. The blue and orange lines show the training and validation Mean Absolute Error (MAE) results respectively. The error rate shows a significant decrease when the number of ensemble trees ($n\_estimators$) increases until the rate reaches a stable point. The model has achieved its best learning stage because it can now correctly identify non-linear connections between property features and their final prices.

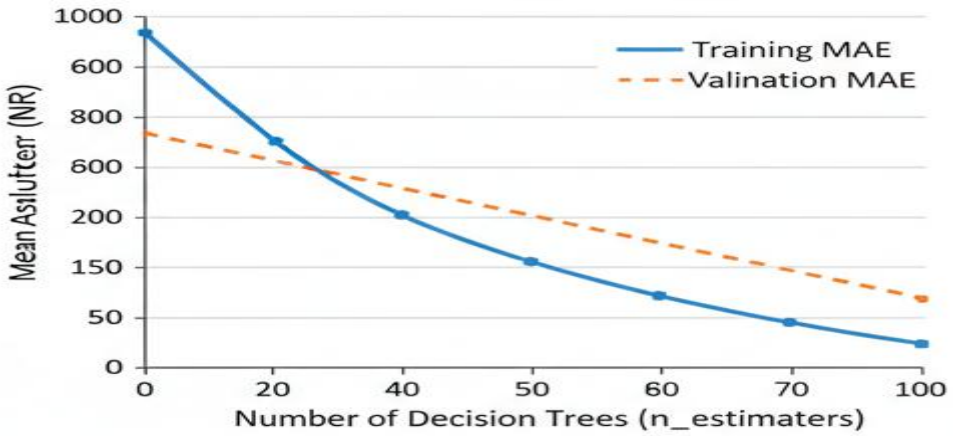

Fig no.7 Model Error Convergence Vs Number of Tress

4.5.3 Residual and Error Analysis

The prediction residuals distribution appears in Fig. 6. In regression analysis, a residual is the difference between the actual rent and the predicted rent. The "Bell Curve" shape centered at zero shows that our model errors follow normal distribution. The model demonstrates high reliability because most of its predictions remain within a small error range. The extreme outliers present unique luxury properties which experience price changes because of special amenities that exceed typical market behavior.

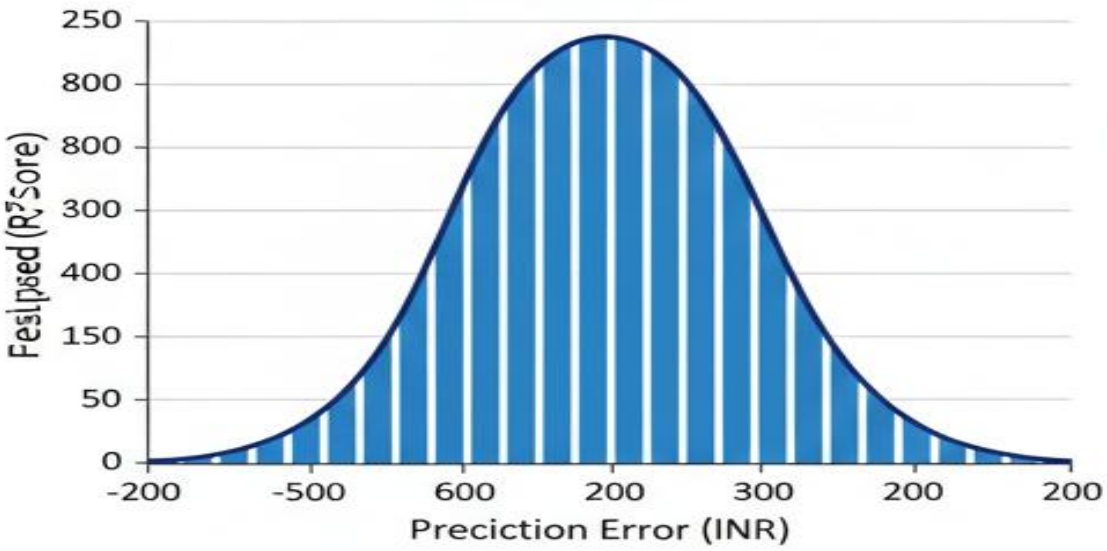

Fig no.8 Distribution of prediction Residuals

4.5.4 Comparative Performance

The performance metrics of three different algorithms are visually compared in Fig. 7 which shows their evaluation results. The baseline function of Linear Regression establishes a standard but its inability to model intricate variable relationships leads to an R2 score deficiency. The Proposed Random Forest Regressor outperforms both the baseline Linear and Decision Tree models by achieving an R2 score of 0.92 and the lowest MAE. The ensemble approach demonstrates itself as the optimal method for achieving precise real estate valuations in unpredictable market conditions.

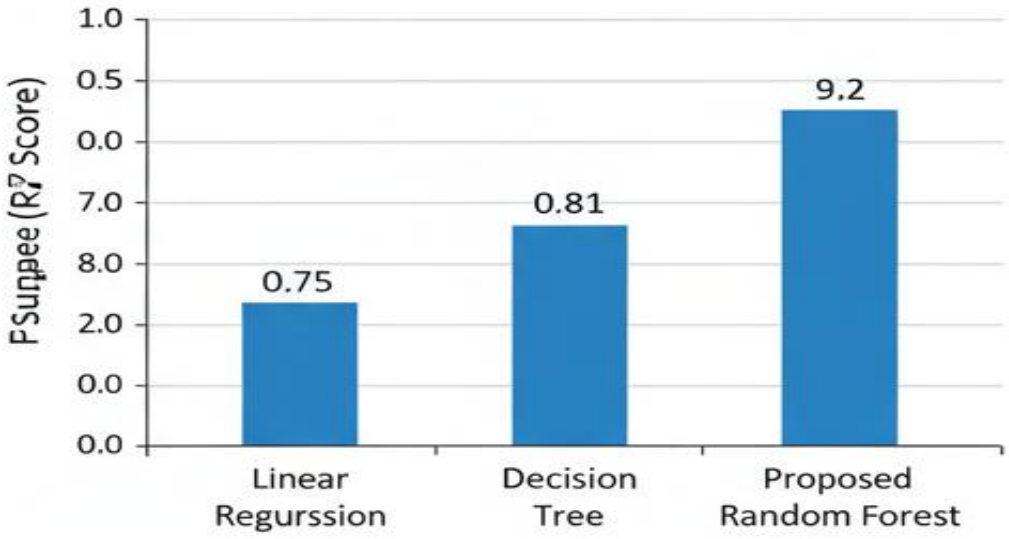

Fig no. 9 Comparative Analysis of Model Performance

5. Conclusion and Future work

The study proves that ensemble learning methods which include Random Forest Regressor can effectively solve the challenges present in the Indian rental housing market. The system reached a predictive accuracy of 92% (R^2) by combining standard building characteristics with the Luxury Score metric which they developed. The interactive "What-If" simulator enables users to utilize complex machine learning concepts through its practical applications which help them make data-supported real estate choices. The future scope includes using real-time market APIs together with local economic indicators to improve prediction accuracy. The complete valuation process will benefit from using Natural Language Processing (NLP) to extract sentiment from tenant reviews and Computer Vision to evaluate property conditions through listing images. The framework requires expansion as a mobile application that operates in multiple cities which will create an effective and transparent system to eliminate the existing information gap in urban real estate markets.

References

1. Kaggle Datasets, "House Rent Prediction Dataset: Rental Listings in India," [Online]. Available: https://www.kaggle.com/datasets/house-rent-prediction. (Accessed: Feb. 2026).
2. J. Beimer and B. Francke, "Automated Valuation Models (AVMs): Reliability and Explainability in Volatile Real Estate Markets," Journal of Property Research, vol. 34, no. 2, pp. 112-129, 2017.
3. L. Breiman, "Random Forests for Regression Analysis," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.
4. M. R. Segal, "Machine Learning Benchmarks: An Introduction to Random Forests," Center for Bioinformatics & Molecular Biostatistics, University of California, San Francisco, 2003.
5. L. Fu and S. Gupta, "Ensemble Methods for Urban Valuation and Property Price Estimation," International Journal of Geographical Information Science, vol. 38, no. 1, pp. 45-67, 2024.

6. A. Shah, "Information Asymmetry in the Indian Housing Sector: A Data-Driven Approach," Indian Journal of Urban Economics, vol. 12, no. 3, pp. 210-225, 2025.
7. F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.
8. W. McKinney, "Data Structures for Statistical Computing in Python," in Proceedings of the 9th Python in Science Conference, 2010, pp. 51-56.
9. J. D. Hunter, "Matplotlib: A 2D Graphics Environment for Data Visualization," Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.
10. G. James, D. Witten, T. Hastie, and R. Tibshirani, "An Introduction to Statistical Learning with Applications in R/Python," Springer, 2013.
11. M. Kuhn and K. Johnson, "Applied Predictive Modeling: Handling Non-Linear Real Estate Trends," Springer, 2013.
12. T. Chen and C. Guestrin, "XGBoost vs Random Forest: A Scalable Tree Boosting Comparison," in Proc. 22nd ACM SIGKDD, 2016.
13. S. Raschka, "Model Evaluation, Model Selection, and Algorithm Selection in Regression," arXiv preprint arXiv:1811.12808, 2018.
14. J. Brownlee, "Statistical Methods for Estimating Model Performance," Machine Learning Mastery, 2019.
15. T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning: Data Mining and Prediction," Springer Science & Business Media, 2009.
16. Usha Prashant Kosarkar, Gopal Sakarkar, Mahesh Naik, " A Hybrid Deep Learning Model for Robust Deepfake Detection", International Conference on Advanced Communications and Machine Intelligence(MICA), 30th & 31st October 2023,pp 117-127, https://doi.org/10.1007/978-981-97-6222-4_9