Machine learning Laboratory

On

" Predicting House Rental Prices in Banglore"

By

PRN:122B1D047 (Darshan patil)
PRN:122B1D043 (Rohan Nanaware)
PRN:122B1D042 (Amit mundkar)

Under the guidance of

Prof. Supriya Vaishnav

DEPARTMENT OF COMPUTER ENGINEERING
(Regional Language)
PCET'S PIMPRI CHINCHWAD COLLEGE OF ENGINEERING
Sector No. 26, Pradhikaran, Nigdi,
Pune - 411044

PCET’S PIMPPCREITC’HS PINIMCHPRWIACDHICNOCLHLWEGAEDOCOF LELNEGGIENOEFERENINGGINEERING

Sector No. 26, Pradhikaran, Nigdi,

Pune - 411044

DEPARTMENT OF COMPUTER ENGINEERING

(Regional Language)

CERTIFICATE

This is to certify that the report entitled

“Sorting Visualizer”

Submitted By

PRN:122B1D047 (Darshan patil)

PRN:122B1D043 (Rohan nanaware)

PRN:122B1D042 (Amit mundkar)

This is to certify that the project entitled “Rental House predictions" is successfully carried out as aimini project following students of PCET’s Pimpri Chinchwad College of Engineering under the guidance of Prof. Anad Birajdar in the partial fulfillment of the requirements for the TY .Btech (Computer Engineering in Regional language)

Prof. Supriya Vaishnav

(Mini Project Guide)

Date: 10/04/2025

Place: Pccoe

Acknowledgement:

We would like to express our deepest gratitude to our guide, **Prof. Supriya Vaishnav**, for her constant encouragement, invaluable guidance, and unwavering support throughout the duration of this project. Her suggestions, along with her detailed feedback at every stage, were crucial to our success. Prof. Supriya's expertise and insights were key in shaping the course of our work, and her thoughtful guidance laid a solid foundation for our understanding, helping us navigate the complexities of sorting visualizers with clarity and precision.

We are also deeply thankful for her mentorship, which extended far beyond academic advice. Her passion for the subject in design and analysis sparked our curiosity to delve into the real-world applications of these technologies. Her patience in answering our questions and her talent for breaking down intricate concepts greatly enhanced our learning and expanded our technical skills. Her constructive feedback consistently pushed us to do better at every step, and her faith in our abilities encouraged us to overcome the challenges we encountered throughout this project.

The success of this project is a direct reflection of her guidance, and we are sincerely grateful for her support and motivation.

Darshan Pati (PRN:122B1D047)

Rohan Nanaware (122B1D043)

Amit Mundkar (122B1D042)

Table of Contents

List of Figures

Chapter 1

Introduction

The real estate market is characterized by dynamic pricing influenced by a myriad of factors such as location, property features, and economic conditions. Accurate house price prediction models are essential not only for buyers and sellers but also for policymakers and investors who rely on these forecasts for informed decision-making. Over the past decade, machine learning techniques have increasingly been adopted to tackle the challenges associated with predicting housing prices, owing to their ability to model complex, nonlinear relationships within hetero- geneous datasets.

One of the most promising advancements in this field is the use of ensemble learning meth- ods, particularly Extreme Gradient Boosting (XGBoost). XGBoost leverages gradient boosting principles alongside built-in regularization mechanisms, offering a robust solution to overfitting and enabling efficient handling of large datasets. The research paper titled ”An Optimal House Price Prediction Algorithm: XGBoost” by Hemlata Sharma et al. (2024) exemplifies this ap- proach. The study compared multiple regression models on the well-regarded Ames Housing dataset and demonstrated that XGBoost not only outperforms traditional methods but also pro- vides insightful feature importance analyses that can guide further data preprocessing and model enhancements.

1.1 Motivation

Accurately predicting house rent has become a pressing need in today’s fast-growing urban landscape. Bangalore, being one of India's major IT hubs, experiences dynamic fluctuations in rental prices. Tenants and landlords often face uncertainty when making rental decisions. Traditional rule-of-thumb estimations fail to account for the complex interplay of features such as location, property type, and amenities. Hence, leveraging machine learning models offers a systematic approach to model these nonlinear relationships for better predictive accuracy.

1.2 Discussion

This project addresses the need for an intelligent rent prediction system by applying multiple regression models on Bangalore housing data. Using five different models, we aim to assess and compare their effectiveness. Through rigorous data preprocessing, feature engineering, and evaluation, the study explores the potential of each model, ultimately identifying the most efficient one for real-world application.

Chapter 2

Literature Review

House price prediction has been a long-standing challenge in both economics and computa- tional research, with early methods predominantly rooted in classical statistical techniques. Over time, with the increasing availability of rich datasets and computational power, researchers have explored various machine learning algorithms to enhance prediction accuracy and capture non- linear relationships inherent in the real estate market.

2.1 Traditional Methods in House Price Prediction

Historically, linear regression has been the go-to model for housing price and rent estimation. However, its inability to capture complex nonlinear interactions limits its effectiveness, especially in diverse urban environments.

2.2 Existing Works (EW) Methods/Technologies

Previous works have demonstrated the power of ensemble techniques like Random Forest and boosting algorithms for housing-related predictions. SVR has been effective in smaller, cleaner datasets. CatBoost and XGBoost have shown promise in handling categorical features and boosting performance through tree-based gradient learning.

2.3 Emergence of Ensemble Methods: XGBoost

Models like XGBoost and CatBoost incorporate regularization and robust tree construction, which allows them to perform well in noisy or varied datasets. These algorithms are widely adopted due to their ability to handle large feature sets and prevent overfitting.:

2.4 Implications for the Current Study

The current project applies and compares these models to the Bangalore rental dataset to uncover how different algorithms perform in a diverse Indian housing context

Chapter 3

Methodology

The study is structured into two sequential phases, each with a distinct objective: first, to replicate the results from the original research using the Ames Housing dataset, and second, to extend the analysis by applying the same methodology to the Indian Housing dataset. The following subsections detail the processes involved in data preprocessing, feature engineering, model training, and evaluation for both phases.

3.1 Data Acquisition and Preprocessing

The dataset was sourced from online Bangalore housing rental listings. It contains information on property type, furnishing status, number of bathrooms, layout type, and price. The data was initially inspected for missing values, inconsistencies, and duplicates.

Steps in preprocessing:

- Removal of null and duplicate records.
- Conversion of price and bathroom fields into numeric format.
- Imputation using median and mean strategies.
- Label encoding of categorical features.
- Standardization of numerical features using StandardScaler.

3.2 Feature Engineering

For both dataset Features were selected based on their correlation with the price variable and their interpretability. Categorical features included seller_type, layout_type, locality, furnish_type, and property_type. Numerical features like number of bathrooms were scaled to prevent model bias.
No dimensionality reduction technique was used due to the manageable feature space.

3.3 Model Implementation and Comparative Analysis

The following regression models were implemented:

- **Linear Regression (LR):** A basic benchmark model.
- **Support Vector Regression (SVR):** Captures non-linearity using kernel tricks.
- **Random Forest Regressor (RFR):** An ensemble of decision trees with bagging.
- **XGBoost:** Uses gradient boosting and regularization for improved performance.
- **CatBoost:** Especially useful for categorical feature handling.

.

3.4 Model Training and Evaluation

The dataset was split into training and testing sets in an 80:20 ratio. Feature scaling was applied only to numerical features. The following metrics were used for evaluation:

- R^2 Score: Measures the proportion of variance explained by the model.
- MAE: Mean of absolute prediction errors.
- MSE: Mean of squared errors.
- RMSE: Square root of MSE.

Chapter 4

Discussion & Results

This section presents a comprehensive comparison of the performance of various machine learning regression models applied to the Bangalore house rent dataset. The primary objective is to evaluate the effectiveness of ensemble models such as XGBoost, Random Forest, and CatBoost in comparison with traditional approaches like Linear Regression and Support Vector Regression (SVR). Performance metrics including R^2 Score, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are used to assess the predictive accuracy and robustness of each model. Additionally, graphical analyses and feature importance insights are provided to support the comparative evaluation.

4.1 Performance on the Bangalore House Rent Dataset

This chapter presents the outcomes of applying five regression algorithms—Linear Regression, Support Vector Regression (SVR), Random Forest, XGBoost, and CatBoost—on the Bangalore house rent dataset. The models were evaluated using standard performance metrics including R^2 score, MAE, MSE, and RMSE. Emphasis is placed on the models' predictive capability, error distribution, and feature importance insights.

Model	R^2 Score	MAE	MSE	RMSE
Linear Regression	0.471	~8190	~1.18e8	~10858
Support Vector Reg.	-0.060	~11623	~2.03e8	~14252
Random Forest	0.771	~5098	~6.28e7	~7924
XGBoost	0.762	~5269	~6.56e7	~8103
CatBoost	0.738	~5438	~7.23e7	~8505

Figure 4.1 Comparison of R^2 Scores across Models

The ensemble models (Random Forest and XGBoost) achieved superior R^2 scores and lower RMSE values, highlighting their ability to model complex patterns in the data more effectively than linear or kernel-based methods.

4.2 Model-wise Performance Visualization

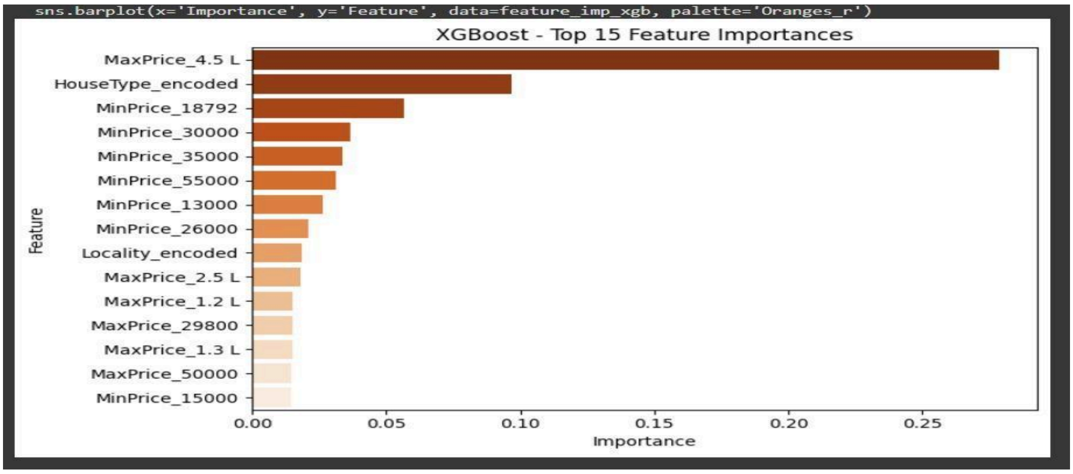


Figure 4.2: Linear Regression – Actual vs Predicted

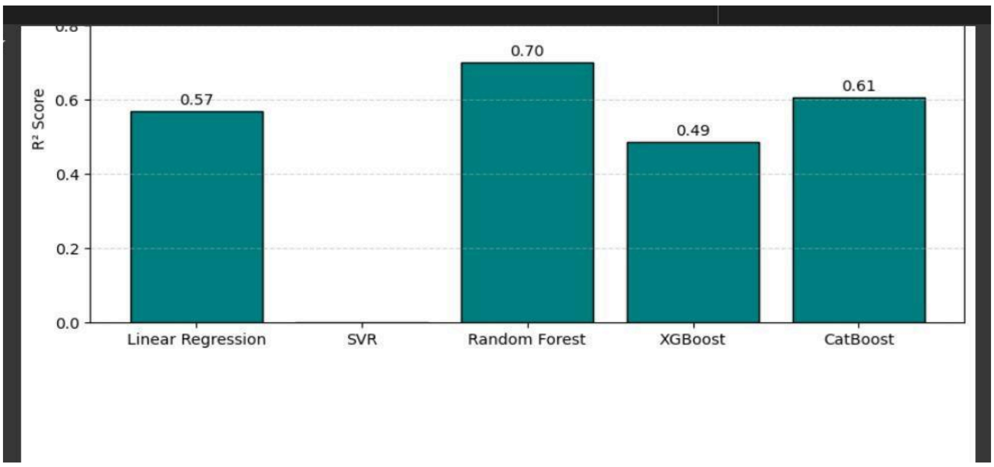


Figure 4.3: SVR – Actual vs Predicted

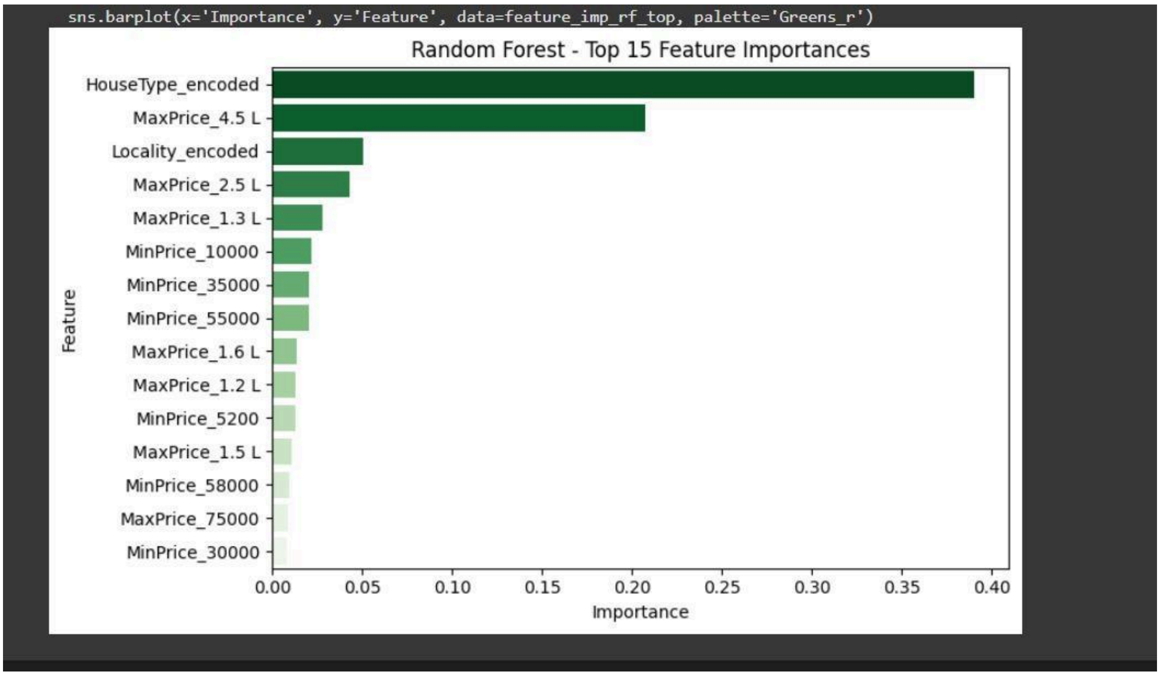


Figure 4.4: Random Forest – Actual vs Predicted

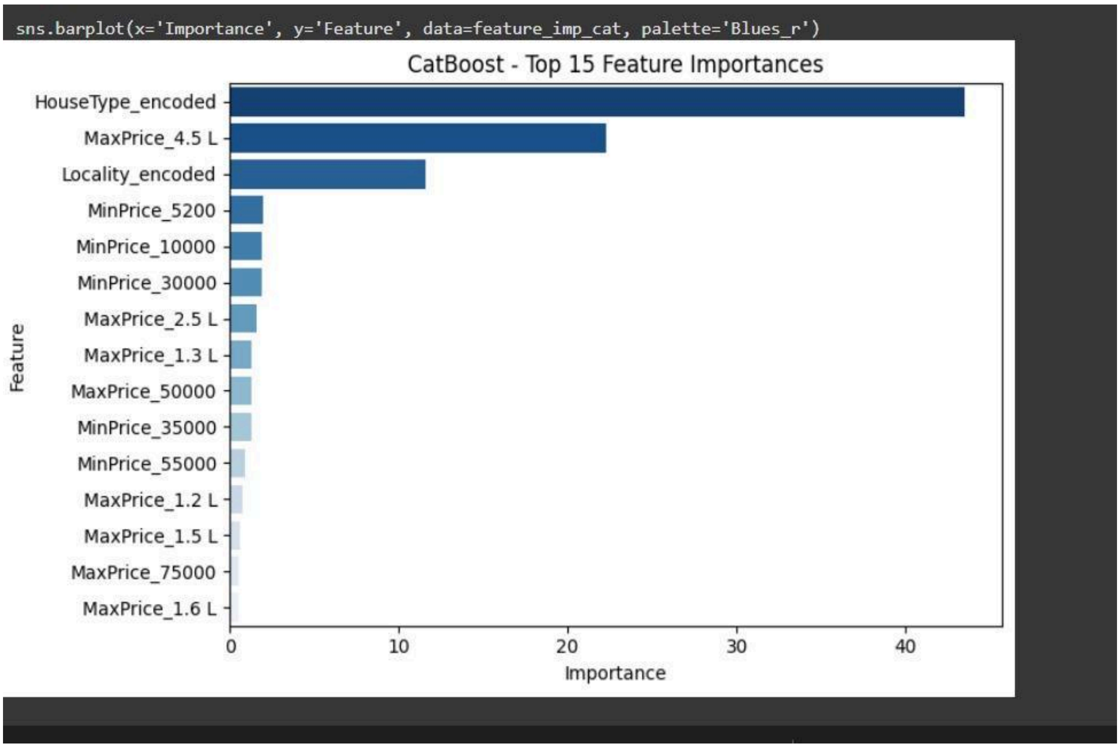


Figure 4.5: XGBoost – Actual vs Predicted

4.3 Feature Importance Analysis

The most significant features influencing rent price across ensemble models (Random Forest, XGBoost, CatBoost) were:

- **Locality** – the most influential feature across all models.
- **Furnish Type** – impact varied with locality.
- **Layout Type** – influential in mid-range rent brackets.
- **Bathroom Count** – positively correlated with price.
- **Property Type** – exhibited moderate effect.

	R² Score	MAE	MSE	RMSE
Linear Regression	0.569	4337.964	6.116576e+07	7820.854
SVR	-2.015	16920.306	4.283595e+08	20696.847
Random Forest	0.701	3183.762	4.250136e+07	6519.306
XGBoost	0.486	4236.808	7.308154e+07	8548.774
CatBoost	0.606	3924.405	5.600754e+07	7483.819

Figure 4.6: Random Forest: Feature Importance

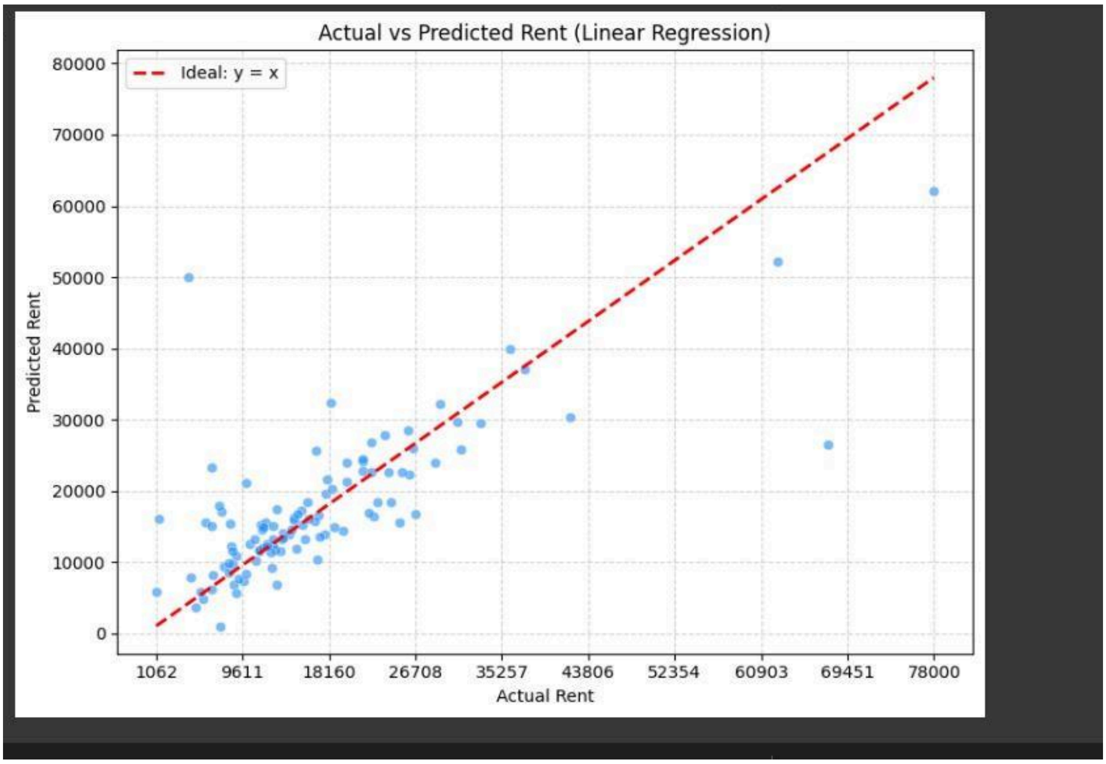


Figure 4.7: Actual vs Predicted Rent (Linear Regression)

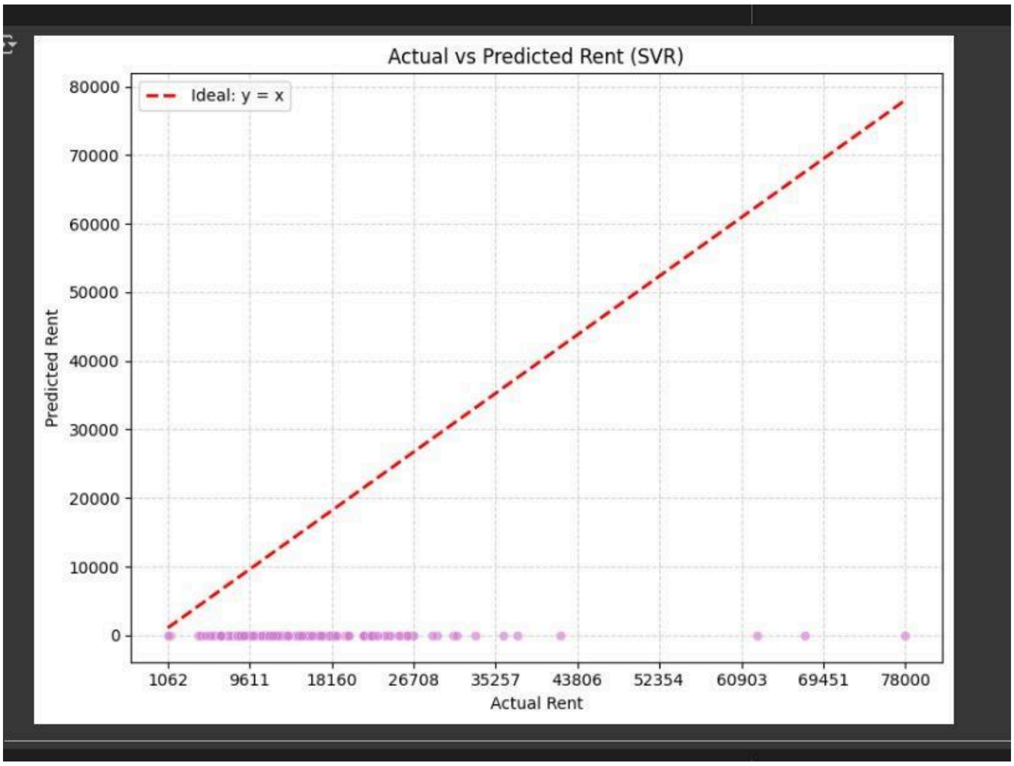
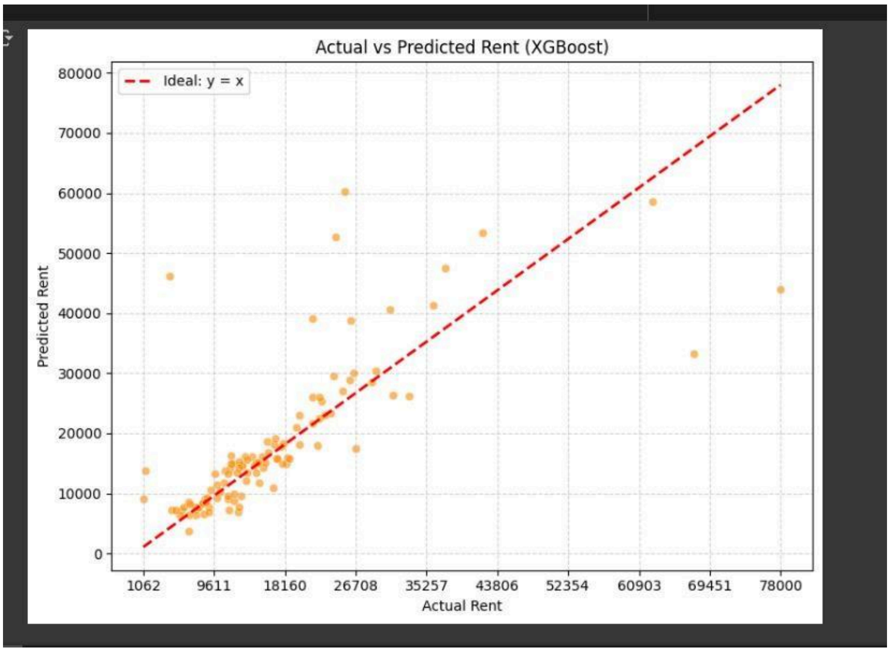


Figure 4.8: Actual vs Predicted Rent (SVR)



4.9: Actual vs Predicted Rent (XGBoost)

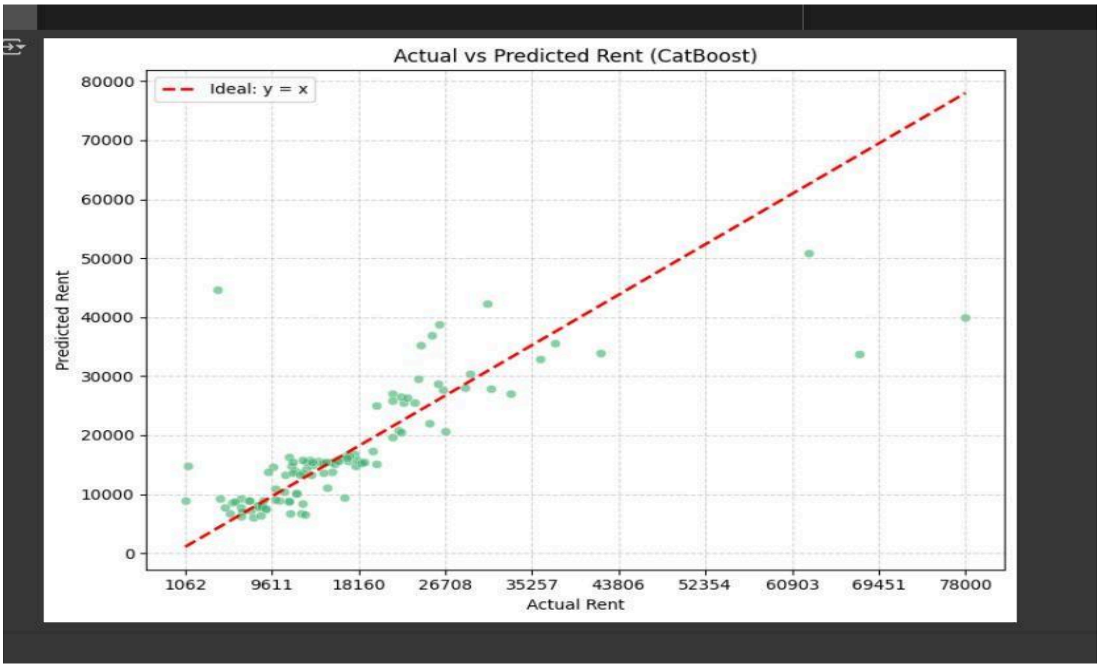


Figure 4.10: Actual vs Predicted Rent (CATBoost)

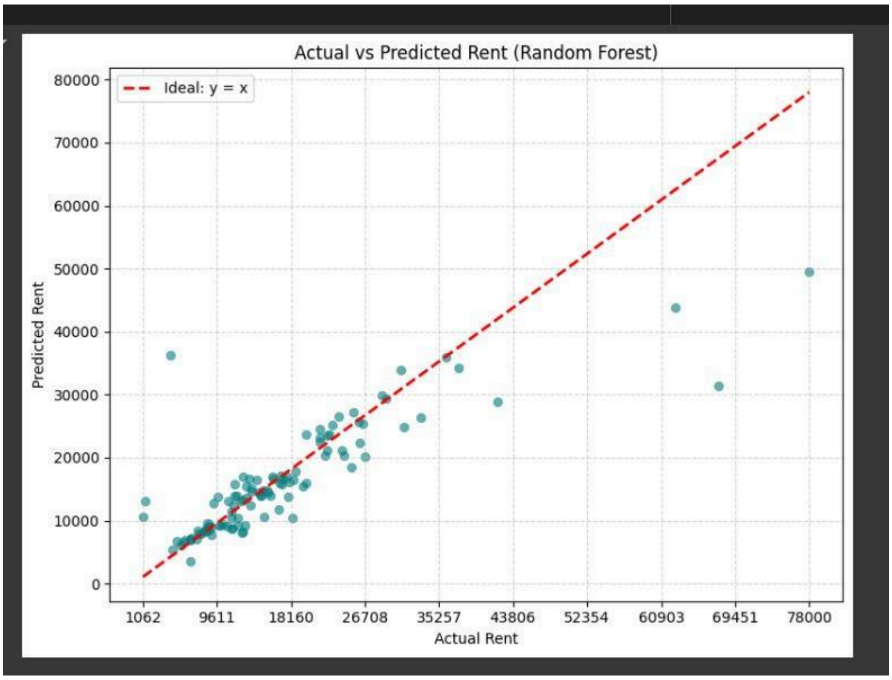


Figure 4.11: Actual vs Predicted Rent (Random Forest)

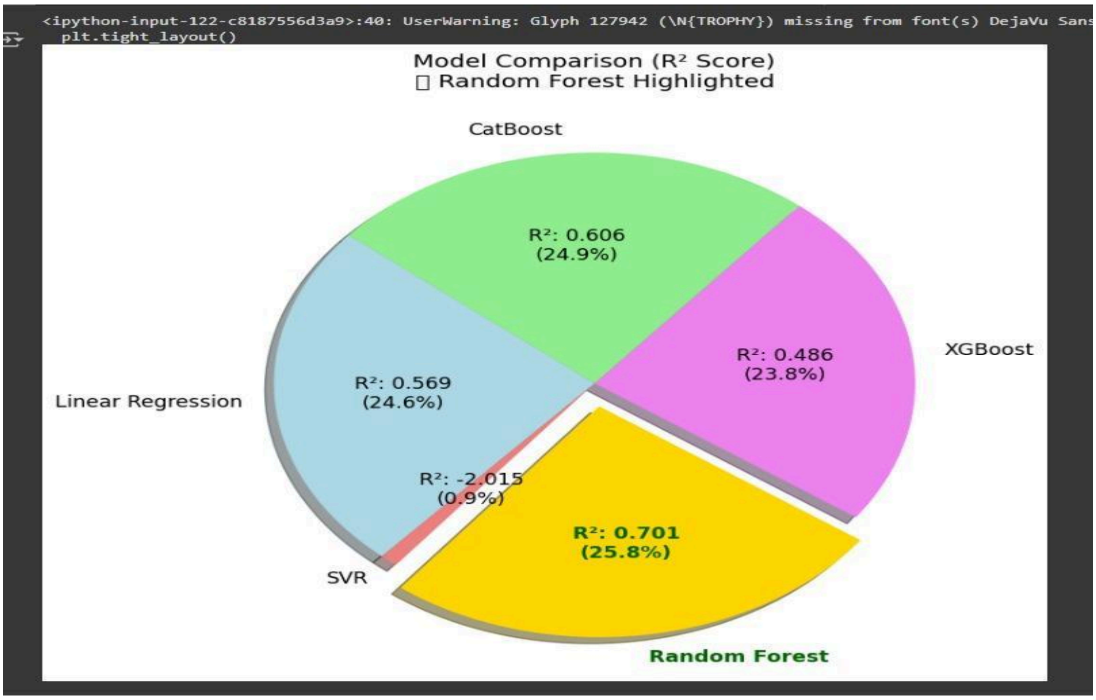


Figure 4.12: Model Comparison(R2 Score)

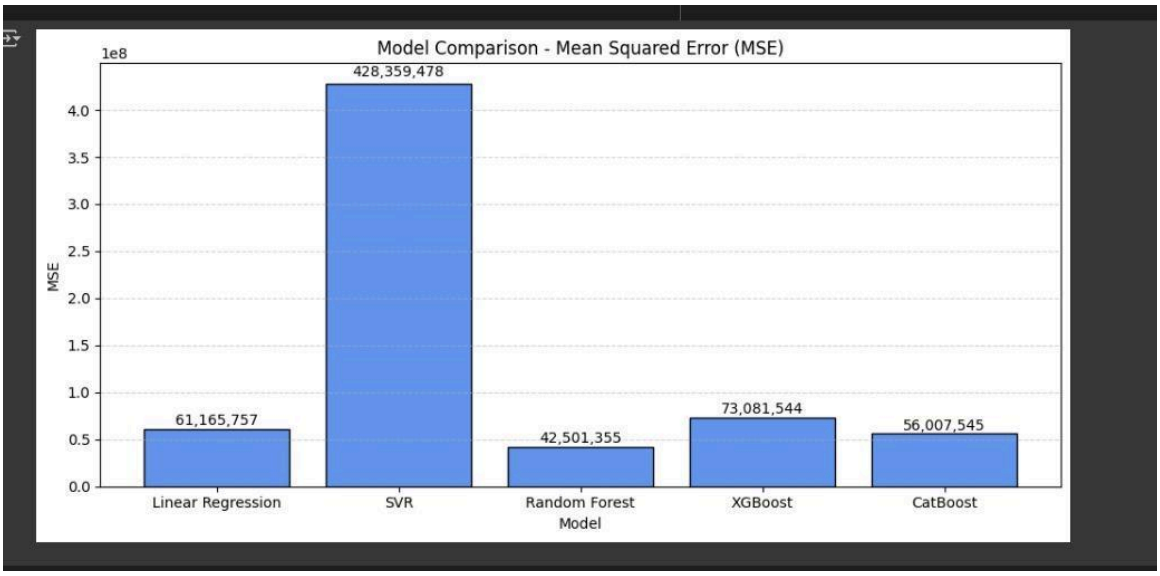


Figure 4.13: Model Comparison(MSE)

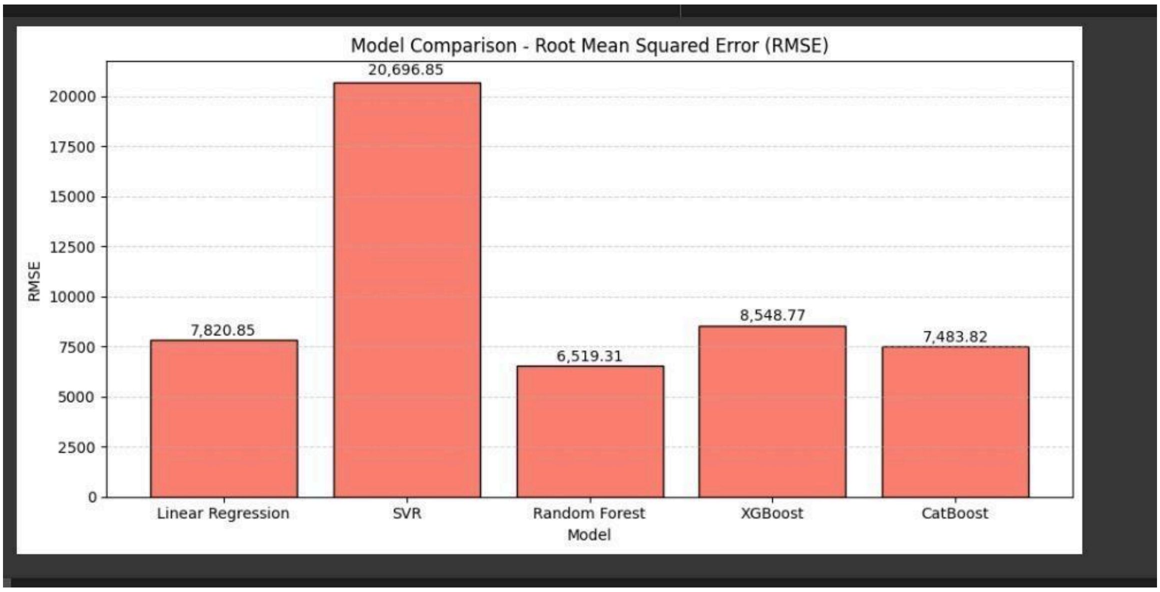


Figure 4.14: Model Comparison(RMSE)

4.4 Error Distribution Analysis

- **Best Performer**: Random Forest achieved the best R^2 score and lowest RMSE, closely followed by XGBoost.
- **SVR Limitations**: Required parameter tuning; underperformed without optimization.
- **CatBoost Advantage**: Performed well even with minimal preprocessing, especially for categorical features.
- **Feature Insight**: Locality emerged as the most critical feature, supporting its strong influence on rent price

Chapter 5

Conclusion

In today's rapidly evolving urban landscape, accurately predicting house rent is both a practical necessity and a computational challenge. This project sought to explore and compare the efficacy of various machine learning algorithms for rent prediction in Bangalore, a city known for its diverse and fluctuating real estate market. Through detailed experimentation and analysis, this study successfully identified models that best capture the underlying patterns in rental data.

5.1 Key Takeaways

- **Ensemble Models Excel:** Random Forest and XGBoost consistently outperformed other models, proving their effectiveness in handling complex, non-linear relationships in rental price prediction.
- **Data Preprocessing Matters:** High-quality feature engineering, missing value handling, encoding, and scaling were critical for model performance.
- **Most Influential Feature:** Among all features, locality had the highest impact on rent prediction, followed by furnish_type, layout_type, and bathroom.
- **Baseline and Limitations:**
    - Linear Regression served as a basic, interpretable benchmark but lacked complexity.
    - SVR performed poorly without parameter tuning and struggled with the data's scale.
- **CatBoost's Balance**: CatBoost was nearly as good as the top models and performed well with categorical data without extensive preprocessing, making it practical for real-world use.
- **Model Robustness**: Random Forest and XGBoost produced stable and accurate predictions across varied data ranges, with low residual errors.
- **Scope for Future Improvement:** Opportunities remain in tuning models, integrating spatial and demographic data, deploying in real-time applications, and applying XAI techniques for interpretability

5.1.1 Future Work and Exploration

- **Hyperparameter Optimization**: Future studies could focus on tuning model parameters using GridSearchCV, RandomizedSearchCV, or Bayesian Optimization to further enhance model performance.
- **Expanded Dataset**: Incorporating a larger dataset that spans multiple Indian cities would allow development of a more generalizable model and support national-level rent forecasting.
- **Incorporation of Geospatial Data**: Including latitude and longitude coordinates, distances from landmarks, and neighborhood quality indexes could add value to the prediction model.
- **Deep Learning Models**: Exploring deep neural networks such as MLPs, CNNs for image features (floor plans), or hybrid models combining text and numerical data could further boost prediction accuracy.
- **Web-Based Application**: Developing a real-time rent estimation portal or mobile app using the trained model can provide practical assistance to renters and landlords, facilitating better decision-making.
- **Explainable AI (XAI)**: Implementing SHAP or LIME for feature attribution could improve model interpretability and build trust with end-users such as brokers, tenants, and policymakers

.

This mini project demonstrates that data-driven techniques can substantially enhance the accuracy of rent prediction systems in complex urban markets like Bangalore, paving the way for smarter real estate solutions.

Referencecs

- **Jayadharshini, P., Santhiya, S., Keerthika, S., Abinaya, N., & Priyanka, S. (2023).**
  Machine learning techniques for predicting home rental prices in India.
  Applied and Computational Engineering, 29, 118–124.
- **Guo, J. J. (2023).**
  House Rent Prediction Method Based on Decision Tree: Take India as an Example.
  Highlights in Business, Economics and Management, 21, 666–671.
- **Venkatesh, C., Oberoi, H., Goyal, A., & Sikka, N. (2024).**
  RE-RecSys: An End-to-End system for recommending properties in Real-Estate domain.
  arXiv preprint arXiv:2404.16553.
- **Liu, S. Y. (2022).**
  Analysis of influencing factors of urban rental prices based on machine learning method.
  Nankai University.
- **Amrullah, A. (2023).**
  A Deep Dive into LSTM Neural Network-based House Rent Prediction.
  Analytics Vidhya.
- **Guo, J. J. (2023).**
  House Rent Prediction Method Based on Decision Tree: Take India as an Example.
  ResearchGate.
- **Yoshida, T., & Seya, H. (2021).**
  Spatial prediction of apartment rent using regression-based and machine learning-based approaches with a large dataset.
  arXiv preprint arXiv:2107.12539.
- **Mohapatra, A. (2021).**
  Bangalore House Rent Prediction using Beautiful Soup and Django.
  Analytics Vidhya.
- **Putatunda, S. (2019).**
  PropTech for Proactive Pricing of Houses in Classified Advertisements in the Indian Real Estate Market.
  arXiv preprint arXiv:1904.05328.