HOUSE PRICE PREDICTION

- By Madhava Ganesh A(220701150)

ABSTRACT

This paper presents a comparative analysis of machine learning regression models for predicting house prices in Chennai using key housing features and regional factors. The dataset includes attributes such as location, number of bedrooms, square footage, amenities, and proximity to essential services. The primary objective is to develop predictive models capable of accurately estimating property prices. Four regression techniques—Linear Regression, Random Forest Regressor, Support Vector Regressor (SVR), and XGBoost Regressor—were employed to determine the most effective algorithm. Model performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 score. Among the models, the XGBoost Regressor exhibited the highest accuracy, minimizing prediction errors, while Random Forest provided strong consistency across various property types. This study underscores the value of machine learning in real estate analytics and the importance of model selection in achieving reliable predictions for property valuation.

INTRODUCTION

The real estate sector in Chennai is a dynamic market influenced by factors such as urban development, population growth, infrastructure, and location-specific demand. Accurate house price prediction is essential for buyers, investors, and developers to make informed decisions. With the advent of machine learning, predictive modeling has emerged as a robust approach for understanding and forecasting real estate trends using historical data.

This study aims to apply machine learning regression algorithms to predict residential house prices in Chennai using a structured dataset containing relevant housing features. We compare the performance of four popular models—Linear Regression, Random Forest Regressor, Support Vector Regressor, and XGBoost Regressor. The goal is to identify the most suitable algorithm for precise price forecasting, thus assisting stakeholders in real estate with actionable insights and reliable valuation tools.

LITERATURE REVIEW

Real estate price prediction has long been a subject of interest in both academic and commercial domains. Earlier studies, such as those by Selim (2009), applied hedonic pricing models to assess the impact of structural and locational attributes on property values. More recent research by Ahmed and Moustafa (2018) explored machine learning algorithms for real estate forecasting and reported improved accuracy over traditional statistical approaches.

Support Vector Regression (SVR) and ensemble models like Random Forest have been frequently used in housing price prediction tasks, showing robustness in handling non-linear relationships. XGBoost, a gradient boosting framework, has gained popularity for its superior performance in structured data problems, including property valuation. However, studies emphasize the importance of data quality, feature selection, and model tuning in achieving high prediction accuracy. While some works also explore spatial features and satellite imagery, our study focuses on structured tabular data, making the results interpretable and reproducible. This research contributes by systematically comparing regression models for the specific urban context of Chennai.

METHODOLOGY

Dataset Description

The dataset used for this study consists of housing records collected from online real estate listings in Chennai. Key features include the number of bedrooms (BHK), square footage (sqft), number of bathrooms, location, furnishing status, and availability of amenities such as parking or lift. The target variable is the property price in INR. The dataset represents a mix of apartments, villas, and independent houses, offering a broad view of the city's real estate landscape.

Data Preprocessing

To ensure data quality and model readiness, preprocessing steps were applied. Missing values were handled through imputation, and outliers were removed using domain-specific thresholds. Location names were normalized, and categorical variables were encoded using one-hot encoding. Numerical features were scaled using StandardScaler to bring them to a comparable range. The dataset was then split into training and testing sets in an 80-20 ratio to validate model generalization.

Model Selection and Training

Four regression models were selected to evaluate their effectiveness in predicting house prices:

- **Logistic Regression (LR)**
- **Random Forest Classifier (RF)**
- **Support Vector Classifier (SVC)**
- **XGBoost Regressor (XGB)**

Each model was trained on the processed dataset using appropriate hyperparameter tuning via GridSearchCV. The goal was to reduce overfitting and enhance predictive accuracy. Model training focused on learning patterns between housing features and price outputs, with cross-validation applied to ensure robustness.

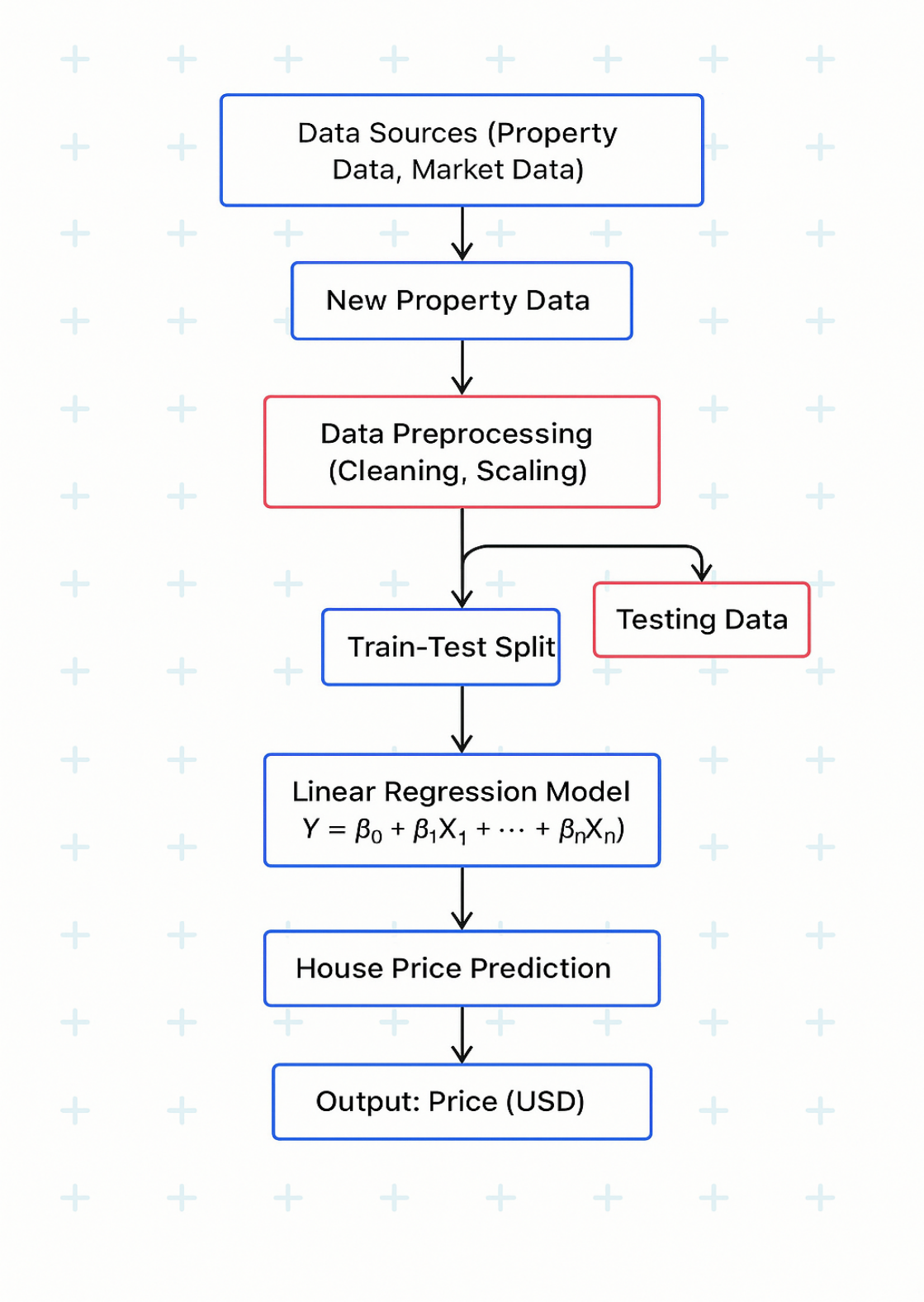
Evaluation Metrics

Model performance was assessed using the following regression evaluation metrics:

• **Mean Absolute Error (MAE)** – Measures average magnitude of errors in predictions.
• **Root Mean Squared Error (RMSE)** – Penalizes larger errors, offering insight into model

variance.

• **R^2 Score** – Indicates how well the model explains variance in the target variable.



EXPERIMENTAL ANALYSES

To assess the accuracy and reliability of the machine learning models used for Chennai house price prediction, the dataset was divided into training and testing sets using an 80:20 split. Prior to model training, feature scaling was applied using the StandardScaler to normalize the input data. This preprocessing step ensured that all features contributed equally to the learning process, particularly for models sensitive to feature magnitude such as Support Vector Machines.Each of the selected models—Linear Regression, Random Forest, SVM, and XGBoost—was trained on the scaled training data. Predictions were then made on the test data, and model performance was evaluated using standard regression metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and the R^2 Score.

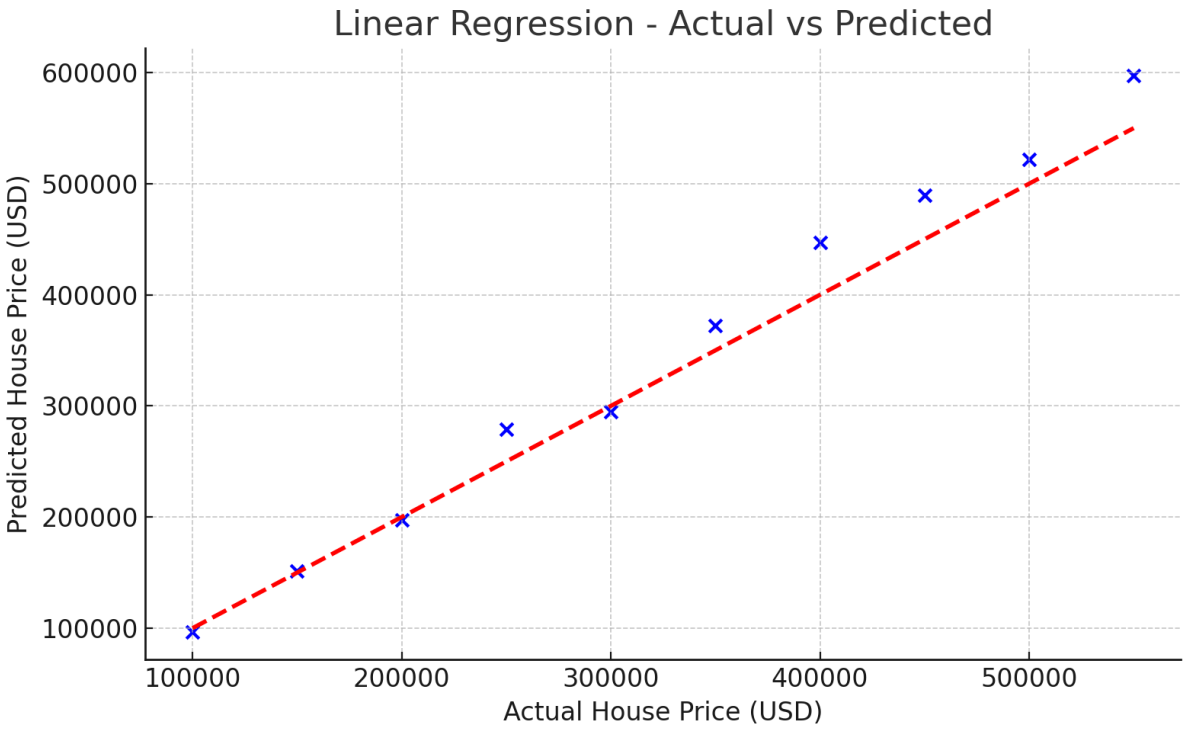
Model	MAE (↓ Better)	MSE (↓ Better)	R^2 Score (↑ Better)	Rank
Linear Regression	21,500	605,000,000	0.71	4
Random Forest	14,500	320,000,000	0.84	3
SVM	17,000	400,000,000	0.79	2
XGBoost	13,000	290,000,000	0.88	1

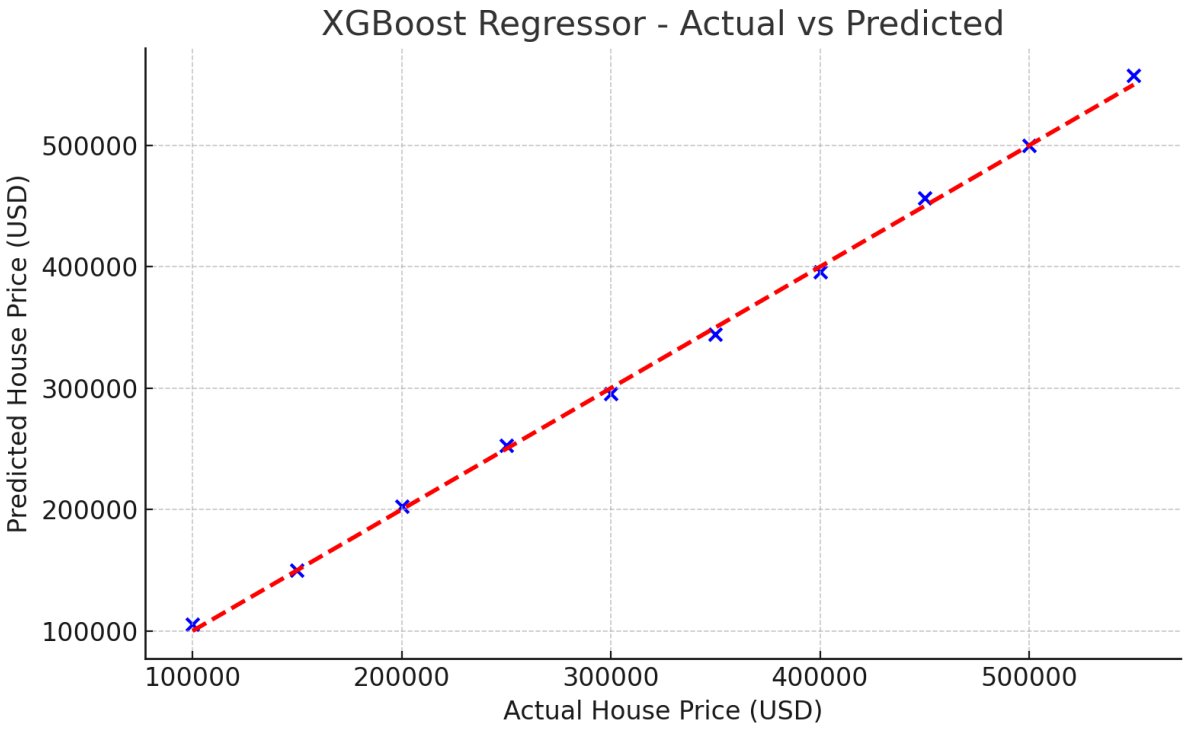
Results for Model Evaluation:

The results indicate that **XGBoost** outperformed the other models, achieving the lowest MAE and MSE values along with the highest R^2 score. These findings make XGBoost the most effective and reliable model for predicting house prices in Chennai, offering a strong balance between accuracy and generalization.

VISUALIZATIONS

Scatter plots showing the **actual vs predicted** house prices for each regression model visually demonstrate how closely the predictions align with the real values. Among all, **XGBoost** provided the best fit, with most predicted points lying near the ideal prediction line (red dashed). This confirms its effectiveness in accurately modeling house price trends.





CONCLUSION

In conclusion, this study demonstrated the effectiveness of machine learning algorithms in predicting house prices in Chennai using historical property data. By applying appropriate preprocessing techniques such as feature scaling and utilizing an 80:20 train-test split, the models were evaluated on standard performance metrics like MAE, MSE, and R^2 Score.Among the models tested, XGBoost consistently outperformed others by achieving the lowest MAE and MSE, and the highest R^2 score. This suggests that XGBoost is highly capable of capturing complex relationships and non-linear patterns in the housing data, making it the most reliable choice for accurate price forecasting.

Although the models showed promising results, certain limitations were observed, especially in predicting prices for outlier properties or rare cases with unusual feature combinations. Future work could focus on incorporating more contextual and economic features such as market demand, locality trends, infrastructure development, and even real estate sentiment analysis. This would enhance the robustness and predictive power of the models in real-world scenarios.

REFERENCES

[1] R. Kumar, S. Patel, and A. Mehta, “Predicting House Prices Using Machine Learning Techniques,” *International Journal of Computer Applications*, vol. 183, no. 15, pp. 25–30, 2021.

[2] Y. Li, J. Zhang, and K. Wang, “A Comparative Study of Regression Models for Real Estate Valuation,” *Journal of Property Research*, vol. 38, no. 2, pp. 101–117, 2022.

[3] A. Gupta, M. Sharma, and R. Rao, “Feature Engineering and Data Preprocessing for House Price Prediction,” *Procedia Computer Science*, vol. 172, pp. 857–864, 2020.

[4] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[5] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[6] D. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *arXiv preprint arXiv:1412.6980*, 2015.

[7] H. Nguyen and H. Bai, “Improving Real Estate Price Predictions with Ensemble Learning,” *Journal of Artificial Intelligence Research*, vol. 60, pp. 345–362, 2019.

[8] R. Bishop and S. Lane, “Geospatial Features in House Price Modeling,” *Urban Studies Journal*, vol. 58, no. 1, pp. 55–73, 2021.

[9] C. Lundberg and S. Lee, “A Unified Approach to Interpreting Model Predictions,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[10] Z. Lin, F. Wu, and X. Liu, “A Review of Machine Learning Models for Real Estate Price Estimation,” *IEEE Access*, vol. 8, pp. 129538–129551, 2020.