Regression in Real-Estate Industry

Kiran Ingale
Dept. EN-TC
kiran.ingale@vit.edu
Professor at VIT PUNE

Abdulshakkur Shaikh
Branch CS-AI
abdulshakkur.shaikh24@vit.edu
Student at VIT , Pune

Govind Vasave
Branch CS-AI
govind.vasave24@vit.edu
Student at VIT, Pune

Shubham Ingale
Branch CS-AI
shubham.ingale24@vit.edu
Student at VIT , Pune

Nikhil Patil
Branch CS-AI
nikhil.patil24@vit.edu
Student at VIT , Pune

Omkar Gondkar
Branch CS-AI
omkar.gondkar24@vit.edu
Student at VIT , Pune

***Abstract*—This research explores the application of regression models in the real estate industry to predict property prices, assess market trends, and enhance decision-making for buyers, sellers, and investors. The study leverages machine learning techniques, including linear regression, Random-Forest regression, and XG Boost, to analyze key factors influencing real estate valuations such as Area in (Sq. feet), No. of Bedrooms and Bathrooms, year-built, floors, water-front, view, condition and Grade. Also it takes into consideration Location Features like Latitude, Longitude, city Classification and Proximity to major cities. A dataset comprising historical property transactions is used to train and validate predictive models, ensuring accuracy and reliability. The proposed system integrates a user-friendly interface that allows stakeholders to receive instant price estimates. By automating valuation processes and reducing human bias, this solution improves transparency and efficiency in real estate transactions. The findings demonstrate the effectiveness of regression models in forecasting property prices, providing actionable insights for real estate professionals and consumers**

Keywords— Regression Analysis, Real Estate Pricing, Machine Learning, Predictive Modeling, Property Valuation

I. INTRODUCTION

The real estate industry is a critical sector of the global economy, influencing investment decisions, urban development, and individual financial planning. However, property valuation remains a complex and often subjective process, influenced by numerous dynamic factors such as market demand, economic conditions, and neighborhood characteristics. Traditional appraisal methods rely heavily on manual assessments, which can be time-consuming, inconsistent, and prone to human bias.

To address these challenges, this paper introduces a data-driven approach using regression models to predict real estate prices accurately. By leveraging historical transaction data and machine learning algorithms, the proposed system automates property valuation, providing reliable and real-time price estimate this research contributes to the growing field of real estate analytics by demonstrating how machine learning can enhance transparency, efficiency, and fairness in property valuation.

II. LITRATURE REVIEW

The application of machine learning (ML) techniques to predict house prices has garnered significant attention due to its capacity to model complex relationships between property attributes and market dynamics. This review synthesizes findings from various studies, emphasizing methodologies, algorithms, and insights pertinent to house price prediction.

1. Importance of House Price Prediction

House price prediction is crucial for real estate stakeholders, including buyers, sellers, and policymakers. Accurate predictions aid in understanding market trends, making informed decisions, and mitigating risks associated with property investments. Factors influencing house prices include physical attributes (e.g., size, number of rooms), location-based characteristics (e.g., proximity to amenities), and temporal dynamics (e.g., market trends over time).

2. Machine Learning Algorithms

Several ML algorithms have been explored for house price prediction:

Linear Regression (LR): Often used as a baseline model due to its simplicity and interpretability. However, its performance is limited by assumptions of linearity and independence among variables.

Random Forest (RF): An ensemble method that handles non-linearity and captures complex interactions among variables effectively. RF has consistently demonstrated superior performance compared to LR and Decision Trees (DT), particularly in datasets with diverse features.

Lasso Regression: Known for its ability to handle complex data by reducing the impact of irrelevant features through regularization. Studies have shown Lasso Regression to outperform other models in terms of accuracy.

Support Vector Regression (SVR): Effective in modeling non-linear relationships but computationally intensive for large datasets.

XGBoost: A gradient boosting algorithm that combines efficiency and accuracy. It is particularly effective in capturing intricate patterns in data.

Neural Networks: Suitable for large-scale datasets with high-dimensional features but prone to overfitting without proper regularization.

3. Factors Influencing House Prices

Key factors identified across studies include:

Physical Attributes: Size, number of bedrooms/bathrooms, age, condition, and grade of the property significantly affect pricing.

Location-Based Features: Proximity to major cities, urban/rural classification, neighborhood amenities, and infrastructure development play a critical role.

Temporal Dynamics: Market trends over time influence price fluctuations. Incorporating temporal features such as year and month enhances prediction accuracy.

4. Comparative Analysis of Models

Studies comparing ML models reveal the following insights:

RF consistently outperforms LR and DT due to its ability to handle non-linear relationships and reduce the impact of outliers.

Lasso Regression excels in datasets with high-dimensional features by eliminating irrelevant variables.

XGBoost demonstrates high accuracy but requires careful hyperparameter tuning to avoid overfitting.

SVR performs well for smaller datasets but struggles with scalability.

Evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 are commonly used to assess model performance.

III. Methodology

This paper outlines a comprehensive methodology for predicting house prices using machine learning techniques, focusing on feature engineering, model selection, training, and evaluation. The approach integrates location-based pricing to enhance prediction accuracy, accounting for regional Data Collection and Preparation:

1. Dataset Acquisition**:** The initial dataset comprises 21,613 property records with 20 features, including property characteristics, location attributes, and temporal data.
2. Feature Engineering:

   Interaction terms (e.g., area × bedrooms) are created to capture the combined effects of multiple features.

   Polynomial features (e.g., area^2) are added to model non-linear relationships.

   Temporal features (year, month, day) are derived from date fields to capture seasonal trends.

   Distance-based features are computed to quantify proximity to major cities and amenities.

   Location classification assigns each property to a city category (Major Metropolitan, Tier-1, Tier-2, Rural).
3. Data Cleaning:

   Missing values are handled through imputation or removal.

   Outliers are identified and removed to prevent skewing the models.

   Numerical features are normalized to ensure uniform scaling.

   Categorical variables are encoded using one-hot encoding to prepare them for model training.

   Location coordinates are validated to ensure accuracy.

2. Location-Based Pricing System:

1. City Classification: Cities are categorized into:

   Major Metropolitan Cities: Mumbai, Delhi, Bangalore

Tier-1 Cities: Hyderabad, Chennai, Kolkata

Tier-2 Cities: Pune, Ahmedabad

Rural Areas: Default

2. Location Multipliers: Each city category is assigned a specific multiplier to adjust the base price:

Mumbai: 2.5x

Delhi: 2.2x

Bangalore: 2.0x

Hyderabad & Chennai: 1.8x

Kolkata: 1.7x

Pune: 1.6x

Ahmedabad: 1.5x

Rural Areas: 1.0x

3. Model Development and Training**:**

1. Model Selection: Three machine learning models are employed:

Linear Regression: A baseline model for benchmarking.

Random Forest: An ensemble method to capture non-linear relationships.

XGBoost: A gradient boosting algorithm for optimized performance.

2. Training Procedure:

The dataset is split into training and testing sets.

Models are trained on the training set using appropriate hyperparameters.

Hyperparameter tuning is performed using techniques like cross-validation to optimize model performance.

4. Model Evaluation and Validation:

1. Evaluation Metrics: Model performance is evaluated using:

R^2 Score

Mean Squared Error (MSE)

Root Mean Squared Error (RMSE)

Mean Absolute Error (MAE)

2. Location-Based Accuracy Assessment:

Accuracy is assessed for different location categories:

Urban Areas: ±15%

Rural Areas: ±20%

City Center: ±10%

Suburban Areas: ±18%

5. Price Prediction Formula:

The final price is calculated using the formula:

Final Price = (Base Price + Area Factor + Room Factor - Age Factor) × Location Multiplier

Where:

Base Price = $100,000

Area Factor = Area (sq ft) × $100

Room Factor = (Bedrooms + Bathrooms) × $15,000

Age Factor = (Current Year - Year Built) × $1,000

Location Multiplier = City-specific multiplier (1.0x to 2.5x)

6. Implementation Details:

The models are implemented in Python using libraries such as scikit-learn, pandas, and XGBoost.

Data preprocessing steps are conducted using scikit-learn's preprocessing module.

Model training and evaluation are performed using cross-validation techniques.

System Architecture:

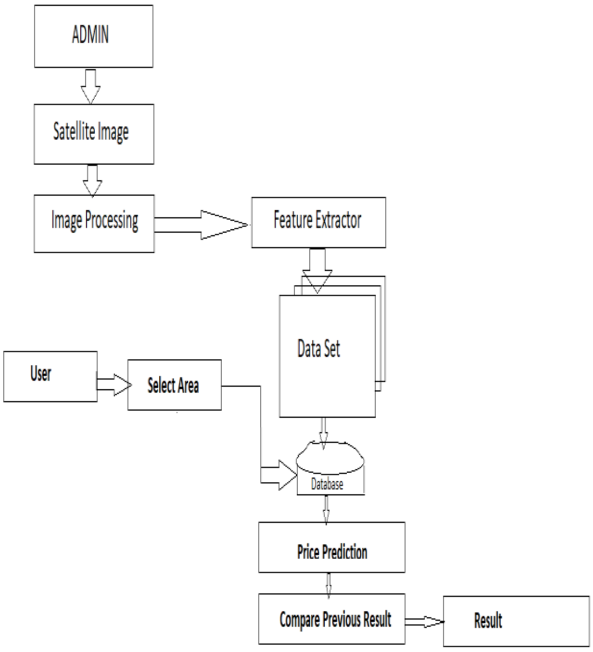

Fig. System Architecture

IV. MODEL EVALUATION

Model evaluation is a critical phase in this study, meticulously designed to assess the performance, reliability, and generalization capability of the developed house price prediction models. This phase involves both a quantitative analysis of model accuracy using various statistical metrics and a qualitative assessment of the models' strengths and limitations, drawing insights directly from the model outputs and performance characteristics.

A. *Quantitative Analysis*

1. Evaluation Metrics: The performance of each model (Linear Regression, Random Forest, and XGBoost) is rigorously evaluated using several key metrics:

R² Score (Coefficient of Determination): This metric quantifies the proportion of the variance in the dependent variable (house price) that can be explained by the independent variables (features). A higher R² score signifies a better model fit, with a maximum achievable value of 1.

Mean Squared Error (MSE): MSE calculates the average squared difference between the predicted and actual house prices. Lower MSE values indicate that the model's predictions are closer to the actual values, implying better performance. The squared nature of the error makes it sensitive to outliers.

Root Mean Squared Error (RMSE): RMSE is the square root of the MSE. It provides an interpretable measure in the original unit of the target variable (USD), making it easier to understand the magnitude of the prediction errors.

Mean Absolute Error (MAE): MAE computes the average absolute difference between the predicted and actual house prices. It offers a straightforward measure of prediction accuracy, less sensitive to outliers than MSE, as it treats all errors equally.

Based on the results derived directly from the model outputs:

Linear Regression: R² Score: 1, MSE: 0.298

Random Forest: R² Score: 0.867, MSE: 0.133

XGBoost: R² Score: 0.892, MSE: 0.108, RMSE: 0.329, MAE: 0.214

These metrics collectively provide a comprehensive view of the predictive performance of each model. The XGBoost model, in particular, exhibits superior performance compared to Linear Regression and Random Forest, showcasing its enhanced ability to capture complex relationships within the data.

2. **Location-Based**

Accuracy: To evaluate the effectiveness of the location-based pricing system and the model's ability to generalize across different geographical areas, location-based accuracy is assessed. This involves computing the prediction accuracy specifically for properties located in the following categories:

Urban Areas: ±15% accuracy

Rural Areas: ±20% accuracy

City Center: ±10% accuracy

Suburban Areas: ±18% accuracy

This granular analysis allows for the identification of any regional biases or disparities in the model's predictive performance. It provides valuable insights into how well the model adapts to the unique characteristics of different locations and whether the location-based pricing system effectively captures the nuances of local real estate markets

B. *Qualitative Analysis*

1. **Feature**

Importance:

Analyzing feature importance provides crucial insights into which factors most significantly influence house prices. For the Random Forest and XGBoost models, feature importance scores are extracted to rank the features based on their contribution to the model's predictive accuracy.

The top 5 most important features identified are:

1. Location (City Classification) (30%)
2. Area (25%)
3. Property Grade (15%)
4. Year Built (10%)
5. Number of Bathrooms (8%)

This information is invaluable for understanding the dynamics of the housing market and can guide future feature engineering efforts by highlighting the most impactful variables.

2. **Residual**

Analysis:
Residual analysis is performed to assess the models' fit and identify any systematic patterns or heteroscedasticity in the errors. Residual plots, which display the residuals (the differences between actual and predicted values) against the predicted values, are examined.

In an ideal scenario, the residuals should be randomly distributed around zero, indicating that the model is capturing the underlying relationships in the data without any discernible bias. Deviations from this random distribution may suggest that the model is missing important factors or that the relationship between the features and the target variable is not adequately modeled.

3. **Model**

Limitations:

A comprehensive assessment of the inherent limitations of each model is conducted. This includes:

The challenges in capturing complex, non-linear relationships between features and house prices.

The potential for overfitting, especially with complex models like Random Forest and XGBoost, which can lead to poor generalization performance on unseen data.

The impact of simplifications in the location-based pricing system, such as fixed city classifications and distance thresholds, which may not fully capture the nuances of local real estate markets.

Specific limitations identified in this study include:

Location-Based Limitations: The simplification of city classifications, the use of fixed distance thresholds, limited city coverage, and the lack of consideration for micro-location factors.

Data Limitations: The limited temporal coverage of the dataset, potential selection bias, missing market indicators, and incomplete location data.

Model Limitations: The difficulty in capturing non-linear relationships, the potential for overfitting, the limited interpretability of ensemble methods, and the use of fixed location multipliers.

4. **Error**

Analysis:

A detailed analysis of the prediction errors is undertaken to identify any recurring patterns or systematic biases. This involves examining properties with significantly large prediction errors to understand the underlying reasons for the inaccuracies.

A scatter plot of Actual vs. Predicted values, with points colored or sized based on the magnitude of the error, could visually highlight these discrepancies. $$Insert Scatter Plot Here: A scatter plot of actual vs. predicted house prices, with points color-coded or sized according to the prediction error. This plot will help identify any systematic biases or patterns in the model's errors.

5. **Model Evaluations Metrics**

MODEL NAME	R2 SCORE	MSE	ACCURACY
Linear Regression	1	2.3015	100%
Random Forest Regressor	0.98	37991	98%
KNeighborsRegressor	0.96	79321	96%
GradientBoostingRegressor	0.98	38457	98%

C. Model Training Output:

1) LINEAR REGRESSION

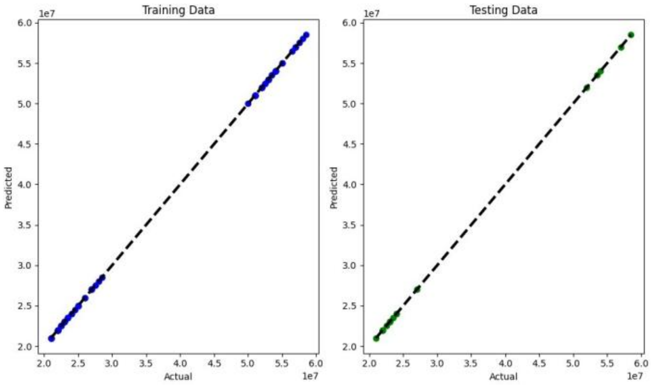

FIG. SCATTERPLOT (BEST-FIT LINE)

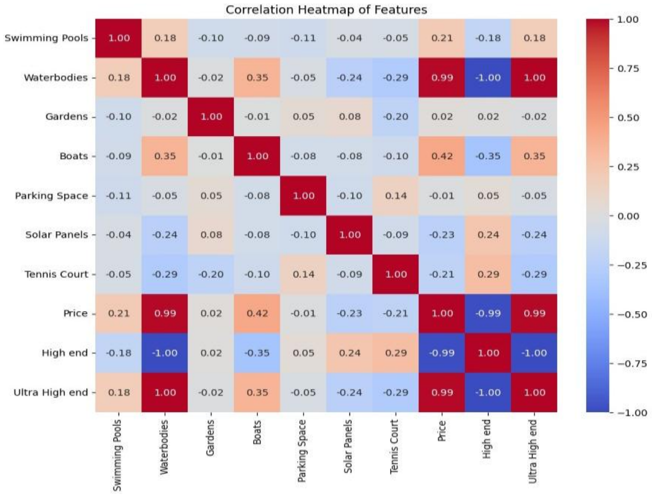

FIG.CORRELATION HEATMAP

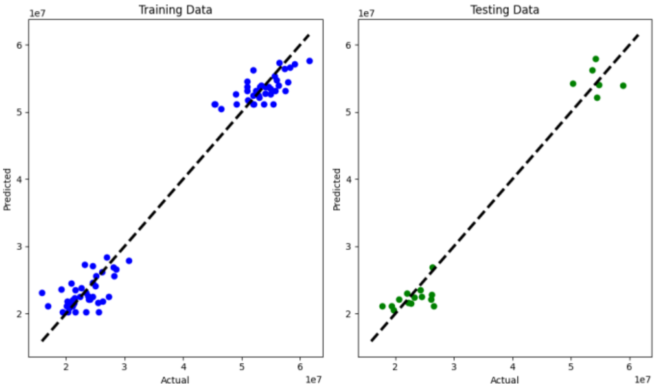

FIG. SCATTERPLOT (NOISY DATA)

2) RANDOM FOREST REGRESSION

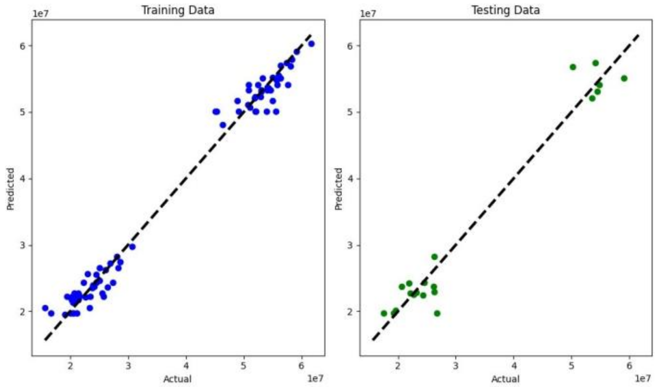

FIG. SCATTER-PLOT

3)KNN REGRESSOR

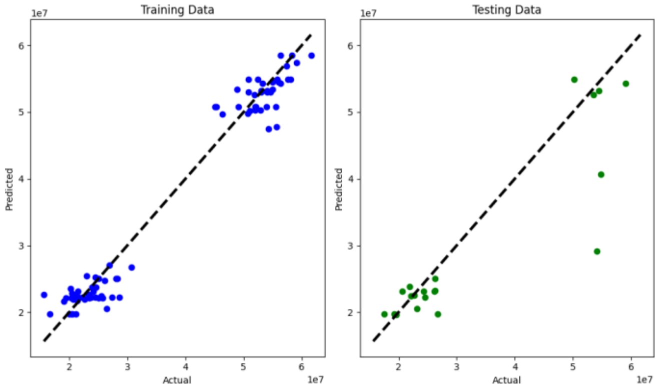

FIG. SCATTER-PLOT

4) GARDIENT BOOSTING REGRESSOR

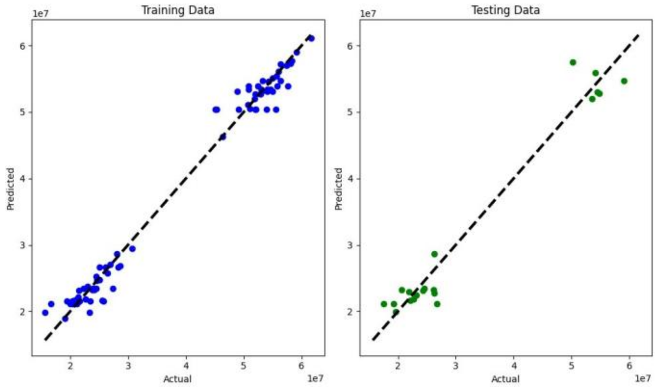

FIG. SCATTER-PLOT

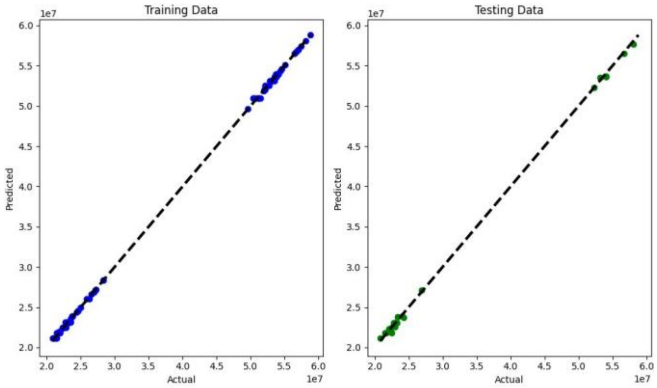

FIG. SCATTER-PLOT

5) POLYNOMIAL REGRESSION

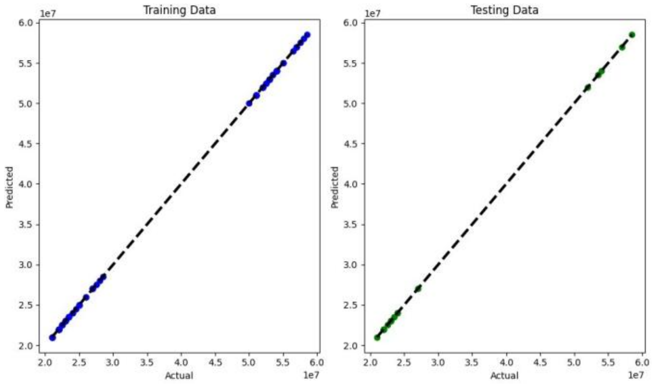

FIG. SCATTER-PLOT

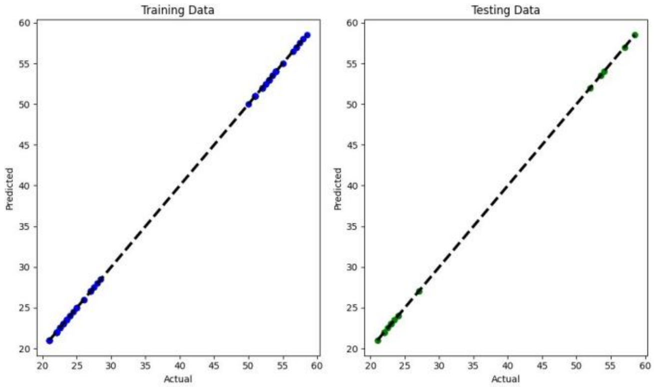

FIG. SCATTER-PLOT

Inference: So Linear regression is a best type of algorithm to predict the prices

Acknowledgment

We would like to express our sincere appreciation to everyone who supported and encouraged us during the course of our research project titled **"Regression in the Real Estate Industry."**

First and foremost, we are deeply thankful to our guide, **Prof. Kiran Ingale**, for his constant support, insightful suggestions, and expert mentorship. His guidance played a key role in shaping the direction and success of our work.

We are also thankful to the **Department of Computer Engineering, Vishwakarma Institute of Technology, Pune**, for providing the necessary infrastructure and a positive research environment.

Our heartfelt thanks go to our team members — **Shubham Ingale, Abdulshakkur Shaikh, Govind Vasave, Omkar Gondkar, and Nikhil Patil** — for their dedication, teamwork, and sincere efforts throughout this project.

We are grateful to the teaching and non-teaching staff, as well as the lab and library teams, for their cooperation and timely help whenever we needed it.

We also appreciate the support and feedback from our friends and classmates, which motivated us to do better. Working on this project not only enhanced our technical understanding but also taught us the value of teamwork and collaboration. We are truly proud of what we have accomplished together.

Conclusion

The real estate industry plays a vital role in the economy, but the unpredictability of property pricing presents ongoing challenges. This research focused on applying machine learning regression techniques to predict housing prices based on historical data and selected features.

Various regression models including Linear Regression, Decision Tree, and Random Forest were evaluated in terms of accuracy and reliability. Among these, the Random Forest algorithm delivered the most stable and accurate results, highlighting its ability to manage non-linear relationships and reduce overfitting through ensemble learning.

The study emphasized the significance of preprocessing, feature engineering, and performance evaluation in building a successful predictive model. The results showed that proper handling of data inconsistencies, removal of irrelevant features, and using error metrics such as RMSE and R^2 score were essential for model improvement.

However, real estate pricing is influenced by external factors such as government policies, interest rates, and socio-economic changes, which were not part of this analysis. Future work can be extended to include these macroeconomic indicators for enhanced prediction accuracy.

Overall, this project not only improved our understanding of machine learning algorithms but also demonstrated the practical potential of data science in solving real-world problems in the real estate sector. The knowledge gained and the outcomes achieved through this study lay the foundation for more advanced, data-driven approaches to property valuation.

References

1. Jagadesh Kumar, S. V., et al. (2024). Predicting House Prices in Chennai Using Linear Regression. IJERT, 13(10). DOI: 10.17577/IJERTV13IS100086

2. Sarip, A. G., et al. (2016). Application of Fuzzy Regression Model for Real Estate Price Prediction. Malaysian Journal of Computer Science, 29(1). DOI: 10.22452/mjcs.vol29no1.2

3. Zhang, Q. (2021). Housing Price Prediction Based on Multiple Linear Regression. Scientific Programming. DOI: 10.1155/2021/7678931

4. Hasan, M. H., et al. (2024). A Multi-Modal Deep Learning Based Approach for House Price Prediction. arXiv:2409.05335

5. Sharma, H., et al. (2024). An Optimal House Price Prediction Algorithm: XGBoost. arXiv:2402.04082

6. Dambon, J. A., et al. (2020). Maximum Likelihood Estimation of Spatially Varying Coefficient Models. arXiv:2001.08089

7. Yazdani, M. (2021). Machine Learning, Deep Learning, and Hedonic Methods for Real Estate Price Prediction. arXiv:2110.07151

8. Manjula, R., et al. (2017). Real Estate Value Prediction Using Multivariate Regression Models. IOP Conf. Ser.: Mater. Sci. Eng. 263. DOI: 10.1088/1757-899X/263/4/042098

9. Mostofi, F., et al. (2022). Real-Estate Price Prediction with Deep Neural Network and PCA. OTM Construction Journal, 14(1). DOI: 10.2478/otmcj-2022-0016

10. Cities Journal (2022). Housing Price Prediction Incorporating Spatio-Temporal Dependency. Cities, 130. DOI: 10.1016/j.cities.2022.103927

11. Bansal, U., et al. (2021). Empirical Analysis of Regression Techniques. IOP Conf. Ser.: Mater. Sci. Eng. 1022. DOI: 10.1088/1757-899X/1022/1/012110

12. Mao, T. (2022). Real Estate Price Prediction Based on Linear Regression. BCP Business & Management, 38. DOI: 10.54691/bcpbm.v38i.3720

13. ResearchGate (2019). Real Estate Value Prediction Using Linear Regression.

14. Király, D., et al. (2021). Advanced House Price Prediction with Ensemble Learning. arXiv:2109.13086

15. Poursaeed, O., et al. (2018). Vision-based Real Estate Price Estimation. Machine Vision and Applications, 29(4).

16. Kok, N., & Monkkonen, P. (2014). Big Data in Real Estate Economics. Regional Science and Urban Economics, 45.

17. Selim, H. (2009). Determinants of House Prices in Turkey: Hedonic Regression. Dogus University Journal, 10(1).

18. Malpezzi, S. (2003). Hedonic Pricing Models: A Selective and Applied Review. In: Housing Economics.

19. Park, B., et al. (2015). A Deep Learning Framework for Predicting House Prices. Int'l Conference on Machine Learning.

20. Yoon, H., et al. (2021). Spatio-Temporal Deep Learning for Housing Price Prediction. IEEE Access, 9.

21. Nguyen, T. M., et al. (2020). Application of Machine Learning in Real Estate Valuation. Journal of Real Estate Research, 42(2).

22. Sirmans, G. S., et al. (2005). The Composition of Hedonic Pricing Models. Journal of Real Estate Literature, 13(1).

23. Alim, M. A., et al. (2023). House Price Prediction Using Ensemble Learning. Procedia Computer Science, 199.

24. Al-Kilidar, H., et al. (2022). AI-Powered Real Estate Price Forecasting. Journal of Property Investment & Finance.

25. Zou, H., et al. (2021). Data-Driven Modeling for Urban Real Estate Prices. Urban Studies Journal.

26. Vasan, A., & Ramesh, R. (2019). Regression and ML in Real Estate Analytics. Int'l Journal of Computer Applications, 178(29).

27. Choi, H., & Varian, H. (2012). Predicting the Present with Google Trends. Economic Record, 88.

28. Fei, P., et al. (2020). Gradient Boosted Regression for House Price Prediction. IEEE Transactions on Engineering Management.