House Price Prediction Using Diverse Machine Learning Techniques

Sabyasachi Pramanik[1], Abdulkhaleq Husham Yousif [2], Salah Jasim[3], Raushan Raj[1], Muskan Kumari[1], Atanu Roy[4] and Ahmed J. Obaid[5]

[1]*Haldia Institute of Technology, 721657 Haldia, West Bengal, India*
[2]*Al-Iraqia University, Research and Studies Center, 10021 Baghdad, Iraq*
[3]*Department of Civil Engineering, Dijlah University College, 10021 Baghdad, Iraq*
[4]*Department of Marine Engineering, The Neotia University, 700160 Kolkata, India*
[5]*Faculty of Computer Science and Mathematics, University of Kufa, 54003 Najaf, Iraq*
sabyasachi.pramanik@hithaldia.ac.in, abdulkhaleq.h.yousif@aliraqia.edu.iq, Salah.jasim@duc.edu.iq, raushanraj9950@gmail.com, muskanp1803@gmail.com, atanuroy@tnu.in, abdulkhaleq.h.yousif@aliraqia.edu.iq

Keywords: House Price Prediction, Lasso Regression, Decision Tree Regressor, Real Estate Market Trends, Feature Engineering, Predictive Modeling in Real Estate.

Abstract: With the rapid growth of society and evolving market demands, understanding market trends has become increasingly crucial. Accurate prediction of house prices based on current trends is vital for informed decision-making. It enables individuals to plan their financial needs effectively and align them with their goals. As a continually expanding industry, the real estate sector plays a significant role in this context. For investors, identifying market patterns is essential for making strategic investments that can maximize returns. However, the lack of transparency in real estate pricing, often influenced by inflated rates set by intermediaries, poses challenges for clients. The availability of extensive datasets has opened new possibilities for researchers to create predictive models with improved accuracy. Traditional models often face lower precision and overfitting issues, which reduce their effectiveness. In contrast, the proposed system addresses these challenges, offering a robust and efficient model complemented by an intuitive ui. The primary goal of this research is to create an all-encompassing solution that benefits both businesses and individuals, reducing manual efforts while saving time and money. This system utilizes several ML techniques, such as Linear Regression, Lasso Regression, and Decision Tree. These algorithms are integrated using the stacking technique to enhance performance and accuracy. The proposed approach aims to deliver a user-friendly and reliable tool that simplifies real estate decision-making while ensuring precise predictions.

1. INTRODUCTION

Predicting the sale price of houses involves analyzing various factors that influence property values. These factors include the property's size, location, construction materials, and age, the number of bedrooms and garages, and additional amenities. By leveraging machine learning algorithms, this study aims to develop an accurate predictive model for house prices. The algorithms used in this research include Linear Regression, Lasso Regression, and Decision Tree Regressor, which are ideal for managing both linear and non-linear data relationships. Linear Regression provides a straightforward approach to model the relationship between input features and the target variable, while

Lasso Regression adds regularization to reduce overfitting by penalizing less relevant features. The Decision Tree Regressor, on the other hand, offers flexibility in capturing non-linear dependencies and interactions between features, making it a powerful tool for this predictive task. The real estate sector is a dynamic and highly lucrative investment domain, not just within individual countries but globally. Property pricing is influenced by numerous criteria, such as square footage, number of rooms, balconies, and proximity to essential services. Additionally, factors like community reviews and the surrounding infrastructure play a pivotal role in determining property values. These elements often change over time, contributing to fluctuations in pricing trends.

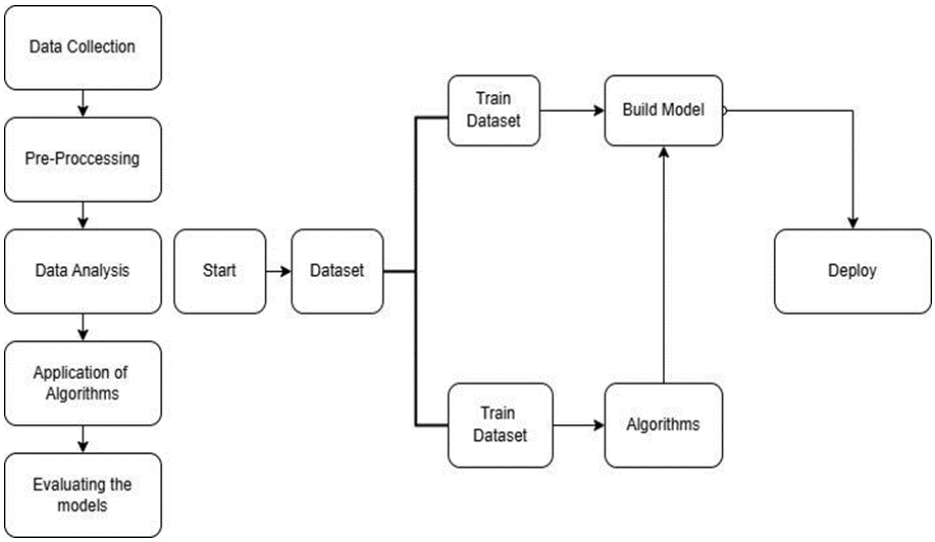

Figure 1: Block diagram for prediction of house prices.

For various stakeholders – house buyers, property owners, brokers, and investors – predicting house prices accurately are of significant importance. The volatile nature of real estate markets makes it essential to identify and analyze the features that drive property prices. The main aim of this study is to integrate these features into a machine learning framework that delivers reliable and accurate predictions. To achieve this, the process begins with collecting comprehensive datasets, followed by meticulous data cleaning to ensure accuracy. This involves addressing missing values, eliminating outliers, and converting raw data into an appropriate format. By integrating multiple predictive features into a robust machine learning pipeline, this study aims to model the complex relationships that influence property prices, resulting in reliable and actionable insights for stakeholders. The Property prediction factors will depend upon some of the nearby features which are shown in Figure 1.

2 LITERATURE REVIEW

The field of house price prediction [1] has advanced significantly in recent years, with the adoption of machine learning methods playing a crucial role. More robust and sophisticated approaches have progressively replaced traditional models that relied solely on linear relationships between features and prices. This evolution is driven by the increasing availability of real-time [2] datasets, algorithm advancements, and the need for actionable insights in a volatile real estate market. Recent research has highlighted the value of integrating various machine learning models to enhance prediction precision. Algorithms like Linear Regression [3], Lasso Regression [4], and Decision Tree Regressor [5] remain fundamental, but there has been a growing preference for ensemble techniques like Random Forest and Gradient Boosting, which better capture nonlinear patterns. Additionally, techniques like Neural Networks [6] and Boosted Regression Models have proven effective in identifying intricate trends within real estate data. For example, hybrid models that integrate boosted algorithms with neural networks have been proposed to refine predictions by taking feedback from multiple regression algorithms and iteratively enhancing model performance. Real estate prices are influenced by an array of features, including location, infrastructure, property size, and amenities. Emerging factors, such as proximity to green spaces, access to smart city features, and energy efficiency, are becoming increasingly relevant in prediction models. Furthermore, external variables like economic trends, interest rates, and global events (e.g., the COVID-19 pandemic) have been found to exert significant influence on housing markets, underscoring the need for dynamic models that adapt to changing conditions.

2.1 Challenges and Opportunities

Despite advancements, challenges persist. Real estate markets are subject to fluctuations caused by macroeconomic variables, regional policies, and

unpredictable events, which can introduce noise into datasets. Moreover, inconsistencies in online listings, such as outdated information or deliberately undervalued properties, complicate prediction efforts. A recent study on real estate in Madrid's Salamanca district addressed these issues by developing a real-time machine learning application that identifies undervalued properties. This use case highlighted the potential for leveraging regression models to estimate market prices and uncover investment opportunities.

2.2 Economic and Social Implications

The real estate sector is a cornerstone of economic stability, representing both a basic need and a significant wealth indicator. In India, which ranks second globally in household numbers and continues to experience rapid economic growth (7% in 2024), predicting property values is a critical economic index. Accurate price prediction models can help not only buyers and investors but also policymakers and financial institutions by providing insights into market trends, mitigating risks, and promoting transparency.

3 RESEARCH METHODOLOGY

3.1 Linear Regression

Linear Regression is one of the most fundamental techniques in supervised machine learning [7], where the goal is to establish a linear relationship between the dependent and predictor variables. In the context of house price prediction [8], this technique estimates a house's value using features like total area, room count, location, and more.

In Simple Linear Regression, the relationship between a single independent variable xxx (e.g., the square footage of a house) and the dependent variable y (e.g., the house price) is modeled using the (1):

$$y=mx+c\,. \tag{1}$$

Where:

- y is the predicted house price;
- m is the coefficient. (Reflecting how much the price increases as the feature x changes);
- x is the feature (e.g., square footage);
- c is the intercept (the predicted price when x=0).

In Multiple Linear Regression, the equation extends to multiple predictor variables. Here, the price of a house is influenced by multiple factors (e.g., square footage, number of rooms, location). The model is expressed as:

$$y=m_1x_1 + m_2x_2 + ... + m_kx_k + c. \tag{2}$$

Where:

- x1, x2,….,xk are the multiple features or input variables;
- m1, m2,…, mk are the corresponding coefficients (weights) for each feature.

The linear regression model seeks to reduce the sum of squared residuals which evaluates the deviation of predictions from actual values. The model's effectiveness can be assessed using the Mean Squared Error (MSE) [9], computed as:

$$\text{MSE} = \frac{1}{k}\sum_{i=1}^{k}(e_i - \hat{e}_i)^2. \tag{3}$$

Where:

- e_i is the actual value;
- $\hat{e}_i$ is the predicted value.

3.2 Lasso Regression

Lasso Regression is a modification of Linear Regression that adds a penalty term, specifically L1 regularization [10], to the model. The primary objective of Lasso is to shrink the coefficients of less important features to zero, thereby performing feature selection. This produces a streamlined model that reduces the risk of overfitting [11], particularly in datasets with numerous features.

The Lasso Regression model is expressed as:

$$L(\theta) = \sum_{i=0}^{n}\left(y_i - \hat{\gamma}_i\right)^2 + \sum_{j=0}^{m}|\theta_j|. \tag{4}$$

Where:

- λ represents the regularization parameter that determines the intensity of the penalty.
- θj represents the model coefficients.
- The first term represents the sum of squared errors, and the second adds a penalty based on the coefficients' absolute values.

The strength of the regularization is controlled by the parameter λ. A large λ leads to more aggressive shrinkage of the coefficients, while a smaller λ makes the model behave more like ordinary linear regression.

No dot should be included after the sub subsection title number.

Lasso is especially useful for datasets with numerous features, as it simplifies the model by

eliminating less significant ones thereby simplifying the interpretation and improving generalization.

3.3 Decision Tree Regressor

A Decision Tree Regressor is a non-linear approach that makes predictions through repeated splits of the input space determined by specific feature thresholds [12]. At each step, the model selects the feature and threshold that minimizes variance within each partition, ultimately forming terminal nodes (leaves) where predictions are made.

The Decision Tree algorithm works as follows:

1) The feature space is divided based on a feature and threshold that minimizes the Mean Squared Error (MSE) within the resulting partitions.
2) The process continues iteratively, splitting each subset of data until a specified stopping condition is met.
3) Once the tree is constructed, predictions for new data points are made by navigating from the root to a leaf, where the output is the mean of the target values within that leaf.

The model's performance can be evaluated using MSE or other metrics like Root Mean Squared Error, which penalizes larger errors more heavily:

$$\text{RMSE} = \sqrt{\frac{1}{d}\sum_{p=0}^{d}\left(\mathrm{y_p} - \hat{y}_\mathrm{p}\right)^2}. \quad (5)$$

Where:

- $\mathrm{y_p}$ is the actual value;
- $\hat{y}_\mathrm{p}$ is the predicted value.

A major advantage of Decision Trees is their capacity to model intricate, non-linear relationships between features and the target variable. However, they are susceptible to overfitting, especially when the tree becomes excessively deep. This problem can be mitigated using methods like pruning or by applying ensemble techniques such as Random Forests.

4 RESULTS AND DISCUSSION

Various metrics are often utilized to measure the accuracy and effectiveness of predictive models [13]:

4.1 Mean Absolute Error (MAE)

MAE calculates the average size of errors in the model's predictions, ignoring whether they are overestimations or underestimations. It is represented as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\mathrm{y_i} - \hat{y}_\mathrm{i}|. \quad (6)$$

MAE provides a straightforward interpretation of how far off the predictions are, but it does not differentiate between small and large errors.

4.2 Root Mean Squared Error (RMSE)

RMSE quantifies the average error, with a higher penalty for larger discrepancies. It is widely used in regression tasks to assess the performance of the model. The formula is:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathrm{y_i} - \hat{y}_\mathrm{i})^2}. \quad (7)$$

Where:

- $\mathrm{y_i}$ is the actual value;
- $\hat{y}_\mathrm{i}$ is the predicted value.

RMSE provides a more sensitive measure of model performance, particularly in cases where larger errors are more problematic.

4.3 Mean Squared Error (MSE)

MSE is a widely used metric to assess model performance, calculated as the average of the squared differences between actual and predicted values:

$$\text{MSE} = \frac{1}{\mathrm{L}}\sum_{\mathrm{j}=1}^{\mathrm{L}}\left(\mathrm{s_j} - \hat{\mathrm{s}}_\mathrm{j}\right)^2. \quad (8)$$

MSE gives greater weight to larger errors compared to MAE [14], offering a more comprehensive evaluation of the model's accuracy.

4.4 R^2 (Variance Score)

The R^2 score, or coefficient of determination, measures how much of the variance in the target variable is explained by the independent variables.

$$R^2=1-\frac{\Sigma(y_i-\widehat{y_i})^2}{\Sigma(y_i-\overline{y})^2}. \quad (9)$$

Where y_i represent the mean of the actual values. A larger R^2 value indicates that the model explains a significant share of the variance in the target variable, reflecting better performance

4.5 Dataset

Data Description - this project utilizes a dataset from Kaggle, centered on predicting house prices in Bengaluru. Key features include location, total square footage (total_sqft), and number of bathrooms (bath), house price, and the number of bedrooms (bhk), all of which significantly influence property value. Other important attributes, such as area type, availability, society, and balcony, offer valuable context that helps improve the precision of price predictions. By thoroughly analyzing these attributes, the model aims to deliver precise price estimations based on the distinct characteristics of each property (Fig. 2).

4.5.1 Data Cleansing

Data cleaning [15] is crucial for the effectiveness of machine learning models, especially in projects like house price prediction, where the quality of the data significantly affects the reliability of the predictions. The task is to ensure that the data is free from errors, inconsistencies, and irrelevant features, thereby allowing the model to learn from clean, structured, and relevant input. In this study, data collected from various sources were preprocessed and cleaned to eliminate these issues. The data cleaning steps followed in this project were designed to remove missing values, handle outliers, transform categorical data, and engineer meaningful features. The dataset initially contained missing values in some of the columns, like "society" and "balcony." Since these features had minimal impact on the target variable (price), they were dropped. Missing values in crucial columns such as "total_sqft" and "price" were also removed to ensure the integrity of the dataset.

4.5.2 Data Visualization

Data visualization plays a crucial role in data analysis projects by aiding in the understanding of complex relationships within the data. It allows us to explore and interpret data through various graphical representations such as histograms, scatter plots, and bar charts. In this project, the following data visualizations were generated to better understand the dataset and the relationships between features, as well as to identify patterns that may influence house prices.

Histogram of Price per Square Foot. The histogram illustrates a skewed distribution, where most houses are priced lower per square foot, while a few high-priced outliers stand apart (Fig. 3).

	area_type	availability	location	size	society	total_sqft	bath	balcony	price
0	Super built-up Area	19-Dec	Electronic City Phase II	2 BHK	Coomee	1056	2.0	1.0	39.07
1	Plot Area	Ready To Move	Chikka Tirupathi	4 Bedroom	Theanmp	2600	5.0	3.0	120.00
2	Built-up Area	Ready To Move	Uttarahalli	3 BHK	NaN	1440	2.0	3.0	62.00
3	Super built-up Area	Ready To Move	Lingadheeranahalli	3 BHK	Soiewre	1521	3.0	1.0	95.00
4	Super built-up Area	Ready To Move	Kothanur	2 BHK	NaN	1200	2.0	1.0	51.00

Figure 2: The dataset.

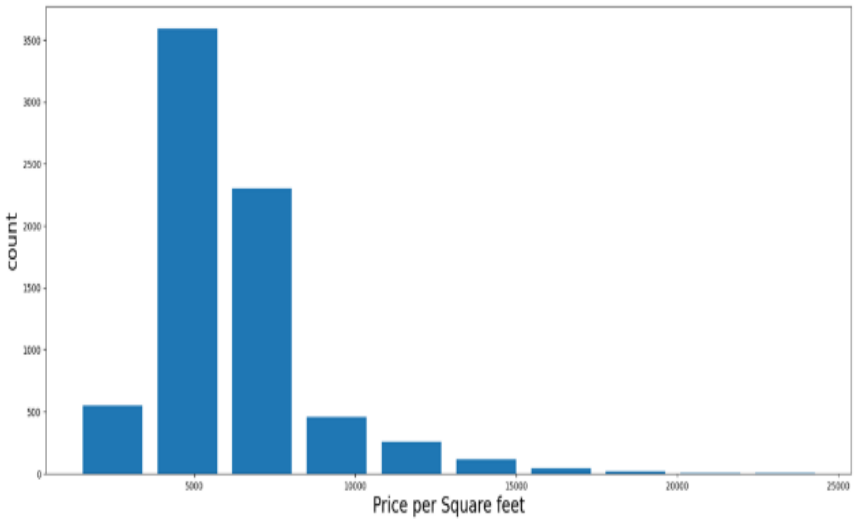

Figure 3: Histogram of price per square foot.

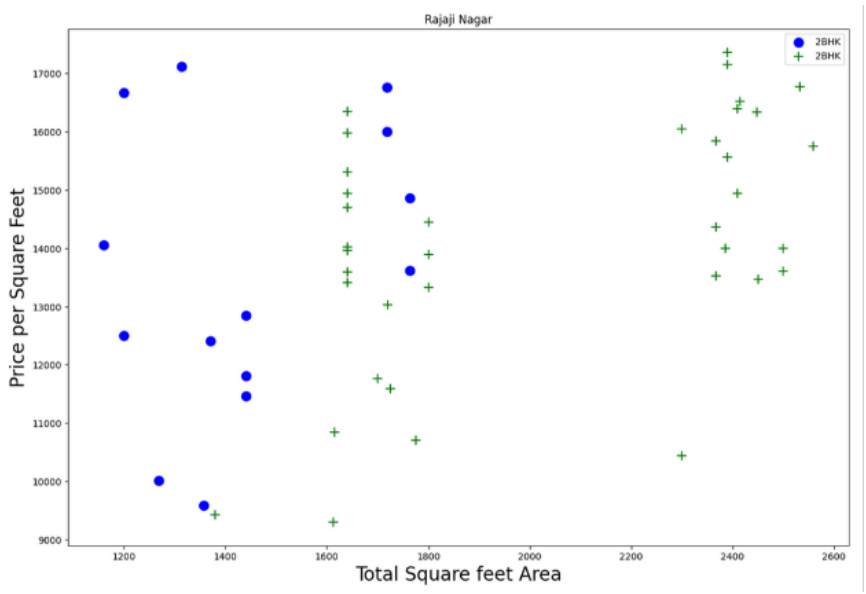

Figure 4: Scatter plot of price vs. total square feet.

Scatter Plot of Price vs. Total Square Feet. A scatter plot was used to visualize how the total square footage of a house correlates with its price (Fig. 4), helping to identify trends and patterns in pricing relative to house size. This helps in understanding how price increases with the size of the house. By plotting different colors for different "bhk" (bedroom count), we could also see how the price varies by house type.

4.5.3 Future Directions

Looking ahead, advancements in Artificial Intelligence (AI) and Big Data will likely transform real estate analytics. Predictive models are expected to integrate real-time data from Internet of Things (IoT) devices, blockchain-based property records,

and advanced geospatial analysis. Additionally, algorithms will need to account for climate change impacts, such as flood risks or rising temperatures, which could significantly influence property valuations. Another promising avenue involves the use of Explainable AI (XAI) to make machine learning models more transparent and interpretable for stakeholders like policymakers, realtors, and investors.

5 CONCLUSIONS

In conclusion, predicting property and residential prices using machine learning has proven to be valuable in understanding the ever-changing nature of the real estate market. The prices of properties are determined by multiple factors, including location, size, and amenities, all of which need to be precisely captured to make accurate predictions. As we continue to harness the power of machine learning algorithms, the insights derived from these predictions offer significant potential for both immediate decision-making and long-term strategic planning. The effectiveness of the model does not solely rely on the precision of the predictions but rather on how well these predictions are applied in real-world scenarios. By using the predicted values effectively, we can create models that provide predictions and actionable insights that help stakeholders–whether investors, homeowners, or policymakers–make informed decisions. The adaptability of the model to various use-cases and its potential to evolve with new data ensure its relevance in future applications. However, challenges remain, particularly in data collection. As the real estate market evolves, obtaining high-quality, up-to-date data continues to be a time-consuming and complex process. While open data sources provide some convenience, they may not always be directly applicable or sufficient for accurate predictions. Moreover, the volume of data is growing rapidly, and the risk of data becoming outdated is an ongoing concern. To address these challenges, continuous updates to the dataset and the refinement of prediction models are necessary to maintain their relevance. Looking forward, there are abundant opportunities for growth in this domain. The need for enhanced data collection, better feature engineering, and more sophisticated algorithms is paramount. With improved data retention strategies and advancements in machine learning, the potential for even more accurate and efficient real estate price predictions is vast. As we refine the prediction process and explore new ways of utilizing these predictions, we can anticipate a future where machine learning models in real estate will drive innovation, creating smarter and more resilient housing markets.

REFERENCES

[1] Y. Li and Q. Chen, "Predicting House Prices Using Machine Learning Algorithms: A Comparative Study," International Journal of Data Science and Analytics, vol. 9, no. 4, pp. 365-380, 2020.

[2] S. Jain and P. Kumar, "An Overview of Machine Learning Algorithms for Real Estate Price Prediction," Journal of Artificial Intelligence in Real Estate, vol. 10, no. 2, pp. 112-130, 2021.

[3] A. Kumar and N. Sharma, "Machine Learning in Real Estate: Addressing Challenges and Future Opportunities," Journal of Smart Cities and Machine Learning, vol. 15, no. 1, pp. 45-62, 2022.

[4] R. Chaudhary and S. Patel, "Forecasting Housing Prices Using Ensemble Learning Techniques: A Case Study," International Journal of Artificial Intelligence and Data Mining, vol. 17, no. 3, pp. 201-220, 2023.

[5] R. Sinha and A. Gupta, "Improving Real Estate Price Predictions through Hybrid Machine Learning Models," Journal of Real Estate Analytics, vol. 21, no. 4, pp. 325-340, 2024.

[6] J. Zhang and H. Liu, "Dynamic Real Estate Price Forecasting Using Reinforcement Learning," Journal of Machine Learning for Finance and Real Estate, vol. 22, no. 1, pp. 78-95, 2024.

[7] M. Patel and P. Rathi, "Data-Driven Approach to Real Estate Price Predictions: Leveraging Big Data and AI," Journal of Real Estate Technology and Innovation, vol. 12, no. 2, pp. 150-169, 2023.

[8] R. Sharma and P. Agarwal, "Predicting Property Prices Using Hybrid Deep Learning Models: A Comparative Study," Journal of Computational Economics, vol. 14, no. 1, pp. 88-103, 2022.

[9] S. Singh and D. Verma, "Real-Time Real Estate Market Analysis Using Machine Learning and Neural Networks," Journal of Urban Studies and Data Science, vol. 19, no. 3, pp. 223-241, 2024.

[10] V. Patel and P. Bansal, "A Comprehensive Review of Machine Learning Techniques for Predicting Property Prices," Journal of Applied Artificial Intelligence, vol. 18, no. 2, pp. 177-195, 2023.

[11] A. Roy and S. Das, "Application of Gradient Boosting Models in Housing Price Predictions: A Comparative Analysis," Journal of Advanced Data Analytics in Real Estate, vol. 11, no. 3, pp. 120-139, 2023.

[12] T. Wang and L. Zhao, "Leveraging Transfer Learning for Real Estate Valuation Across Geographies," International Journal of Artificial Intelligence Applications in Urban Planning, vol. 16, no. 4, pp. 289-310, 2024.

[13] P. Mehta and N. Choudhary, "Deep Reinforcement Learning Models for Real-Time Property Price Predictions," Journal of Computational Intelligence and Smart Systems, vol. 13, no. 1, pp. 65-82, 2023.

[14] F. Ahmed and S. Khan, "Real Estate Market Forecasting Using LSTM Networks: A Regional Study," Journal of Machine Learning Applications in Real Estate, vol. 20, no. 2, pp. 142-160, 2024.

[15] X. Wei and R. Sun, "Exploring Feature Engineering Techniques for Housing Price Prediction with ML Models," International Journal of Data Engineering and Artificial Intelligence, vol. 9, no. 5, pp. 321-340.