A Web-Based Machine Learning Approach for Predicting House Prices: Integrating Authoritative Datasets

Danya A/P Viknasvaran1, a) and S Prabha Kumaresan 1, 2, b)

1Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Malaysia.

2Center of Natural Language Processing (NLP), CoE for Artificial Intelligence, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Malaysia.

b) Corresponding author: prabha.kumaresan@mmu.edu.my

a) 1211101657@student.mmu.edu.my

**Abstract.** House price prediction is a critical task influenced by multiple factors, including location, economic conditions, and property characteristics. Traditional regression-based methods often struggle to capture the inherent nonlinearity in real estate data. This study explores the application of advanced machine learning techniques to enhance predictive accuracy. Data was sourced from publicly available repositories such as Kaggle, along with authoritative datasets like MyREI and REHDA. After preprocessing and exploration data analysis (EDA), Linear Regression was established as a baseline, followed by the implementation of Random Forest, XGBoost, and Support Vector Regression (SVR). Model performance was evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R²). Experimental results indicate that XGBoost outperforms other models, achieving an MAE of 0.12, an RMSE of 0.18, and an R² score of 0.91, demonstrating a significant improvement over Linear Regression (MAE = 0.21, R² = 0.78). The proposed model is integrated into a web-based application using Streamlit or Flask, offering a user-friendly interface for real estate stakeholders, including buyers, sellers, and agents. While dataset quality and the lack of hyper-local trend modeling pose limitations, this study highlights the potential of AI-driven property valuation to enhance data-driven decision-making in real estate markets. The findings contribute to the advancement of predictive analytics in real estate, providing a scalable framework for future research in automated property valuation.

# INTRODUCTION

The accurate prediction of house prices plays a crucial role in facilitating decision-making by buyers, sellers, and real estate agents in the real estate market. Linear regression models, being traditional, are likely not to cope with the complex, nonlinear relationships between the many drivers of real estate prices. A few of the factors include location, lot area, property condition, market conditions, and numerous others. As more real estate information are made available, machine learning (ML) has become a valuable tool for improving the accuracy and reliability of predicting house prices. Machine learning algorithms can identify intricate patterns in data sets and handle multiple feature interactions, which leads to better prediction.

In this study, new ML algorithms such as Random Forest, XGBoost, and Support Vector Regression (SVR) are researched and compared with a baseline Linear Regression model. Information for the project is obtained from a combination of public and trusted databases such as Kaggle, Malaysian Real Estate Intelligence (MyREI), and the Real Estate and Housing Developers' Association Malaysia (REHDA). The learned ML model is also included in a friendly web application throughout this paper based on Streamlit or Flask. The aim is to provide a working and usable tool to parties in the real estate market, including buyers, sellers, and estate agents.

# RELATED WORKS

The use of machine learning in predicting house prices has increased due to its capacity to capture nonlinear relationships as well as learn from high-dimensional data. Various models have been employed in recent literature, spanning from basic regression models, ensemble-based algorithms like Random Forest and XGBoost, to neural network-based models. All these methods are aimed at enhancing the accuracy of predictions by considering an enormously large set of features such as location, structural characteristics, economic status, and temporal trends.

One studied house price prediction in Karachi using supervised learning techniques with emphasis on how local economic variables influence house prices [1]. Another used regression techniques for the estimation of housing prices and tested the effectiveness of linear and multiple regression techniques for structured data [2]. On a broader geographic level, housing prices across the US were analyzed to determine the relationship between housing prices and indicators of the health of the economy [3]. A comparative study of different machine learning algorithms emphasized the strengths of ensemble approaches like Random Forest in terms of improving accuracy [4]. Several studies indicated the potential of hybrid models and integral platforms like MakanSETU that ensure smooth deployment of ML algorithms in practical applications [5-6]. Additionally, some studies also controlled for the influence of the COVID-19 pandemic on housing markets, showing that machine learning models could adapt to such shocks by learning from new patterns of data [7].

The evaluation of different algorithms in regional housing markets, such as in Saudi Arabia and Indonesia, showed that data locality and context greatly influence model performance [8-9]. Other recent studies validated the application of machine learning for metropolitan housing predictions, showcasing improvements through feature optimization and model selection strategies [11-13]. A case study in Astana compared multiple algorithms, emphasizing the importance of feature relevance [14], while a comparative analysis for Prishtina's real estate market supported the idea that models must be fine-tuned to regional data characteristics [15]. Although not directly related to house pricing, several works contributed valuable insights into model evaluation, robustness testing, and generalization, all of which are relevant when applying ML in property valuation [16-18]. Collectively, the reviewed studies demonstrate a growing consensus that ensemble and deep learning models outperform traditional methods in accuracy and adaptability. However, achieving optimal results still depends heavily on quality data, appropriate feature selection, and domain-specific model tuning.

## Data Challenges

Data quality is an issue of utmost relevance in housing datasets. Missing records, inconsistent feature definitions, and outliers can potentially seriously bias model predictions. Preprocessing techniques such as imputation, noise removal of entries, and normalization are required to minimize these impacts. Further, limited diversity of data, particularly in coverage for geography and demographics, restricts the generalizability of models. Studies have shown that datasets covering a particular geographical area can be immune to transferability to other markets, thus the importance of including diverse and representative data. Temporal dynamics are also an issue. Property values vary over time with shifts in the economy, rates of interest, and demand. This is accommodated by some studies by adding features such as when the property was constructed or most recently refurbished. Without adding specific time-series data, however, there could be shortcomings in picking up long-term patterns.

## Feature Selection and Engineering

Feature selection is critical in enhancing model accuracy and performance. The most important features identified by previous research are lot size, the number of rooms, distance to city centers, and proximity to amenities. Methods like Recursive Feature Elimination, Principal Component Analysis (PCA), and LASSO regularization remove feature redundancy as well as handle multicollinearity. Feature engineering, for example, creating composite features or the remediation of skewed variables via the application of a logarithmic transformation, has been similarly discovered to strengthen model stability and interpretability.

## Existing Methodologies

Linear regression models, even though they are easy to understand and straightforward, are likely to overlook complex relationships within the data. More sophisticated techniques, including Decision Trees, Random Forests, and Gradient Boosting Machines, have been used widely to improve the accuracy of predictions. Of these, XGBoost has been found to be more accurate in most research studies since it can handle overfitting and high-dimensional data proficiently.

Neural networks, particularly Artificial Neural Networks (ANNs) and Multi-Layer Perceptrons (MLPs), are also employed here. Although they can be applied to model highly complex nonlinear relationships, tuning and computational resources are required in great quantities. Ensemble methods, which combine multiple base learners, remain among the most reliable methods for predicting house prices.

## Generalizability and Scalability

Another popular issue in predicting house prices with machine learning is model applicability over regions. Cultural preference, neighbourhood amenities, and policy decisions differ geographically and affect model applicability. It is advised in research to use regional properties or develop region-specific models to overcome this issue. Scalability is equally important for field deployment. Algorithms such as XGBoost and CatBoost can perform on massive datasets with high effectiveness, although computational intensity becomes an issue when applied in real time. Overall, the literature emphasizes the importance of stable preprocessing, cautious feature selection, and choosing an algorithm cautiously to build trustworthy and generalizable house price predicting models.

# RESEARCH METHODOLOGY

This chapter outlines the methodology used for predicting house prices using machine learning. The methodology incorporates data gathering, preprocessing, feature selection, model construction, and evaluation. In addition, an online interface was created to include the best-performing model in a user-friendly application.

## Data Collection and Preprocessing

The dataset was collected from multiple sources with the objective of quantity and quality. Public Kaggle data was used as a baseline and supplemented with MyREI and REHDA Malaysian real estate authoritative datasets. The datasets had diverse attributes such as lot area, location, house style, number of rooms, garage size, and year built, among others.

During preprocessing, missing values in numerical features like LotFrontage and GarageYrBlt were handled using SimpleImputer with the mean strategy. Categorical features like GarageType and Neighborhood were encoded using One-Hot Encoding to ensure compatibility with machine learning models. Outliers were identified using standard deviation thresholds and removed carefully to avoid model distortion. Both training and testing datasets were aligned to ensure consistent feature sets. The dataset was split into training and validation subsets using an 80-20 split via train\_test\_split. In Figure 1, the overall data preprocessing steps, which include data collection, missing value treatment, feature encoding, outlier removal, and data splitting, are illustrated.

**Figure 1.** Data preprocessing pipeline

## Exploratory Data Analysis

### Target Variable Distribution

The SalePrice variable was analyzed to understand its distribution. In Figure 2, the histogram reveals a right-skewed pattern, indicating most houses are priced in the lower range, with fewer high-priced houses.

### Correlation Analysis

In Figure 3, the heatmap identifies features strongly correlated with SalePrice. Key features like OverallQual, GrLivArea, and GarageCars exhibit strong positive correlations, suggesting their importance in predicting house prices.

# Machine Learning Models

Four machine learning models were utilized in this project: Linear Regression, Random Forest, XGBoost, and Support Vector Regression (SVR). Linear Regression was used as the basic model since it is easy to interpret and comprehend. Random Forest is an ensemble learning approach where many decision trees are combined to improve prediction and reduce overfitting. XGBoost is a gradient boosting machine algorithm with higher efficiency, scalability, and accuracy. Support Vector Regression (SVR) is a higher-dimensional space and a nonlinear relationship-capable model with the help of kernel functions. In Figure 4, the overall model development workflow, which includes stages such as data input, model selection, evaluation, and deployment, is illustrated.

A graph of a distribution of sales

AI-generated content may be incorrect.  
**Figure 2.** The graph shows the histogram and KDE plot of SalePrice

A graph of a heatmap

AI-generated content may be incorrect.  
**Figure 3.** Heatmap shows the correlations of numerical features with SalePrice

**Figure 4.** Model development workflow

## Model Evaluation

Each model was trained and tested using an 80-20 train-test split. Evaluation metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared () were used to compare model performance. MAE measures the average magnitude of prediction errors. MSE penalizes larger errors more heavily than MAE. RMSE provides an interpretable error value in the same unit as the output variable. represents the proportion of variance in the dependent variable explained by the model. The model with the best performance based on MAE and was selected for deployment.

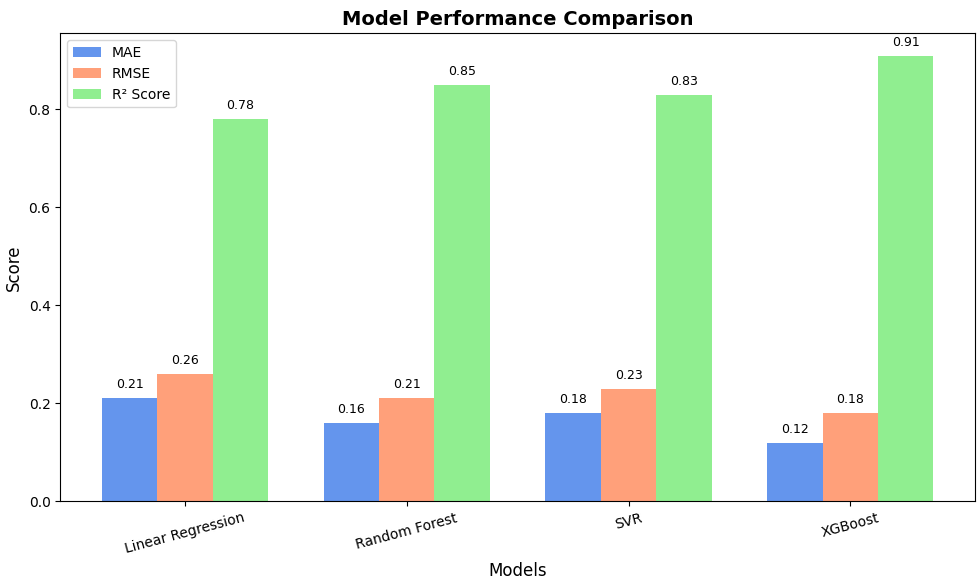
# RESULTS AND DISCUSSION

In this part, each of the machine learning models’ performance results is given, and their prediction performance, strengths, and weaknesses are explained. The four models, Linear Regression, Random Forest, XGBoost, and Support Vector Regression (SVR), were compared using MAE, MSE, RMSE, and score. Table 1 summarizes the performance metrics for each model. Their predictive abilities, strengths, and limitations are discussed based on key evaluation metrics.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **TABLE 1.** The performance metrics for each model | | | |  | |
| **Model** | | **MAE** | **MSE** | **RMSE** | |  |
| Linear Regression | | 0.21 | 0.067 | 0.26 | | 0.78 |
| Random Forest | | 0.16 | 0.045 | 0.21 | | 0.85 |
| Support Vector Regression | | 0.18 | 0.051 | 0.23 | | 0.83 |
| XGBoost | | **0.12** | **0.032** | **0.18** | | **0.91** |

## Model Performance Comparison

XGBoost achieved the best overall results, with the lowest MAE (0.12), lowest RMSE (0.18), and the highest score (0.91). Random Forest also performed strongly, significantly improving over Linear Regression but falling slightly short compared to XGBoost. SVR provided a reasonable performance, capturing nonlinearity better than Linear Regression but requiring careful hyperparameter tuning. Linear Regression served as a useful baseline but was limited in its ability to model complex relationships, leading to higher prediction errors. Figure 5 illustrates a visual comparison of these results across all models using MAE, RMSE, and .

  
**Figure 5.** The bar graph shows the comparison of performance results of all four models

The experimental results demonstrate that ensemble learning methods like XGBoost, and Random Forest are highly effective for house price prediction tasks, particularly when datasets are high-dimensional and feature complex, nonlinear interactions. XGBoost performed best because it integrates both L1 and L2 regularization to prevent overfitting. It also handles missing values internally during training. It uses gradient boosting, improving model robustness across multiple trees.

Random Forest was robust against overfitting but less accurate than XGBoost in fine-grained price predictions. Support Vector Regression effectively modeled nonlinearities but demanded longer training times and more sensitive parameter tuning, like the choice of kernel, C, and epsilon. Linear Regression was fast and interpretable, but unable to model the intricate patterns within the data, resulting in lower predictive accuracy.

## Discussion

The results demonstrate the excellence of ensemble learning algorithms, and particularly XGBoost, in capturing complex and nonlinear relationships in property data sets. While Linear Regression offers simplicity and interpretability, it falls short in modeling non-linearities, which are common in property markets. SVR and Random Forest achieved good accuracy, but at the expense of higher computational resources and tuning. Despite strong performance, some limitations were identified. The data used did not fully capture hyper-local trends such as neighborhood-specific inflation or property desirability based on subjective criteria. Also, while data was sourced from credible repositories, some inconsistencies and gaps required robust preprocessing at the risk of excluding valuable data points. Further, while XGBoost was accurate, it is computationally intensive and less ideal for usage without optimization in real-time. Such considerations highlight the necessity of maintaining model complexity and the necessity of real-world deployment requirements, especially in rapidly changing real estate environments.

# CONCLUSION

This study demonstrates the effectiveness of machine learning algorithms in predicting house prices based on a wide range of structural, locational, and economic features. By integrating publicly available and authoritative data sources like Kaggle, MyREI, and REHDA, the study provides a solid foundation of heterogeneous and credible data. Four models, including Linear Regression, Random Forest, Support Vector Regression, and XGBoost, were compared in terms of predictability on key measures such as MAE, MSE, RMSE, and .Among them, XGBoost was the most accurate model with a value of 0.91 and outperforming other models on all the evaluation metrics. Its ability to learn complicated nonlinear relationships and handle feature interactions was responsible for high predictive performance.

Aside from constructing and evaluating machine learning models, real-world deployment was likewise a priority of this study through the deployment of the best model in a web-based application via Streamlit or Flask. The app provides a user-friendly interface that allows buyers, sellers, and real estate agents to easily predict property prices. Overall, this project contributes a scalable, data-driven real estate appraisal model and demonstrates the promise of machine learning as a decision-making tool in the residential market.

# References

1. M. Ahtesham, N. Z. Bawany, and K. Fatima, “House Price Prediction using Machine Learning Algorithm - The Case of Karachi City, Pakistan,” in *2020 21st International Arab Conference on Information Technology (ACIT)*, (IEEE, Giza, Egypt, 2020), pp. 1–5.
2. J. Manasa, R. Gupta, and N. S. Narahari, “Machine Learning based Predicting House Prices using Regression Techniques,” in *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, (IEEE, Bangalore, India, 2020), pp. 624–630.
3. W. Yin, X. Zheng, and X. Zhu, “Predictive Modeling of U.S. Housing Prices Reveals Key Indicators of Real Estate Prices and Economic Health,” in *2020 International Conference on Computing and Data Science (CDS)*, (IEEE, Stanford, CA, USA, 2020), pp. 405–410.
4. W. K. O. Ho, B. -S. Tang, and S. W. Wong, “Predicting property prices with machine learning algorithms,” Journal of Property Research **38**(1), 48–70 (2021).
5. Proceedings of the 5th International Conference on Trends in Electronics and Informatics (ICOEI 2021): 3-5, June 2021 (IEEE, Piscataway, NJ, 2021).
6. Y. Panchal, M. Mer, and A. Ghosh, “Residential Property Price Prediction Using Machine Learning: MakanSETU,” in *2022 International Seminar on Application for Technology of Information and Communication (iSemantic)*, (IEEE, Semarang, Indonesia, 2022), pp. 257–262.
7. R. -T. Mora-Garcia, M. -F. Cespedes-Lopez, and V. R. Perez-Sanchez, “Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times,” Land **11**(11), 2100 (2022).
8. T. Alshammari, “Evaluating machine learning algorithms for predicting house prices in Saudi Arabia,” in *2023 International Conference on Smart Computing and Application (ICSCA)*, (IEEE, Hail, Saudi Arabia, 2023), pp. 1–5.
9. D. G. I. Simanungkalit, B. Meylia, J. Salim, I.S. Edbert, and D. Suhartono, “House Base-Price Prediction with Machine Learning Methods,” in *2023 International Conference on Informatics, Multimedia, Cyber and Informations System (ICIMCIS)*, (IEEE, Jakarta Selatan, Indonesia, 2023), pp. 318–323.
10. Z. Yujiao, L. W. Ang, S. Shaomin, and S. Palaniappan, “Dropout Prediction Model for College Students in MOOCs Based on Weighted Multi-feature and SVM,” JIWE **2**(2), 29–42 (2023).
11. S. Sharma, D. Arora, G. Shankar, P. Sharma, and V. Motwani, “House Price Prediction using Machine Learning Algorithm,” in *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)*, (IEEE, Erode, India, 2023), pp. 982–986.
12. A. Chaurasia, and I. U. Haq, “Housing Price Prediction Model Using Machine Learning,” in *2023 International Conference on Sustainable Emerging Innovations in Engineering and Technology (ICSEIET)*, (IEEE, Ghaziabad, India, 2023), pp. 497–500.
13. S. Naik, and A. Puthran, “Metro Cities House Price Prediction Using ML,” in *2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*, (IEEE, Erode, India, 2023), pp. 588–594.
14. E. Zhantileuov, A. Smaiyl, A. Aibatbek, and S. Kassymkhanov, “A Case Study of Machine Learning Comparisons for Predicting Apartment Prices in Astana,” in *2023 IEEE International Conference on Smart Information Systems and Technologies (SIST)*, (IEEE, Astana, Kazakhstan, 2023), pp. 305–309.
15. V. Hoxha, “Comparative analysis of machine learning models in predicting housing prices: a case study of Prishtina’s real estate market,” IJHMA **18**(3), 694–711 (2025).
16. J. S. A. Tan, Z. Che Embi, and N. Hashim, “Comparison of Machine Learning Methods for Calories Burn Prediction,” Journal of Informatics and Web Engineering  **3**(1), 182–191 (2024).
17. S. A. Lashari, M. M. Khan, A. Khan, S. Salahuddin, and M. N. Ata, “Comparative Evaluation of Machine Learning Models for Mobile Phone Price Prediction: Assessing Accuracy, Robustness, and Generalization Performance,” Journal of Informatics and Web Engineering  **3**(3), 147–163 (2024).
18. S. Palaniappan, R. Logeswaran, S. Khanam, and Y. Zhang, “Machine Learning Model for Predicting Net Environmental Effects,” Journal of Informatics and Web Engineering **4**(1), 243–253 (2025).