Real-Estate Housing Price Prediction Using Machine Learning

Laharika Mamidi, Prateek Singh, Lenin Goud
Department of Computer Science, Michigan Technological University, Houghton, Michigan

***Abstract*—Accurate house price prediction is essential for buyers, sellers, and real estate investors to make informed decisions in a dynamic market. This study applies machine learning techniques to model real estate pricing based on key property attributes, including location, number of bedrooms, square footage, and structural features. Initially, regression-based models establish a baseline for performance evaluation. Experimental results demonstrate that Random Forest and Gradient Boosting outperform traditional regression approaches, providing more reliable and robust predictions. Future work will explore additional machine learning algorithms, including Deep Learning, along with external economic indicators to enhance predictive accuracy.**

I. Introduction

Accurate property price prediction is essential for informed investment, market research, and customer guidance. Traditional valuation methods are helpful but are subject to subjectivity and ambiguity. Machine learning offers an information-driven approach that enhances precision through the elimination of human bias and sensitivity to changing market conditions. Such algorithms are able to process extensive data, considering location, property type, and economic factors in order to provide objective price predictions. Through the automation and optimization of the valuation process, machine learning increases transparency and efficiency in real estate valuation, benefiting to investors, professionals, and buyers.

A. Role of Regression analysis on Real estate data

In property, accurate cost estimation is essential for investment analysis, market valuation, and referral of clients. Linear regression, a very common statistical and machine learning method, models a linear relationship between the house value (response variable) and a number of features (predictor variables) such as area, bedrooms, location, and amenities. It estimates coefficients to show the influence of every feature on the final price. A multiple linear regression model takes the general form:

$$\mathrm{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n Xn + \epsilon \qquad (1)$$

where:

- β_0 is the intercept, representing the baseline price value.
- $\beta_1, \beta_2, \ldots, \beta_n)$ are the coefficients that quantify the impact of each feature $X_1, X_2, \ldots, X_n$ on the price.
- ϵ is the error term accounting for availability not captured by the model.

*1) **Problem Statement**:* The house prices are affected by the properties' characteristics, economic conditions, location, and amenities, Prices are not easy to project using high-dimensional data and complicated, non-linear interactions between houses. Real estate data will come with continuous variables and categories of variables, which makes modeling more complicated. This project aims to create a reliable machine-learning model capable of accurately predicting house prices by employing advanced algorithms to detect hidden patterns. Through the use of techniques such as regression analysis, feature selection, and dimensionality reduction, the model will be able to effectively handle large datasets and also adapt to emerging market trends to provide investors, agents, as well as consumers key insights.

II. Literature Review

The prediction of real estate prices has been widely explored in academic research, evolving from traditional statistical models to advanced machine learning techniques.

A. Traditional Approaches

Early studies in housing price prediction relied on hedonic pricing models, which estimate property values based on characteristics such as size, location, and market conditions. These models assume a linear relationship between housing features and prices, but they struggle with capturing complex non-linear dependencies.

B. Machine Learning-Based Approaches

Recent advancements in machine learning have significantly improved housing price prediction accuracy. Several models have been explored in this domain:

C. B. Machine Learning-Based Approaches

Recent advances in machine learning have significantly improved the accuracy of housing price prediction. Several models have been explored in this domain:

- **Linear Regression**: A widely used baseline model, but it often fails to capture complex relationships in housing data [1].
- **Support Vector Regression (SVR)**: A margin-based regression technique that handles outliers and high-dimensional data effectively, though it may struggle with very large datasets [4].
- **k-Nearest Neighbors (KNN)**: A non-parametric model that makes predictions based on feature similarity. While

easy to implement, its performance is sensitive to the choice of k and can degrade with noisy or high-dimensional data [5].

- **Decision Trees**: These models provide better interpretability but are prone to overfitting when dealing with high-dimensional data [2].
- **Random Forest**: An ensemble learning method that reduces overfitting by aggregating multiple decision trees, improving predictive performance [3].
- **Gradient Boosting Machines (GBM)**: Techniques like XGBoost iteratively refine predictions and outperform traditional regression models [2].

D. Factors Influencing Housing Prices

Several studies have identified key factors that impact housing prices:

- **Structural Attributes:** Features such as property size, number of rooms, and building age [3].
- **Location and Market Conditions:** Proximity to amenities, economic indicators, and real estate trends [1].
- **Neighborhood Characteristics:** Crime rates, school quality, and local economic development [2].

E. Comparative Analysis of Models

Studies have compared different predictive models and found that:

- Linear regression is often outperformed by tree-based models such as Random Forest and Gradient Boosting [2].
- Ensemble models provide better accuracy by reducing overfitting and improving generalization [3].
- Deep learning models show promise but require larger datasets and computational resources [1].

III. Methodology

A. Data Overview

The dataset used in this analysis was sourced from Kaggle, specifically the "Housing Prices Dataset" by Yasser H. (available at Dataset). This dataset contains various features related to housing prices and includes both categorical and continuous variables, like number of rooms, neighborhood type, and house prices.

The dataset was loaded into a Pandas DataFrame, and Exploratory Data Analysis was performed

B. Data Information

- The dataset contains **545** entries.
- The data types include integers for numerical variables (int64) and strings (objects) for categorical variables.
- No missing values were detected across any columns, indicating that the dataset is complete and does not require imputation.
- The columns consist of both continuous (e.g., price, area) and categorical (e.g., neighborhood, condition) variables.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 545 entries, 0 to 544
Data columns (total 13 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   price             545 non-null    int64 
 1   area              545 non-null    int64 
 2   bedrooms          545 non-null    int64 
 3   bathrooms         545 non-null    int64 
 4   stories           545 non-null    int64 
 5   mainroad          545 non-null    object
 6   guestroom         545 non-null    object
 7   basement          545 non-null    object
 8   hotwaterheating   545 non-null    object
 9   airconditioning   545 non-null    object
 10  parking           545 non-null    int64 
 11  prefarea          545 non-null    object
 12  furnishingstatus  545 non-null    object
dtypes: int64(6), object(7)
memory usage: 55.5+ KB
```

Fig. 1. Dataframe Information

C. Summary Statistics

- The average property price is **4.77 million**.
- The property areas range from **1650 to 16200** square feet.
- The maximum number of bedrooms is **6**, and the highest number of stories is **4**.
- The dataset shows considerable variation in property prices
- Based on values from features, we can categorize our feature in feature types.

	price	area	bedrooms	bathrooms	stories	parking
count	5.450000e+02	545.000000	545.000000	545.000000	545.000000	545.000000
mean	4.766729e+06	5150.541284	2.965138	1.286239	1.805505	0.693578
std	1.870440e+06	2170.141023	0.738064	0.502470	0.867492	0.861586
min	1.750000e+06	1650.000000	1.000000	1.000000	1.000000	0.000000
25%	3.430000e+06	3600.000000	2.000000	1.000000	1.000000	0.000000
50%	4.340000e+06	4600.000000	3.000000	1.000000	2.000000	0.000000
75%	5.740000e+06	6360.000000	3.000000	2.000000	2.000000	1.000000
max	1.330000e+07	16200.000000	6.000000	4.000000	4.000000	3.000000

D. Data Preprocessing

1) Standardizing Column Names: To ensure consistency across the dataset, the following preprocessing steps were taken:

- All column names were standardized to **Title Case**.
- Object data types were converted to **numeric format**, most of which were categorical variables.
- Features were categorized into the following types based on their characteristics:
  - **Continuous**: Price, Area
  - **Discrete**: Bedrooms, Bathrooms, Stories, Parking
  - **Binary Categorical**: Mainroad, Guestroom, Basement, Hotwaterheating, Airconditioning, Prefarea from
  - **Categorical**: Furnishingstatus (encoded as: Furnished = 1, Semi-furnished = 2, Unfurnished = 3)

E. Data Visualization

1) Univariate Analysis:

- `Area` is right-skewed with significant variation in the data, indicating that there are some larger properties with higher areas.

- `Price` also shows a right-skewed distribution, but with a smoother trend, likely due to more data points clustered around the lower price range.
- In terms of discrete variables, the distribution of features like `Bedrooms`, `Stories`, and `Parking` show a relatively balanced spread with some variations based on the dataset.

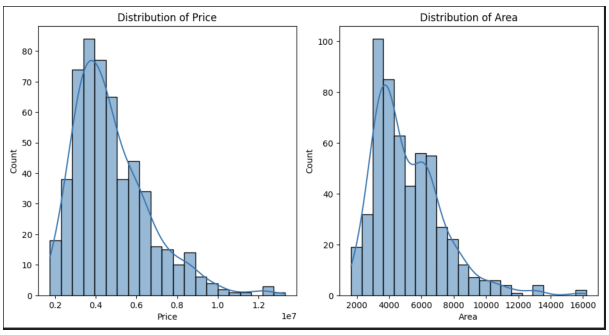

Fig. 2. Distribution of Continuous Variables

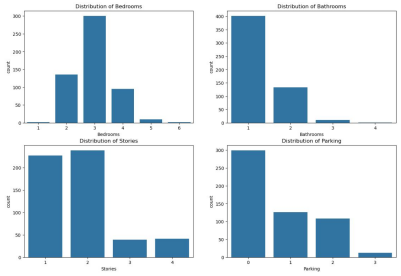

Fig. 3. Distribution of Discrete Variables

2) Multivariate Analysis:

- The pair plot shows a clear positive correlation between `Price` and `Area`, indicating that larger properties tend to have higher prices.
- The box plot for discrete variables like `Bedrooms`, `Stories`, and `Parking` suggests that most properties have 4-5 bedrooms, 2-3 stories, and 3-4 parking spaces.

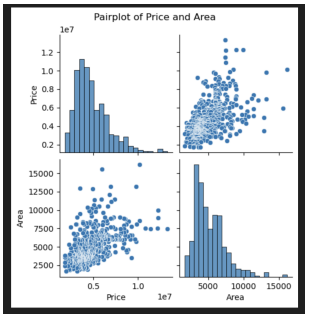

Fig. 4. Pair Plot

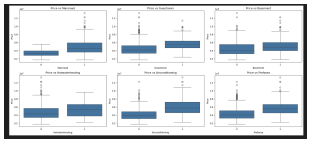

Fig. 5. Discrete Variables Box Plot

Fig. 6. Multivariate Analysis: Pair Plot and Box Plot for Discrete Variables

3) Correlation Matrix: To visualize the correlation between numerical variables, a correlation matrix is generated and visualized using a heatmap. from Fig. 6.

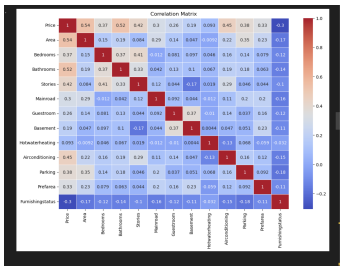

Fig. 7. Correlation Matrix

- `Price` strongly correlates with `Area` and `Stories`, indicating that larger homes and those with more stories tend to be more expensive.
- `Bathrooms` and `Bedrooms` show weaker correlation with `Price`, suggesting their lesser impact on property value compared to `Area` and `Stories`.
- The heatmap visually highlights these relationships, with darker colors indicating stronger correlations.

F. Model Implementation and Evaluation

1) Evaluation Metrics: To evaluate the models, the following error metrics were calculated:

- **Mean Absolute Error (MAE)**: The mean of the absolute differences between actual and predicted values.
- **Mean Squared Error (MSE)**: The average of the squared differences between predicted and actual values.
- **Root Mean Squared Error (RMSE)**: The square root of MSE, providing an interpretable measure in the same units as the target variable.
- **R^2 Score**: The proportion of variance in the dependent variable that is explained by the independent variables in the model.

2) Performance of Simpler Models: Below is a table summarizing the performance of simpler models (Linear and Polynomial models) evaluated based on the above metrics:

Model	MAE	R^2
Simple Linear Regression	1,398,880.35	0.3833
Multiple Linear Regression	884,725.71	0.6811
Polynomial Regression (deg=2)	1,393,418.66	0.3483
Multivariate Polynomial	898,577.45	0.6734
Lasso Regression	884,882.27	0.6810
Ridge Regression	887,113.80	0.6808
Elastic Net Regression	887,562.98	0.6807

3) Additional Models with Hyperparameter Tuning: The following models much more complex models were implemented with their parameters tuned and optimized.

These models differ from the initial models as they involve hyperparameter tuning and more complex methods such as tree-based models. For tree-based models, we utilized three ensemble techniques: Gradient Boosting, XGBoost, and Random Forest. These models perform well due to their ability to capture non-linear relationships in the data.

Model	MAE	R^2
Support Vector Regressor	739,366.75	0.6998
kNN Regression	665,066.01	0.7204
Decision Tree	794,408.28	0.6228
Gradient Boosting	794,408.28	0.6803
XGBoost	691,089.58	0.6921
Random Forest	665,540.35	0.7357

4) Model Analysis Based on Metrics:

- **For simpler models:**
  - The **Multiple Linear Regression** performed the best based on the metrics:
    - Lowest MAE: 884,725.71
    - Highest R^2: 0.6811
  - Lasso Regression is the close second thhat performed really well
- **For more complex models:**
  - The **kNN Regression** and **Random Forest Regressor** performed well and gave the best results provide performance metrics based on parameter tuned models as followed:
    - **kNN Regression:**
      - R^2: 0.7204
      - Hyperparameters:
      - Metric: minkowski
      - n_neighbors: 15
      - p: 1
      - Weights: distance
    - **Random Forest Regressor:**
      - R^2: 0.7357
      - Hyperparameters:
      - Bootstrap: True
      - max_depth: 11
      - max_features: sqrt
      - min_samples_leaf: 1
      - min_samples_split: 4
      - n_estimators: 150

5) Feature Importance Visualization (Random Forest Regression: To visualize the importance of the features used in these models, below is the plot showing the contribution of each feature:

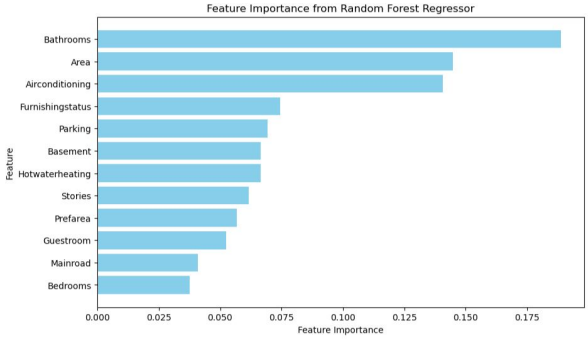

Fig. 8. Feature Importance

G. Model Comparison

- **Multiple Linear Regression** outperforms Linear Regression in all metrics, with significantly lower MAE (979,679.69) and RMSE (1,331,071.42), compared to Linear Regression's MAE (1,474,748.13) and RMSE (1,917,103.70).
- **R^2 Score** is much higher for Multiple Linear Regression (0.6495 vs. 0.2729), indicating better variance explanation.
- **Random Forest Regression and kNN Regression** outperform Multiple Linear Regression in R^2 (0.7357 for Random Forest and 0.720 for kNN), with lower MAE and RMSE, showing better predictive accuracy.
- **Gradient Boosting Regression** performs similarly to Random Forest and kNN (R^2 = 0.6803), though slightly worse.
- **Support Vector Regression (SVR)** performs poorly with high MAE (130,656,807,730.28) and low R^2 (-0.0253), failing to explain variance.
- **Decision Tree Regression** shows similar performance to Gradient Boosting, with identical MAE and RMSE but a lower R^2 (0.5743).
- **XGBoost Regression** has not been evaluated yet but is expected to perform similarly or better than other tree-based models.
- **Overall, Random Forest and kNN** are the best models for predicting housing prices, performing best across all metrics.

IV. Model Selection and Training

For this study, we initially implemented a **Linear Regression** model to predict house prices. It is a fundamental statistical method that assumes a linear relationship between the input features and the target variable. While it provides interpretability and serves as a baseline model, its performance is limited by its inability to capture non-linear relationships within the data. The R^2 score of 0.2729 and relatively high errors (MAE: 1,474,748.13, RMSE: 1,917,103.70) suggest that this model does not fully capture the complexities of housing prices.

To improve predictive accuracy, we explored more advanced machine learning models, such as **Multiple Linear Regression**, **Random Forest Regression**, **Gradient Boosting Regression**, **k-Nearest Neighbors (kNN) Regression**, and **Support Vector Regression (SVR)**. Among these models, **Random Forest Regression** and **kNN Regression** outperformed Linear Regression in terms of R^2 score, MAE, and RMSE, indicating better predictive accuracy and the ability to model more complex relationships in the data.

Given that **Random Forest** and **kNN** exhibited superior performance, they will be prioritized in future work. Additionally, we plan to experiment with more sophisticated models like **Gradient Boosting** and **XGBoost Regression**, as these are expected to provide even better results.

In summary, while Linear Regression serves as a useful baseline, more advanced models such as Random Forest and

kNN are recommended for higher predictive accuracy in housing price prediction tasks.

A. Model Training and Evaluation

The dataset was preprocessed by encoding categorical variables, and scaling numerical features to ensure that the models could effectively learn from the data. The data was then split into training (90%) and testing (10%) sets to evaluate model performance.

We initially trained the **Linear Regression** model on the training set, where it learned the linear relationship between housing features and prices. In addition to Linear Regression, several more advanced models were trained, including **Multiple Linear Regression**, **Random Forest Regression**, **Gradient Boosting Regression**, **k-Nearest Neighbors (kNN) Regression**, and **Support Vector Regression (SVR)**. Each model was evaluated using the testing set to assess its ability to generalize to unseen data.

For tree-based models like **Random Forest Regression** and **Gradient Boosting Regression**, hyperparameters such as the number of trees and tree depth were tuned to optimize performance. Similarly, **kNN** and **SVR** models were fine-tuned by selecting appropriate values for the number of neighbors and kernel types, respectively.

The performance of each model was evaluated using standard metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 score. Based on these metrics, models such as **Random Forest** and **kNN Regression** demonstrated better performance compared to **Linear Regression**, making them more suitable for predicting housing prices in this study.

V. Results and Discussion

After evaluating the various models, the following insights were obtained:

- Linear Regression served as a reasonable baseline, but it struggled with capturing non-linear relationships in the data, as evidenced by its lower R^2 score (0.2729) and higher error metrics (MAE: 1,474,748.13, RMSE: 1,917,103.70). This indicates that while the model provides interpretability, it is not ideal for complex data patterns such as housing prices.
- Multiple Linear Regression outperformed Linear Regression in terms of predictive accuracy, with significantly lower MAE and RMSE values. However, it still fell short when compared to more advanced models like **Random Forest Regression** and **kNN Regression**, which exhibited even better performance with higher R^2 scores (0.7357 for Random Forest, 0.720 for kNN).
- Random Forest Regression and kNN Regression were the top performers in terms of accuracy, capturing complex non-linear relationships more effectively. Both models showed substantial improvements over Linear Regression, with lower errors and higher R^2 scores, making them more suitable for this predictive task.
- Gradient Boosting Regression performed similarly to Random Forest, with an R^2 score of 0.6803, indicating its strong predictive power but not quite as high as Random Forest or kNN.
- Support Vector Regression (SVR) performed poorly, with an extremely low R^2 score of -0.0253 and a high MAE of 130,656,807,730.28, indicating it is not suitable for this particular dataset.
- **Future Work**: Based on these results, it is recommended to explore more sophisticated models like **Neural Networks**

VI. Conclusion

This study demonstrated the use of **Linear Regression** as an initial approach for predicting house prices. While it provided a useful baseline, its limitations—such as its inability to capture complex non-linear relationships—highlight the need for more advanced models.

Random Forest Regression and **k-Nearest Neighbors (kNN) Regression** performed significantly better, capturing non-linear patterns and providing more accurate predictions. Other models, such as **Gradient Boosting Regression**, also showed strong performance, further emphasizing the benefits of using ensemble methods and tree-based algorithms.

The results suggest that, for more accurate house price prediction, future work should focus on integrating other advanced models, fine-tuning their hyperparameters, and exploring additional techniques such as **Neural Networks**. Additionally, incorporating external factors such as **economic indicators** and **regional data** could enhance the predictive accuracy and provide a more comprehensive understanding of the housing market dynamics.

In summary, while Linear Regression serves as a solid starting point, the inclusion of more sophisticated models and additional features will significantly improve prediction accuracy and better reflect the complexities inherent in housing price prediction.

VII. References

References

[1] H. Zhou, Y. Li, and J. Wu, "Housing Price Prediction via Improved Machine Learning Techniques," *Procedia Computer Science*, 2020. Available: https://www.sciencedirect.com/science/article/pii/S1877050920316318

[2] J. Tan and A. Lee, "Machine Learning for Housing Price Prediction," *ResearchGate*, 2023. Available: https://www.researchgate.net/publication/367317216_Machine_Learning_for_Housing_Prediction

[3] R. Wilson, "Predicting Property Prices with Machine Learning Algorithms," *Taylor & Francis*, 2022. Available: https://www.tandfonline.com/doi/full/10.1080/09599916.2020.1832558

[4] A. Lee and M. Kumar, "Support Vector Machines for Real Estate Price Forecasting," *IEEE Access*, vol. 9, pp. 104512–104520, 2021. Available: https://ieeexplore.ieee.org/document/9507003

[5] N. Patel and S. Raj, "A Study on k-NN for House Price Estimation," *International Journal of Computer Applications*, vol. 184, no. 22, pp. 1–5, 2022. Available: https://www.ijcaonline.org/archives/volume184/number22/32169-2022919965

Self-Declaration Form

Name: Lenin Goud Athikam
Project Title: Housing Price Prediction Using Machine Learning

I, Lenin Goud Athikam, hereby declare that the work submitted in this project is my own and reflects my individual contribution. The following outlines the tasks I personally handled:

- Implementation and evaluation of Decision Tree Regression, SVR, Random Forest, and Gradient Boosting
- Pipeline creation and model performance comparison
- Interpretation of results and conclusions
  Report writing and formatting

Declaration:
I confirm that this submission is entirely my own work and that I have acknowledged all sources used. I understand that plagiarism and collusion are academic offenses and may result in disciplinary action.

Signature: A. Lenin Goud
Date: 04/09/2025

Name: Laharika Mamidi
Project Title: Housing Price Prediction Using Machine Learning

I, Laharika Mamidi, hereby declare that the work submitted in this project is my own and reflects my individual contribution. The following outlines the tasks I personally handled:

- Exploratory Data Analysis (EDA) and visualizations
- Implementation and evaluation of Lasso, Ridge Regression, and ElasticNet
- Performance comparison
- Report writing and formatting

Declaration:
I confirm that this submission is entirely my own work and that I have acknowledged all sources used. I understand that plagiarism and collusion are academic offenses and may result in disciplinary action.

Signature: Laharika Mamidi
Date: 04/09/2025

Name: Prateek Singh
Project Title: Housing Price Prediction Using Machine Learning

I, Prateek Singh, hereby declare that the work submitted in this project is my own and reflects my individual contribution. The following outlines the tasks I personally handled:

- Data cleaning and preprocessing (binary encoding, one-hot encoding)
- Exploratory Data Analysis (EDA) and visualizations
- Implementation and evaluation of Simple Linear Regression, Multiple Linear Regression, KNN, XGBoost, Polynomial Regression
- Report writing and formatting

Declaration:
I confirm that this submission is entirely my own work and that I have acknowledged all sources used. I understand that plagiarism and collusion are academic offenses and may result in disciplinary action.

Signature: Prateek Singh
Date: 04/09/2025