House Price Prediction Using Machine Learning

DIPAK JADHAV[1], AADESH GHULE[2], PROF. M. D. SARJARE
Navsahyadri Education Society Group of Institute, Pune

Abstract- ***House price prediction is an essential problem in the real estate industry. Accurately forecasting housing prices helps buyers, sellers, and investors make informed decisions. Traditional valuation methods lack precision and adaptability to dynamic market factors. This research paper explores machine learning techniques to predict house prices using multiple features like location, area, number of bedrooms, and amenities. Various regression algorithms, including Linear Regression, Decision Tree, and Random Forest, are implemented and compared. The Random Forest algorithm emerged as the best-performing model with high accuracy and low error rates.***

Index Terms- ***House Price Prediction, Machine Learning, Regression, Random Forest, Feature Engineering, Real Estate***

I. INTRODUCTION

The real estate market significantly impacts economic development, and predicting house prices accurately is vital for various stakeholders. Over the years, machine learning has revolutionized predictive analytics across industries, including real estate. This paper focuses on building predictive models for house price estimation using supervised learning algorithms. The models are trained on publicly available datasets containing historical data on house sales. The main objectives include pre-processing data, training multiple models, evaluating their performance, and selecting the best one.

II. LITERATURE REVIEW

Previous research has utilized statistical methods for price prediction, such as hedonic pricing models and multiple linear regression. However, these techniques often struggle with non-linear relationships. Studies have shown that tree-based ensemble methods like Random Forest and Gradient Boosting perform better due to their ability to capture complex patterns. Feature engineering and data pre-processing play a crucial role in enhancing model performance. Recent advancements include the use of deep learning, although its computational complexity is higher.

III. PROBLEM STATEMENT

The real estate industry often struggles with the challenge of accurately predicting house prices due to the complex interplay of numerous factors such as location, size, amenities, neighborhood quality, and market trends. Traditional valuation methods rely heavily on manual assessment and linear models, which fail to capture non-linear relationships and dynamic market behaviors.

This project aims to address this issue by leveraging machine learning techniques to develop a predictive model that can estimate house prices with high accuracy. By using a rich dataset of historical housing data and applying various regression algorithms, the objective is to identify the most effective model for predicting property prices. The ultimate goal is to create a data-driven, scalable solution that can assist real estate professionals, buyers, and sellers in making informed decisions.

This project aims to address this issue by leveraging machine learning techniques to develop a predictive model that can estimate house prices with high accuracy. By using a rich dataset of historical housing data and applying various regression algorithms, the objective is to identify the most effective model for predicting property prices. The ultimate goal is to create a data-driven, scalable solution that can assist real estate professionals, buyers, and sellers in making informed decisions.

IV. PROPOSED METHODOLOGY

3.1. Dataset Description

The dataset used in this study is sourced from the Kaggle competition "House Prices - Advanced Regression Techniques". It includes 1460 observations and 81 features describing different aspects of residential homes in Ames, Iowa.

3.2. Data Pre-processing

- Handling missing values using mean, median, or mode imputation
- Encoding categorical features using one-hot encoding
- Feature scaling using Min-Max normalization
- Outlier detection and removal

3.3. Feature Selection

Feature importance is calculated using correlation matrices and model-based techniques. Irrelevant and highly collinear features are dropped to reduce overfitting.

3.4. Machine Learning Models

- Linear Regression: A basic model for comparison
- Decision Tree Regressor: Captures non-linear relationships
- Random Forest Regressor: An ensemble method for improved accuracy
- Gradient Boosting Regressor: Combines weak learners to form a strong predictor

3.5. Evaluation Metrics

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R-squared (R^2) Score

V. IMPLEMENTATION

The implementation is carried out using Python programming language with libraries like Pandas, Numpy, Scikit-learn, and Matplotlib. Jupyter Notebook is used as the development environment.

Steps:

1. Load and explore dataset
2. Pre-process data
3. Split dataset into training and testing sets
4. Train models using training data
5. Evaluate models on test data
6. Compare model performance

VI. RESULITS AND ANALYSIS

The Random Forest and Gradient Boosting models demonstrated superior performance. Gradient Boosting slightly outperformed Random Forest in all metrics but at a higher computational cost.

VII. DISCUSSION

The results indicate that ensemble models significantly improve prediction accuracy. Feature engineering and hyper parameter tuning played a critical role in achieving better results. The study also highlights the limitations of linear models in capturing non-linear interactions among variables. While deep learning models may offer further improvements, their complexity and data requirements make them less feasible for smaller datasets.

VIII. CONCLUSIONS

This research successfully implements and evaluates multiple machine learning models for house price prediction. Random Forest and Gradient Boosting were found to be the most effective. These models can be integrated into real-world applications to assist stakeholders in decision-making. Future work can focus on incorporating geographic and temporal data, as well as experimenting with deep learning architectures.

REFERENCES

[1] Kaggle - House Prices: Advanced Regression Techniques. https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques

[2] Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.

[3] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.

[4] Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer.

[5] Tan, P. N., Steinbach, M., & Kumar, V. (2005). Introduction to Data Mining. Pearson