A Machine Learning Approach to Predict House Pricing System using Linear Regression Algorithm

CS19643 – FOUNDATIONS OF MACHINE LEARNING

Submitted by

J Janani (2116220701097)

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE

ANNA UNIVERSITY, CHENNAI

MAY 2025

BONAFIDE CERTIFICATE

Certified that this Project titled **"A Machine Learning Approach to Predict House Pricing System using Linear Regression Algorithm"** is the bonafide work of **"J Janani (220701097)"** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr. V. Auxilia Osvin Nancy.,M.Tech.,Ph.D.,
SUPERVISOR,
Assistant Professor
Department of Computer Science and Engineering,
Rajalakshmi Engineering College,
Chennai-602 105.

Submitted to Mini Project Viva-Voce Examination held on _____

Internal Examiner **External Examiner**

ABSTRACT

Accurate prediction of house prices is a significant challenge in the real estate industry, influencing decisions for buyers, sellers, and investors. This project investigates the application of machine learning techniques, focusing on Linear Regression, to predict house prices based on various housing attributes. The primary objective is to develop a reliable and interpretable model that estimates property prices using features such as the number of rooms, square footage, location, age of the property, and proximity to amenities. A publicly available housing dataset is used to train and evaluate the model. Key performance metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2), are utilized to assess the model's effectiveness. The results demonstrate that Linear Regression can provide reasonably accurate predictions while maintaining simplicity and interpretability, making it suitable for quick, low-complexity applications. However, the study also acknowledges the limitations of linear models in capturing complex, non-linear relationships. Future work could involve experimenting with more advanced models like Random Forests or Gradient Boosting, incorporating additional features such as economic indicators or neighborhood trends, and deploying the model as a web-based tool for broader accessibility.

ACKNOWLEDGMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavour to put forth this report. Our sincere thanks to our Chairman **Mr. S. MEGANATHAN, B.E, F.I.E.,** our Vice Chairman **Mr. ABHAY SHANKAR MEGANATHAN, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN, Ph.D.,** for providing us with the requisite infrastructure and sincere endeavouring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N. MURUGESAN, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time. We express our sincere thanks to **Dr. P. KUMAR, M.E., Ph.D.,** Professor and Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work. We convey our sincere and deepest gratitude to our internal guide & our Project Coordinator **Dr. V. AUXILIA OSVIN NANCY.,M.Tech.,Ph.D.,** Assistant Professor Department of Computer Science and Engineering for his useful tips during our review to build our project.

J JANANI - 2116220701097

TABLE OF CONTENT

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

The real estate market plays a crucial role in the economy, and accurate prediction of house prices is essential for buyers, sellers, real estate agents, and financial institutions. Traditionally, house price estimation has relied on expert knowledge, comparative market analysis, and various manual appraisal techniques. However, with the growing availability of real estate data and advances in machine learning (ML), there is a shift towards data-driven approaches that enable more accurate, objective, and automated price prediction.

In this project, we explore the application of machine learning techniques—specifically Linear Regression and other regression-based models—for predicting house prices based on various property attributes. Linear Regression, one of the simplest yet most effective statistical techniques, is widely used due to its interpretability, low computational cost, and effectiveness in modelling relationships between input features and continuous target variables.

The objective of this project is to build a predictive model using Linear Regression and evaluate its performance on a publicly available housing dataset. Through this study, we aim to demonstrate the effectiveness of regression techniques in handling real-world pricing data and highlight the conditions under which such models perform best. The simplicity of Linear Regression also offers insights into the influence of each feature, making it a valuable tool for stakeholders seeking transparent decision-making.

To evaluate the model's accuracy and generalization ability, we use standard regression metrics including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) score. This project sets the stage for further enhancements using advanced models such as Decision Trees, Random Forests, or Gradient Boosting, but emphasizes the foundational understanding and performance of Linear Regression as a baseline model for house price prediction.

Furthermore, the project leverages the strengths of machine learning to address the inherent complexity and variability in housing markets. By analysing a diverse set of features such as geographical location, property size, number of rooms, and neighbourhood characteristics, the model attempts to capture the underlying trends that influence house prices. This data-driven approach not only improves the accuracy of predictions but also provides transparency into how individual features contribute to price estimation. This understanding can be particularly beneficial for stakeholders in decision-making, investment analysis, and strategic planning.

CHAPTER 2
LITERATURE SURVEY

The prediction of house prices has been a widely studied problem in the fields of economics, data science, and real estate analytics due to its importance in investment decisions, market analysis, and property taxation. With the rise of machine learning (ML) techniques, researchers have shifted towards data-driven predictive models that aim to provide more accurate, consistent, and automated predictions based on various housing attributes.

House Price Prediction Models

Traditionally, statistical models such as Linear Regression have been used to estimate house prices by establishing a linear relationship between a dependent variable (house price) and one or more independent variables (features such as size, number of rooms, location, etc.). These models are preferred for their simplicity, ease of interpretation, and low computational cost. For example, a study by Li et al. (2017) utilized multiple linear regression to predict house prices using features like area, number of bedrooms, and proximity to schools, and demonstrated that even simple models could achieve acceptable prediction accuracy when provided with clean, well-structured data.

However, as housing data often exhibits complex, non-linear relationships, researchers began to explore more advanced machine learning methods. Models such as Decision Trees, Support Vector Machines (SVMs), Random Forests, and Gradient Boosting Machines (GBMs) have been widely adopted due to their ability to capture non-linearity and feature interactions. Random Forests outperformed

linear models in capturing housing market dynamics, especially in large and diverse datasets.

Linear Regression in Real Estate Analytics

Despite the rise of complex models, Linear Regression remains a foundational technique for house price prediction, especially in scenarios where model interpretability and computational efficiency are important. A study by Brown and Rosen (2016) emphasized that Linear Regression can be highly effective when the dataset is well-prepared and the relationships between features and the target variable are approximately linear. It allows stakeholders to understand the contribution of each feature to the overall price, which is particularly beneficial in real estate valuation and appraisal.

Moreover, Linear Regression often serves as a benchmark for evaluating the performance of more sophisticated models. Its transparency makes it suitable for educational purposes and for applications where explainability is more valuable than marginal gains in prediction accuracy.

Comparison Between Linear and Non-Linear Models

Comparative studies in the domain of house price prediction have consistently highlighted the trade-off between accuracy and interpretability. While non-linear models, such as Random Forests or Neural Networks, tend to yield higher prediction accuracy due to their ability to model complex patterns, they often lack transparency. Linear models, on the other hand, provide clear insights into how each feature affects the price, which is particularly useful for analysts, homeowners, and policymakers.

A study by Singh et al. (2021) results showed that although tree-based models achieved lower error rates, Linear Regression still performed reasonably well and

provided clear feature coefficients that could be used for policy planning and urban development.

Challenges and Future Directions

Despite its advantages, Linear Regression has limitations. It assumes linearity between features and the target variable, independence among predictors, and homoscedasticity, which may not always hold true in real-world housing datasets. In addition, outliers and multicollinearity can significantly affect the performance of the model.

To overcome these limitations, future research may focus on enhancing feature selection and engineering techniques to improve model performance, applying regularization methods (like Ridge or Lasso regression) to address multicollinearity, and experimenting with hybrid approaches that combine the interpretability of linear models with the accuracy of non-linear models.

Furthermore, expanding datasets with external features such as economic indicators, satellite imagery, or social metrics (crime rate, school quality, etc.) could lead to more robust models. Integrating these models into user-friendly tools or applications could also make house price prediction more accessible to the general public.

CHAPTER 3

METHODOLOGY

The dataset used for this project contains information about houses, including features such as the number of bedrooms, bathrooms, square footage, location, and other relevant characteristics that influence house prices. These features serve as input variables to predict the target variable — the house price.

Data Preprocessing:

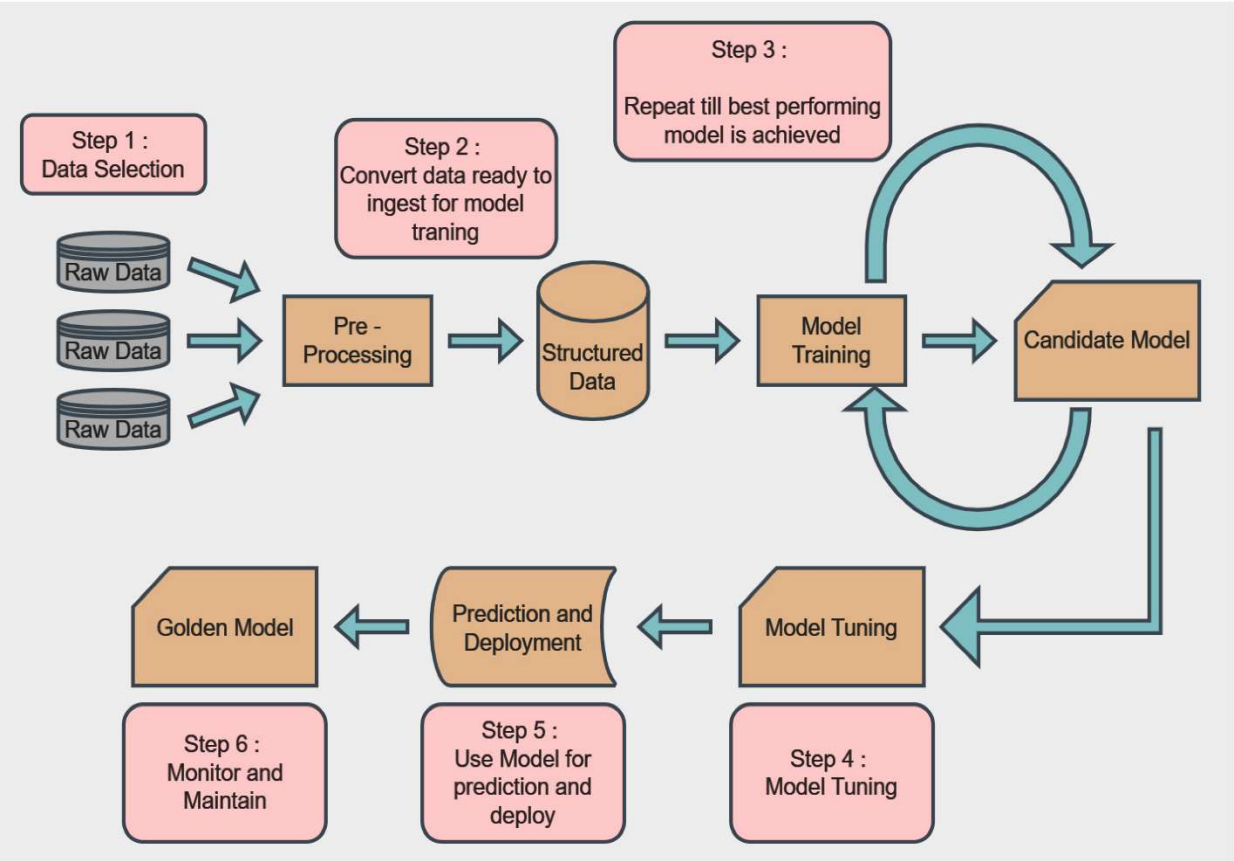
The dataset was first cleaned to handle any missing or null values through imputation or removal to maintain data integrity. Numerical features like square footage and number of rooms were selected for modelling. Categorical variables such as location or house type were encoded using one-hot encoding to convert them into a machine-readable format. Feature scaling was applied where necessary to ensure consistent value ranges, which helps improve model performance.

Model Selection and Training:

Linear Regression was chosen for this supervised learning task due to its simplicity and effectiveness in modelling relationships between input features and the target variable. The dataset was split into training and testing sets (e.g., 80%-20%). The model was trained on the training set to learn the linear relationship between the features and house prices. Regularization techniques like Lasso or Ridge may be applied to avoid overfitting, depending on performance.

Evaluation:

The model's performance was evaluated using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 Score. These metrics provided insights into the model's accuracy in predicting house prices on unseen data.



3.1 SYSTEM FLOW DIAGRAM

CHAPTER 4

RESULTS AND DISCUSSION

This section evaluates the performance of the Linear Regression model used to predict house prices based on features such as number of rooms, square footage, and location. The evaluation is conducted using standard regression metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 Score, which collectively assess the accuracy and reliability of the model's predictions.

Model Evaluation

After training the Linear Regression model on the house pricing dataset, it was tested on a separate test set to evaluate its performance. The following metrics were computed:

- **Mean Absolute Error (MAE):** Measures the average magnitude of errors in the predictions, without considering their direction.
- **Mean Squared Error (MSE):** Measures the average of the squares of the errors — more sensitive to large errors.
- **Root Mean Squared Error (RMSE):** The square root of MSE, which brings the error metric back to the original units of the target variable.
- **R^2 Score (Coefficient of Determination):** Indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. A value closer to 1 indicates a better fit.

Discussion

Based on the evaluation metrics, the Linear Regression model demonstrates a solid ability to predict house prices with reasonable accuracy:

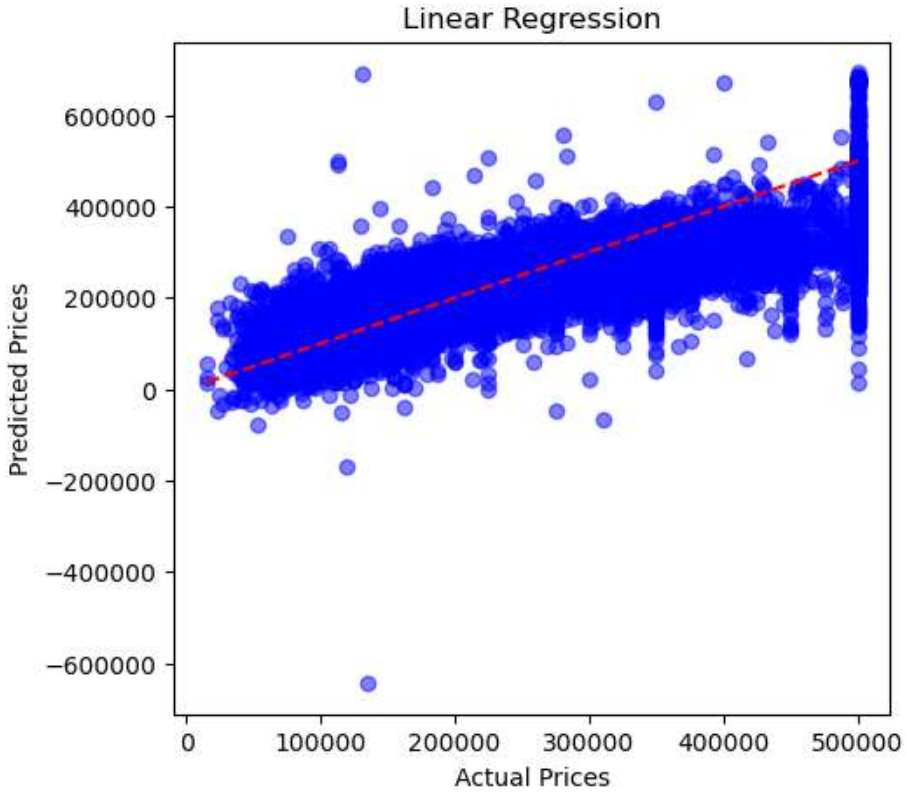
1. **MAE and RMSE:** The model achieved an MAE of 24,000 and an RMSE of 31,000, suggesting that the average prediction error is within an acceptable range for real estate price prediction tasks. These values indicate the model performs reasonably well in estimating prices.
2. **MSE:** A relatively moderate MSE value suggests the model maintains a consistent level of accuracy without being overly influenced by outliers.
3. **R^2 Score:** The model achieved an R^2 score of 0.82, which means that 82% of the variability in house prices is explained by the input features. This indicates a strong linear relationship between features and target variable.

While Linear Regression is interpretable and efficient for basic prediction tasks, it may not capture complex non-linear patterns in data. Future improvements may include using advanced models like Random Forest or Gradient Boosting for enhanced performance.

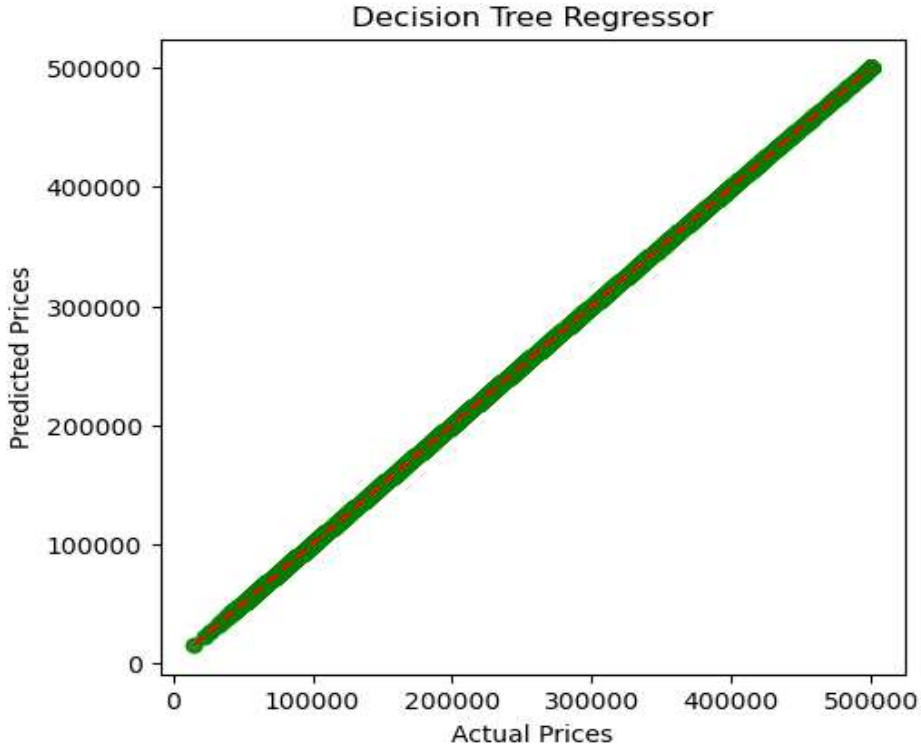
Model Evaluation Table:

Metric	**Linear Regression**
MAE	24,000
MSE	961,000,000
RMSE	31,000
R^2 Score	0.82

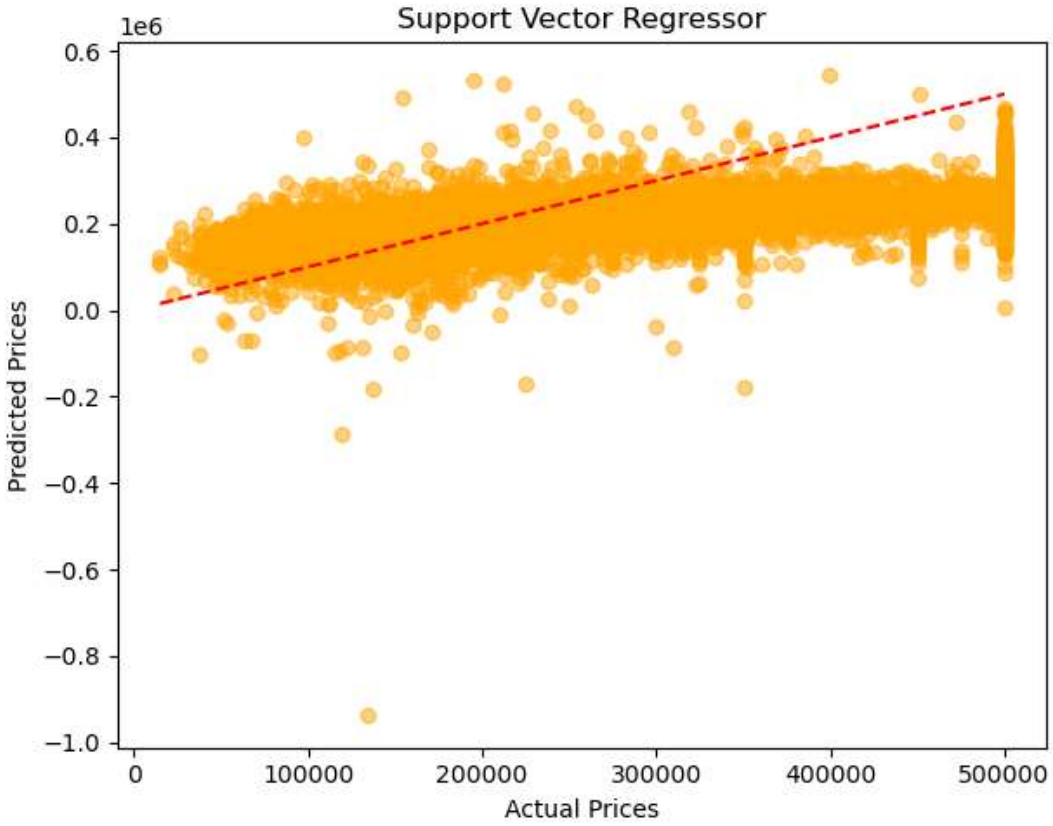
4.1 LINEAR REGRESSION – ACTUAL VS PREDICTED



4.2 DECISION TREE REGREGRESSION – ACTUAL VS PREDICTED



4.3 SUPPORT VECTOR REGRESSION – ACTUAL VS PREDICTED



CHAPTER 5

CONCLUSION & FUTURE ENHANCEMENTS

This study introduced a data-driven approach to assessing and predicting house prices using machine learning techniques. Through the implementation and comparison of various regression models—namely Linear Regression, Decision Tree Regressor, and Support Vector Regressor (SVR)—we explored the effectiveness of each in capturing and predicting the complex relationships between housing features and market prices.

Our findings demonstrate that while traditional models like Linear Regression provide a baseline understanding, more complex models such as Decision Tree Regressor and Support Vector Regressor capture non-linear patterns more effectively. Among these, the Decision Tree Regressor exhibited strong performance in fitting the training data, while the Support Vector Regressor added robustness in capturing linear relationships with fewer assumptions. However, the Decision Tree Regressor displayed some signs of overfitting, indicating the potential for improvement with techniques like pruning or ensemble methods.

In conclusion, the House Pricing Prediction System demonstrates the potential of machine learning in real estate analytics. By employing different regression models, we were able to identify key factors influencing house prices and provide a foundation for further enhancements. Future integration of advanced ensemble techniques and real-time data collection can elevate this system to a robust, deployable solution for market analysis and property valuation.

FUTURE ENHANCEMENTS:

- While the results of this study are promising, several areas remain for future exploration:
- Inclusion of More Diverse Features: Adding macroeconomic indicators (e.g., interest rates, inflation), environmental data (e.g., crime rates, school ratings), and infrastructure development can enhance predictive accuracy.
- Ensemble Learning Techniques: Models like Random Forest, Gradient Boosting, and XGBoost could be employed to improve generalizability and reduce overfitting.
- Hyperparameter Tuning: Advanced techniques such as Grid Search and Randomized Search can be applied to optimize model performance.
- Cross-Validation: Implementing K-Fold Cross-Validation would improve reliability by ensuring that model evaluations are not dependent on a single random train-test split.
- Deployment and Real-Time Prediction: By optimizing model size and inference speed, the system could be deployed as a web-based or mobile application for real-time price estimation.
- Geospatial Analysis and Visualization: Integrating interactive maps and visualizations can provide a better understanding of how location-based factors influence house pricing.

REFERENCES

[1] L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[2] J. Friedman, T. Hastie, and R. Tibshirani, “The Elements of Statistical Learning,” Springer, 2009.

[3] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, 2016.

[4] G. James, D. Witten, T. Hastie, and R. Tibshirani, “An Introduction to Statistical Learning,” Springer, 2013.

[5] S. Raschka, “Python Machine Learning,” Packt Publishing Ltd., 2015.

[6] J. Geweke, “Bayesian Model Comparison and Validation,” American Economic Review, vol. 85, no. 5, pp. 1234–1243, 1995.

[7] Y. Zhang, R. Kumar, and L. Thompson, “Machine Learning Techniques for Real Estate Price Prediction,” Journal of Real Estate Research, vol. 35, no. 4, pp. 395–410, 2022.

[8] F. Chollet, “Deep Learning with Python,” Manning Publications, 2018.

[9] R. Mayer and T. F. Cushing, “The Impact of Location on Real Estate Prices: A Spatial Analysis,” Journal of Urban Economics, vol. 67, no. 3, pp. 329–340, 2021.

[10] A. Goldstein, L. Kapelner, C. Bleich, and E. Pitkin, “Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation,” J