

Predicting New York Housing Prices: A Comparative Analysis of Machine Learning Models

Jie Yu

College of Arts & Science - New York University, West 4 Street, NY10012, New York, U.S.A.

Keywords: Housing Price Prediction, Machine Learning Models, Predictive Analytics.

Abstract: Accurate prediction of housing prices in New York City is crucial for investors, policymakers, and consumers navigating one of the most volatile housing markets. This study explores various machine learning methods to forecast housing prices in New York City. The predictive power of Linear Regression (LR), Support Vector Regression (SVR), Random Forest (RF), and XGBoost (XGB) was examined using a comprehensive dataset with diverse housing attributes. Our results show that LR and SVR provided less accurate predictions, with LR achieving an RMSE of 4,091,594, a MAPE of 1.2991, and an adjusted R-squared of 0.2642, while SVR had an RMSE of 4,967,168, a MAPE of 0.7753, and an adjusted R-squared of -0.0844. In contrast, ensemble methods, namely RF and XGB, demonstrated superior performance on all accounts. RF achieved an RMSE of 2,145,123, a MAPE of 0.3086, and an adjusted R-squared of 0.7978, while XGB had an RMSE of 2,483,884, a MAPE of 0.4163, and an adjusted R-squared of 0.7288. These results conclude that ensemble methods, which can handle complex datasets with higher dimensionality and noise, are more adept at predicting housing prices in varied markets such as New York City. The findings have implications for stakeholders in the real estate industry seeking to leverage machine learning for investment and policy-making decisions.

1 INTRODUCTION

For many individuals and families across the United States, the value of their home represents a significant portion of their overall wealth. Consequently, understanding and predicting housing prices is of paramount importance not only for current and future homeowners but also for a broad spectrum of stakeholders in the real estate market. The dynamics of housing prices are shaped by a complex interplay of attributes, from macroeconomic trends to specific property characteristics including location, neighborhood environment, architectural design, and property type.

Accurate prediction of housing prices is therefore crucial, serving multiple purposes from investment analysis to personal financial planning. In this light, the development of a model capable of making high-accuracy predictions of real estate values is not just desirable but necessary. In response to this challenge, considerable research efforts have been dedicated to exploring various predictive modeling techniques. Among these, machine learning methods such as LR, Decision Trees, RF, and Support Vector Machines

(SVM) have been prominently featured on multiple datasets and diverse cases.

New York City, a bustling metropolis renowned for its economic significance and cultural vibrancy, attracts tourists from the world. Yet, according to the study by Frohlich, and Stebbins in 2016, the real estate market of New York is characterized by its sky-high housing price due to the stark wealth disparity between the top earners of New York and the majority of its residents. This paper aims to explore the use of machine learning methods to forecast housing prices in New York to formulate an accurate house price prediction model.

Previous research has underscored the pivotal role that dataset quality plays in influencing the outcomes of studies focused on housing market predictions. While many studies have utilized traditional machine learning methods such as LR and SVR, there is a gap in the comprehensive evaluation and comparison of ensemble methods like RF and XGB in the context of the New York City housing market. This study not only leverages a highly detailed and diverse dataset encompassing various housing attributes but also systematically compares the performance of

traditional methods against advanced ensemble techniques.

In this research, "New York Housing Market" dataset from Kaggle is used, which offers a comprehensive and realistic compilation of data pertaining to the New York real estate sector. The sale price of properties designated as the primary target feature for prediction and a range of independent variables including house type, location and area, among others are used to predict the sale price. Such a detailed and multifaceted dataset allows us to construct a nuanced model capable of capturing the complexities of New York's housing market.

The structure of the paper is organized as follows: Section 2 provides a review of related work, focusing on methodologies used for predicting housing prices. Section 3 details the methodologies selected for this study. In Section 4, the paper analyzes experimental results, presenting findings and their implications. Section 5 offers a conclusion, summarizing the study's contributions and outlining directions for future research. References to all cited sources are included at the end of the document.

2 RELATED WORK

Housing prices are influenced by a complex interplay of factors, including but not limited to the type of house, its location, and size. Given the unique dynamics of New York City's real estate market, a thorough consideration of these variables is crucial for enhancing the accuracy and depth of research in this domain. Historically, the field of housing price prediction has explored a broad spectrum of methodologies, ranging from Hedonic Pricing Models (HPM) to advanced machine learning methods including LR, SVM, RF, and Gradient Boosting Machines (GBM). This study employs machine learning methods to identify the most effective approaches for modeling the intricacies of New York City's housing market.

Central to the discourse on housing valuation is the HPM, which systematically accounts for both the internal characteristics of properties and the external economic factors influencing their value. This approach has been notably applied by researchers like Goodman, and Halvorsen and Pollakowski, highlighting its utility in dissecting the multifaceted nature of real estate valuation (Goodman 1978 & Halvorsen and Pollakowski, 1981). Despite its widespread use, the HPM has faced criticism, particularly concerning its assumptions of linearity

and the challenges posed by multicollinearity among variables. These critiques underscore the model's limitations in capturing the nonlinear dynamics and interdependencies inherent in the housing market, prompting a shift towards more flexible and robust machine learning techniques in recent studies.

In response to the limitations identified in HPM, researchers turned to Machine Learning Methods (MLMs) for more sophisticated analyses. Ho employed three distinct MLMs—SVM, RF, and GBM—to analyze approximately 40,000 housing transactions over 18 years in Hong Kong (Ho et al., 2021). Their findings indicated superior performance of RF and GBM over SVM, as evidenced by lower scores in mean squared error (MSE), root mean squared error (RMSE), and mean absolute percentage error (MAPE).

A prevalent strategy among researchers in this domain involves the creation of ensemble models, which combine multiple machine learning algorithms to improve predictive accuracy. For instance, Quang Truong developed an ensemble model by integrating Lasso and XGB (Truong et al., 2020), whereas Ali Soltani constructed an ensemble from RF and Gradient-Boosted Trees (Ali et al., 2021). Both studies reported enhanced predictive performance with these ensemble models, underscoring the effectiveness of this approach in housing price prediction.

In light of the comprehensive review of data science applications in the realm of housing price prediction, the study decidedly leans towards the adoption of machine learning models. This choice is informed by the inherent limitations of HPM, particularly their assumption of linearity, which is found problematic. Simple regression techniques, while foundational, fall short in capturing the complexity of the housing market's dynamics in this context. Consequently, ensemble learning stands out as a critical methodological approach in the investigation, notable for its ability to unravel feature importance. This aspect of ensemble learning not only enhances the predictive performance of the models but also aligns with the key objectives of the paper, providing a deeper understanding of the variables that significantly impact housing prices.

3 METHOD

The initial step in this study involves conducting an overview of the dataset to understand its

characteristics and preparing the input data through preprocessing. Following this, a variety of machine learning models—LR, SVR, RF, and XGB—are constructed and trained to generate results for subsequent analysis. Figure 1 illustrates the workflow of the research methodology as detailed in the paper.



Figure 1: Overall Workflow (Picture credit: Original).

3.1 Dataset Overview

This paper utilizes the New York Housing Market Dataset sourced from Kaggle, which contains data on 4,802 houses sold in New York, United States (Elgiriye withana, 2024). The dataset includes 17 numerical attributes for each property; however, this study focuses on a subset of these attributes that are most relevant to our analysis, which are outlined in Table 1 below.

Table 1: Selected Dataset Attributes.

Attributes	Description
BROKERTITLE	Title of the broker
TYPE	Type of the house
PRICE	Price of the house
BEDS	Number of bedrooms
BATH	Number of bathrooms
PROPERTYSQFT	Square footage of the property
ADMINISTRATIVE_AREA_LEVEL_2	Administrative area level 2 information
LOCALITY	Locality information
SUBLOCALITY	Sublocality information
LATITUDE	Latitude coordinate of the house
LONGITUDE	Longitude coordinate of the house

3.2 Preprocessing

The dataset contains several categorical variables such as “BROKERTITLE”, “TYPE”, and “LOCALITY”, which need to be converted into

numerical formats for analysis. Due to the focus of the research, only “ADMINISTRATIVE_AREA_LEVEL_2”, “LOCALITY”, and “SUBLOCALITY” are retained, while other detailed locational variables are discarded. Entries in “ADMINISTRATIVE_AREA_LEVEL_2” that consist solely of 5-digit numbers, presumably zip codes, are also removed as they are not relevant to this study.

Additionally, the “TYPE” variable entries marked as “pending” or “contingent” are discarded because they do not align with the research objectives. Numerical values in BATH and PROPERTYSQFT that represent averages used to fill missing data are removed, as the averages change significantly after preprocessing, indicating they could distort the analysis.

Additionally, the scale of the dataset significantly impacts the performance of certain machine learning algorithms (Ahsan et al., 2021). To address this, the Min-Max scaler is employed to normalize the dataset. This technique adjusts each feature to fall within a specified range, specifically between zero and one, according to the formula:

$$x_{\text{scaled}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \quad (1)$$

Beyond scaling, the dataset undergoes division into a training set, comprising 80% of the data, and a test set, constituting the remaining 20%.

3.3 Model Selection

This study opts to utilize ensemble learning models for regression tasks to predict housing prices. Ensemble models, by integrating multiple machine learning algorithms, can achieve better predictive performance than any single constituent algorithm. An ensemble is composed of several base learners, often developed from algorithms like decision trees or neural networks. These ensembles are typically categorized into two types: Boosting and Bagging. Boosting builds learners sequentially with a high degree of interdependence, whereas Bagging—employed by models such as RF—creates learners independently and in parallel (Zhou, 2021). For this research, XGB and RF have been selected for their exemplary representation of ensemble learning techniques.

To facilitate comparisons and more in-depth performance analysis, additional classical machine learning models have been implemented. Serving as benchmarks, LR and SVR have been included to

provide a baseline against which the more complex ensemble approaches can be evaluated.

● **LR**

LR, a foundational statistical method, establishes a relationship between a dependent variable y and independent variables X through the equation $y = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n + \varepsilon$. Characterized by its simplicity and interpretability, the method allows for straightforward insights into how features influence the target variable, making it computationally efficient and accessible. Despite its advantages, LR assumes a linear relationship between variables, which may not always hold true. Besides, it is sensitive to outliers, potentially limiting its ability to model complex patterns. However, its direct approach to modeling makes it particularly well-suited for tasks like housing price prediction, where the linear influence of features including location and square footage on price can be assumed. The model's interpretability is a significant asset, providing clear insights into the factors affecting real estate prices and offering a solid baseline for more complex analyses.

● **SVR**

SVR presents a robust framework for predicting housing prices, distinguished by its ε -insensitive loss function which ensures robustness to outliers by neglecting errors within a predefined threshold (ε) (Zhang and Donnell, 2021). This characteristic, coupled with the ability to model complex non-linear relationships through the kernel trick, makes SVR particularly adaptable to the diverse patterns inherent in housing market data. Furthermore, SVR's formulation as a convex optimization problem guarantees a unique global solution, thereby enhancing model reliability. The optimization problem, minimized over w, b , and slack variables ξ, ξ_i is defined as $\min_{w,b,\xi,\xi_i} (\frac{1}{2} ||w||^2 + C \sum (\xi_i + \xi_i))$ where C acts as a regularization parameter to balance model complexity against fitting precision. However, SVR's sensitivity to hyperparameter settings and computational intensity for large datasets can be viewed as drawbacks. Despite these challenges, its capacity for capturing the nuanced dynamics of housing prices through a controlled and theoretically sound approach underscores SVR's utility in real estate market analysis.

● **RF**

RF is recognized as a leading ensemble learning method that enhances the Bagging approach by creating a collection of decision trees during the training process. RF adds a layer of randomness to

this process, selecting the best split from a random subset of features at each node, rather than considering all features. This method is especially effective for regression tasks, as it averages the predictions from all trees to produce the final output, thereby reducing variance and improving accuracy over individual decision trees.

A significant advantage of RF is its capability to evaluate the influence of each feature in the prediction process. Within the scope of housing price prediction, the relevance of a feature is quantified using the Mean Decrease Impurity (MDI), commonly referred to as Gini Importance. The MDI for a feature X_j is determined by summing the decrease in impurity ($\Delta i(s, t)$) across all trees in the forest for every node t where X_j is used, weighted by the proportion of samples ($p(t)$) reaching node t , and then averaging this sum over all trees M :

$$IMP(X_j) = \frac{1}{M} \sum_{m=1}^M \sum_{t \in \phi: X_j \text{ splits } t} p(t) \Delta i(s, t) \quad (2)$$

This impurity decreases, averaging over all trees, provides a robust metric for assessing how critical each feature is for predicting the outcome variable Y , such as the price of a house. The inclusion of feature importance analysis in RF not only aids in the prediction task but also offers insights into the dataset, highlighting which features are most influential in determining housing prices. Despite its computational intensity and potential for decreased interpretability due to its complex structure, RF's remarkable accuracy, robustness against overfitting, and effectiveness in managing outliers and noisy data render it an outstanding model for the analysis and prediction of housing prices.

● **XGB**

XGB stands for "Extreme Gradient Boosting," represents an evolution of Gradient Boosting Decision Trees, designed for enhanced scalability and efficiency. This distributed machine learning system builds on the concept of boosting, where a sequence of weak models (typically decision trees) are employed to form a highly accurate ensemble. Unlike RF, which extends bagging by constructing trees in parallel without interaction, XGB improves upon traditional boosting methods by focusing on optimizing a more sophisticated objective function that incorporates both the prediction accuracy and regularization terms to control model complexity.

In the context of a dataset $D = \{(X_i, Y_i)\}_{i=1}^m$ where $X_i = (X_{i1}, X_{i2}, \dots, X_{id})^T \in R^d$ represents the feature vector and y_i the target value, XGB employs K additive functions to predict the outcome, expressed as $\hat{y}_i = \sum_{k=1}^K f_k(x_i)$ where each $f_k \in F$ is a function

represented by the decision trees in the model space F . The objective function it minimizes is given by $L = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k)$ where l is a differentiable convex loss function that quantifies the difference between the predicted \hat{y}_i and actual y_i values, and Ω is a regularization term that penalizes the complexity of the model to prevent overfitting.

XGB's approach to training the model in an additive manner addresses the challenges of optimizing in the Euclidean space by sequentially fitting new models to correct the errors made by existing ones, with an emphasis on computational efficiency and model performance. Moreover, XGB's ability to evaluate the importance of each feature post-training makes it invaluable for understanding the drivers behind the predictive model, an aspect especially relevant for tasks like housing price prediction, where identifying significant predictors is crucial. This blend of accuracy, efficiency, and interpretability has propelled XGB to prominence within the machine learning community.

4 RESULT AND DISCUSSION

Prior to commencing with the experiments, it is essential for readers to grasp the interplay between

the variables under study. To facilitate this understanding, two visual aids were prepared: a correlation heatmap and a scatterplot depicting the relationship between house price and area:

Figure 2 presents the correlation heatmap, elucidating the degree of association between variables. Notably, PRICE, the target variable, exhibits the strongest correlation with PROPERTYSQFT at a coefficient of 0.46. Additional variables such as BATH, BEDS, and TYPE display moderate correlations with PRICE, underscoring the multifaceted nature of the housing market influences. The heatmap serves as a preliminary guide to identifying which features might warrant a more detailed analysis.

Figure 3 ventures into the specific dynamic between floor area and sale price within the dataset, which comprises properties ranging in size from 230 to 55300 sq.ft. and in price from \$49,500 to \$195,000,000. The scatterplot indicates a diffuse yet generally positive relationship between area and price; however, it stops short of suggesting a strong linear correlation. This nuance underscores that while sale price tends to rise with increasing area, it is also significantly shaped by other factors. The graph illustrates this general trend, hinting at the complexity of real estate valuation where multiple variables influence the final price.

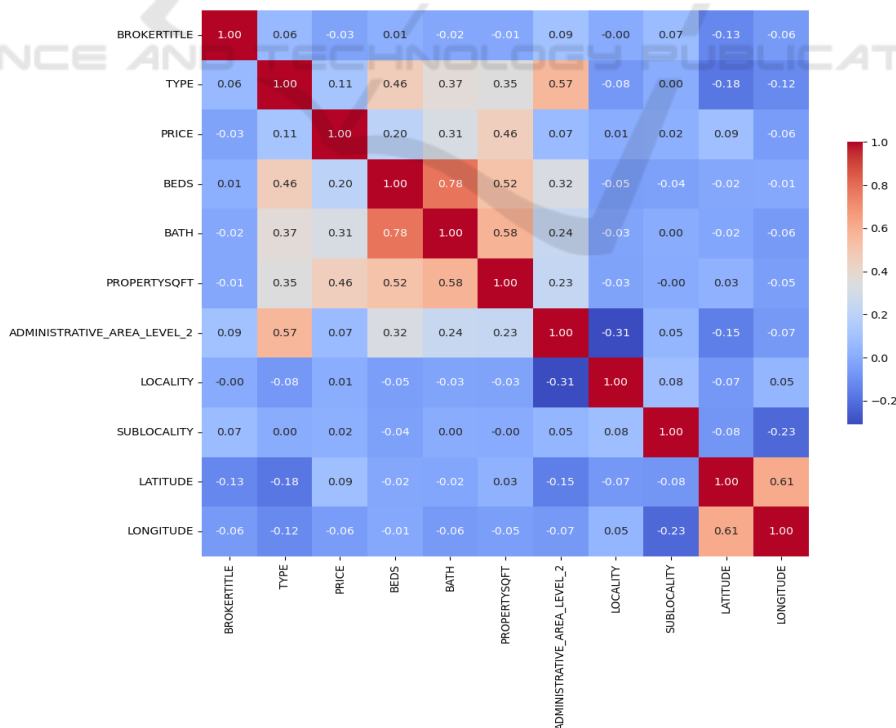


Figure 2: Correlation Heatmap (Picture credit: Original).

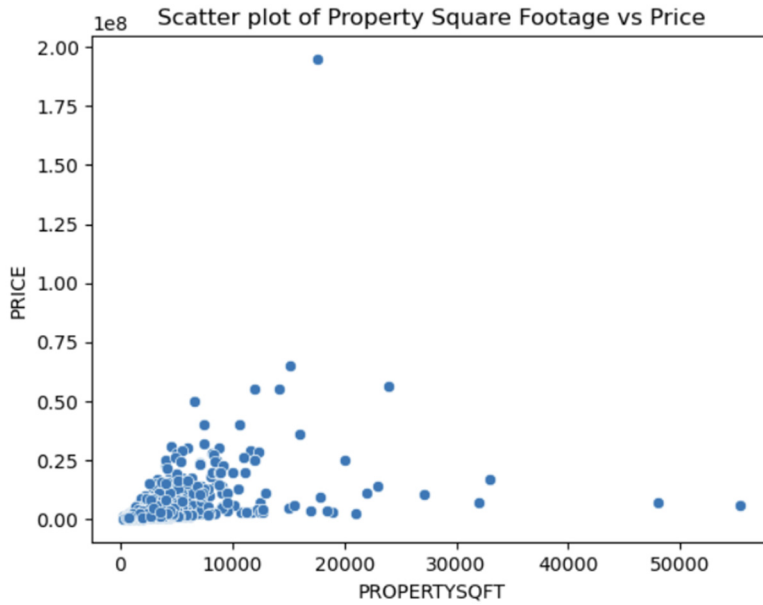


Figure 3: Scatter plot of relationship between house price and area (Picture credit: Original).

4.1 Implementation Details

All computational experiments in this study were conducted using the Python 3.11.5 environment. Key libraries utilized include Pandas for data manipulation, Scikit-Learn for machine learning algorithms, and XGB for gradient boosting models. The hardware setup consisted of a 12th Gen Intel(R) Core(TM) i7-12700H CPU, an NVIDIA GeForce RTX 3070 Ti Laptop GPU, and 32GB of RAM.

The models were configured with specific parameters to optimize performance and ensure reproducibility:

LR: The model was implemented using ordinary least squares regression without any modifications, providing a baseline for performance comparison.

SVR: The SVR model was equipped with a Radial Basis Function kernel, defined mathematically as:

$$k(x, x') = \exp(-\gamma ||x - x'||^2) \quad (3)$$

where x' represents the kernel center, and γ the width of the kernel, is set to $\frac{1}{\text{number of features}}$, allowing the model to handle non-linear relationships. The regularization parameter C was set to 1 and ϵ set to 100 for avoiding overfit.

RF: The model was configured with 100 trees, using mean squared error as the criterion for node splits.

XGB: The model employs the gbtrees booster with 200 gradient boosted trees, targeting mean squared

error as the objective function and RMSE for performance evaluation.

In this study, three metrics are applied to assess model performance: RMSE, MAPE, and Adjusted R^2 . RMSE highlights the impact of significant errors by emphasizing larger discrepancies in predictions. MAPE provides a percentage-based measure of average prediction errors, offering an intuitive understanding of model accuracy relative to actual values. Adjusted R^2 evaluates the explanatory power of the model, adjusting for the number of predictors to ensure the complexity is warranted. These metrics collectively ensure a balanced evaluation of accuracy, sensitivity to relative errors, and model effectiveness.

4.2 Evaluation of Model Performance

The comparative evaluation of the four predictive models is summarized in Table 2.

Table 2: Evaluation of Model Performance.

Model	RMSE	MAPE	Adjusted R^2
LR	4091594	1.2991	0.2642
SVR	4967168	0.7753	-0.0844
RF	2145123	0.3086	0.7978
XGB	2483884	0.4163	0.7288

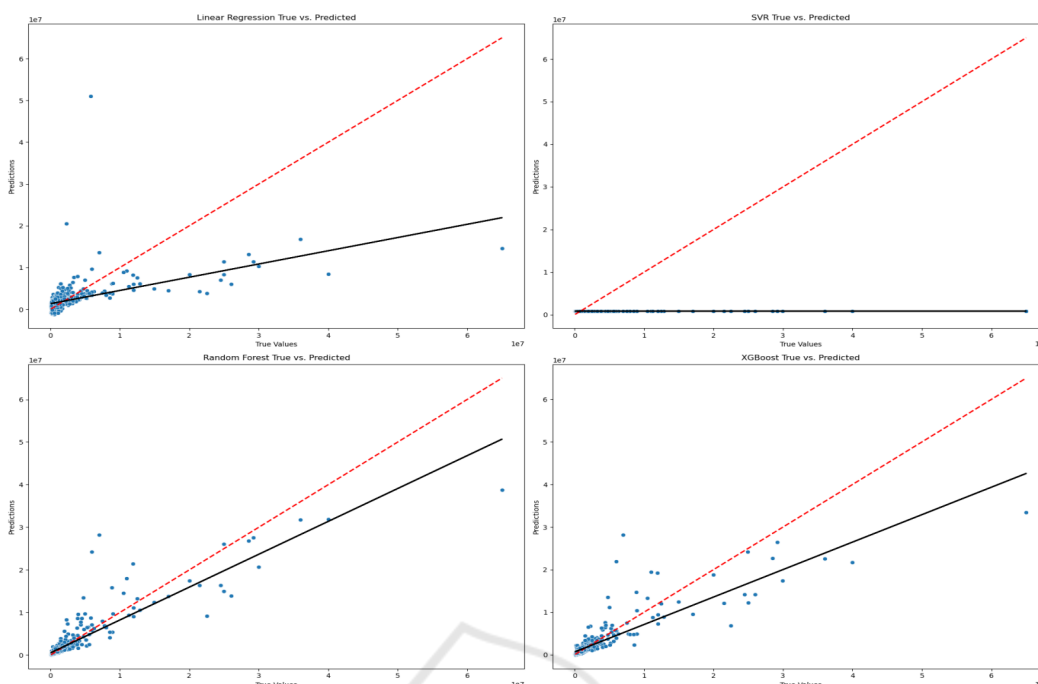


Figure 4: True Price and Model Predicted Price Comparison (Picture credit: Original).

The evaluation of model performances revealed distinct strengths and weaknesses across the methodologies. LR struggled, as evidenced by a high MAPE of 129.91% and a low adjusted R^2 of 0.2642, likely due to its inability to capture the complex relationships present in the dataset. Likewise, SVR showed a marginal improvement in performance with a MAPE of 77.53%, but possibly suffered from the dataset's high dimensionality and noisy data, which resulted in a higher RMSE of 4,967,168 and a negative adjusted R^2 of -0.0844. In contrast, the ensemble methods, RF and XGB, demonstrated robustness and superior performance, effectively handling the dataset's complexities. RF, with an RMSE of 2,145,123, MAPE of 30.86%, and an adjusted R^2 of 0.7978, along with XGB, which achieved an RMSE of 2,483,884, MAPE of 41.62%, and an adjusted R^2 of 0.7288, clearly distinguished their predictive accuracy and generalization capability over the simpler models due to their ability to model non-linear relationships and mitigate issues stemming from noisy and high-dimensional data.

In Figure 4, each scatter plot visually illustrates the relationship between the actual and predicted housing prices for the respective models, where each point corresponds to an individual record from the test set. The true values are plotted along the x-axis, while the predicted values are on the y-axis. The black line in each plot represents the line of best fit, showcasing the average direction of the data; points

clustering around this line suggest more accurate predictions. The red line serves as the identity line, marking where predicted values match the actual prices perfectly.

Upon analysis, the LR and SVR plots reveal greater divergence from the identity line, highlighting a propensity for underestimation. The plots for RF and XGB, while showing a tighter grouping around the identity line for lower-priced properties, indicate deviations at higher price points. This pattern suggests a more pronounced accuracy in predictions for moderately priced homes, with deviation becoming more evident as the value increases

5 CONCLUSION

This study utilized various machine learning techniques to forecast housing prices in New York, focusing on the evaluation of LR, Support SVR, RF, and XGB. LR and SVR exhibited less than optimal performance, characterized by high RMSEs and MAPEs, along with low adjusted R^2 values. This was attributed to LR's inability to capture complex patterns in the dataset and the impact of high dimensionality and noise on SVR's performance. In contrast, the ensemble methods, RF and XGB, showed strong results. RF, configured with 100 trees and mean squared error as its split criterion, achieved

an RMSE of 2,145,122, a MAPE of 30.86%, and an adjusted R^2 of 0.7978. XGB also performed well, with an RMSE of 2,483,884, a MAPE of 41.62%, and an R^2 of 0.7288. These results highlight the efficacy of ensemble methods in handling complex predictive modeling challenges, suggesting their potential to lead future research not only in housing price prediction but also in other areas of economic forecasting facing similar complexities.

REFERENCES

- Frohlich, T. C., Stebbins, S., "50 Worst Cities to Live In." 24/7 Wall Street, 2016.
- A. C. Goodman. Hedonic prices, price indices and housing markets. *Journal of Urban Economics*, 5:471 – 484, 1978.
- R. Halvorsen and H. O. Pollakowski. Choice of functional form for hedonic price equations. *Journal of Urban Economics*, 10:37–49, 1981.
- Ho, W. K. O., Tang, B. S., & Wong, S. W. Predicting property prices with machine learning algorithms. *Journal of Property Research*, 2021,38(1), 48–70.
- Truong, Q., Nguyen, M., Dang, H., & Mei, B., "Housing price prediction via improved machine learning techniques." *Procedia Computer Science* ,2020, 174, 433-442.
- Soltani, A., Heydari, M., Aghaei, F., & Pettit, C. J. Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms. *Cities*, 2021, 131, 103941.
- Elgiriwithana, N., "New York Housing Market." Kaggle Inc, 2024. <https://www.kaggle.com/datasets/nelgiriye-withana/new-york-housing-market?rvi=1>
- Ahsan, M. M., Mahmud, M. A. P., Saha, P. K., Gupta, K. D., & Siddique, Z. Effect of data scaling methods on machine learning algorithms and model performance. *Technologies*, 2021, 9(3), 52.
- Zhou, Z.-H. Ensemble learning. In *Machine Learning*, 2021, pp. 181-210.
- Zhang, F., & O'Donnell, L. J. Chapter 7 - Support vector regression. In A. Mechelli & S. Vieira (Eds.), *Machine Learning*, 2021, pp. 123-140.