Predicting New York Housing Prices: A Machine Learning Approach Incorporating School, Living facilities and Real Estate Market Factors

Chen Liang *

Department of Computer Science and Technology, Shanghai University of Electricity Power, Shanghai, China

* Corresponding author: 1811571105@mail.sit.edu.cn

Abstract. This study aims to predict housing prices in New York by utilizing machine learning methods that incorporate factors such as schools, living facilities, and the real estate market. This paper collected extensive data, including metrics on school quality, assessments of living facility convenience, and real estate market data. Employing a regression-based machine learning algorithm, the study incorporated these factors into a predictive model. Through training and testing the model, this study discovered that school quantity and the convenience of living facilities significantly impact housing prices. The predictive model demonstrated good accuracy and predictive capability on the test set, validating the effectiveness of this approach. The findings of this study provide valuable insights for real estate market participants, policymakers, and investors to better understand and forecast housing price trends in New York.

Keywords: Real Estate, Machine Learning, Prediction.

1. Introduction

The real estate market has long been a focal point, holding significant relevance for households, investors, and the broader economy. Understanding and accurately predicting fluctuations in house prices have become imperative for real estate professionals and decision-makers alike. An essential part of the global economy is the real estate sector, with its fluctuations and trends exerting profound impacts on household finances, investment decisions, and macroeconomic stability. However, the price of houses is influenced by a multitude of complex factors, including geographical location, market supply and demand dynamics, economic conditions, among others, rendering prediction a formidable challenge [1].

Prior research has made strides in the field of house price prediction, yet significant challenges persist. Traditional statistical methods may struggle to capture non-linear relationships within the data, and their efficiency in handling large-scale datasets is limited. In this context, the introduction of machine learning techniques provides a powerful toolset [2]. Machine learning can process intricate data patterns and deliver more accurate predictions. Moreover, in recent years, the real estate market in New York City has undergone rapid transformations, marked by a surging demand from homebuyers, regional price disparities, and emerging housing development projects. Consequently, precise house price predictions are vital for making informed investments and property purchases in this competitive market. The significance of this research lies in providing stakeholders with a methodology for better comprehending and forecasting the dynamics of house prices in New York City through the application of machine learning [3].

The primary focus of this study revolves around house price prediction analysis using machine learning techniques. This paper intends to collect extensive data from the real estate market in New York City and leverage advanced machine learning algorithms to construct predictive models. These models will consider various critical factors such as housing characteristics, geographical location and market trends to forecast future variations in house prices [4].

The main techniques used in this study are data collecting, data preprocessing, feature engineering and feature selection, as well as the creation and assessment of machine learning models. The overarching goal is to provide accurate house price predictions and furnish valuable information for

participants and decision-makers in the real estate market, thereby assisting them in making prudent investment and property purchase decisions [5]. This study aspires not only to enhance the accuracy of house price predictions but also to contribute to the real estate market's resilience and stability. This endeavor aims to better address the ever-growing housing demand in New York City while improving market transparency and efficiency.

2. Data and Approaches

2.1. Data Origin

This study's dataset was obtained from kaggle.com and covers the period from 2017 to 2020, focusing on housing sales data in New York. The dataset provides detailed information about the housing market in New York, which is essential for our machine learning prediction approach.

The dataset includes various factors that influence housing prices. Firstly, considering the characteristics of the properties themselves, such as the number of bedrooms and bathrooms, as these factors have a significant impact on housing prices. Secondly, this paper also takes into account the availability of amenities, such as the number of restaurants and supermarkets in the vicinity, which also influence housing prices. Additionally, this paper incorporates factors like the number of schools nearby and the crime rate to provide a comprehensive evaluation of housing prices.

To guarantee the accuracy and reliability of the data, this study conducted rigorous data cleaning and processing. This involved addressing missing values, outliers, and inconsistencies to enhance the reliability and accuracy of the dataset.

In summary, our dataset provides detailed records of housing sales in New York, including factors such as the number of bedrooms and bathrooms, availability of amenities like restaurants and supermarkets, as well as the number of nearby schools and crime rate. This dataset will serve as a robust foundation for our machine learning model to predict housing prices in the New York area.

2.2. Project Preparations

2.2.1. Preprocessing and Feature Selection

Data Cleaning: The raw dataset underwent rigorous data cleaning to address missing values, outliers, and inconsistencies. Cleaning was essential to ensure the quality and integrity of the dataset.

Normalization and Scaling: Continuous numerical features were normalized to bring them within a common scale, preventing certain variables from dominating the analysis. Standardization techniques were employed to ensure that features with different units did not bias the model.

Categorical Feature Encoding: One-Hot Encoding: Categorical variables, such as neighborhood names, were transformed using one-hot encoding to represent them as binary values. This facilitated the integration of categorical information into machine learning models.

2.2.2. Machine Learning Models

Linear Regression: As a baseline model, linear regression was employed to understand how well a simple linear relationship could predict house prices.

Random Forest Regression: An integrated approach, was chosen for the data's capacity to capture complicated interactions and handle both numerical and categorical features.

Model Evaluation: To train and assess the performance of the models, the dataset was split into training and testing sets. It was customary to divide the data up into 20% for testing and 80% for training.

Cross-Validation: K-fold cross-validation enhances the reliability of results and reduces the risk of overfitting by dividing the dataset into k subsets. The model is then trained and tested k times using different combinations of these subsets.

Hyperparameter Tuning: Grid Search: Grid search was used for hyperparameter tuning to determine the ideal hyperparameters for each model, such as the number of estimators. and maximum depth for random forest [6].

2.2.3. Evaluation Metrics

To assess model performance, this paper used important measures like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R2) to evaluate the effectiveness of the model. These measures gave information on prediction accuracy and the capacity to capture price variation in homes.

MSE [7]: MSE stands for Mean Squared Error. It is a statistical metric used to measure the difference between predicted values and actual values.A smaller MSE indicates that the predicted values are closer to the actual values, indicating better predictive performance of the model. Mathematically, it is calculated as:

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(True - predicted)^2 \quad (1)$$

R^2 [8]: R-squared is a statistical metric used to measure the goodness-of-fit of a regression model. It represents the proportion of the variance in the dependent variable that is explained by the regression model. R^2 ranges from 0 to 1, where a value closer to 1 indicates that the model can explain a larger amount of the variability in the dependent variable, while a value closer to 0 suggests a poorer fit of the model. The equation is as follows:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \quad (2)$$

SSE: SSE stands for Sum of Squared Errors.It is used to calculate the sum of the squared differences between the predicted values and the actual values in a regression model. A smaller SSE indicates that the predicted values are closer to the actual values, suggesting a better fit of the regression model to the data.

2.3. Training and Testing

2.3.1. Model Types

Paper Study selects many models, including linear regression, decision trees, random forests, and support vector machines, which were trained and evaluated.[9].

2.3.2. Training and Tuning

Following the methodology suggested by Francisco et al. (2020), an 80/20 train/test split was adopted [10]. Additionally, the dataset was divided into K equally-sized subgroups, or "folds," using the k-fold cross-validation method. The mean of K validation scores served as the final evaluation metric, ensuring a robust and reliable assessment, and aiding in the prevention of overfitting.

The models were ranked according to their Mean Squared Error (MSE), with random forests exhibiting the lowest error and emerging as the most promising model. The performance of linear regression and decision trees was slightly inferior, suggesting potential for hyperparameter tuning.

This paper used grid search to optimize the hyperparameters, such as the number of trees, maximum features, tree depth, minimum samples for splitting, and minimum leaf samples, after deciding that the Random Forest was the best model to use. This exhaustive search approach systematically assesses the defined hyperparameter space, identifying the combination that maximizes predictive performance. While effective, the complexity of grid search necessitates careful consideration of computational resources.

For Random Forest, common hyperparameters to tune include:

1. n_estimators: Specifies the number of decision trees in the random forest.
2. max_features: Indicates the maximum number of features considered when splitting a node in each decision tree.
3. max_depth: Limits the maximum depth or number of levels in each decision tree.

4. min_samples_split: Specifies the minimum number of samples that must be present in a node for further splitting to occur.

5. min_samples_leaf: Specifies the minimum number of samples required to be in a leaf node.

This study will define some potential values for these hyperparameters. This enables us to investigate various combinations systematically and identify the optimal hyperparameter configuration.

This procedure can substantially enhance the efficacy of the model, particularly for complex models and large datasets. Nonetheless, it can be very time-consuming, particularly if the dataset is large and there are numerous hyperparameters.

3. Results and Discussion

The accuracy of the prediction made by the Random Forest model in this study was evaluated by comparing the true housing prices with the predicted prices through a scatter plot (Fig.1). In an ideal scenario, the data points would align perfectly along a line, indicating a perfect agreement among the actual and anticipated values. As depicted in Fig.1, the majority of data points cluster around this line, demonstrating a strong correlation between the predicted and true prices. However, it is worth noting that there are some outliers presents, indicating occasional significant deviations between the predicted and true prices.

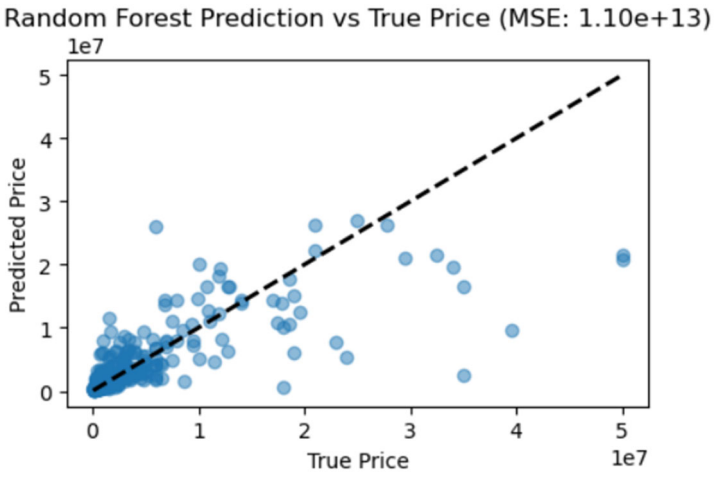

Figure 1. Random Forest Prediction vs True Price

The feature importance analysis reveals the relative significance of various attributes in the housing price prediction model. As shown in Fig2, Specifically, bathrooms emerge as the most salient feature, followed by factors such as life facilities and number of restaurants. Conversely, spatial characteristics like groceries and nightlife hold moderate importance, while the number of schools are found to be the least influential.

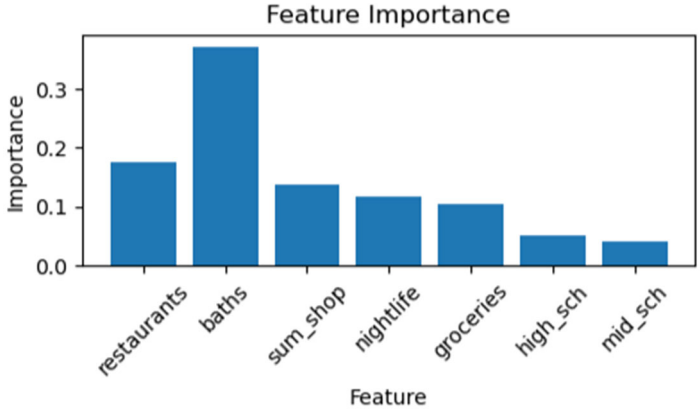

Figure 2. Feature Importance

This hierarchical organization of feature importance not only sheds light on the underlying dynamics that influence housing prices but also contributes to the academic discourse surrounding real estate economics and market behavior. Such understanding holds broader implications, assisting in the development of real estate strategies, urban planning policies, and even informing investment decisions. Hence, the analysis represents a valuable intersection between computational modeling and socio-economic insights, affirming the relevance of machine learning in contemporary economic studies. Additionally, the predictive accuracy of the Random Forest model may vary depending on the specific context and dataset.

Looking ahead, further research could explore additional factors that affect housing prices, such as transportation accessibility or proximity to recreational areas. Additionally, incorporating more advanced machine learning techniques or exploring different models could potentially improve the predictive accuracy.

4. Conclusion

This study primarily utilized a Random Forest regressor, which was fine-tuned through meticulous hyperparameter tuning to create a well-fitting model for the dataset. Through meticulous hyperparameter tuning, this study has synthesized a model that demonstrates a robust fit to the underlying data structure. This Random Forest model transcends mere academic application; it can be pragmatically deployed within the real estate market to furnish more precise price evaluations for various stakeholders, including real estate agents, investors, and potential homeowners.

However, it is imperative to recognize that the deployment of such a model in a dynamically evolving market necessitates continuous monitoring and iterative refinement. This involves not only retraining on contemporary data to capture shifting trends but also vigilant scrutiny for potential model biases or disparities that might inadvertently arise. Such a proactive approach to model maintenance ensures that the predictive tool retains its relevance, accuracy, and fairness, thus aligning with the ethical imperatives and methodological rigor demanded by modern data-driven decision-making in the business environment.

References

[1] Pai, P.-F., & Wang, W.-C. Using machine learning models and actual transaction data for predicting real estate prices. Applied Sciences-Basel, 2020, 10 (17), 5832. doi: 10.3390/app10175832.

[2] Bilgilioglu, S. S., & Yilmaz, H. M. Comparison of different machine learning models for mass appraisal of real estate. Survey Review, 2023, 55 (388), 32 - 43. doi: 10.1080/00396265.2021.1996799.

[3] Xu, D. D. Comparative study of Xi'an house price prediction based on multiple linear regression model and BP neural network. Real Estate World, 2022, (08), 11 - 13.

[4] Zhu, H. Y., Wang, Z. J., & Ye, C. C. House price prediction of urban hotspot areas based on XGBoost algorithm: A case study of Nanjing Jiangbei New District. Construction Economics, 2022, 43 (S2), 433 - 437. DOI: 10.14181/j.cnki.1002-851x.2022S2433.

[5] MDPI and ACS Style Mora-Garcia, R.-T.; Cespedes-Lopez, M.-F.; Perez-Sanchez, V.R. Housing price prediction using machine learning algorithms in COVID-19 times. Land 2022, 11, 2100. https: //doi.org/10.3390/land11112100.

[6] Sun, Y., Gong, H., Li, Y., & Zhang, D. Hyperparameter importance analysis based on n-rrelieff algorithm. international journal of computers, Communications & Control, 2019, 14 (4), 557 - 573. DOI: 10.15837/ijccc.2019. 4. 3593.

[7] Mean squared error 2021, April 1, Mean Squared Error - Wikipedia. https: //en.wikipedia.org/wiki/Mean_squared_error.

[8] Coefficient of determination 2021, April 23, Coefficient of Determination https: //en.wikipedia.org/wiki/Coefficient_of_determination.

[9] Kim, J., Lee, Y., Lee, MH., & Hong, SY. A comparative study of machine learning and spatial interpolation methods for predicting house prices. Sustainability, 2022, 14 (15), Article 9056. DOI: 10.3390/su14159056.

[10] Francisco L. Deep learning-based computer vision to recognize and classify suturing gestures in robot-assisted surgery. Artificial intelligence, 2020, 169 (5), 1240 - 1244. https: //doi.org/10.1016/j.surg.2020. 08. 016.