REGULAR ARTICLE

The House of Us: Metro Manila Housing Price Prediction Using Machine Learning

Alyanna Mae Bulatao, Lorenzo Emmanuel de la Paz, Godfrey Bryan Satiada, and Jeremiah Dominic Soliman

Abstract

This study addresses the challenge of accurately predicting housing prices in Metro Manila, an area where property listings often feature inaccuracies or inflated values, and where traditional pricing methodologies fall short. Leveraging machine learning techniques, the researchers aim to provide buyers and sellers with more precise pricing insights to facilitate informed decision-making and optimize pricing strategies. Utilizing a dataset from Kaggle comprising properties listed on lamudi.com.ph in 2022, and incorporating data from OpenStreetMap (OSM) for additional contextual factors like nearby amenities, the researchers explore the impact of various property features and amenities on housing prices. Gradient Boosting Regressor model achieved the highest mean test accuracy score for the combined dataset of property listings and surrounding amenities. The results indicate that including amenity data improved predictions significantly in comparison to the dataset without amenities as features, highlighting as well that the researchers were able to extract which amenities contributed most to property prices. SHAP analysis revealed that floor area, location, and financial amenities are the most influential factors in property pricing. This research contributes to the field by offering a novel, data-driven approach to real estate pricing in Metro Manila. By incorporating additional features beyond the traditional property attributes, the study provides a comprehensive analysis that enhances understanding and predictive accuracy in the real estate market.

Keywords:

property pricing; price prediction; real estate; Metro Manila

Highlights

1. Feature engineering can lead to enhanced insights for existing studies
2. Financial Amenities are the highest contributing amenity type in property pricing.
3. SHAP enables better understanding factors that drive property prices.
4. The advantage of the value of open-source datasets should be maximized.
5. Ensemble tree-based models performed best in this study.

1. Introduction

Determining property prices presents considerable challenges. Frequently, property listings exhibit inaccuracies or inflated values. Accessing reliable information regarding market or zonal values proves to be difficult, often generalized over larger regions. Moreover, appraisal values tend to be subjective and commonly spread through informal channels. Pricing methodologies often resort to simplistic approaches, such as multiplying property area by an arbitrary price per square meter. This inadequacy underscores the researchers' impetus to devise a data-driven solution for predicting property prices.

The primary objective of this research is to empower both buyers and sellers with more accurate pricing insights. By leveraging data-driven methodologies, the researchers endeavor to facilitate informed decision-making for buyers and enable sellers to optimize their pricing strategies effectively.

2. Related Study

Past studies have tried to predict property prices, with some focusing on linear regression models for price prediction in Metro Manila. However, these discussions primarily revolve around analyzing price variations over time, particularly during the pandemic (Abellana et al., 2021).

To the researchers' knowledge, there have been minimal efforts in studies aimed at predicting property prices on an instance-based level. Instance-based analysis refers to evaluating property prices at a specific point in time. The present study seeks to address queries such as "What is the current price of a property in Barangay Ungong, Pasig?" through the application of machine learning techniques.

3. Data and Methods

To accomplish the objective of this study, the researchers utilized a dataset sourced from Kaggle (Blanco, n.d.), which comprises web scraped data from lamudi.com.ph from the year

2022. Lamudi serves as an online real estate marketplace facilitating property transactions, including buying, selling, and renting. Acting as a platform for individuals, agents, and developers, Lamudi allows users to list properties and engage with potential buyers or renters. The dataset encompasses various property attributes, including a brief description, location details such as barangay and city, latitude and longitude coordinates, number of bedrooms and bathrooms, floor or land area, a weblink to the respective listing on lamudi.com.ph, and the corresponding selling price. For this study, the researchers limited it to the pricing of condominiums within Metro Manila. The rationale behind this is that they believed that pricing for condominiums is not the same for pricing for land or lots.

In addition to the dataset from Kaggle, the researchers incorporated data from OpenStreetMap (OSM) into their study. OSM operates as a collaborative platform where users worldwide contribute to the collection and updating of information on various amenity points, including restaurants, schools, and parks. This crowd-sourced data, freely accessible to the public, undergoes continuous updates by a global community and finds application in diverse fields such as navigation and urban planning. The OSM dataset available to the researchers was from 2018.

From the OSM dataset, the researchers extracted amenity counts for the top amenities in Metro Manila. These amenities were then grouped into six broader categories: Healthcare Amenities, Food Amenities, Transportation Amenities, Financial Amenities, Public Services Amenities, and Education Amenities. Furthermore, these amenity types were aggregated at the barangay level for this study.

Typically, property prices are influenced by property features like the number of bedrooms, bathrooms, and floor area. However, the researchers aimed to investigate whether additional factors contribute to property pricing. Specifically, they sought to explore the impact of various amenities on property prices.

The researchers combined the property features extracted from the Lamudi web-scraped data with amenity information sourced from the OSM database to create a comprehensive dataset.

Data preprocessing primarily focused on the Lamudi dataset. Null values in the 'Price (PHP)' column, representing the target variable, were removed. Additionally, listings with zero values for the number of bedrooms and bathrooms were adjusted to have a minimum value of 1, under the assumption that each listing should include at least one room.

Following preprocessing, the final dataset comprised 969 columns and 11 features, with 'Price (PHP)' serving as the target variable.

Table 1. Summary of dataset

#	Column	Non-Null	Dtype	Description
1	Latitude	949	float64	Latitude of the Property
2	Longitude	949	float64	Longitude of the Property
3	Bedrooms	949	int64	No. of Bedrooms of the Property
4	Bath	949	int64	No. of Bathrooms of the Property
5	Floor_area (sqm)	949	float64	Size of the Property
6	Food_Count	949	float64	No. of Food Amenities Present
7	Education_Count	949	float64	No. of Education Amenities Present
8	Healthcare_Count	949	float64	No. of Healthcare Amenities Present
9	Public_Services_Count	949	float64	No. of Public Service Amenities Present
10	Finance_Count	949	float64	No. of Finance Amenities Present
11	Transportation_Count	949	float64	No. of Transportation Amenities Present
12	**Price (PHP)**	**949**	**float64**	**Price of the property**

After completing data cleaning and preprocessing, the researchers proceeded with model training. To establish a baseline comparison, models were trained on both the Lamudi dataset alone and the combined dataset integrating information from both Lamudi and OpenStreetMap (OSM).

The chosen regression models for this study include: Gradient Boosting, Random Forest, XGBoost, AdaBoost, Linear Regression without regularization, Linear Regression with L1 and L2 regularizations, and K-Nearest Neighbors. These models are trained to optimize mean absolute errors.

The summary of the model trainings are as follows:

Table 2. Summary of training performance on dataset without amenities

Model	Mean Train Score	Mean Validation Score
Random Forest	883,765.90	1,286,228.00
XGBRegressor	700,386.60	1,356,071.00
Gradient Boosting	1,089,899.00	1,490,472.00
Knearest Neighbor	1,292,744.00	1,641,638.00
Linear Regression (L2)	2,538,230.00	2,570,945.00
Linear Regression (L1)	2,530,133.00	2,571,898.00
Linear Regression	2,540,134.00	2,571,898.00
AdaBoost	2,918,678.00	3,219,392.00

Table 3. Summary of training performance on dataset with amenities

Model	Mean Train Score	Mean Validation Score
Gradient Boosting	272,093.00	762,716.10
XGBRegressor	272,093.80	864,976.70
Random Forest	516,634.20	874,820.90
KNearest Neighbor	339,899.60	1,010,460.00
Linear Regression (L1)	2,396,472.00	2,480,025.50
Linear Regression	2,396,478.00	2,480,032.00
Linear Regression (L2)	2,396,478.00	2,480,033.00
AdaBoost	2,142,613.00	2,525,853.00

4. Results and Discussion

The comparison revealed an improvement in R^2 scores from 98.50 to 99.86 in the train-validation set and from 97.18 to 98.97 in the test or holdout set. It's worth highlighting that Random Forest and Gradient Boosting emerged as the top performers among the selected models. These findings align with the results reported by Chowhaan et al. (2021), suggesting that tree-based models exhibit superior performance in similar real estate predictive modeling scenarios.

To provide a more practical understanding of the model's performance, the researchers also prepared a table illustrating the improvement in mean absolute error (MAE) and mean absolute percent error (MAPE) across three property price ranges. The table includes median values, as indicated below:

Table 4. Summary of MAE across different property price ranges

Price Range (Php)	Base (Random Forest)	with Amenities (Gradient Boosting)
High (27,431,000)	PHP 1,124,528	PHP 565,536
Medium (15,492,000)	PHP 756,233	PHP 473,412
Low (7,667,973)	PHP 933,943	PHP 261,425

Table 5. Summary of MAPE across different property price ranges

Price Range (Php)	Base (Random Forest)	with Amenities (Gradient Boosting)
High (27,431,000)	(3.96%)	(1.98%)
Medium (15,492,000)	(5.44%)	(1.91%)
Low (7,667,973)	(15.09%)	(3.39%)

The improvements in the mean absolute errors were subjected to a one sample T-test and it provided the researchers with 95% confidence that the improvements are significant

After being satisfied with the results of the model training, the researchers proceeded with the study's interpretability. To gain a better understanding of the interactions between the features and their impact on the outcome, the researchers utilized the Python library: SHAP (SHapley Additive exPlanations). SHAP is a Python library commonly used for explaining the output of machine learning models. It provides a unified approach to explaining the output of any machine learning model by computing Shapley values, a concept derived from cooperative game theory. Shapley values offer a fair distribution of the "credit" of a model's prediction among its features.

The researchers utilized SHAP's global feature importance method, enabling them to observe the effects of features on predicting property prices.

Below is the summary plot of mean SHAP values.

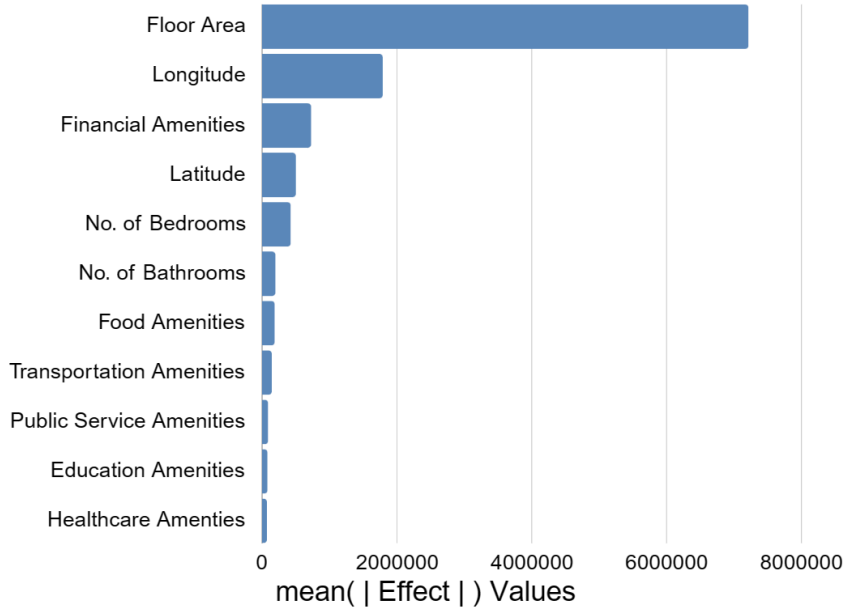

Figure 1. Mean effect of different features in price prediction.

The plot reveals that the most significant driving factor for property prices remains to be floor area, which aligns with intuitive expectations and supports the common practice of estimating average property prices based on floor area. Longitude and latitude also rank highly among the top factors, indicating that location plays a substantial role in property prices. This supports the notion that proximity to certain areas, such as business districts, influences property prices.

However, what surprised the researchers was the high importance of 'Financial Amenities' as the third-highest driving factor for property prices. This unexpected finding suggests that access to financial amenities, such as banks or ATMs and foreign exchange kiosks, significantly impacts property prices. This underscores the complexity of factors influencing property prices beyond traditional metrics like size and location, highlighting the importance of considering diverse variables in predictive modeling and real estate analysis.

5. Conclusion

The researchers have successfully developed an accurate predictive model for property prices. Aside from having notable improvement from the baseline, the novelty of their work lies in their enhanced interpretability of how prices vary based on the amenities present at the barangay level.

By incorporating amenity data from OpenStreetMap and utilizing techniques like SHAP to analyze feature importance, the researchers have provided valuable insights into the factors influencing property prices beyond traditional metrics. Specifically, their analysis sheds light on the impact of amenities on property prices at a localized level, offering a more nuanced understanding of real estate dynamics.

6. Recommendations

Firstly, normalize price to price per square meter. Given that floor area emerged as the primary driving factor for property prices, the researchers propose normalizing prices to price per square meter. This adjustment would help mitigate the influence of floor area on property prices and provide a more standardized metric for comparison across different properties. By doing so, the researchers aim to enhance the accuracy and interpretability of property price predictions.

Second, enhance amenity counts by proximity: Instead of relying solely on amenity counts at the barangay level, the researchers recommend incorporating amenity counts within a certain distance from the property units. By considering amenities within a specified radius (e.g., 500m, 1000m, or 2000m), the analysis can capture the immediate neighborhood environment more accurately, providing a more nuanced understanding of how amenities influence property prices.

Third, the researchers recommend future studies to explore additional amenities, While the study focused on commonly found amenities in cities and barangays, future research could delve into exploring fewer common amenities. Investigating the impact of unique or less frequently encountered amenities on property prices could uncover new insights and refine

predictive models further. Moreover, exploring the gravity of "unwanted" amenities, such as prisons, recycling plants, mortuaries, and similar facilities, could provide valuable insights into their effects on property prices and inform urban planning and development strategies.

Lastly, the researchers advise incorporating additional features beyond amenities into property price prediction models. These features could encompass factors like population density at the barangay or city level.

Acknowledgement

The researchers extend their sincere gratitude to their mentors, Prof. Chris Monterola, Prof. K-Ann Carandang, and Prof. Leo Lorenzo II, for their unwavering support and guidance throughout the entire Machine Learning 2 course, and especially during the completion of this project.

Additionally, a special mention goes to ChatGPT-3.5 and ChatGPT-4 by OpenAI, which the researchers utilized as augmenting tools to assist in this project.

Appendix

Please see attached jupyter notebooks and other files contained in the zip file.

References

Abellana, J. A., & Devaraj, M. (2021). Hedonic Modeling for Predicting House Prices during COVID-19 Pandemic in the Philippines. In *Proceedings of the 2021 3rd International Conference on Management Science and Industrial Engineering* (pp. 21–26). Retrieved from https://doi.org/10.1145/3460824.3460828

Jagan Chowhaan, M., Nitish, D., Akash, G., Nelli, S., & Shaik, S. (2023). Machine Learning Approach for House Price Prediction. *Asian Journal of Research in Computer Science, 16,* 54-61. Retrieved from https://doi.org/10.9734/ajrcos/2023/v16i2339

Arlo Blanco. (n.d.). Philippine Real Estate. Kaggle. Retrieved from https://www.kaggle.com/datasets/arloblanco/philippine-real-estate/data