Utilization of Machine Learning, Government-Based and Non-Conventional Indicators for Property Value Prediction in the Philippines

Gabriel Isaac L. Ramolete[1], Bryan Bramaskara, Dustin A. Reyes, and Adrienne Heinrich

Aboitiz Data Innovation, 47 Scotts Road #16-01/02. Goldbell Towers Singapore 228233

ABSTRACT

Property appraisal and value estimation in the Philippines are prone to human errors and bias, due to price subjectivity and the general difficulty in properly quantifying the impact of factors beyond the property itself. Predictive models for property valuation typically involve conventional features of the house (e.g., number of bathrooms) and market prices of nearby properties. This paper investigates the value of incorporating alternative data to account for deviations in true market value and improve property value predictions in the Philippines and other developing countries. The study considers public data and anchors socio-economic indicators to assess its relevance to property value prediction in the Philippines. By utilizing the Department of Trade and Industry's 2021 National Competitiveness Index Rating, this research also investigates the significance of a Local Government Unit's competitiveness based on their economic dynamism, government efficiency, infrastructure, and resiliency. Different commonly used Machine Learning (ML) methods and features from various data sources are compared and it is found that the inclusion of government indicators has substantial positive effect on the model performance on top of conventional indicators that can be globally replicated. A Mean Average Percentage Error (MAPE) of 10.7-21% is obtained which is competitive compared to the performance ranges of other reported models. A property segment (personalized) approach is proposed to achieve lower error rates in Philippine appraisal (in 87.5% of cases), better access and transparency for populations outside the real estate network, and minimally biased assessments, all of which are also relevant for other developing countries.

Keywords: property appraisal, spatial analysis, city competitiveness, clustering

1. INTRODUCTION

Assessing the value of real estate properties can often be a problematic and iterative process for buyers and sellers. While interested parties may rely on local market information, valuations of similar properties, and the experience of professional appraisers, the number of variables to consider when determining the value of a property is often a source of contestation. Location, home size, usable space, and neighborhood comparisons are common factors considered by most professionals during the appraisal process (Naqvi, 2017; The Danh Phan, 2018; Nallathiga, Upadhyay et al., 2019). Other alternative externalities have also been explored both in research and operationalized services – accessibility, public service facilities, commercial places of interest, safety, and livability have all been quantified and explored as possible additional indicators to determine a property's true market value (Wittowsky et al., 2020; Zhang et al., 2018; Chen et al., 2020; Santos and Jiang, 2020; Buyukkaracigan, 2021).

Traditionally, methods such as the Hedonic Pricing Model, Sales Comparison Approach, the Cost Approach, the Income Capitalization Approach, the Discounted Cash Flow (DCF) Method, and the Gross Rent Multiplier Method have been used to predict property values (Adetiloye and Eke, 2014). The authors of the book entitled '*Modern Methods Approach in Real Estate Valuation'* note that these methods are preferred in practice as sales information is easily attainable or because future income can be determined (Buyukkaracigan, 2021). The Philippine Valuation Standards Manual and Malaysian Valuation Standards mention the usage of such methods as those recognized by valuers and users of valuation. (Bureau of Local Government Finance, 2018; Board of Valuers, Appraisers and Property Managers, 2019)

[1] Corresponding Author: gabriel.ramolete@aboitiz.com

However, some market analyses over periods of time have shown that utilizing these methods can be met with difficulties (Chaphalkar and Sandbhor, 2013; Kershaw and Rossini, 1999; Adetiloye and Eke, 2014). Due to the sheer number of factors to examine, human error and unconscious bias are likely to affect appraised property prices, potentially causing valuation variation (Howard, 2004; Evans et al., 2019; Yiu, et al., 2006; Tidwell and Gallimore, 2014). Traditional methods such as the income approach also do not take non-pecuniary values into account or may change drastically due to capitalization rates or for urban fringes (Buyukkaracigan, 2021; Tanrivermis, 2016). Other macroeconomic factors affecting property valuations and prices include the country's employment rate, inflation rate, interest rate, income of the local government unit, and poverty incidence of the area (Naqvi, 2017). Most significantly, appraisal bias can occur due to professionals' differing methods and views in the appraisal and valuation of properties. Given that property valuation is a human activity, judgment bias may occur in the form of random and systematic errors which can have a great effect on an investor's decision (Evans et al., 2019).

In a 3rd world country such as the Philippines, the difficulty of evaluating the price of properties is exacerbated due to the presence of multiple valuation systems imposed by different government entities (Mandani Bay, 2018). Achieving consistently accurate prices can be hindered by the lack of updated zonal-based fair market value (FMV) and the presence of multiple valuation systems executed by local government authorities. While most land valuation standards in the Philippines adopt International Valuation Standards (IVS) published by the IVSC, the inadequacies of common valuation methods in the Philippines can still lead to undervalued properties and misinformed decisions on both the appraisers' and buyers' ends (Domingo and Fulleros, 2002).

One major factor of appraisal variation lies in the zonal valuation system in the Philippines, spearheaded by two main entities: The Bureau of Internal Revenue (BIR) and the Local Government Unit (LGU) of which the property is located. According to the Philippines' 1997 Tax Code, the fair market value of a property is prominently assessed by the BIR. As amended by the TRAIN Law in 2017, the BIR helps LGUs assess the FMV of real properties in each zone or area upon mandatory consultation with competent appraisers (Congress of the Philippines, 2017). These values are subject to automatic adjustment every three (3) years. However, only 60% of LGUs have updated their zonal values in 2017-2020. Similarly, only 37% of LGUs have been able to submit updated schedules of market values during the same timeframe (Unciano, 2020).

Why do LGUs find it difficult to update zonal values on time? A study from the German Institute for Development Evaluation suggests that the current issues of Philippine land planning and management system can negatively affect property valuation (Lech and Gerald, 2018). LGUs are required to develop Comprehensive Land Use Plans (CLUPs) – a basis for handling the allocation of land resources and properties of an LGU's territory. However, accomplishing the CLUP is highly dependent on the cooperation of different agencies and their often-overlapping mandates; horizontal and vertical frictions occur when dealing with provincial development plans, budget planning, municipal budgeting, barangay development, and other frameworks to be developed in parallel. Due to often outdated property valuation references, it is common that taxpayers and administrators employ their own strategies and methods of property valuation; often, the same traditional methods mentioned above.

As shown, the Philippines has deep-rooted issues in its property valuation system which may make appraisals vary in accuracy and reliability. Other countries have addressed similar concerns by relying on statistical and AI-driven methods and decision support systems to aid interested parties in determining more accurate values. Studies in Dortmund, Kuala Lumpur, Guangzhou, London, and Shanghai showcase the usage of multiple regression, boosted regression, spatial lag, and geographically weighted regression models as methods to achieve price predictions with reliable accuracy (Wittowsky et al., 2020; Nallathiga et al., 2019; Santos and Jiang, 2020; Huang et al., 2017; McCluskey et al., 2014). Other machine learning (ML) techniques, such as Random Forest, Gradient Boosting Machine, LightGBM, and XGBoost, have also been utilized in identifying real estate opportunities around the world while attempting to understand spatial dependence (Gao, et al., 2022; Chou, et al., 2022; Zhao, et al., 2019; Baldominos, et al., 2018; The Dank Phan, 2018). Neural networks and fuzzy logic have been used in sales prices of apartments and residential housing values, performing better than traditional methods (Chaphalkar and Sandbhor, 2013; Nguyen and Cripps, 2001; Pi-ying, 2011; Krzystanek et al.,

2009; Lughofer et al., 2011). Unsupervised techniques have also been utilized to aid these techniques and pre-group properties with similar characteristics (Azimlu et al., 2021).

Compared to other countries, the Philippines has not been the subject of such experiments – a hedonic model for house prices affected by COVID-19 infected individuals (Abellana and Devaraj, 2021), and an analysis of determinants of land values in Cebu City (Agosto, 2017) only provide limited insight in a Philippine context. Thus, the opportunity of using Machine Learning for property valuation beckons in developing countries like the Philippines, as it may help address the shortcomings of traditional approaches currently being utilized. In combination with using alternative data sources, the outputs of objective ML-based valuations may allow for more accurate and explainable conclusions for buyers, sellers, and appraisers to interpret (Angrick, et al., 2022; van der Hoeven, 2022; Joy, 2021; Rico-Juan and de La Paz, 2021), which is arguably of higher importance in developing countries such as the Philippines.

This paper aims to evaluate the effectiveness of utilizing commonly used ML techniques for the valuation of properties in the Philippines. In addition, as some conventional indicators such as zonal values are not readily available, the paper also seeks to verify the usefulness of alternative data not commonly used by Philippine appraisers and real estate agents. This includes geolocation data sourced from free mapping initiatives like OpenStreetMap, as well as other socio-economic indicators obtained from the Philippine Statistics Authority (PSA), BIR, and other government resources.

The research questions the paper seeks to address are as follows:

- RQ1: Are commonly used ML techniques found in similar property prediction publications also effective under a Philippine context?
- RQ2: Does incorporating socio-economic indicators and geolocation data provide predictive power in the estimation of property prices in areas from the Philippines?
- RQ3: Will the use of indicators measured by government entities have a substantial effect in increasing model performance related to machine learning-based property valuations?
- RQ4: What is the benefit of a segmented approach and personalized ML models for property valuations?

2. MATERIALS AND METHODS

2.1. Data Sources

Four main data sources were used for the study, namely:

- Property listings from Lamudi, a popular online estate listing marketplace in the Philippines
- Department of Trade and Industry's Cities and Municipalities Competitive Index (CMCI)
- Amenities and buildings listed in OpenStreetMap.
- Selected socio-economic datasets developed by the Philippine Statistics Authority (PSA)

The study consists of information gathered for two primary locations in the Philippines: the province of Cavite in Region IV-A, and Metro Manila, also known as the National Capital Region. These locations were chosen mainly due to their prevalence in Lamudi, a popular real estate listing website in the Philippines.

The models aim to predict the average price per square meter of a property utilizing a combination of factors sourced from these data sources. This will be derived by dividing the given price with the property's given floor area, to lessen the variation of performance metric outputs such as Mean Absolute Error (Mean AE) and Mean Absolute Percentage Error (MAPE), as prices of different properties in the Philippines do tend to flare up to huge numbers. While price alone is commonly used in a variety of property valuation papers with ML approaches, price per square meter is an alternative target variable used by other experiments (Gao, et al., 2022; Xiao and Yan, 2019; Ahlfeldt, 2013; Sommervoll and Sommervoll, 2018) and is also commonly used in appraisal of mass real estate (Antipov and Pokryshevskaya, 2012; Thanasi, 2016; Beimer and Francke, 2019; Hau, 2020).

Lamudi

Lamudi-based property listings from Cavite and Metro Manila were collected via web-scraping in Python. Only houses were considered and scraped, as other house types such as apartments, condominiums, and lots may be characterized or evaluated differently. While Lamudi is considered as a premier online marketplace in the country (Primer, 2021; Similarweb, 2022; Camella, 2022), the limitations of the scraped data include slight inaccuracies with the coordinates, unlisted amenities and furnishments, and distributions skewed to higher-end real estate developers. It is assumed that the scraped data reflected the current state of the housing market during the second half of 2022. The variables extracted from the Lamudi site can be found in Table 1. All prices listed in Lamudi are in Philippine Peso.

Table 1. Variables from Lamudi

Variable Group	**Description**	**Features**
Location	Features pertaining to spatial characteristics of the property	Longitude, latitude, postcode, LGU, region, subdivision
Amenity	Amenities found within the property and its vicinity	# of AC units, balconies, decks, fences, fireplaces, fitness centers, garages, gates, grass areas, libraries / bookstores, airports, parking lots, meeting rooms, parks, pools, basketball courts, tennis courts, volleyball courts, warehouses. If the property had security, was smoke-free, or was fully or partially furnished
Property Specification	Includes features detailing a property's structural specifications	# of bedrooms, # of bathrooms, floor area (m^2), land size (m^2), total rooms, property classification, car spaces
Price	Market price of property	Price, Agency Name
Total Number of Variables		**39**

Cities and Municipalities Competitive Index

The Department of Industry and Trade (DTI)'s CMCI, developed by the National Competitiveness Council, is an annual ranking of the competitiveness of all provinces, cities, and municipalities in the Philippines (Department of Trade and Industry, n.d.). The overall competitiveness of an LGU every year is composed of four (4) main pillars of equal weights, namely: Economic Dynamism, Government Efficiency, Infrastructure, and Resiliency. A list of sub-indicators per pillar, each with their own score, is added to create the pillar's final score. Ranks of pillars and sub-indicators were provided. The 2021 rankings of LGUs from Cavite and Metro Manila were scraped from the site; LGU ranks instead of base scores were utilized for the models. Figure 1 exhibits pillar and sub-indicator scores of Pasig City, an LGU in Metro Manila, while Table 2 details the features extracted from the CMCI website.

Figure 1. Sample 2021 CMCI Score and Ranking of Pasig City (Department of Trade and Industry, n.d.)

Table 2. Variables from DTI's CMCI 2021

Variable Group	Description	Features
Pillar Indicators	Ranks scores for the four key indicators	Economic Dynamism, Government Efficiency, Infrastructure, Resiliency
Economic Dynamism	Rank scores for Economic Dynamism sub-indicators	Size of the Local Economy, Growth of the Local Economy, Capacity to Generate Employment, Cost of Living, Cost of Doing Business, Financial Deepening, Productivity, Presence of Business and Professional Organizations
Government Efficiency	Rank scores for Government Efficiency sub-indicators	Capacity of Health Services, Capacity of Schools, Security, Business Registration Efficiency, Compliance to BPLS standards, Presence of Investment Promotions Unit, Compliance to National Directives for LGUs, Ratio of LGU collected tax to LGU revenues, Most Competitive LGU awardee, Social Protection
Infrastructure	Rank scores for Infrastructure sub-indicators	Existing Road Network, Distance from City/Municipality Center to Major Ports, DOT-Accredited Accommodations, Availability of Basic Utilities, Annual Investments in Infrastructure, Connection of ICT, Number of Public Transportation Vehicles, Health Infrastructure, Education Infrastructure, Number of ATMs
Resiliency	Rank scores for Resiliency sub-indicators	Land Use Plan, Disaster Risk Reduction Plan, Annual Disaster Drill, Early Warning System, Budget for DRRMP, Local Risk Assessments, Emergency Infrastructure, Utilities, Employed Population, Sanitary System
Total Number of Variables		**42**

OpenStreetMap

OpenStreetMap (OSM) was used to scrape and count varying types of amenities and buildings within a walking distance of 500 meters, 1, 3, and 5 kilometers away from each Lamudi-scraped property. The locations of interests as found in Table 3 were determined based on indicators indicative of property prices in other publications (Agosto, 2017; Chen et al., 2020; Nguyen and Cripps, 2001; Huang et al., 2017; The Danh Phan, 2018; Wittowsky et al., 2020; van der Hoeven, 2022).

Table 3. Variables from OpenStreetMap

Variable Group	Description	Features
Neighborhood Amenities	Count of amenities within walking distance of 0.5, 1, 3, and 5 kilometers (km)	# of Cafés, Fast Food, Pubs, Restaurants, Colleges, Kindergarten Facilities, Schools, Universities, Gas Stations, Parking Areas, ATMs, Banks, Clinics, Hospitals, Pharmacies, Police Stations, Townhalls, Marketplaces
Neighborhood Buildings	Count of buildings within walking distance of 0.5, 1, 3, and 5 kilometers (km)	# of Residentials, Commercials, Industrials, Retail Stores, Supermarket, Fire Stations, Government Buildings
Public Transportation	Count of public transportation spots within walking distance of 0.5, 1, 3, and 5 kilometers (km)	# of Platforms, Stations, and Stop Positions
Shops	Count of shops within walking distance of 0.5, 1, 3, and 5 kilometers (km)	# of Convenience Stores, Malls, Dry Cleaning areas, and Laundry areas
Tourist Attractions	Count of tourist attractions within walking distance of 0.5, 1, 3, 5 kilometers (km)	# of Hostels, Hotels, Motels, and Viewpoints
Total Number of Variables		**148**

Philippine Statistics Authority

The study also leverages on the use of statistical data published by the Philippine Statistics Authority (PSA). These datasets include the LGU's income class, type, annual regular income, total capital expenditures, total social services expenditures, poverty level estimates, population, and population growth in five (5) and ten (10) years (National Quickstat for 2022, n.d.; Census of Population and Housing, n.d.; Statistics, n.d.). The variables used are found in Table 4. The expenditures and income variables are in Philippine Peso.

Table 4. LGU Socio-Economic Variables

Variable Group	Description	Features
LGU Expenditures and Income	Data regarding the LGUs' annual regular income and capital expenditures	Total capital expenditures (2021), Total social services expenditures (2021), Annual regular income (2021)
Population and Population Growth	Data regarding the LGU's population and growth rate in 5 and 10 years	LGU 2022 population, 5- and 10-year population growth rate
Poverty Indicators	Data regarding the poverty incidence of the different LGUs	LGU Poverty Incidence Rate (2021), LGU Subsistence Rate (2021)
Total Number of Variables		**8**

2.2. Methodology

The study made use of the Python programming language to conduct the data collection, processing, analysis, modeling, and evaluation. Datasets were collected and stored in CSV format. Python libraries used include but are not limited to Scikit-learn, Pandas, Numpy, OSMnx, Seaborn, Matplotlib, Yellowbrick, and Geopandas in data wrangling and modelling. Figure 2 shows an overview of the approach used throughout the study.

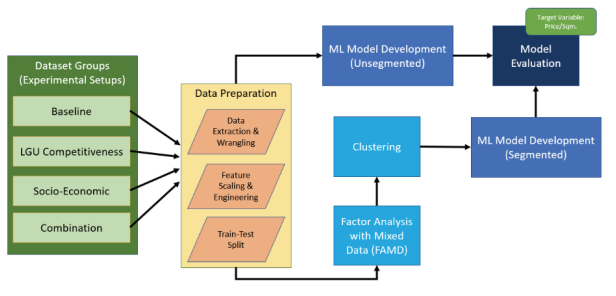


Figure 2. Summary of model development.

2.2.1. Data Extraction and Preprocessing

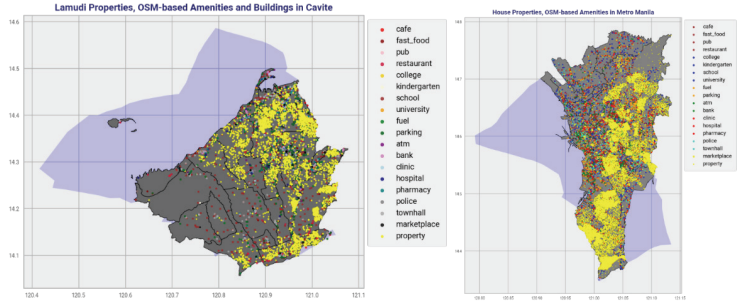
Data for property listings in Lamudi were collected via the BeautifulSoup library in Python. Search pages for 'House and Lot for Sale' were filtered to LGUs within Cavite and Metro Manila. Duplicates on base price, relative location, and other property specifications were removed for a final total of 3,212 and 9,657 houses for the two locations, respectively. Some amenity objects mentioned in the specific listing pages were removed due to repetitiveness or specificity. To create the target variable, the original price of each property was divided by its floor area. The original price variable was removed during modelling, and the base datasets of Cavite and Metro Manila were divided into 80/20 train-test splits, also preserving this ratio for each LGU.

Neighborhood features were extracted via the OSMnx library to query information in the OpenStreetMap database. The two areas of focus were set as input locations wherein all amenities and buildings were extracted. These were then overlayed with the collected property listings to which the walking distance to the amenities and buildings were computed. Counting of values was done and summarized within the vicinity of 0.5, 1, 3 and 5 kilometers via the K-Nearest Neighbors algorithm. Note that properties near the boundaries of the said areas that neighbored other land areas may lack true counts of amenities and buildings.

Reverse geocoding was done via the GeoPy Python library to extract postcode and regional data. In the Philippines, some LGUs have multiple postcodes designated to specific areas. This convention is particularly effective in segmenting areas that may have different demographics and jurisdiction of local authorities. From the dataset, this convention was considered through the aggregation of postcodes to its LGU. Additionally, Philippine Statistics Authority data which did not have 2021 nor 2022 values were forecasted using Holt-Winters' method (Chatfield, 1978), due to a lack of observed trends or seasonal variations.

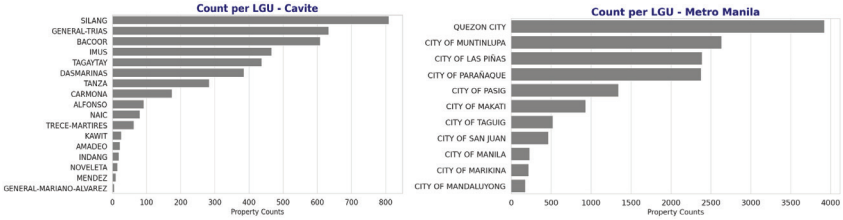
2.2.2. Exploratory Data Analysis

Exploratory data analysis was performed separately on the two locations to gain further understanding on the properties scraped in Lamudi and identify potential features to be removed before the modelling phase. The analysis was done on the whole datasets before splitting. Figures 3a and 3b show plots of Cavite and Metro Manila with the Lamudi-scraped property listings and the OSM-extracted amenities and buildings in those locations. Denoted on the two maps are the Lamudi locations in yellow.



Figures 3a and 3b: Visualizations of OSM-extracted amenities and buildings with property listings in Cavite (a) and Metro Manila (b); the legend denotes the different entities on the map.

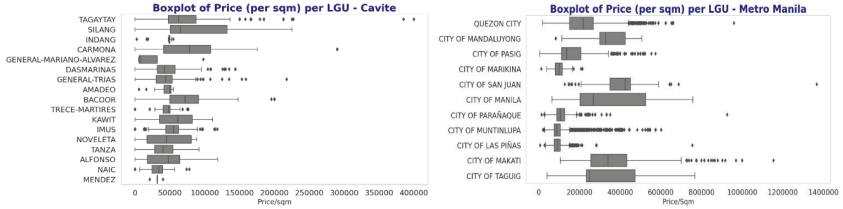
The Province of Cavite contains 16 municipalities and 7 cities; of those, 17 LGUs are present in the Lamudi-scraped dataset. The majority of the 3,212 houses in the province are found in Silang, General Trias, Bacoor, Imus, and Tagaytay, which are either component cities or highly populated municipalities. Metro Manila contains 16 cities and 1 municipality; of those only 11 cities are present. The majority of the 9,657 houses are found in Quezon City, with Muntinlupa and Las Pinas also having a sizable number of real estate properties.



Figures 4a and 4b. Counts of houses scraped in (a) Cavite and (b) Metro Manila

The average prices per LGU in both areas do follow a similar trend in Cavite. The LGUs of Silang, Carmona, Bacoor, and Tagaytay have the most expensive houses on average. Component cities which do not feature as highly in Figure 6a, such as General Trias and Dasmarinas, do have many outlier houses, which may skew modelling results.

In Metro Manila, the cities of San Juan, Makati, and Quezon City contain more expensive houses. This could be attributed to the presence of business districts and commercial areas in some barangays within these LGUs. Generally, houses from Metro Manila are more expensive than Cavite in terms of the target variable.

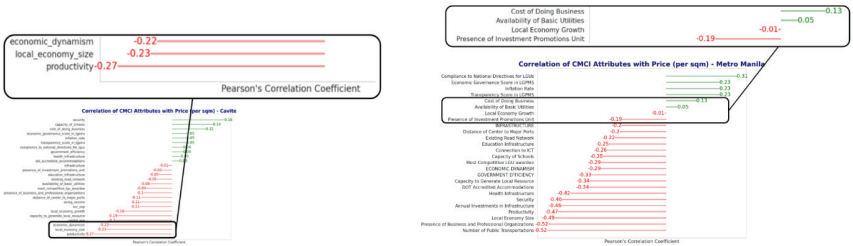


Figures 5a and 5b. Boxplots of Lamudi-Scraped House Prices per LGU in (a) Cavite and (b) Metro Manila

2.2.2.1. CMCI Pillars and Sub-Indicators

To recall, ranks of each LGU were scraped and provided for modelling purposes – as such, the lower the rank an LGU has for a pillar or sub-indicator, the better that LGU performs in that aspect. With this, correlation analysis was performed on both locations.

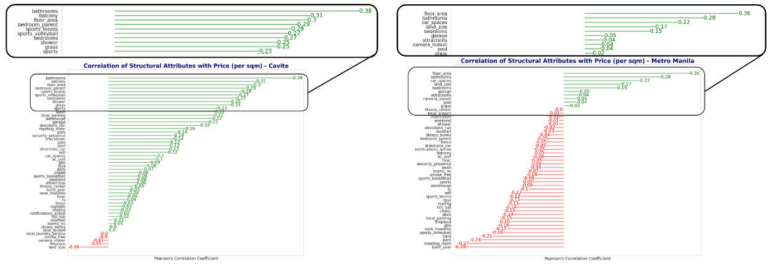
In Figure 6a, it can be seen that 'Productivity', 'Local Economy Growth', and 'Local Economy Size' are slightly important indicators in increasing the target variable in Cavite. In Figure 6b, it can be inferred that infrastructure and economy-related indicators such as "Number of Public Transportations", "Presence of Business and Professional Organizations", and "Local Economy Size" have higher correlation in Metro Manila.



Figures 6a and 6b. Correlation of 2021 CMCI Sub-Indicators to Price/Sqm. in (a) Cavite and (b) Metro Manila

2.2.2.2. Lamudi Amenities and Structural Attributes

In Cavite, some structural attributes and Lamudi-based amenities play bigger roles in influencing the price/sq. meter of a real estate property. Standard indicators such as '# of Bathrooms" and "Floor Area" positively influence the target variable. It can also be seen that the presence of balconies, parent bedrooms, sports areas, and grass areas are considered more important structural attributes or surrounding amenities, as found in Figure 7a.



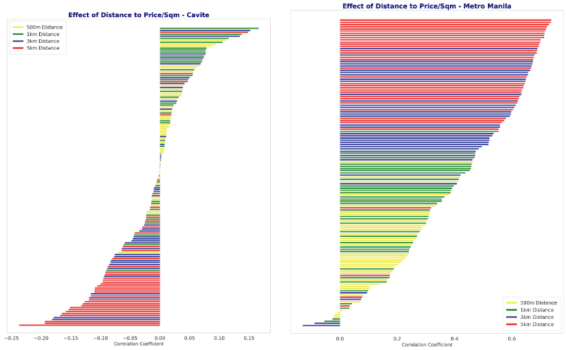
Figures 7a and 7b. Correlation of Structural Attributes to Price/Sqm of Real Estate Properties in (a) Cavite and (b) Metro Manila

For Metro Manila, structural attributes do still positively correlate, but not at the same intensity as with Cavite. Alternatively, it can also be seen that the presence of other nearby amenities negatively correlates; these could be attributed in relation with other more positive variables, such as those found in geospatial attributes in Figure 8b.

2.2.2.3. Geospatial Attributes from OSM

The characteristics of correlations of the OSM-based geospatial attributes to the target variable greatly differ from Cavite and Metro Manila, as found in Figures 8a and 8b. In Cavite, the presence of many amenities relatively far from houses, such as 5 kilometers, negatively correlate with the target variable.

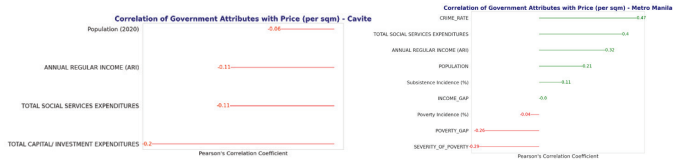
Compared to Cavite, geospatial attributes in Metro Manila are more positively correlated with the target variable, evidenced by coefficients reaching a 0.5-0.6 threshold. Distance between amenities means much more in this highly urbanized context, with 3-km and 5-km variables dominating both ends of the correlation spectrum. An example of this would be banks within 3-km and 5-km are relatively more correlated than other geospatial variables; the presence of banks or other similar financial institutions are likely well placed as to cater houses with inhabitants of higher income or who are more likely to transact.



Figures 8a and 8b. Correlation Analysis on OSM-based Geospatial Attributes to Price/Sq. Meters of Real Estate Properties in (a) Cavite and (b) Metro Manila, grouped by distance

2.2.2.4. Socio-Economic Indicators from other Government Sources

Socio-economic indicators from other government sources do not seem to have much correlation with the target variable. With both locations as found in Figures 9a and 9b, increasing poverty incidence slightly negatively correlates with average price/sq. meter; lower prices may be positioned to entice buyers who may not be able to afford more expensive homes. Population in Cavite seems to lower average prices, which may also be correlated with the annual regular income – the higher the population a Cavite LGU has, the lower their average income may be. Social services expenditures have the most positive correlation with the target variable.



Figures 9a and 9b. Correlation Analysis on Socio-Economic Government Attributes to Price/Sq. Meters of Real Estate Properties in (a) Cavite and (b) Metro Manila

2.2.3. Experimental Setup for Machine Learning Models

The study implemented two main experimental setups: a non-segmented approach where all houses under a single location were considered in training and testing models, and a segmented approach where unsupervised learning techniques were utilized to segment houses with similar characteristics. Separate baseline models with commonly used features in other publications were set up for the two investigated locations. The baselines were compared with models with combinations of different data. Model performances of clusters in each segmented experiment were evaluated against the baselines and non-segmented counterparts.

Table 5. Experimental Setups

Approach	Experiment	Datasets Used
Non-Segmented	Baseline	Lamudi + OSM
	LGU Competitiveness	Lamudi + OSM + CMCI
	Socio-Economic	Lamudi + OSM + Government
	Combination	Lamudi + OSM + CMCI + Government
Segmented	Segmented Baseline	Lamudi + OSM
	LGU Competitiveness	Lamudi + OSM + CMCI
	Socio-Economic	Lamudi + OSM + Government
	Combination	Lamudi + OSM + CMCI + Government

As previously mentioned, the base datasets of Cavite and Metro Manila were divided into 80/20 train-test splits, also preserving this ratio for each LGU. These were utilized throughout all iterations of the two experimental setups.

2.2.4. Feature Design and Selection

A summary of variables extracted and engineered, as detailed in Tables 1-4, can be found in Table 6. Their data sources would be the basis of differentiating the experiments mentioned in Table 5.

Table 6. Variables considered in the study.

Variable	Data Source
Location, Amenities, Property Specification	Lamudi
Pillar Indicators, Economic Dynamism Sub-Indicators, Government Efficiency Sub-Indicators, Infrastructure Sub-Indicators, Resiliency Sub-Indicators	CMCI
Location, Neighborhood Amenities, Neighborhood Buildings, Shops, Public Transportation, Tourist Attractions	OSM
LGU Expenditures and Income, Population and Population Growth, Poverty Indicators	PSA
Target Variable – Price/sqm.	**Lamudi**

As 230+ variables were initially available, a set of feature selection processes were conducted before modelling to improve model performance. Pearson and Spearman correlation analysis were done to identify numerical variables with no correlation, which were either dropped or kept note of during modelling. One hot and ordinal encoding was performed on categorical variables. As doing so would further increase the dimensionality of the inputs, variance thresholding and mutual information regression methods were used to decrease the final number of columns used for the machine learning models. Multicollinearity checks via variance inflation factor and a homoscedasticity test were also performed to identify removable or transformable variables. To prepare for property segmentation, a Factor Analysis of Mixed Data (FAMD) method was used to create principal components usable for clustering, as a variety of qualitative and quantitative features were present. Guided by rules of thumb (Cangelosi and Goriely, 2007), a 90% variance threshold was used to determine the optimal number of principal components.

2.2.5. Machine Learning Modeling

For each experiment, a comparative analysis was conducted on sets of tree-based machine learning models. Two clustering algorithms were utilized for property segmentation. A summary of the model development and comparison of segmented and non-segmented approaches is found in Figure 10, while the list of models used for prediction and clustering is found in Table 7.

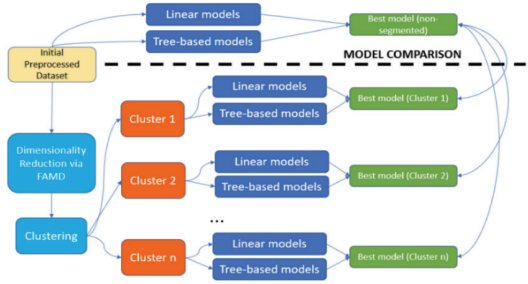


Figure 10. Model development and comparison of segmented and non-segmented approaches.

Table 7. Machine learning models utilized by the study.

Model Group	Model Name
Tree-Based Machine Learning Models	(1) Decision Tree Regressor, (2) AdaBoost estimator on Decision Tree Regressor, (3) Gradient Boosting Machine Regressor, (4) Random Forest Regressor, (5) Extremely Randomized Trees Regressor, (6) Bagging estimator on Support Vector Regression, (7) Stacking Regressor, (8) XGBoost Regressor, (9) LightGBM Regressor
Clustering Algorithms	(1) K-Means, (2) Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH)

Machine Learning Models

A set of tree-based machine learning algorithms were utilized to train property prediction models. Tree-based models utilize tree-like structures for deciding target variable classes or values and may be useful when input and output variables do not exhibit linear relationships. These were preferred over linear models due to the high dimensionality and complexity of the data setups, and over neural networks to provide a higher layer of intrinsic explainability. While tree-based models are commonly used for classification problems, regression trees can obtain numerical values in their terminal nodes by selecting splits that minimize the sum of squared deviations from the mean. Ensemble methods can also be used to produce optimal predictive results through considering weighted scores from sets of weaker classifiers. The models in this study were picked mainly due to their prevalence in other ML-based price prediction papers. The selected models are a Decision Tree regressor, an AdaBoost estimator on a Decision Tree regressor (Freund and Robert, 1995), a Gradient Boosting Machine regressor (Freidman, 2001), a Random Forest regressor (Ho T. , 1995), an Extremely Randomized Trees regressor (Geurts et al., 2006), a Bagging estimator on a Support Vector Regression model, a Stacking regressor (Breiman, 1996), an XGBoost Regressor (Chen and Carlos, 2016), and a Light Gradient Boosting Machine regressor (Ke et al., 2017).

Assumptions for Tree-based Machine Learning Models and Clustering Algorithms

As compared to linear-based models, tree-based machine learning techniques are not held as strongly to assumptions of linearity and normality for outputs to be valid (Chowdhury et al., 2021). Despite this, four other common assumptions are discussed below:

1. Independence of errors: While Linear-based models assumes error independence, Tree-based models does not due to their reliance on feature values for data partitioning, rather than residuals (Chowdhury et al., 2021).
2. Homoscedasticity: Tree-based methods are less constrained compared to linear regression regarding uniform residual variance across independent variable levels (Chowdhury et al., 2021).
3. Normality of residuals: This is not a strict requirement for ensemble and tree-based models such as Random Forests and Gradient Boosting Machines (GBMs). Tree-based models do not assume normality of residuals, as they do not involve estimating coefficients that require such assumptions (Das, 2019).
4. Multicollinearity: Tree-based models handle this differently due to their ability to accommodate non-linearity, inherent feature selection, and decisions based on individual variables. (Mane, 2021).

Apart from these assumptions, other factors like outliers and sampling bias must be considered. Tree-based models exhibit resilience to outliers and to address sampling bias, utilizing machine learning evaluations like cross-validation can help understand the generalization capabilities of these models (Kotu and Deshpande, 2019).

Clustering Algorithms

After restructuring the datasets using FAMD, a set of unsupervised techniques - K-Means, a centroid-based algorithm portioning *n* observations into *k* clusters (Likas et al., 2003), and Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH), a hierarchical-based algorithm effective over large datasets (Zhang et al., 1996) - were used to create property segmentations with which individual models could be tested. These two techniques were used and are deemed suitable due to the nature of the data: having high dimensionality, mixed types of features, or not linearly separable (Rokach and Maimon, 2005; Kaushik and Mathur, 2014).

2.3. Model Evaluation

Each data setup was split into an 80% training set and 20% test set. Five-fold Grid Search Cross Validation was performed on the training set to derive best performing hyperparameters and reduce sampling error. These were then applied on the held-out test set. Performance metrics used in regression models were used. Mean Absolute Percentage Error (MAPE) measures the accuracy of regression methods by computing the mean ratio between the actual value A_t and forecasted value, F_t. As it is usually used for measuring regression model quality (de Myttenaere, Golden et al., 2016; McKenzie, 2011), it will likely be a better indication of true model performance as compared to Mean Absolute Error (Mean AE), which outputs the mean of taking the absolute differences between A_t and F_t. . However, the Mean AE and Median Absolute Error (Median AE), which gives the median of the same procedure, will give more practical interpretations when read through the units of measurement used: Philippine Peso / square meters. The R^2 score, while not sufficient alone in judging a model's regression performance nor for non-linear relationships (Dunn, 2021), can be utilized for comparisons between other similar publications.

Separate metrics were used for determining the optimal number of clusters used for property segmentation. The inertia or sum-of-squares error method measures the squared average distance between all cluster centroids (Chavent, 1998) was the primary metric observed, while the Calinski-Harabasz index and Silhouette Score were also measured (Calinski and Harabasz, 1974; Wang and Xu, 2019; Rousseeuw, 1987). Utilizing these techniques with the elbow method, it was found that having four clusters was optimal for both Cavite and Metro Manila contexts.

3. RESULTS AND DISCUSSION

Sections 3.1 and 3.2 capture the results and findings of our study. Section 3.3 discusses the insights of this study in relation to the four research questions introduced in *Section 1*.

3.1. Results from Non-Segmented (NS) Approach

For the non-segmented set-up, the study aimed to identify the best performing model when applied with incremental additions of data. The results were summarized in Table 8 wherein the best performing models per data setup according to MAPE for Cavite and Metro Manila were highlighted in green. The best model for each area and data setup according to MAPE is marked in green, while the second and third best models are in blue. The averages and standard deviation of the performance metrics from the Top 3 Models of each data set up ranked by MAPE were taken. A similar procedure was performed for the rest of the models. As observed in Table 8, the AdaBoost algorithm performed most reliably as compared to other algorithms.

In Cavite, the AdaBoost algorithm consistently performed the best out of all algorithms and all data setups, achieving MAPE values of 22-23%. In Metro Manila, AdaBoost and GBM algorithms performed the best across the four data set ups, achieving MAPE values of 20.2-20.4%. The additional information added by features from LGU Competitiveness, Socio-Economic and a combination thereof do not notably improve the baseline performance. This finding means that a generalized approach or

the data used are not sufficiently discriminating. With this, the segmented approach becomes more interesting (see the corresponding results in Section 3.2). A post-model analysis was implemented on the best-performing models of each experimental setup, as seen in Figures 11a, 11b, 12a, and 12b. It was observed that the results from the measured metrics may have been skewed by a few significantly misvalued properties. While most houses have good predictions, as verified by their close distances to the generated red line, a few have predictions misclassified as far as Php 75,000-120,000/sqm.

Feature importance of the best-performing non-segmented models for Cavite and Metro Manila are shown in Tables 9 and 10. Property specification variables from Lamudi and different OSM-engineered features dominate the influence of the two models. In general, government-based variables do not account for much importance in both best-performing models. In checking the Top 5 features of each data source on the best models as seen in Tables 11 and 12, it is observed that PSA or Socio-Economic variables are not influential in the model's results. Other features which may act as proxies to other variables, such as proximity to airports or commercial areas for noise pollution and traffic, do not appear in these tables and are not considered as influential variables.

Figures 11a and 11b. Predicted vs Actual Plots of Best-Performing NS Model – Cavite (AdaBoost – LGU Competitiveness)

Figures 12a and 12b. Predicted vs Actual Plots of Best-Performing NS Model – Metro Manila (GBM – Socio-Economic)

Table 8. Non-Segmented Approach – Summary of Results. The best model for each area and data setup according to MAPE is marked in green, while the 2nd and 3rd best models are in blue. The averages and standard deviation of the performance metrics from the Top 3 Models (and lesser models) of each data set up ranked by MAPE were taken.

Non-Segmented Approach			Cavite				Metro Manila			
Data Setup	Row #	Algorithm	MAPE (%)	Mean AE	Median AE	R^2 Score	MAPE (%)	Mean AE	Median AE	R^2 Score
Baseline (Lamudi + OSM)	1	Decision Tree	31.3932%	18473.16	9550.68	0.1085	28.5033%	37902.90	13117.44	0.6162
	2	AdaBoost	23.4270%	13363.16	7601.62	0.6220	21.2203%	28096.27	11037.02	0.8116
	3	GBM	24.0699%	13629.85	7554.50	0.5742	20.3580%	26845.53	9491.957	0.8220
	4	RF	25.8864%	14517.09	9481.17	0.5961	27.1850%	36087.52	11891.42	0.6653
	5	ERT	31.9112%	19594.83	13331.45	0.1974	21.8022%	27393.35	10629.89	0.8287
	6	Bagging	28.2325%	17448.15	8720.93	0.2390	25.6182%	34632.80	15328.59	0.7283
	7	Stacking	25.6754%	15268.52	9466.53	0.5435	24.7852%	31529.04	12784.01	0.7992
	8	XGBoost	24.0923%	13548.69	8095.92	0.6344	21.6611%	27980.58	10833.34	0.8054
	9	LightGBM	24.8399%	13880.40	8491.49	0.6002	22.6544%	27695.55	12727.03	0.8362
	10	Top 3 Models Avg ± Std	23.8631% ± 0.3778	13513.90 ± 136.70	7750.68 ± 299.91	0.6102 ± 0.0317	21.0798% ± 0.6628	27640.79 ± 691.14	10454.11 ± 839.45	0.8130 ± 0.0084
	11	Lesser Models Avg ± Std	27.9898% ± 3.0545	16530.36 ± 2309.78	9840.38 ± 1767.08	0.3808 ± 0.2230	25.0914% ± 2.5750	32540.19 ± 4397.82	12746.40 ± 1550.34	0.7457 ± 0.0911
LGU Competitiveness (Lamudi + OSM + CMCI)	12	Decision Tree	30.1962%	18423.71	9350.65	0.1892	28.9858%	39392.90	13111.11	0.6024
	13	AdaBoost	22.3423%	12700.51	7416.78	0.6520	20.2280%	26909.87	10625.00	0.8223
	14	GBM	23.8282%	13491.50	7197.34	0.5766	20.5629%	27116.26	9656.17	0.8208
	15	RF	28.2900%	17141.40	8217.32	0.2484	27.8690%	36972.21	13147.90	0.6529
	16	ERT	26.3722%	14714.55	9239.74	0.5870	22.2869%	27519.28	11344.70	0.8292
	17	Bagging	31.2647%	19483.11	13047.56	0.2082	27.4421%	36771.51	15281.27	0.6945
	18	Stacking	26.9910%	16192.43	9473.21	0.4617	22.7913%	28969.61	12301.70	0.8156
	19	XGBoost	23.9271%	13547.27	8223.60	0.6263	21.3950%	28053.72	10412.50	0.8071
	20	LightGBM	25.2141%	14465.46	9056.11	0.5738	22.2822%	27875.28	12726.49	0.8318
	21	Top 3 Models Avg ± Std	23.3659% ± 0.8905	13246.43 ± 473.60	7612.57 ± 540.41	0.6183 ± 0.0380	20.7286% ± 0.6009	27359.95 ± 609.92	10231.22 ± 509.22	0.8167 ± 0.0084
	22	Lesser Models Avg ± Std	28.0547% ± 2.3236	16736.78 ± 2004.81	9730.77 ± 1685.15	0.3800 ± 0.1845	25.2762% ± 3.1385	32916.80 ± 5354.91	12985.53 ± 1306.87	0.7377 ± 0.1007
Socio-Economic (Lamudi + OSM + Government)	23	Decision Tree	31.9300%	19428.38	9644.99	0.0711	28.8959%	38611.74	13508.44	0.5999
	24	AdaBoost	23.2356%	13288.70	7677.78	0.6126	21.1155%	28006.97	10538.75	0.8013
	25	GBM	23.8339%	13528.47	6868.24	0.5721	20.1791%	26422.00	9346.681	0.8277
	26	RF	28.6685%	17385.97	8213.20	0.2225	26.5142%	35164.16	11054.84	0.6855
	27	ERT	26.4510%	14688.26	9350.15	0.5942	21.5955%	26903.95	10258.52	0.8296
	28	Bagging	31.1146%	19468.33	13296.62	0.2063	25.5614%	34498.34	15286.59	0.7299
	29	Stacking	25.8116%	15487.65	9121.74	0.5173	24.6730%	30959.00	12585.32	0.8103
	30	XGBoost	23.9880%	13688.26	8689.69	0.6268	22.0971%	28592.94	10858.43	0.7987
	31	LightGBM	25.0320%	14222.15	8565.38	0.5750	22.3367%	27529.25	12971.47	0.8412
	32	Top 3 Models Avg ± Std	23.6858% ± 0.3950	13501.81 ± 201.10	7745.24 ± 912.59	0.6038 ± 0.0283	20.9634% ± 0.7204	27110.97 ± 812.51	10047.98 ± 623.30	0.8195 ± 0.0158
	33	Lesser Models Avg ±Std	28.1680% ± 2.8781	16780.12 ± 2332.22	9698.68 ± 1838.11	0.3600 ± 0.2244	25.0131% ± 2.5850	32559.24 ± 4260.79	12710.85 ± 1644.88	0.7443 ± 0.0908
Combination (Lamudi + OSM + CMCI + Government)	34	Decision Tree	30.7643%	17899.24	9337.73	0.2592	28.4003%	38546.25	13272.60	0.6031
	35	AdaBoost	22.7354%	13266.14	7704.27	0.5956	20.4441%	26830.61	10186.51	0.8213
	36	GBM	23.2044%	12996.69	6842.83	0.6316	20.9442%	27144.96	9872.93	0.8200
	37	RF	28.3535%	17042.06	8893.61	0.3827	27.3649%	36487.06	11831.94	0.6609
	38	ERT	26.2374%	14626.69	9245.87	0.5931	21.6713%	27513.14	10634.38	0.8256
	39	Bagging	30.8683%	19369.48	12629.50	0.2237	25.4799%	34503.06	15152.51	0.7317
	40	Stacking	25.2055%	15494.37	9676.71	0.5172	21.9530%	28702.07	12185.88	0.8219
	41	XGBoost	23.7940%	13636.47	8790.02	0.6367	21.4470%	27633.21	10880.01	0.8142
	42	LightGBM	24.4304%	14054.80	8724.32	0.5895	22.5955%	27932.34	12880.44	0.8375
	43	Top 3 Models Avg ± Std	23.3659% ± 0.5263	13299.77 ± 321.21	7779.04 ± 975.74	0.6213 ± 0.0224	20.9451% ± 0.5015	27202.93 ± 404.43	10313.15 ± 515.35	0.8185 ± 0.0038
	44	Lesser Models Avg ± Std	27.6432% ± 2.7871	16414.44 ± 2046.52	9751.29 ± 1449.58	0.4300 ± 0.1634	24.5775% ± 2.9141	32280.65 ± 4823.57	12659.63 ± 1526.44	0.7468 ± 0.0983

Table 9. Top 15 Important Features for Best-Performing NS Model – Cavite (AdaBoost – Combination)

Feature	Category	Feature Importance
Floor Area	Lamudi – Property Specification	0.26535
Land Size	Lamudi – Property Specification	0.21464
No. of Public Transportations (Rank)	CMCI – Infrastructure	0.02659
# of ATMs within 5km	OSM – Neighborhood Amenities	0.01921
# of Viewpoints within 3km	OSM – Tourist Attractions	0.01826
# of Gas Stations within 5km	OSM – Neighborhood Amenities	0.01549
# of Hotels within 3km	OSM – Tourist Attractions	0.01490
# of Residential Areas within 5km	OSM – Neighborhood Buildings	0.01302
# of Bedrooms	Lamudi – Property Specification	0.01189
Infrastructure (Rank)	CMCI – Pillar Indicators	0.01064
Availability of Basic Utilities (Rank)	CMCI – Infrastructure	0.01019
# of Bathrooms	Lamudi – Property Specifications	0.00974
# of Residential Areas within 1km	OSM – Neighborhood Buildings	0.00912
# of Kindergarten Schools within 5km	CMCI – Neighborhood Amenities	0.00910

Table 10. Top 15 Important Features for Best-Performing NS Model – Metro Manila (GBM – Socio-Economic)

Feature	Category	Feature Importance
LGU – Makati	Lamudi – Location	0.22987
Poverty Incidence Rate	PSA – Poverty Indicators	0.22761
Land Size	Lamudi – Property Specification	0.19053
Floor Area	Lamudi – Property Specification	0.05669
# of Malls within 5km	OSM – Shops	0.01167
# of Stations within 5km	OSM – Public Transportation	0.01166
# of Bedrooms	Lamudi – Property Specification	0.00874
# of Retail Stores within 3km	OSM – Neighborhood Buildings	0.00832
Agent Name – GREAT SUCCESS REALTY	Lamudi – Price	0.00786
# of Industrial Areas within 5km	OSM – Neighborhood Buildings	0.00599
# of Apartments within 3km	OSM – Neighborhood Amenities	0.00566
# of ATMs within 3km	OSM – Neighborhood Amenities	0.00551
# of Marketplaces within 5km	OSM – Neighborhood Amenities	0.00521
# of Car Spaces	Lamudi - Amenity	0.00491

Table 11. Top 5 Features per Data Source for Best-Performing NS Model – Cavite (AdaBoost – Combination)

Lamudi			**CMCI**		
Feature	**Rank**	**Importance**	**Feature**	**Rank**	**Importance**
Floor Area	1	0.26535	No. of Public Transportations	3	0.02659
Land Size	2	0.21464	Infrastructure	10	0.01064
# of Bedrooms	9	0.01189	Availability of Basic Utilities	11	0.01019
# of Bathrooms	13	0.00974	Presence of Investment Promo Unit	16	0.00873
Subdivision Name – NO DATA	20	0.00753	Government Efficiency	17	0.00817
OSM			**PSA**		
Feature	**Rank**	**Importance**	**Feature**	**Rank**	**Importance**
# of ATMs within 5km	4	0.01921	5-year Growth	101	0.00140
# of Viewpoints within 3km	5	0.01826	Annual Regular Income (2021)	173	0.00045
# of Gas Stations within 5km	6	0.01549	Poverty Incidence Rate (2021)	189	0.00034
# of Hotels within 3km	7	0.01490	Social Expenditure (2021)	199	0.00027
# of Residential Areas within 5km	8	0.01302	10-year Growth	200	0.00026

Table 12. Top 5 Features per Data Source for Best-Performing NS Model – Metro Manila (GBM–Socio-Economic)

Lamudi			**CMCI**		
Feature	**Rank**	**Importance**	**Feature**	**Rank**	**Importance**
LGU – Makati	1	0.22987	N/A		
Land Size	3	0.19053	N/A		
Floor Area	4	0.05669	N/A		
# of Bedrooms	7	0.00874	N/A		
Agent Name – GREAT SUCCESS REALTY	9	0.00786	N/A		
OSM			**PSA**		
Feature	**Rank**	**Importance**	**Feature**	**Rank**	**Importance**
# of Malls within 5km	5	0.01167	Poverty Incidence Rate	2	0.22761
# of Stations within 5km	6	0.01166	10-year Growth	27	0.00298
# of Retail Stores within 3km	8	0.00832	Annual Regular Income (2021)	95	0.00119
# of Industrial Areas within 5km	10	0.00599	Population (2022)	99	0.00109
# of Apartments within 3km	11	0.00566	Capital Expenditure (2021)	183	0.00015

3.2. Results from Segmented Approach

Building on the initial approach, it was found that segmentation generally provided substantial reduction in minimizing the MAPE. Comparing the best-performing non-segmented models in Table 8 to the best-performing segmented models per cluster in Table 13, this is particularly noticeable in the data setups wherein the city competitiveness index, the socio-economic variables, or a combination of both were used. Highlighted in blue in Table 13 are clusters which performed better than their best non-segmented model counterpart of the same data setup, and highlighted in green are metrics which on average beat the same metric of the best non-segmented model counterpart of the same data setup.

For both locations, only the cluster averages for MAPE and Mean AE for Cavite's LGU Competitiveness data setup, MAPE, Mean AE, and R^2 score for Cavite's Socio-Economic data setup were better than the averages of the best non-segmented and top 3 non-segmented counterparts, as shown in rows 5, 12, 19, and 26 of Table 8. However, as found in Table 14, Clusters 2, 3, and 4 of both Cavite and Metro Manila had models that were much more effective than the best overall non-segmented model of that area (see bold MAPE results). Cavite's models show extra promise, as all their best data setups per model per cluster utilized data setups that combined additional government data. The prevalence of Baseline models in Metro Manila may suggest that further granularity in data sources should be considered during model development.

In general, some of the errors in model performances could be attributed to the lack of more granular data utilized outside of OSM-based data sources. Despite Cavite and Metro Manila being in urbanized settings, other factors such as zonal values and average income may highly vary across barangays and populated areas within a single LGU. For example, images from a 2020 article discussing geospatial divides in Metro Manila display the disproportions visually, as found in differences in Eastwood and Santolan (Figure 13a, Marikina and eastern Quezon City), Pembo, Brgy. Rizal, and Bonifacio Global City (Figure 13b, Taguig and Makati), and Culiat and Lower Puroks 4-8 (Figure 13c, Central Quezon City) (Commoner, 2020).

Recent and trustworthy socio-economic data which may capture these disparities, such as the CMCI and PSA-based datasets, are only readily available at an LGU level, explaining the less-than-optimal results of the models in the non-segmented approach. Clustering allowed the models to further differentiate between certain commercial, residential, high-income, and low-income areas. Houses that were less expensive and had common attributes may have been grouped together, enabling performances that recorded MAPEs as low as 10.7%.

Cluster characteristics were verified with the same feature importance method used in the non-segmented approach. This was done on two Cavite clusters with best-performing models coming from the same experimental setup (LGU Competitiveness), and the overall best-performing model of Metro Manila, as seen in Tables 15, 16, and 17. Highlighted in orange in these tables are features which do not appear on other tables. As expected, the top 15 features vary on each case – while some property specification features remained present, a diverse collection of features from Lamudi and OSM were observed. Cavite's Cluster 1 seems to be influenced by residential areas nearing essentials such as education, convenience stores, and public transportation. Cavite's Cluster 2 is more influenced by property amenities such as pools, security, and number of rooms, which suggests that these are for ones in closed-off subdivisions. Metro Manila's Cluster 3 is significantly different from the Cavite clusters, as noted with the presence of malls, parking spaces, hotels and other commercial amenities deemed influential.

Figures 13a, 13b, and 13c. Satellite images of certain residential and commercial areas across Metro Manila (Commoner, 2020)

Table 13. Segmented Approach – Summary of Results. Highlighted in blue are clusters which performed better than their best non-segmented model counterpart of the same data setup, while those highlighted in green are metrics which on average beat the same metric of the best non-segmented model counterpart of the same data setup).

Segmented (vs. Non-Segmented)			Cavite					Metro Manila				
Data Setup	**Row #**	**Model Type**	**Best Model**	**MAPE (%)**	**Mean AE**	**Median AE**	**R^2 Score**	**Best Model**	**MAPE (%)**	**Mean AE**	**Median AE**	**R^2 Score**
Baseline (Lamudi + OSM)	1	Cluster 1	AdaBoost	23.3823%	10519.56	6167.51	0.4587	LGBM	32.0401%	125675.60	93270.19	0.3163
	2	Cluster 2	XGBoost	29.0398%	21871.93	14683.53	0.5379	GBM	**18.2149%**	19275.53	8154.00	0.6797
	3	Cluster 3	Stacking	34.3261%	22125.20	13385.91	0.4489	AdaBoost	**10.6869%**	8796.90	4706.73	0.7529
	4	Cluster 4	Stacking	23.7349%	15699.12	10797.54	0.4724	GBM	23.6141%	26285.66	13860.68	0.8000
	5	Cluster Avg.		27.6208%	17553.95	11258.62	0.4795		21.1390%	45008.42	29997.90	0.6372
	6	Best Non-Segmented	AdaBoost	23.4270%	13363.16	7601.62	0.6220	GBM	20.3580%	26845.53	9491.96	0.8220
	7	(Avg. ± Std). Top 3 Non-Segmented		23.8631% ± 0.0038	13513.90 ± 136.71	7750.68 ± 299.91	0.6102 ± 0.032		21.0798% ± 0.0066	27640.79 ± 691.14	10454.11 ± 839.45	0.813 ± 0.01
LGU Competitivenesss (Lamudi + OSM + CMCI)	8	Cluster 1	LGBM	**21.0005%**	9740.08	6654.05	0.5572	LGBM	35.3714%	121447.30	98464.56	0.4030
	9	Cluster 2	AdaBoost	**11.5174%**	6771.86	4296.18	0.7300	ERT	24.6956%	28709.10	14116.17	0.6755
	10	Cluster 3	AdaBoost	29.2997%	16965.63	10714.29	0.6176	AdaBoost	16.8305%	15834.53	7361.96	0.7706
	11	Cluster 4	XGBoost	21.5638%	13615.95	9718.539	0.5966	AdaBoost	**12.1389%**	9430.27	4705.50	0.5357
	12	Cluster Avg.		20.8454%	11773.38	7845.76	0.6254		22.2591%	43855.30	31162.05	0.5962
	13	Best Non-Segmented	AdaBoost	22.3423%	12700.51	7416.78	0.6520	AdaBoost	20.2280%	26909.87	10625.00	0.8223
	14	(Avg. ± Std). Top 3 Non-Segmented		23.37% ± 0.0089	13246.43 ± 473.60	7612.57 ± 540.42	0.6183 ± 0.038		20.7286% ± 0.0060	27359.9 ± 609.62	10231.22 ± 509.22)	0.8167 ± 0.001
Socio-Economic (Lamudi + OSM + Government)	15	Cluster 1	GBM	22.3434%	13603.46	6938.58	0.7308	AdaBoost	36.1526%	126216.90	98986.22	0.3436
	16	Cluster 2	LGBM	26.3540%	16628.93	11304.08	0.7716	GBM	23.4519%	27402.41	13449.32	0.7640
	17	Cluster 3	GBM	**15.9263%**	9988.85	7063.91	0.8732	AdaBoost	18.1130%	16778.28	8482.99	0.7601
	18	Cluster 4	GBM	23.0537%	16500.47	11391.67	0.7444	GBM	**12.7322%**	9849.31	4899.36	0.5292
	19	Cluster Avg.		21.9170%	14180.43	9174.56	0.7800		22.6124%	45061.73	31454.47	0.5992
	20	Best Non-Segmented	AdaBoost	23.2356%	13288.70	7677.78	0.6126	GBM	20.1791%	26422.00	9346.68	0.8277
	21	(Avg. ± Std). Top 3 Non-Segmented		23.6858% ± 0.0040	13501.81 ± 201.11	7745.24 ± 912.60	0.6038 ± 0.028		20.9634% ± 0.0072	27110.97 ± 812.51	10047.98 ± 623.30	0.8195 ± 0.016
Combination (Lamudi + OSM + CMCI + Government)	22	Cluster 1	GBM	27.7516%	17529.84	12639.45	0.6556	GBM	20.7717%	2050.76	8311.44	0.7491
	23	Cluster 2	AdaBoost	22.8916%	10313.00	7013.11	0.5241	AdaBoost	23.4643%	28389.32	14827.96	0.7711
	24	Cluster 3	ERT	25.6070%	20814.21	7420.38	0.4267	ERT	**10.7307%**	7441.242	3731.788	0.7513
	25	Cluster 4	XGBoost	**14.1173%**	7087.45	4453.50	0.5157	AdaBoost	31.8970%	126304.60	102344.60	0.3071
	26	Cluster Avg.		27.7516%	17529.84	12639.45	0.5305		21.7159%	41046.48	32303.95	0.6447
	27	Best Non-Segmented	AdaBoost	22.7354%	13266.14	7704.27	0.5956	AdaBoost	20.4441%	26830.61	10186.51	0.8213
	28	(Avg. ± Std). Top 3 Non-Segmented		23.2446% ± 0.0053	13299.77 ± 321.21	7779.04 ± 975.75	0.6213 ± 0.022		20.9451% ± 0.0050	27202.9 ± 404.43	10313.15 ± 515.35	0.8185 ± 0.004

Table 14. Comparison of Best Models per Cluster across all Data Setups to Best Non-Segmented Model.

Area	**Cavite**		**Metro Manila**	
Best Overall Non-Segmented Model	LGU Competitiveness – AdaBoost – 22.3423% MAPE, Php 12770.51/sqm Mean AE, Php 7416.78/sqm Median AE		Socio-Economic – GBM – 20.1791% MAPE, Php 26422.00/sqm Mean AE, Php 9436.68/sqm Median AE	
Cluster	**Best Data Setup + Model**	**Better than Overall Non-Segmented Model (MAPE)?**	**Best Data Setup + Model**	**Better than Overall Non-Segmented Model (MAPE)?**
Cluster 1	LGU Competitiveness – LGBM – 21.0005% MAPE	Yes	Combination – GBM – 20.7717% MAPE	No
Cluster 2	LGU Competitiveness – AdaBoost – 11.5174% MAPE	Yes	Baseline – GBM – 18.2149% MAPE	Yes
Cluster 3	Socio-Economic – GBM – 15.9263% MAPE	Yes	Baseline – AdaBoost – 10.6869% MAPE	Yes
Cluster 4	Combination – XGBoost – 14.1173% MAPE	Yes	LGU Competitiveness – XGBoost – 12.1389% MAPE	Yes

Table 15. Top 15 Features for Best-Performing Seg. Model (Cluster 1) – Cavite (LGBM – LGU Competitiveness).

Feature	**Category**	**Feature Importance**
Land Size	Lamudi – Property Specification	0.07754
Floor Area	Lamudi – Property Specification	0.06952
# of Residential Areas within 3km	OSM – Neighborhood Buildings	0.02373
# of Residential Areas within 5km	OSM – Neighborhood Buildings	0.02306
# of Industrial Areas within 5km	OSM – Neighborhood Buildings	0.02139
# of Convenience Stores within 3km	OSM – Shops	0.01972
# of Schools within 5km	OSM – Neighborhood Amenities	0.01705
# of Residential Areas within 1km	OSM – Neighborhood Buildings	0.01504
# of Industrial Areas within 3km	OSM – Neighborhood Buildings	0.01370
# of Convenience Stores within 5km	OSM – Shops	0.01370
# of Terminal Platforms within 3km	OSM – Public Transportation	0.01337
# of Commercial Areas within 3km	OSM – Neighborhood Buildings	0.01303
# of Bedrooms	Lamudi – Property Specification	0.01270
# of Schools within 3km	OSM – Neighborhood Amenities	0.01237

Table 16. Top 15 Features for Best-Performing Seg. Model (Cluster 2) – Cavite (AdaBoost – LGU Competitiveness)

Feature	Category	Feature Importance
Floor Area	Lamudi – Property Specification	0.12658
Land Size	Lamudi – Property Specification	0.11849
# of Schools within 5km	OSM – Neighborhood Amenities	0.02860
# of Convenience Stores within 3km	OSM – Shops	0.02838
# of Pools in Vicinity	Lamudi - Amenities	0.02705
# of Bedrooms	Lamudi – Property Specification	0.01720
Presence of Security	Lamudi – Amenities	0.01406
# of Residential Areas within 5km	OSM – Neighborhood Buildings	0.01343
# of Bathrooms	Lamudi – Property Specification	0.01310
# of Industrial Areas within 5km	OSM – Neighborhood Buildings	0.01188
# of Residential Areas within 3km	OSM – Neighborhood Buildings	0.01182
# of Schools within 3km	OSM – Neighborhood Amenities	0.01123
# of Warehouses in Vicinity	Lamudi – Amenities	0.01111
Presence of Smoke-Free Areas	Lamudi – Amenities	0.01102

Table 17. Top 15 Features for Best-Performing Seg. Model (Cluster 3) – Metro Manila (AdaBoost – Baseline).

Feature	Category	Feature Importance
Land Size	Lamudi – Property Specification	0.14087
# of Townhalls within 3km	OSM – Neighborhood Buildings	0.06625
Floor Area	Lamudi – Property Specification	0.04983
# of Hotels within 5km	OSM – Tourist Attractions	0.04590
# of Bathrooms	Lamudi – Property Specification	0.04540
# of Apartments within 3km	OSM – Neighborhood Buildings	0.03716
# of Hotels within 3km	OSM – Tourist Attractions	0.03652
# of Government Buildings within 0.5km	OSM – Neighborhood Buildings	0.03236
# of Schools within 5km	OSM – Neighborhood Amenities	0.02962
# of Car Spaces	Lamudi – Property Specification	0.02822
# of Malls within 3km	OSM – Shops	0.02286
# of Apartments within 5km	OSM – Neighborhood Buildings	0.02123
# of Government Buildings within 5km	OSM – Neighborhood Buildings	0.01908
# of Fast Food Restaurants within 3km	OSM – Neighborhood Amenities	0.01795

3.3. Discussion of Research Questions

In the following, our findings are discussed in relation to the four research questions mentioned in *Section 1*.

RQ1: Are commonly used ML techniques found in similar property prediction publications also effective under a Philippine context?

Our findings indicate that utilizing ML techniques under a Philippine context can be similarly effective when including alternative data such as socioeconomic, geospatial, and government-based indicators. As the papers found over the globe differ in currency and in temporal valuation, the models' performances were compared in terms of MAPE and R^2 score. Within the region, regression models on Kuala Lumpur had MAPEs ranging from 11.3-20.9% and R^2 scores from 0.74-0.91 (McCluskey et al., 2014). A Hong Kong study utilizing three ML algorithms outputted MAPEs ranging from 32-54%, but higher R^2 scores of 0.83-0.90 (Ho et al., 2020). Other studies on Shanghai and Xi'an utilize other performance metrics to highlight their results, but provide best R^2 scores of 0.70 and 0.89, respectively (Xue et al., 2020). Outside Asia, studies mostly provide RMSE or Mean AE as primary metrics of comparison, which made R^2 scores as an unreliable basis. A study in Santiago, Chile tested ML-powered models which had scores ranging from 0.74-0.96 (Masias et al., 2016). A Dortmund study which utilized OLS and Spatial Lag models had adjusted R^2 scores of 0.35-0.60 (Wittowsky et al., 2020). A spatial analysis on London real estate prices achieved an R^2 score of 0.7116 (Santos and Jiang, 2020). The MAPEs in the study at hand range from 11.5-21% (Cavite) and 10.7-20.7% (Metro Manila)

and the R^2 scores are within 0.65-0.87 (Cavite) and 0.53-0.82 (Metro Manila). This seems to indicate that also for the Philippines the gap in data quality and availability can – at least to a certain degree - be overcome with Machine Learning algorithms and alternative data sources.

RQ2: Does incorporating socio-economic indicators and geolocation data provide predictive power in the estimation of property prices in areas from the Philippines?

From the highly varying prediction results across property clusters, it was seen that for some clusters, socio-economic indicators and geolocation data are highly effective for a good model performance. For Cavite, Cluster #3 can achieve a MAPE of 15.9% whereas other clusters report MAPEs of at least 22.3%. For Metro Manila, Cluster #4 reports a MAPE of 12.7% versus the performance of at least 18.1% for other clusters. These results suggest that a differentiated look and approach are needed per property cluster. In the Philippines, a personalized approach appears to be beneficial where some clusters profit more from socioeconomic and geolocation data compared to other clusters.

RQ3: Will the use of indicators measured by government entities have a substantial effect in increasing model performance related to machine learning-based property valuations?

On average, the data setups which included government-based data (i.e., CMCI and socio-economic datasets) beat their baseline models. The implications mentioned in the discussion of RQ2 become similarly evident for RQ3. For Cavite, government data focusing on competitiveness are highly beneficial for Cluster #2 (MAPE of 11.5%), government data focusing on socio-economic indicators are beneficial for Cluster #3 as discussed in RQ2, and the combination of the two types of government data becomes strikingly effective for Cluster #4 (MAPE of 14.1%). For Metro Manila, the combination approach becomes effective for Cluster #1 (MAPE of 20.8%) and the data setup focusing on competitiveness showcases improved performance for Cluster #4 (MAPE of 12.1%). These findings emphasize the need for a personalized approach to different property clusters, where different types of government data play a varying role for the property value prediction performance.

RQ4: What is the benefit of a segmented approach and personalized ML models for property valuations?

For both cities, there is added value in a personalized approach where individual ML models are developed for different property segments. Different segments benefit from different types of data, suggesting that the segments themselves differentiate to a sufficiently large degree that for some the competitiveness aspect becomes invaluable where for others it is the socio-economic characteristics or also the combination of both (such as Table 13, rows #3, #9, #11, #17, #18, #24, #25). The results also suggest that for some segments, the additional alternative data introduces noise and even decreases model performance (such as Table 13, for Cavite, row #9 vs. #23 for Cluster #2 a MAPE of 11.5% vs. 22.9%, respectively). Overall, 87.5% of all properties in the test set (100% of Cavite properties and 75% of Metro Manila properties) benefited from a personalized segment-based approach.

4. CONCLUSION

4.1. Summary

The effectiveness of utilizing alternative data not commonly used by similar publications nor Philippine appraisers was evaluated in the context of property valuation. To improve the performance, alternative data was paired with typically high performing ML techniques for the valuation of properties in the Philippines.

The study considered two experimental set-ups consisting of a variety of ML models and combinations of data sources as inputs – a non-segmented approach considering all house listings during modelling, and a segmented approach where data points were clustered according to their structural attributes. Tree-based machine learning methods were compared in finding the best models for each setup. The data sources included geolocation data from OpenStreetMap (OSM), rankings from Department of

Trade and Industry's Cities and Municipalities Competitiveness Index (CMCI), and other socio-economic indicators obtained from the Philippine Statistics Authority (PSA), BIR, and other government resources. House property listings from Cavite and Metro Manila were scraped from a popular Philippine real estate listing site and were used in separate models.

A MAPE range of 10.7-21% has been obtained in this study which is competitive against Kuala Lumpur ML models with 11.3-20.9% MAPE and beat Hong Kong ML models of 32-54% MAPE. For the non-segmented approach, experimental data setups were able to beat the general performance of baseline Lamudi models. The best performing Cavite model utilizing additional OSM and CMCI features resulted in a 22.34% MAPE, Php 12,700/sqm Mean AE, and Php 7,416/sqm Median AE. The best performing Metro Manila model utilizing additional OSM and socio-economic features resulted in a 20.18% MAPE, Php 26,422/sqm Mean AE, and a Php 9,346/sqm Median AE. In general, boosting algorithms such as AdaBoost and Gradient Boosting Machine perform better with the high-dimensionality datasets. Feature importance analysis showed that while property specification variables are still important, geographically-engineered and government-based variables did significantly improve model performance.

Additional experiments utilizing clustering techniques for property segmentation showcased the benefit of personalized approaches (model + data types) per property cluster and the importance of available government data. Overall, 87.5% of properties benefited from a personalized segment-based approach. For Cavite, government data focusing on competitiveness are highly beneficial for Cluster #2 (MAPE of 11.5%), socio-economic indicators are valuable for Cluster #3 (MAPE of 15.9%), and with a MAPE of 14.1%, Cluster #4 surpasses other clusters by combining the two types of government data. For Metro Manila, the combination approach is effective for Cluster #1 (MAPE of 20.8%), while utilizing socio-economic features greatly benefits Cluster #4 (MAPE of 12.1%).

4.2. Recommendations

This paper hopes to spark discussion and further research on more objective and transparent approaches regarding Philippine property valuation, and push for updated and readily available data for appraisers, home buyers, and home sellers to utilize during their property valuation process in the country. To achieve performances that can match the standards needed for operationalization, the following are recommended for further study: comparison with spatial econometric models; finer granularity and availability of other government-based indicators, such as those in the barangay level and other types of indicators; the usage of mapping initiatives with more documentation of amenities and buildings, such as a Google Maps API; further hyperparameter tuning and usage of deep learning techniques; personalized approaches for property segments in different regions across the country; usage of satellite imagery for capturing in detail the granular features of a location; and an increase in and variety of data points found in the specified locations. In addition, while other environmental indicators such as noise pollution can be proxied by points of interest (such as proximity to airports and commercial areas), additional work to acquire direct indicators will only benefit the performance and explainability of these models.

5. LITERATURE CITED

Abellana, J. A., and Devaraj, M. (2021). Hedonic Modeling for Predicting House Prices during CoVid19 Pandemic in the Philippines. *3rd International Conference on Management Science and Industrial Engineering* (pp. 21-26). Association for Computing Machinery.

Adetiloye, K., and Eke, P. (2014). A Review of Real Estate Valuation and Optimal Pricing Techniques. *Asian Economic and Financial Review*, 1878-1893.

Agosto, A. B. (2017). Determinants of Land Values in Cebu City, Philippines. *International Conference on Business and Economy.*

Ahlfeldt, G. M. (2013). If We Build it, Will They Pay? Predicting Property Price Effects of Transport Innovations. *Environment and Planning A: Economy and Space, 45(8)*, 1977-1994.

Angrick, S., Bals, B., Niko, H., Kleissl, M., Schmidt, J., Vanja, D., . . . Friedrich, T. (2022). Towards Explainable Real Estate Valuation via Evolutionary Algorithms. *Genetic and Evolutionary Computation Conference* (pp. 1130-1138). Association of Computing Machinery.

Antipov, E. A., and Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications, 39(2)*, 1772-1778.

Azimlu, F., Rahnamayan, S., and Makrehchi, M. (2021). House price prediction using clustering and genetic programming along with conducting a comparative study. *Genetic and Evolutionary Computation Conference Companion*, (pp. 1809-1816).

Baldominos, A., Blanco, I., Moreno, A., Iturrarte, R., Bernardez, O., and Afonso, C. (2018). Identifying Real Estate Opportunities Using Machine Learning. *Applied Sciences, 8(11)*.

Beimer, J., and Francke, M. (2019). Out-of-Sample House Price Prediction by Hedonic Price Models and Machine Learning Algorithms. *Real Estate Research Quarterly, 18(2)*, 13-20.

Board of Valuers, Appraisers, Estate Agents and Property Managers. (2019). *Malaysian Valuation Standards, 6th ed.*

Breiman, L. (1996). Stacked Regressions. *Machine Learning, 24*, 49-64.

Bureau of Local Government Finance. (2018). *Philippine Valuation Standards.*

Buyukkaracigan, N. (2021). *Modern Methods Approach in Real Estate Valuation.* Iksad Publications.

Calinski, T., and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics, 3(1)*, 1-27.

Camella. (2022, March 2). *The Best Real Estate Websites In The Philippines.* Retrieved from Camella PH: https://www.camella.com.ph/real-estate-websites-philippines/

Cangelosi, R., and Goriely, A. (2007). Component retention in principal component analysis with application to cDNA microarray data. *Biology Direct*.

Cellmer, R., Cichulska, A., and Belej, M. (2020). Spatial Analysis of Housing Prices and Market Activity with the Geographically Weighted Regression. *International Journal of Geo-Information*.

Census of Population and Housing. (n.d.). Retrieved from Philippine Statistics Authority: https://psa.gov.ph/population-and-housing

Chaphalkar, N., and Sandbhor, S. (2013). Use of Artificial Intelligence in Real Property Valuation. *International Journal of Engineering and Technology (IJET)*, 2334-2337.

Chatfield, C. (1978). The Holt-Winters Forecasting Procedure. *Applied Statistics, 27(3)*, 264-279.

Chavent, M. (1998). A monothetic clustering method. *Pattern Recognition Letters, 19(11)*, 989-996.

Chen, S., Zhuang, D., and Zhang, H. (2020). GIS-Based Spatial Autocorrelation Analysis of Housing Prices Oriented towards a View of Spatiotemporal Homogeneity and Nonstationarity: A Case Study of Guangzhou, China. *Complexity*.

Chen, T., and Carlos, G. (2016). XGBoost: A Scalable Tree Boosting System. *International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). New York: Association of Computing Machinery.

Chou, J.-S., Fleshman, D.-B., and Truong, D.-N. (2022). Comparison of machine learning models to provide preliminary forecasts of real estate prices. *Journal of Housing and the Built Environment*.

Chowdhury, S., Lin, Y., and Kerby, L. (2021). Evaluation of Tree Based Regression over Multiple Linear Regression for Non-normally Distributed Data in Battery Performance. 16.

Commoner. (2020, June 24). *The Divide in Our Cities.* Retrieved from Medium: https://mediacommoner.medium.com/the-divide-in-our-cities-bff743e1584

Congress of the Philippines. (2017). *Republic Act No. 10963.*

Das, P. (2019, March 4). *https://www.researchgate.net/post/Is_normality_checking_of_residual_important_for_machine_learning*. Retrieved from ResearchGate: https://www.researchgate.net/post/Is_normality_checking_of_residual_important_for_machine_learning

de Myttenaere, A., Golden, B., Le Grand, B., and Rossi, F. (2016). Mean Absolute Percentage Error for regression models. *Neurocomputing, 192*, 38-48.

Department of Trade and Industry. (n.d.). Retrieved from Cities and Municipalities Competitiveness Index: https://cmci.dti.gov.ph/

Dickinson, C. (2021, February 19). *Inside the 'Wikipedia of Maps,' Tensions Grow Over Corporate Influence.* Retrieved from Bloomberg: https://www.bloomberg.com/news/articles/2021-02-19/openstreetmap-charts-a-controversial-new-direction

Domingo, E. V., and Fulleros, R. F. (2002). *Real estate price index: a model for the Philippines.* Bank of International Settlements.

Dunn, K. (2021, April 6). *Avoid R-squared to judge regression model performance.* Retrieved from Towards Data Science: https://towardsdatascience.com/avoid-r-squared-to-judge-regression-model-performance-5c2bc53c8e2e

Evans, K., Lausberg, C., and Sui Sang How, J. (2019). Reducing Property Appraisal Bias with Decision Support Systems: An Experimental Investigation in the South African Property Market. *Journal of African Real Estate Research, 4(1)*, 108-138.

Freidman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29(5)*, 1189-1232.

Freund, Y., and Robert, S. E. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. *European Conference on Computational Learning Theory* (pp. 23-37). Berlin: Springer.

Gao, G., Bao, Z., Cao, J., Oin, A., and Sellis, T. (2022). Location-Centered House Price Prediction: A Multi-Task Learning Approach. *ACM Transactions on Intelligent Systems and Technology, 13(2)*, 1-25.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning, 63*, 3-42.

Hau, K.-C. (2020). House Prices in the Peripheries of Mass Rapid Transit Stations Using the Contingent Valuation Method. *Sustainability, 12(20)*.

Ho, T. (1995). Random Decision Forests. *3rd International Conference on Document Analaysis and Recognition*, (pp. 278-282). Montreal.

Ho, W. K., Tang, B.-S., and Wong, S. (2020). Predicting property prices with machine learning algorithms. *Journal of Property Research*, 1-23.

Howard, C. (2004). *Is There Assessor Bias in the Real Estate Market?* Illinois Wesleyan University.

Huang, Z., Chen, R., Xu, D., and Zhou, W. (2017). Spatial and hedonic analysis of housing prices in Shanghai. *Habitat International, 67*, 69-78.

Joy, C. (2021, May 1). *Explainable AI for Property Valuation.* Retrieved from Medium: https://medium.com/clear-capital-engineering/explainable-ai-for-property-valuation-cc110381106b

Kaushik, M., and Mathur, B. (2014). Comparative study of K-means and hierarchical clustering techniques. *International Journal of Software and Hardware Research in Engineering*, 93-98.

Ke, G., Meng, Q., Finley, T., Wang, T., Chan, W., Ma, W., . . . Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting. *International Conference on Neural Information Processing Systems* (pp. 3149-3157). New York: Association of Computing Machinery.

Kershaw, P., and Rossini, P. (1999). Using Neural Networks to Estimate Constant Quality. *Pacific-Rim Real Estate Society Conference.*

Kotu, V., and Deshpande, B. (2019). *Data Science (Second Edition).* Morgan Kaufmann.

Krzystanek, M., Lasota, T., and Trawinski, B. (2009). Comparative Analysis of Evolutionary Fuzzy Models for Premises Valuation Using KEEL. *International Conference on Computational Collective Intelligence*, (pp. 838-849).

Lech, M., and Gerald, L. (2018). *Current issues of the Philippine land use planning and management system.* German Institute for Development Evaluation.

Likas, A., Vlassis, N., and Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern Recognition, 36(2)*, 451-461.

Lughofer, E., Trawinski, B., Trawinski, K., and Lasota, T. (2011). On-Line Valuation of Residential Premises with Evolving Fuzzy Models. *International Conference on Hybrid Artificial Intelligence Systems*, (pp. 107-115).

Mandani Bay. (2018). *An expert's manual for real estate valuation in the philippines*. Retrieved from https://www.mandanibay.com/blog/experts-manual-real-estate-valuation-philippines/

Mane, P. (2021, September 30). *Multicollinearity in Tree Based Models*. Retrieved from Medium: https://medium.com/@manepriyanka48/multicollinearity-in-tree-based-models-b971292db140

Masias, V., Crespo, F., Valle, M., and Crespo, R. (2016). Property Valuation using Machine Learning Algorithms: A Study in a Metropolitan-Area of Chile. In C. Berger, *Lectures on Modelling and Simulation* (pp. 98-105). AMSE.

McCluskey, W. J., Daud, D., and Kamarudin, N. (2014). Boosted regression trees: An application for the mass appraisal of residential property in Malaysia. *Journal of Financial Management of Property and Construction, 19(2)*, 152-167.

McKenzie, J. (2011). Mean absolute percentage error and bias in economic forecasting. *Economics Letters, 113(3)*, 259-262.

Nallathiga, R., Upadhyay, A., Karmarkar, P., and Acharya, K. (2019). Tenure-Wise Determinants of Residential Property Value: An Application of Hedonic Pricing Model in Balewadi, Pune, India. *Theoretical and Empirical Researches in Urban Management, 14(4)*, 70-85.

Naqvi, S. (2017). *THE IMPACT OF MACROECONOMIC FACTORS ON THE REAL ESTATE PRICES IN USA.* North Carolina, Wilmington.

National Quickstat for 2022. (n.d.). Retrieved from Philippine Statistics Authority: https://psa.gov.ph/statistics/quickstat/national-quickstat/all/*

Nguyen, N., and Cripps, A. (2001). Predicting Housing Value: A Comparison of Multiple Regression Analysis and Artificial Neural Networks. *Journal of Real Estate Research, American Real Estate Society, 22(3)*, 313-336.

OpenStreetMap. (n.d.). *Map Features*. Retrieved from OpenStreetMap: https://wiki.openstreetmap.org/wiki/Map_features

Pi-ying, L. (2011). Analysis of the Mass Appraisal Model by Using Artificial Neural Network in Kaohsiung City. *Journal of Modern Accounting and Auditing, 7(10)*, 1080-1089.

Primer. (2021, June 19). *8 Real Estate Websites and Apps You Can Rely On in the Philippines*. Retrieved from Primer PH: https://primer.com.ph/tips-guides/2021/06/19/list-6-real-estate-websites-and-apps-you-can-rely-on/

Rico-Juan, J., and de La Paz, P. (2021). Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain. *Expert Systems with Applications, 171*.

Rokach, L., and Maimon, O. (2005). Clustering Methods. In L. Rokach, and O. Maimon, *Data Mining and Knowledge Discovery Handbook* (pp. 321-352). Boston, MA: Springer.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53-65.

Santos, L. J., and Jiang, R. (2020). *Spatial Analysis of House Price Determinants: A Greater London Case Study.* University College London.

Sario, M. S. (2019). A Spatial Econometric Model for Household Electricity Consumption in the Philippines. *14th National Convention on Statistics.*

Similarweb. (2022). *Top Real Estate Websites in Philippines Ranking Analysis for July 2022*. Retrieved from Similarweb: https://www.similarweb.com/top-websites/philippines/category/business-and-consumer-services/real-estate/

Sommervoll, D., and Sommervoll, A. (2018). *Learning from man or machine: Spatial aggregation and house price prediction.* Norwegian University of Life Sciences.

Statistics. (n.d.). Retrieved from Bureau of Local Government Finance: https://blgf.gov.ph/lgu-fiscal-data/

Tanrivermis, H. (2016). *Real Estate Valuation Principles.* Ankara.

Thanasi, M. (2016). Hedonic appraisal of apartments in Tirana. *International Journal of Housing Markets and Analysis, 9(2)*, 239-255.

The Danh Phan. (2018). Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia. *2018 International Conference on Machine Learning and Data Engineering (iCMLDE*, 35-42.

Tidwell, O., and Gallimore, P. (2014). The influence of a decision support tool on real estate valuations. *Journal of Property Research, 31(1)*, 45-63.

Unciano, R. C. (2020, February 25). *Reforming the real-property valuation system in the Philippines.* Retrieved from Business Mirror Philippines: https://businessmirror.com.ph/2020/02/25/reforming-the-real-property-valuation-system-in-the-philippines/

van der Hoeven, D. (2022). *Appraiser-based automated valuation: A case study of valuing buy-to-let properties in the Netherlands.* University of Groningen.

Wang, X., and Xu, Y. (2019). An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. *IOP Conference Series: Materials Science and Engineering.* IOP Publishing Ltd.

Wittowsky, D., Hoekveld, J., Welsch, J., and Steier, M. (2020). Residential housing prices: impact of housing characteristics, accessibility and neighbouring apartments – a case study of Dortmund, Germany. *Urban, Planning and Transport Research*, 44-70.

Xiao, L., and Yan, T. (2019). Prediction of House Price Based on RBF Neural Network Algorithms of Principal Component Analysis. *International Conference on Intelligent Informatics and Biomedical Sciences* (pp. 315-319). IEEE.

Xue, C., Ju , Y., Li , S., Zhou, Q., and Liu, Q. (2020). Research on Accurate House Price Analysis by Using GIS Technology and Transport Accessibility: A Case Study of Xi'an, China. *mdpi*.

Yiu, C., Tang, B., Chiang, Y., and Choy, L. T. (2006). Alternative Theories of Appraisal Bias. *Journal of Real Estate Literature, 14(3)*, 321-344.

Zhang, L., Zhou, J., Hui, E., and Wen, H. (2018). The effects of a shopping mall on housing prices: A case study in Hangzhou. *International Journal of Strategic Property Management*, 65-80.

Zhang, T., Ramakrishnan, R., and Livny, M. (1996). BIRCH: an efficient data clustering method for very large databases. *ACM SIGMOD Record, 25(2)*, 103-114.

Zhao, Y., Chetty, G., and Tran, D. (2019). Deep Learning with XGBoost for Real Estate Appraisal. *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, (pp. 1396-1401).