Residential Property Price Forecasting Model for Central Pangasinan

Fernando S. Viray, Jr

Abstract

This quantitative-experimental study aims to create a price forecasting model for the residential real properties of the fourteen municipalities and cities of central Pangasinan using supervised learning classification algorithms (linear regression and decision tree) to predict whether the real property’s value will increase or decrease in the future and classic statistical forecasting techniques (straight line, moving average, simple linear regression, and multiple linear regression) to predict the rate of increase or decrease using a ±5% margin of error. Data used were derived from the Residential Real Estate Price Index (RREPI) of the Banko Sentral ng Pilipinas (BSP) from 2016 to 2021, Zonal Valuations (ZV) from the Bureau of Internal Revenue (BIR) from 1990 to 2023, and the Housing Cost Construction Index (HCCI) from the Philippine Statistics Authority (PSA) from 2006 to 2021 following an 80:20 training-testing data split ratio. The resulting model utilizing the RandomForest algorithm shows a significant 93% accuracy rate and 93% precision rate. Results further showed that machine learning-based algorithms performed better than the four classic statistical forecasting techniques evaluated with multiple linear regression obtaining the lowest average prediction distance point of 12.46% as compared to Random Forest which achieved 4.32%.

Keywords: residential property price forecasting model, supervised learning, RandomForest

Introduction

Investing in residential real properties has been a surefire way of growing one’s equity be it among personal, business or corporate capitalization. Over the years, we have seen the long-term and consistent rise of real property values among rich nations and developing countries, hence, more and more individuals and corporations have considered real property investment to be always on top of their investment diversification portfolio. However, investing in residential real properties is not easy and initially requires sufficient residential pricing information in order for one’s capital to be invested securely and properly.

Residential property price information in the Philippines varies from one geopolitical subdivision of our government to another due to several factors. Currently, not one website, government

1 School of Economics and Management, Zhengzhou Normal University, Zhengzhou, Henan, 450044, China
Corresponding author: Xin Liu (LXPY20160808@zznu.edu.cn)
agency, information system or computer application can be found containing ALL key information about price information that is comprehensive, reliable and updated. One of the root causes of the lack of a centralized residential property price information in the Philippines is the departmentalized nature of the Philippine government functions based on the 1987 Constitution which mandates that the three branches of the government, namely, the executive, the legislative, and the judiciary, must co-exist with co-equal independence in their functions to meet their respective departmental objectives (Philippine Government Official Gazette, 2022). Likewise, the same constitution provides stature and recognition on the role of local government units. Hence, each branch of the government has its own established departments from national down to local municipal or city level (others down to barangay level), each with unique sphere of constituencies and stakeholders given their unique geopolitical subdivisions to manage (Dayanan, 2023).

Although the aim of the Philippine Constitution is for the departments and agencies to focus on providing quality services, the encompassing information which transcends between departments and government offices can be observed to have been departmentalized also instead of being centralized, thus, in order to get timely, reliable and sufficient public information, particularly the residential property price information, one needs to visit several bureaus and offices such as: the Banko Sentral ng Pilipinas (BSP) for the Residential Real Estate Price Index for the average price change of the different types of residential properties in the country (Banko Sentral ng Pilipinas, 2022); the Bureau of Internal Revenue (BIR) for Zonal Values of real properties per specific location (Department of Finance, 2021); the Philippine Statistics Agency (PSA) for the average housing cost index which provides the average cost per square meter of residential units by type and by location (Philippine Statistics Agency, 2022); the Securities and Exchange Commission (SEC), and Housing and Land Use Regulatory Board (HLURB) to check if a company is, indeed, registered and has valid permits to publicly offer real estate properties such as condominium units, subdivision lots and residential houses; and the local or provincial Land Registration Authority (LRA), Registry of Deeds (RD), local City or Municipal Assessor’s Office, Treasury Office and Mayor's Office, to check if an individual or entity has, indeed, the rights and possession of land and title (Conoza, 2022).

Another root case of the lack of centralized information for residential real property pricing information is the absence of a one-stop hub for residential property pricing information and real property transfer of registration. Although Republic Act No. 11032 which is commonly known as the Ease of Doing Business and Efficient Government Service Delivery Act of 2018 was signed into law that aims to reorganize and simplify the provision of government services to improve the establishment, registration and conduct of doing business in the country and lessen or totally irradicate Red Tape along the process (BusinessWorld, 2022), no one-stop hub was created solely where the public can get reliable, updated and consolidated information regarding residential property pricing, centralized and faster checking, registration and transfer of real properties.
Internal and external factors affecting the price of a residential property are also a major contributary issue to the root causes of the lack of a centralized residential property pricing information in the Philippines. There are several factors that are involved in residential property pricing which includes zonal values, housing construction material costs, labor and professional services costs, demand and availability of stock housing, population and inflation to name a few (Bual, 2023; Hoppler Editorial Board, 2018). With so many parameters applicable for different locations, it is certainly difficult to determine the actual residential property price especially when you have incomplete parameter information (Uy, 2021).

Meanwhile, in the sector of information and communications technology (ICT), several apps and online websites have been tapping the power of the Internet to provide residential property information among the netizens. However, these websites and apps offer only residential real property price information based on what the seller, landlord or lessor has intentionally indicated and not based on standardized pricing model. On the other hand, several recent studies conducted have shown that relevant real estate data can be aggregated to come up with a real property pricing model (Alam & Isikal, 2021; Kannan & Zhang, 2021; Chen, et al., 2022; and Lu & Zhang, 2022).

In October 2022, the group of Mora, Cespedes, and Perez conducted a research regarding housing price prediction using machine learning algorithms among 49,875 residential properties for sale, lease and rent in Alicante City, Spain. Mora, et al. (2022) utilized six ensemble learning algorithms namely gradient boosting, light gradient boosting, extreme gradient boosting, RandomForest and extra-trees regression in a Python program to predict housing price with 28 features. Mora and co-researchers found out that both gradient boosting and random forest algorithms performed better as compared to other algorithms when evaluated using error metrics mean absolute error (MAE), mean squared error (MSE), median absolute error (MedAE), and coefficient of determination (R2) (Mora, et al, 2022).

Meanwhile, a similar study was conducted by Ihre and Engström (2019) using Ames, Iowa housing dataset which is composed of 3,000 entries of houses with 80 variables. The researchers utilized k-Nearest Neighbor (k-NN) and Random Forests algorithm to predict future housing prices. Results revealed that RandomForest better at predicting future house prices than k-NN. The researchers also utilized the error metrics in evaluating the prediction results.

Given the current situation, it is the ultimate goal of this research to establish a real property pricing model out of available data from several related government agencies such as the Residential Real Estate Pricing Index (RREPI) from the Banko Sentral ng Pilipinas (BSP), Zonal Valuations (ZV) per political (barangay) subdivisions from the Department of Finance and Bureau of Internal Revenue, and the Housing Construction Cost Index (HCCI) from the Philippine Statistics Authority (PSA) for residential areas within the cities and municipalities of Central Pangasinan which includes Dagupan City, San Carlos City, Aguilar, Basista, Bayambang,
Binmaley, Bugallon, Calasiao, Lingayen, Malasiqui, Mangaldan, Mangatarem, Sta. Barbara, and Urbiztondo. Data mining techniques, classic statistical forecasting methods and machine learning classification methods will be employed to predict the future value of residential properties which can be deployed later via web or mobile apps for data consumption.

This research aims to benefit individuals or corporations eyeing to invest, sell, or acquire real properties within Central Pangasinan by establishing itself as a current or future residential property pricing information reference so that stakeholders will get a gist whether the price of residential property is within its pricing range, has a good future value on it, and will be a worthwhile investment.

**Methods**

**Research Design**

The Quantitative-Experimental method of research was utilized by the researcher for this study as machine learning uses quantitative research methods with experimental research design being the de facto research approach (Abu-El-Haija & Al-Khateeb, 2022).

**Data Collection and Pre-Processing**

The data for this study came from official government data releases which are readily downloadable from their respective official website. The data are divided into three sources: the Residential Real Estate Price Index (RREPI) of the Banko Sentral ng Pilipinas (BSP) from 2016 to 2021; the Zonal Valuations (ZV) from the Bureau of Internal Revenue (BIR) from 1990 to 2023 composed of 12 valuations per city/municipality; and the Housing Cost Construction Index (HCCI) from the Philippine Statistics Authority (PSA) based on its average construction cost per square meter based on approved building permits aggregated from 2006 to 2021.

The fourteen cities and municipalities of Central Pangasinan which this study covered include Dagupan City, San Carlos City, Aguilar, Basista, Bayambang, Binmaley, Bugallon, Calasiao, Lingayen, Malasiqui, Mangaldan, Mangatarem, Sta. Barbara, and Urbiztondo. Table 1 presents the distribution of collected data for each covered location.

**Table 1: Distribution of Collected Data per Covered Location**

<table>
<thead>
<tr>
<th>#</th>
<th>Location</th>
<th>Barangays</th>
<th>RREPI</th>
<th>ZV</th>
<th>HCCI</th>
<th>Total Data Rows</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>San Carlos</td>
<td>86</td>
<td>602</td>
<td>1,032</td>
<td>1,376</td>
<td>3,010</td>
</tr>
<tr>
<td></td>
<td>Dagupan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>City</td>
<td>31</td>
<td>217</td>
<td>372</td>
<td>496</td>
<td>1,085</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Aguilar</td>
<td></td>
<td>112</td>
<td>192</td>
<td>256</td>
<td>560</td>
</tr>
<tr>
<td></td>
<td>Basista</td>
<td>13</td>
<td></td>
<td>91</td>
<td></td>
<td>455</td>
</tr>
</tbody>
</table>
The researcher merged the three datasets into one final dataset by referencing each location and data year of RREPI, ZV, and HCCI per row. Data transformation was applied to location by instituting Location ID. Meanwhile, for rows where RREPI, ZV and HCCI do not have the same year, the most recent RREPI, ZV or HCCI value prior to that year was encoded to avoid zero values which will indicate a total decrease in value.

A Delta Value(Delta) attribute was added by the researcher to the final dataset which is computed based on the difference of the current(\(c\)) and previous(\(p\)) RREPI(\(R\)), ZV(\(Z\)), HCCI(\(H\)) values of the same Barangay(\(b\)). RREPI(\(R\)), ZV(\(Z\)) and HCCI values are each given a 1/3 weight. The Delta Value(Delta) is computed using the following equation:

\[
\Delta b = \left( \frac{R}{3} + \frac{Z}{3} + \frac{H}{3} \right)_c - \left( \frac{R}{3} + \frac{Z}{3} + \frac{H}{3} \right)_p
\]  

(1)

Likewise, a Movement attribute was added by the researcher which indicates whether the Delta Value of the Barangay increased (value is 1), remained the same (value is 0.5) or decreased (value
In the end, a final dataset composed of seven (7) attributes (Barangay, City, RREPI, ZV, HCCI, Delta, Movement) and 19,550 rows was created and was saved as a comma-separated values (CSV) format file.

**Data Modelling**

The researcher used the Python programming language to develop the model by utilizing PyCharm as the integrated development environment (IDE). The final dataset was imported and was split to two: 80% or 20,125 rows of data or observations as the training dataset and 20% or 4,025 as the test dataset.

Two algorithms were used to evaluate which output model can better predict Delta Value which include linear regression and RandomForest. Linear regression is considered as one of the commonly used technique in forecasting in different fields such as computer science, economics, and social sciences as it is among the simplest regression techniques to implement and its results are easier to interpret (Kumar & Srivastava, 2020; and Ismail & Al-Ghamdi, 2021). Meanwhile, RandomForest is an ensemble machine learning algorithm which is widely known in the computing industry for its high accuracy in forecasting by combining multiple decision trees in order to render predictions which significantly helps in reducing the variance of the predictions, hence, improve overall precision and accuracy (Alam & Siddiqui, 2021; and Zhang, et al., 2022). The researcher utilized the default parameters for each algorithm to avoid bias.

**Performance Evaluation**

For the Linear Regression algorithm, the researcher utilized the Ordinary Least Squares (OLS) assumptions to measure accuracy. The formula for OLS is as follows:

\[ Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon \]  

(2)

where \( Y \) equals the dependent variable, \( \beta_0 \) corresponds the intersection point or intercept of the model, \( X_j \) refers to the \( j \)th or maximum explanatory instance alterable value of the model from \( j=1 \) to \( p \), and \( \epsilon \) relates to the random error value with expectation value of between 0 and variance \( \sigma^2 \).

Conversely, for the RandomForest algorithm, the confusion matrix derivatives such as accuracy, precision, recall and F1-score metrics were used to evaluate the performance of the model. The number of accurate and inaccurate predictions made by the classification model in relation to the actual results (target value) in the data are displayed in a Confusion Matrix or Table. \( N \) is the number of target values (classes), and the matrix is \( N \times N \). The data in the matrix is frequently used to assess the performance of classification models (Duggal, 2023).

The confusion matrix usually provides the base values for Accuracy, Precision and Recall

Accuracy refers the proportion of the total number of predictions that were correct and is defined as:
Accuracy = \frac{TP+TN}{n} \quad (3)

where TP refers to True Positives, TN represents True Negatives and n is the total number predictions (Biecek & Burzykowski, 2020). Meanwhile, precision, which is also referred to as the positive predictive value, is the number of correct predictions among the predicted successes and is characterized to be of high value if false positives are low. Precision is mathematically defined (Berrar, 2019) as:

Precision = \frac{TP}{TP+FP} \quad (4)

Recall, which is also referred to as sensitivity or the true-positive rate, is the fraction of portion of correct predictions among the true successes and is also characterized to have a high value if there are few false negatives. It can be written using the following formula (Harrel, 2018):

\text{Recall} = \frac{TP}{TP+FN} \quad (5)

Finally, F1-score is the harmonic average or mean of Precision and Recall and has a direct similar effect Precision and Recall, that is, it has a higher value if Precision or recall is high and F1-score tend to be low if either Precision or Recall is low. It can be mathematically written (Biecek & Burzykowski, 2020) as:

F1-score = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)

To check the overall prediction performance and allow visualization, the researcher employed a Prediction Distance for each algorithm based on its predicted Delta value. Prediction Distance is the absolute Euclidean distance of the algorithm’s predicted Delta value less than that of the observed or truth value, hence, the smaller the prediction distance, the nearer the prediction value is to the observed value and the larger the prediction distance, the farther away the prediction value is to the observed value. The Prediction Distance formula is as follows:

Prediction Distance = |OV - PV| \quad (7)

where \(OV\) equals to the observed value and \(PV\) corresponds to the algorithm’s predicted value.

The researcher also implemented four classic statistical forecasting techniques over the final dataset to compare with the outcomes of the two machine learning-based algorithms. These techniques include straight line, moving average, simple linear regression, and multiple linear regression.

Results and Discussion

Figure 1 exhibits the evaluation accuracy result of the residential property pricing model created with Python programming language implementing Linear Regression algorithm. The Ordinary Least Squares (OLS) measure was utilized to evaluate the accuracy of the resulting model over its
Delta attribute as the dependent variable which obtained 94.3% R-squared accuracy and 92.8% adjusted R-squared value.

![Table: OLS Regression Results](image)

**Figure 1:** Evaluation Accuracy Result of Model using Linear Regression Algorithm

As expected, the linear regression algorithm got a significantly high accuracy rating of 92.8% considering that the dataset’s main attributes (RREPI, ZV, HCCI, and Delta) are mostly numeric in nature, hence, can be characterized as well-behaved data. The linear regression-based model has achieved a very high accuracy since the algorithm employed the least squares method which has been proven to minimize the error between the predicted and truth values.

Figure 2 displays the evaluation accuracy result of the residential property pricing model developed with the Python programming language and implementing Random Forest algorithm.

![Table: Model Performance Report Using RandomForest](image)

**Figure 2:** Evaluation Accuracy Result of Model using Random Forest Algorithm

The confusion matrix derivates, such as accuracy, precision, recall, and F1-score measures were employed to evaluate the performance of the output model among the 4,025 test dataset which represents data that the model has not yet encountered or seen. An average evaluation accuracy rate of 93% was obtained and a weighted average precision rate of also 93% was achieved which signifies a significantly higher level of prediction accuracy and precision.

In comparison with the linear regression-based model, the Random Forest-based model is
significantly superior, although in just a very small difference. However, with the RandomForest-based model, we are assured of the same significant weighted average accuracy precision, recall and F1-score (all 93%) in terms of individual classes.

Figure 3 shows the comparative Prediction Distance line graphs of techniques used.

![Figure 3: Prediction Distance Results of Forecasting Techniques Implemented](image)

Since the goal of the Prediction Distance is to empirically identify how near or far the prediction values of the algorithm than that of the actual values, lower values of prediction distance are desired. Based on the line graphs shown on Figure 3, Random Forest outperforms all the forecasting techniques used in this study on majority of the cities and municipalities and followed by the Linear Regression. Meanwhile, the Straight Line and Moving Average forecasting techniques rank the least among the techniques utilized in the study having the highest Prediction Distance among the majority of the subject cities and municipalities of the research.

The data on Figure 3 also proves that forecasting using machine learning-based prediction techniques performs better than the four classic statistical forecasting methods.

Figure 4 highlights the overall average comparative Prediction Distance obtained by each forecasting technique used.

![Figure 4: Overall Average Prediction Distance Results of Forecasting Techniques Used](image)
The Random Forest algorithm ranks first among the forecasting techniques utilized in this research with an overall average predicted distance of 4.32 points closer among all actual values of the evaluation dataset and which was followed by the Linear Regression with an overall average predicted distance of 8.75 points, Multiple Linear Regression with 12.46 points, and Simple Linear Regression with 15.11 points. Meanwhile, the Straight Line and Moving Average techniques ranked the lowest among the forecasting methods with an overall average distance point of 26.47 and 21.40, respectively. The data highlighted on Figure 4 again proves that predicting using machine learning-based forecasting techniques accomplishes better results than the four classic statistical forecasting methods.

Conclusion

Based on the results of the data modelling and evaluation conducted, a residential property pricing model was developed with a significantly excellent accuracy rate of 93% and precision rate of 93% by utilizing Random Forest machine learning algorithm that can accurately predict future values of real properties within the cities and municipalities of Central Pangasinan, Philippines based on historical records of residential real estate pricing index, zonal valuations and housing construction cost index. Further, the Random Forest algorithm outperforms the rest of the forecasting techniques used in this study in terms of the lowest overall average Prediction Distance value with only 4.32 points signifying that majority of its prediction values are almost near the actual values.

References


