International Journal of Recent Advances in Multidisciplinary Topics
Volume 6, Issue 1, January 2025
https://www.ijramt.com | ISSN (Online): 2582-7839

Comparative Analysis of Parametric, Semi-Parametric and Non-Parametric Methods on Real-Estate Data of the Kathmandu Valley

Sachin Kafle[1*], Kiran Kandel[2], Nawaraj Paudel[3]
[1]*Lecturer, School of Mathematical Sciences, Tribhuvan University, Kathmandu, Nepal*
[2]*Engineer, Software Tech., Arthasoft Solutions, Kathmandu, Nepal*
[3]*Director, School of Mathematical Sciences, Tribhuvan University, Kathmandu, Nepal*

***Abstract*: Identifying the factors and variables that affect the prices of the real estate housing through the online real estate websites is a challenging task. Along with the thorough study of variables such as land size, number of storied of house, road length, and location, it also depends upon metrics such as people's likes and dislikes indicated by the number of views on these websites for any real-estate property. The main aim of this research is to collect the relevant data (which is collected over 1 month of period) from the online real estate website called as Ghar Jagga Bazaar and apply different machine learning technologies to determine the price of the houses of Kathmandu valley and factors affecting it. After thorough preprocessing of data, the initial test were done to check the linearity of the data and found that several assumptions of the linearity of data such as homoscedasticity, independence of errors, normality of residuals, and multicollinearity test were violated. Even with the log and square root (sqrt) transformation of the data, there is no sign of significant improvement. Therefore, more robust semi parametric models with lenient assumptions such as Gradient Boosting method is experimented and its results are reported. Similarly, non-parametric methods such as Gaussian Processes, and Generative Adversarial Networks is explored. The optimum result is found with state of art non-parametric Gaussian Processes model with mean absolute error of 0.522, Jensen-Shannon distance of 12.79 and negative log-likelihood of -1071.91.**

***Keywords*: Gaussian Processes, Generative Adversarial Network, Gradient Boosting Method, Homoscedasticity, Jensen-Shannon Distance, Multi-Collinearity.**

1. Introduction

In the recent years, researchers from the various fields have been fascinated with the idea of predicting the prices of real estate with minimum error using limited number of predictors. Some Nepalese research predicts that financial systems such as Banks, private investment companies are highly correlated with the changing nature of real estate businesses. An unwarranted credit surge to real estate sector can generate asset price bubbles which can lead to financial fragility and financial crisis [12].

Understanding the nature of real-estate business helps the government control the property market prices and achieve the stability in national economy and growth. Furthermore, it also helps investors, market pundits and policy-makers to take data aided decisions on their investment and growth plan and helps to mitigate risks. Nepal relatively lacks in real estate market experiences and its policy is rather nonflexible which makes abroad investors very hard to invest in real-estate. Thus, appropriate real estate market policy can be constructed by understanding the price structure and factors affecting it. At present, both empirical and theoretical researches have been conducted showing the relationship between national economy and real estate policy. But very few research has been conducted to predict the real estate market prices based on both objective and subjective predictors. Therefore, the appropriate way of collecting both subjective and objective predictors which affects the pricing of real estate market is through online websites. The objective predictors such as size of the land, road length, number of rooms, parking, property type, number of storied of the house can be collected. Similarly, the subjective predictors such as location of the real estate, number of views on that real estate post on the online website can be also collected. Using such statistical data of real estate, this study tries to build the robust model to predict and analyze the factor affecting the prices of real estate housing market of Kathmandu valley.

Ordinary Least Squares (OLS) method is a classical model of regression that tries to obtain the linear relationship between independent and target variables. It has wider application in statistics, machine learning model and economics; investment equation [1] being a prime example. However, the OLS model has too many non-realistic assumptions which make it harder to implement in real world datasets which leaves a scope for transformation of data such as log-transformation, sqrt-transformation etc. Sometimes, even doing transformation of the data, the performance of OLS model is not significantly improved. In such case, we can either use more robust regression model such as Generalized Linear Model. If such point estimates model are not enough to capture the underlying trend of the data, semi-parametric regression or non-parametric regression models comes into the picture.

Gaussian Process regression (GPR) is a non-parametric

*Corresponding author: sachinkafle365@gmail.com

model that has become a practical and reliable Bayesian method in the field of machine learning [2]. [3] proved that Gaussian processes can be used to model different real-world scenarios and predictions from it were better than some classical state of art models. Encouraged with this paper, several authors used Gaussian process in different applications such as time series forecast [4], dynamic system model identification [5] and control system design [6], and combinations with Bayesian filtering [7]. But, no matter how simple and powerful GPR and OLS models are, their performance degrades when the training dataset contains corrupted data in the form of noise and outlier. In such case, Adversarial networks such as Generative Adversarial Network (GAN) can be used for regression problems. Similarly, semi-parametric approaches such as Gradient Boosting tree methods are also able to capture non-linear relationships between dependent and independent variables better than OLS and GLM methods.

2. Literature Review

At present, several researchers use machine learning methodologies such as hedonic pricing linear regression model, ensemble of regression trees, k-nearest neighbors and support vector machines for predicting real estate housing price. Hedonic model used in traditional real estate pricing estimator predicts the value of the real estate by diving its asset into different parts and estimating the price of the assets independently. Thus, professional intervention and correction may be required. [8] studied sales data of 2014 in Singapore real estate properties. They used properties such as location, property type, and district and ownership types to construct the hedonic pricing model. This model was able to help cover the fluctuation in price and aid government in macro control policies.

Subjective models that use predictors such as news and articles were used by some researchers to predict the price and nature of the real estate pricing. This research incorporated price index as a way to determine the real estate market condition. [9] found that index of search and trending queries are correlated to the nature of real estate market of the United States. These search queries and news were collected using Google search. The researchers used Auto regressive model to estimate the relationship between trend of the real estate market and its price. Although the model was good fit with the data obtained online, it could not accurately predict the real estate housing prices. Similar type of research was conducted by [10]. They used text-based sentiments and comments of people online collected from the search engine named Baidu to predict the house price index. They used several machine learning model such as support vector regression, Neural network, and these tests were done in different places of China. They validated their model by comparing non-integrated model using original inputs with integrated model using actual weighted inputs. The integrated model had lower residuals than non-integrated model.

[11] studied about different machine learning methods to build real estate appraisals for small sample data, which is appropriate in poorly developed real estate market. The study tests the hypothesis by building non parametric model against the traditional multiple regression model. Four different types of regression methods were employed such as Ridge regression, random forest, multiple regression model and k-nearest neighbor model. The research found that ridge regression performed best among all the model and they even have higher performance in small training datasets.

3. Methodology

The research design implored by the author consists of Pearson correlation test, descriptive measures, linearity test such as Lagrange Multiplier test, multi-collinearity test, Jarque Bera test for homoscedasticity etc. The statistical measures such as Pearson correlation coefficient, regression coefficient, log-likelihood is used for the model assessment. The descriptive research design is done by the researcher to find the appropriate distribution of each predictor and gather the information about the requirement of transformation for those predictors. The study is based on the primary data. It includes 3176 rows of data defining several attributes of real estate property. After proper transformation of the data, several parametric and non-parametric methods are investigated and its performances is compared.

In the study, the features with higher importance scores, such as the number of views and the presence of parking and land area, have a stronger influence on predicting the prices of the property. Features with lower importance scores, such as the number of bedrooms and bathrooms, contribute less to predicting the prices of the property. This information is vital for the study as the relative importance of each feature can help the process of feature selection and implement better and informed decision making for model training and testing.

4. Results

Upon initial examination of the dataset collected from the website "Gharjaggabazaar.com" encompassing attributes relevant to real estate housing prices, including bathroom count, bedroom count, total rooms, parking availability, views, land area in square feet, number of stories, road size, encoded property type, and distance to reference point, the primary objective of the study was to assess the either linear or non-linear relationships between these attributes and housing prices. In response to the violation of assumptions of linearity in the initial dataset, a log transformation was applied to the predictor variables in an attempt to induce linearity and enhance model fit.

Between non-parametric method, parametric method and semi-parametric methods, GLM came out to be least performed model as shown in table 7, and its result is rather more ambiguous as its assumptions are violated in non-linear data. Even after transformed data is applied, it does not show any significant improvement. Moreover, another robust semi parametric method named as Gradient Boosting method performed better in non-linear data with 0.745 value of R-squared as shown in table 4. We further explored non-parametric method such as Generative Adversarial Network

(GAN) and Gaussian Processes (GP). Out of all these models, GP performed best with least hyper parameter tuning depicted in table 7. Thus, we can conclude that in the case of applying non-linear data such as real estate data for regression, it is better to use non parametric methods.

Table 1
Hyper-parameters of gradient boosting method

Number of estimators/parameters	[100, 200, 300, 400, 500]
Maximum depth of tree	[10, 20, 30, 40]
Minimum sample split	[2, 5]
Minimum samples leaf	[1,2,3]

Table 2
Hyper-parameters for generative adversarial network architectures

Architecture	Total number of neurons on each layer	epochs
First	15	200
Second	40	200
Third	50-100	500
Fourth	100	500
Fifth	100-150	500

Table 3
Generalized linear model r-squared value on log-transformed data

Types of transformation	Linear-Log	Log-Linear	Log-Log
R-squared	0.31	0.21	0.48

Table 4
Performance metrics of gradient boosting method

Pseudo R-squared	Log-likelihood	MSE	MAE	JS divergence	JS distance
0.7454	710.10	1.99	3.05	-	-

Note. MSE: Mean Squared Error, MAE: Mean Absolute Error, JS divergence: Jensen Shannon divergence.

Table 5
Comparison of multiple comparable GAN architecture

Architecture Number	MSE	MAE	JS-divergence	JS-distance
First	0.78	0.6393	259.581 bits	16.112
Second	0.8845	0.635	325.41 bits	18.031
Third	0.78	0.5971	300.96 bits	17.348
Fourth	1.0728	0.5970	292.551 bits	17.104
Fifth	0.7722	0.5881	240.453 bits	15.507

Table 6
Performance of gaussian processes model

MSE	MAE	JS divergence	JS distance
0.522	0.5544	163.616 bits	12.791

Note. MSE: Mean Squared Error, MAE: Mean Absolute Error, JS divergence: Jensen Shannon divergence

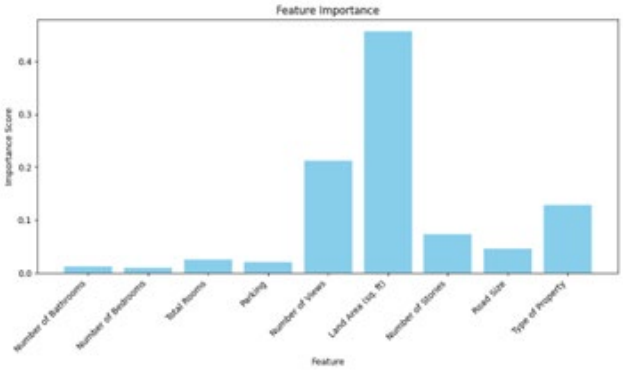

Fig. 1. Feature importance plot

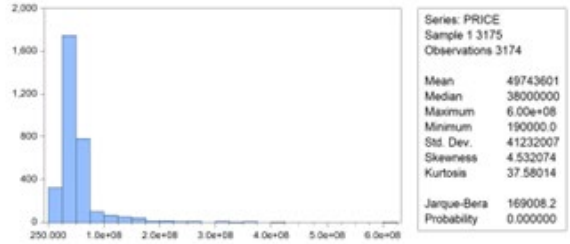

Fig. 2. Histogram of price

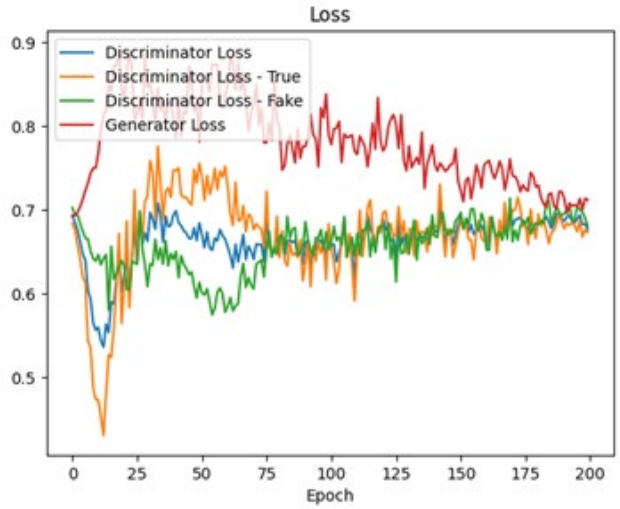

Fig. 3. Loss plot of architecture 3 of generative adversarial network

Table 7
Comparison of Parametric, Semi-parametric and non-parametric methods

Model name	Log-Likelihood	MSE	MAE	J-S divergence	J-S distance
Generalized Linear Model	-1095.5	2.1359	0.6599	310.787 bits	17.629
Gaussian Processes	-1071.91	0.522	0.5544	163.616 bits	12.791
Gradient Boosting Method	-710.10	1.99	3.05	-	-

Note. MSE: Mean Squared Error, MAE: Mean Absolute Error, JS divergence: Jensen Shannon divergence

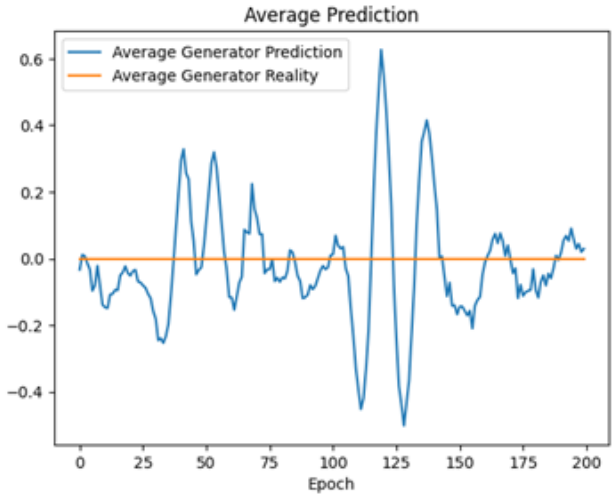

Fig. 4. Average prediction of architecture 3 of generative adversarial network

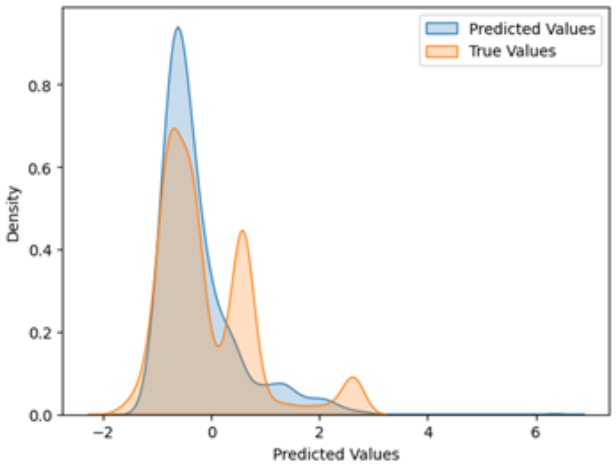

Fig. 5. Posterior distribution of architecture 3 of generative adversarial network

5. Conclusion

This paper presented a comparative analysis of parametric, semi-parametric and non-parametric methods on real-estate data of the Kathmandu valley.

References

[1] W. H. Greene, *Econometric Analysis*, 8th ed. Upper Saddle River, NJ, USA: Prentice Hall, 2017.

[2] C. E. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA: The MIT Press, 2006.

[3] C. E. Rasmussen, "Evaluation of Gaussian processes and other methods for non-linear regression," Ph.D. dissertation, Dept. Computer Science, Univ. Toronto, Toronto, ON, Canada, 1997.

[4] D. van der Meer, M. Shepero, A. Svensson, J. Widén, and J. Munkhammar, "Probabilistic forecasting of electricity consumption, photovoltaic power generation, and net demand of an individual building using Gaussian processes," *Applied Energy*, vol. 213, pp. 195–207, 2018.

[5] W. Ni, "Moving-window GPR for nonlinear dynamic system modeling with dual updating and dual preprocessing," *Industrial & Engineering Chemistry Research*, vol. 51, pp. 6416–6428, 2012.

[6] B. Likar and J. Kocijan, "Predictive control of a gas–liquid separation plant based on a Gaussian process model," *Computers & Chemical Engineering*, vol. 31, pp. 142–152, Jan. 2007.

[7] M. Lundgren, H. Hjalmarsson, and T. McKelvey, "Driver-gaze zone estimation using Bayesian filtering and Gaussian processes," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 10, pp. 2739–2750, Oct. 2016.

[8] L. Jiang, P. C. B. Phillips, and J. Yu, "A new hedonic regression for real estate prices applied to the Singapore residential market," *SSRN Electronic Journal*, 2014.

[9] G. Avi, M. G. Shane, and C. E. T. Catherine, *Economic Analysis of the Digital Economy*. Chicago, IL, USA: University of Chicago Press, 2015.

[10] D. Sun, Y. Du, W. Xu, M. Zuo, C. Zhang, and J. Zhou, "Combining online news articles and web search to predict the fluctuation of real estate market in big data context," *Pacific Asia Journal of the Association for Information Systems*, vol. 5, no. 1, pp. 19–37, 2013.

[11] G. Sebastian and D. Mariusz, "Parametric and non-parametric methods in mass appraisal on poorly developed real estate markets," *European Research Studies Journal*, vol. 23, no. 4, pp. 1230–1245, 2020.

[12] Nepal Rastra Bank, *A Report on Real Estate Financing in Nepal: A Case Study of Kathmandu Valley*. Kathmandu, Nepal: Nepal Rastra Bank, 2011.