Journal of Science and Technology
ISSN: 2362-1443 (Print) 3059-9490 (Online)2024,
Vol. 4, No. 2 (Special Issue), 45-50
DOI : https://doi.org/10.3126/jost.v4i2.78950

Valuation Model using Regression Algorithms for Nepalese Real Estate Marketplace

Brihat Thapa
Nepal College of Information Technology
contact.brihat@gmail.com

Rishav Bhandari
Nepal College of Information Technology
mr.bhandari5406@gmail.com

Tulsiram Ghimire
Nepal College of Information Technology
bibekghimire058.np@gmail.com

Subash Manandhar
Nepal College of Information Technology
subashmanandhar@ncit.edu.np

Roshan Chitrakar*
Nepal College of Information Technology
roshanchi@ncit.edu.np

* Corresponding author

Article History:
Received: 31 July 2024
Revised: 8 August 2024
Accepted: 17 December 2024

***Keywords*—Real Estate, Inefficient process, Property features, Valuation, Regression,Market Dynamics**

Abstract— **Real estate market is a cornerstone to the Nepalese economy, and is the largest alternative asset class. It is a prime field to apply a systematic model of price calculation to sustain the economic growth. The current property valuation process concerns with the mandated minimum price set by the government and the prices of similar nearby properties only. This process does not account for value-based cost increment and personal biases, which further increases the valuation asymmetry in the Nepalese real estate market. The main objective of this study is to address the inefficient valuation process by examining the effects of property features like property area, property location on the price of the property. Having a better understanding of the impact of each property feature on the property pricing will provide the stakeholders such as property brokers, mortgage lenders, property appraiser and investors better insights in their decision making. This examination uses regression methods and is trained on property size and location as input using 80 percent data split. The model is then tested on unseen data and its performance is compared. The loss function is minimized using gradient descent. The result from this research shows that the property latitude and longitude provided very small sensitivity to the change in property price. The large discrepancies in the property price in the same area for similar property size suggests that more comprehensive models are needed to capture the complex market dynamics of the Nepalese Real Estate Marketplace.**

I. Introduction

Real Estate is the business related to land, property and other fixed assets. In the context of Nepal, real estate is one of the main sectors of income source and stable investments [1]. Nepal's positioning as an agricultural country and the government's policies focusing on growing the agriculture sector has led to land being considered an important asset irrelevant of social class. Throughout history, real estate investments as a method of generational wealth preservation has been widely accepted in Nepal. Hence the real estate market in Nepal is experiencing steady growth due to increasing demand from both domestic and international investors [2]. The Real Estate segment is expected to grow at an annual rate of 3.6% resulting in a market volume of US $513.30 billion by 2029 [2].

In 2022, Nepal Rastra bank launched a report about the problems related to real estate investments and refinancing in Nepal. The report highlighted that increase in the prices of land creates problems in land acquisition for mega projects, production sectors, crisis of liquid cash flow, inability of lowincome families to have purchasing capacity of land and house [1].

With the rise in demand for properties, it becomes important to understand the factors that affect the property price. In Nepal, property is sold as per the wish of the seller. For sustainable growth of the economy fuelled by rising property demands, property valuations should be made based on the inherent value of the property. The factors that affect the land prices have to be studied and their impact on price has also to be modelled [3].

In Nepal the property values are obtained through various standards and methods and no single streamlined valuation process exists. The existence of various values such as tax value, market value, compensation value, mortgage value etc. for the same parcel creates social- economic, juridical and technical problems [4].

The approach suggested by this study allows the decision makers and stakeholders to have clear insights on the property features to be used and collected for property estimations with more noteworthy precision and least mistake. Since less research is done on the property valuation calculation in Nepal, we might want to manufacture a framework. The framework should allow to test the impact of certain property features on the property price. The results of this framework then gives the property appraisers freedom to only collect and analyse the property features that actually affects the property price.

This study applies the framework to test the importance of understanding the impact of property size, property location on the price of the property.

II. Literature Review

The research paper titled “Locally Regularized Linear Regression in the Valuation of Real Estate” by author Kubus, M., explores a novel approach to real estate valuation using locally regularized linear regression. In this study, the author challenges the conventional practice of building a single global regression model for property valuation. Instead, Kubus proposes the use of local regression models, which are tailored to specific subsets of data, offering greater flexibility and potentially improved accuracy in capturing local market variations [5]. Kubus conducted empirical research using three real estate market datasets to test the effectiveness of locally regularized linear regression. The results showed promising outcomes, with the local models outperforming global models in terms of prediction accuracy.

The research paper titled “Comparative evaluation of the performance of different regression models in land valuation” by authors Yalpir, S. & Yalpir, E. present a comparative analysis of various regression models for land valuation, emphasizing the growing importance of machine learning (ML) algorithms in this domain. The study focuses on predicting land values in central neighborhoods of Konya, Turkey, using Lasso, Elastic Net, ML.Net, and Ordinary Least Squares (OLS) regression models. The results reveal that ML.Net, a Microsoft-developed framework for building custom ML models, exhibits superior performance compared to other models, including the commonly used OLS [6]. However, the study is limited to a specific geographic location and dataset, raising questions about the generalizability of its findings to other contexts. Further research across diverse regions and datasets is needed to validate the superior performance of ML.Net and explore its potential in broader land valuation applications.

The research paper “Land Valuation and Management Issues in Nepal” by authors Ghimire, S. and Tuladhar, A. presented at the FIG (International Federation of Surveyors) Workshop in Nepal, addresses the critical issues surrounding land valuation and management in the country. It highlights the absence of a unified, standardized system for land valuation, leading to inconsistencies and inaccuracies in property assessments. The current practice often relies on subjective judgment and outdated methods, resulting in disputes and undervaluation of land, which can hinder economic development and equitable land distribution [4]. It highlights the urgent need for reforms and adoption of modern technologies to establish a robust and equitable land valuation system in the country.

Laalpurja Nepal is a prominent online real estate marketplace operating in Nepal, catering to buyers, sellers, and renters across the country. The platform offers a comprehensive listing of properties, including residential, commercial, and land options, along with detailed descriptions, images, and contact information for agents or owners [7]. While Laalpurja Nepal has established itself as a popular platform, its impact on the Nepali real estate market and its overall effectiveness in addressing the challenges of the industry warrant further research. A comprehensive analysis of user experiences, transaction success rates, and the platform's influence on property prices and market trends would provide valuable insights into its role and potential for improvement.

The research paper titled “Ridge Regression: Biased Estimation for Nonorthogonal Problem” by author Hoerl, A. and Kennard, R. address the issue of ordinary least square (OLS) method, which encounters challenges when predictor variables exhibit multicollinearity, a condition where predictors are highly correlated. This can lead to inflated variances of estimate coefficients making them unstable and unreliable for prediction and inference. To address this alternative estimation techniques has been proposed which is ridge regression. This method introduces a bias into the estimation process by adding a small positive constant to the diagonal of the predictor variable correlation matrix. This augmentation known as the ridge parameter, effectively reduces the impact of multicollinearity, leading to more stable coefficient estimates [8]. They introduced the concept of the "ridge trace," a graphical tool that aids in selecting an appropriate ridge parameter value. The ridge trace illustrates the changes in coefficient estimates as the ridge parameter varies, allowing researchers to identify a value that balances bias and variance for improved overall estimation accuracy [8].

The research paper titled “Regression Shrinkage and Selection via the Lasso” by author Tibshirani, R., introduces the Least absolute shrinkage and selection operator (Lasso) as a novel approach to bridge the gap between subset selection and ridge regression. Lasso implements both Subset selection and Ridge regression which have their own strengths and weakness. Subset selection offers interpretable models by selecting a subset of predictors, but it can be unstable due to its discrete nature. Ridge regression, on the other hand, shrinks coefficients for improved stability but does not eliminate any predictors, affecting interpretability. The Lasso achieves this by minimizing the residual sum of squares subject to a constraint on the absolute sum of the coefficients [9]. This constraint encourages sparsity in the solution, effectively setting some coefficients to zero and shrinking others. The degree of shrinkage and selection is controlled by a tuning parameter, allowing for a balance between model complexity and prediction accuracy. The Lasso's unique formulation offers several advantages. It produces interpretable models by selecting a subset of predictors, similar to subset selection. Simultaneously, it exhibits the stability of ridge regression due to its continuous shrinkage process [9]. This combination of interpretability and stability makes the Lasso a compelling alternative for regression analysis in high-dimensional settings.

III. EXISTING AND PROPOSED SYSTEM

The problem is creating a framework that can effectively map important property feature to property valuation. This framework should be based on hypothesis which shows that change in a property feature should result in change in the accurate property valuation that reflects market dynamics.

Currently, the most common method of valuation used in the Nepalese real estate market is a proportional method. The method uses the minimum property price based on its characteristics that is mandated by the government and the price of similar properties nearby. These two factors are combined in some proportion to get the final property valuation. The proportion used is not constant and may depend on the appraiser of the property. Such a method of valuation exposes the market pricing to biases and propagates an unsustainable increase in property prices. Ideally the property pricing should be driven by the inherent property features.

This study proposes using regression techniques to find the property price based on property features like property area, property location. Regression techniques like Ridge and Lasso regression add regularization terms hence supporting continuous shrinkage and removing multicollinearity among property features. This allows us to verify the importance of a property feature to predict the property price.

This method provides the stakeholders like online real estate marketplace access to better insights to make informed decisions in collecting property data for listings, appraisers to select a limited set of property features to consider during property assessments, mortgage lenders to maintain risk and loss aversion through access to accurate property price.

IV. SYSTEM DESIGN AND ARCHITECTURE

The design of the framework is separated into three stages: Initial, Middle and Last stage. The initial stage is associated with Data collection. The middle stage is the Data cleaning and pre-processing stage which includes various sub stages like Data consolidation, Feature selection, Handling missing values, feature extraction, feature engineering. The last stage of the framework consists of Data analysis and modelling.

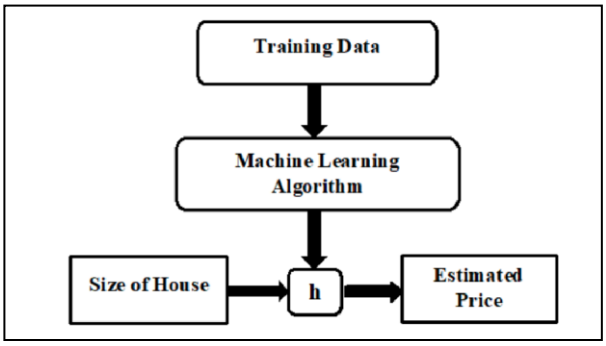


Fig. 1. System flow Model

A. Data Collection

Collection of data is the process in which information is gathered. We will be gathering quantitative data of property prices, property area and property location. The data for this project was acquired through aggregation of publicly available records. The lack of a centralized data source for property transactions in the Nepalese real estate market makes it necessary to check the validity of the dataset. The executive director of Economic Research Development of Nepal Rastra Bank has said that the NRB has stopped publication of survey on real estate due to lack of credible statistics [3].

B. Data Cleaning and Pre-processing

The data resulted from the collection efforts have amassed some amount of data from diverse sources that are heterogeneous in format and quality. Further, the data cleaning and pre-processing process that reveals inconsistencies, missing values, and outliers in the raw data. Therefore, a systematic cleaning and pre-processing pipeline has to be implemented to ensure data consistency, accuracy, and suitability for analysis and model training.

1) Data Consolidation: All aggregated data from various sources should be consolidated into a single CSV (Comma seperated values) file, for data manipulation and analysis

2) Feature Selection: Feature columns on the basis of high level of validity and their potential impacts on property valuation and market analysis were selected. Feature columns with significant amount of null values were not used. On the basis of these conditions the feature columns included: Property price, total property area, property location.

3) Handling Missing values: For the data rows with all selected feature columns empty were removed as they provided no information. Feature coluns with a small number of null values were imputed. The imputation method used the mean of the respective columns.

4) Feature Extraction: The geopy python framework was utilized to extract latitude and longitude coordinates from the location data (city name and address). This enabled spatial analysis and visualization of property locations. *5) Feature Engineering:*

a) Price Conversion: Price data was standardized by converting all values to crore to ensure uniformity.

b) Area Conversion: The total area was converted to a standardized unit of measurement aana which was originally presented in various units like ropani, square feet etc.

C. Data Analysis

Data analysis is a process that involves a series of steps to derive meaningful insights from the pre-processed data.

1) Exploratory Data Analysis:

a) Data Profiling: Understanding the dataset's sturcture, size and variable types through summary statistics and frequency distribution.

b) Data Visualization: Utilizing scatter, contour plots to visualize the relationships between property features to identify patterns and potential outliers

c) Outlier Detection and Treatment: Employing statistical Inter quartile range technique to identify and address outliers to prevent skewness in modelling results.

2) Statistical Analysis:

a) Correlation Analysis: Calculating correlation coeffiecients to assess the strength and direction of linear relationships between different features.

b) Hypothesis testing: Formulating and testing hypothesis to determine the statistical significance of observed relationships and make inferences about the model results.

3) Neighbourhoods Identification:

a) K-means clustering: K-means clustering of property location data gives insights into the spatial distribution of property. Clustering of other relevant features also allows the grouping of similar properties into distinct clusters. In the presented approach clustering is merely the starting point for the actual generation of knowledge. Useful clusters are ones that help spatial planners, politicians and decision makers in their actions [15].

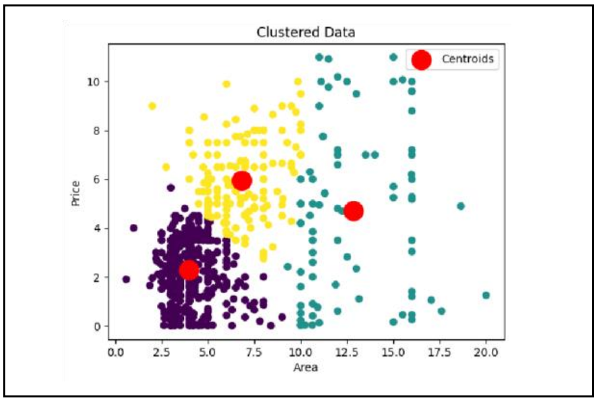


Fig. 2. K means cluster of Area vs Price

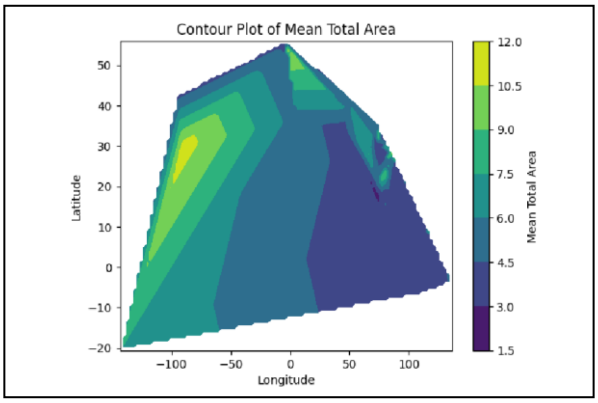


Fig. 3. Contour Plot of Longitude vs Latitude with Mean Total Area

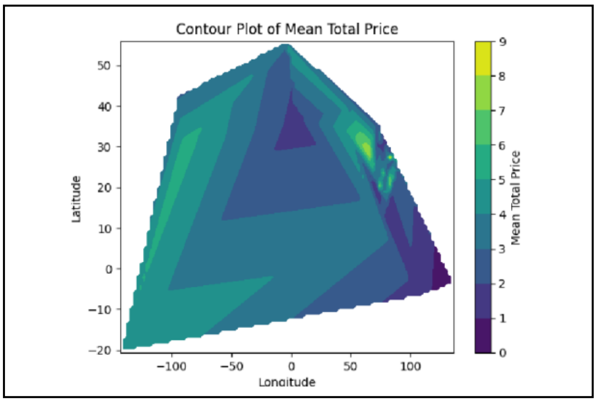


Fig. 4. Contour Plot of Longitude vs Latitude with Mean Total Price

V. Methodology

The afore mentioned system provides the basis for the development of an accurate and reliable valuation model for the Nepalese real estate market. This study details the results obtained in each stage of model development to test the impact of property size and property location on property price.

A. Linear Regression

A basic linear regression model was trained using the preprocessed and scaled dataset with a 80 20 split of training and testing set. It accept that the connection between subordinate variable and regressors is direct. The aggravation in anticipated esteem and the watched esteem is named as blunder [13]. The model’s performance was evaluated using following metrics.

Metric	Training Set	Test Set
Mean Squared Error (MSE)	0.794458735215897	0.923371854795044
Mean Absolute Error (MAE)	0.691504948040444	0.691504948040444
R-squared	0.197766080800876	0.083551861234759

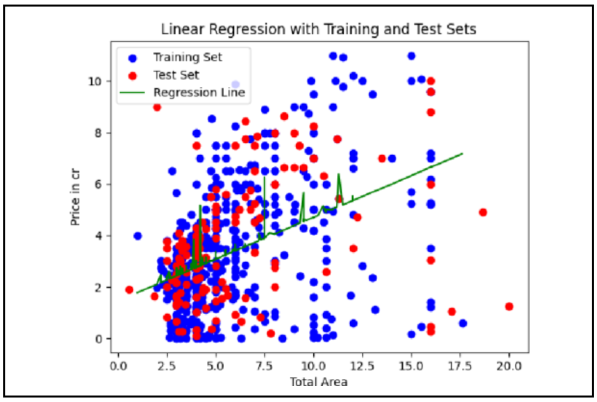


Fig. 5. Linear regression model fit for the training and testing set

While the model achieved a relatively lower MSE (0.794) on the training set, it exhibited a higher MSE (0.923) on the test set, indicating a degree of overfitting. Similarly, the Rsquared value was higher on the training set (0.198) compared to the test set (0.084), suggesting that the model's ability to explain the variance in property prices is weaker on unseen data.

The MAE values for both training and test sets were identical (0.692). This metric provides a more interpretable measure of the average absolute difference between predicted and actual prices. In this case, the model's predictions were off by an average of 0.692 crores on both the training and test sets.

The model's tendency to overfit and its relatively low explanatory power on unseen data underscore the need for exploring alternative models that can better generalize to new property listings.

B. Gradient Descent Optimization

The plot of the cost function of MSE with respect to the model’s coefficients was generated. This visualization aided in understanding the error landscape and the trajectory of gradient descent.

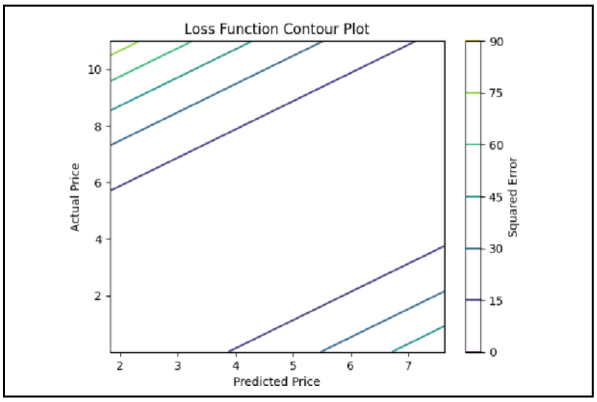
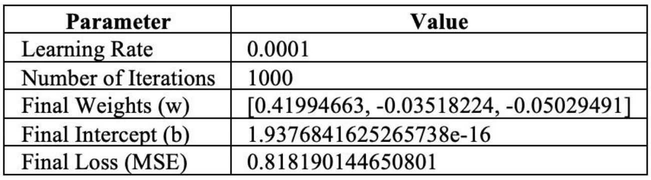


Fig. 6. Plot of Loss function of the regression mode the

The gradient descent algorithm was applied iteratively to optimize the model’s parameters and minimize the MSE loss of testing data.

TABLE II. GRADIENT DESCENT WEIGHTS AND MSE



Parameter	Value
Learning Rate	0.0001
Number of Iterations	1000
Final Weights (w)	[0.41994663, -0.03518224, -0.05029491]
Final Intercept (b)	1.9376841625265738e-16
Final Loss (MSE)	0.818190144650801

The linear regression model was trained using gradient descent optimization with a learning rate of 0.0001 and 1000 iterations. The resulting model achieved a final Mean Squared Error (MSE) of 0.818. This indicates that, on average, the squared difference between the predicted property values and the actual values was 0.818. The model's weights (w) and intercept (b) represent the learned coefficients of the linear equation that best fits the relationship between property features and prices.

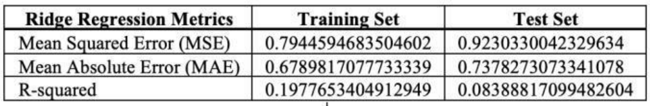
C. Regularized Regression

Both ridge and lasso regression, as regularization techniques, were applied to address potential overfitting in the linear regression model.

1) Ridge Regression Analysis: Ridge regression demonstrated a negligible improvement in R-squared on the test set compared to linear regression. While the MSE on the test set is slightly lower for ridge regression, the MAE is higher.

This suggests that ridge regression, while potentially reducing some overfitting compared to linear regression, does not significantly enhance the model's predictive performance or generalizability in this context.

TABLE III. RIDGE REGRESSION METRICS

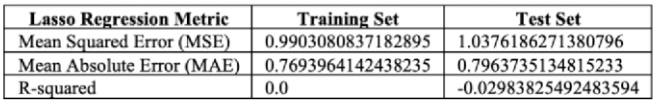


Ridge Regression Metrics	Training Set	Test Set
Mean Squared Error (MSE)	0.7944594683504602	0.9230330042329634
Mean Absolute Error (MAE)	0.6789817077733339	0.7378273073341078
R-squared	0.1977653404912949	0.08388817099482604

2) Lasso Regression Analysis: Lasso regression performs worse than both linear and ridge regression on all metrics. The R-squared values are 0 on the training set and negative on the test set, indicating that the model does not explain any of the variance in property prices and may even be performing worse than simply predicting the mean.

The higher MAE and MSE values further confirm the poor performance of lasso regression in this scenario.

TABLE IV. LASSO REGRESSION METRICS



Lasso Regression Metric	Training Set	Test Set
Mean Squared Error (MSE)	0.9903080837182895	1.0376186271380796
Mean Absolute Error (MAE)	0.7693964142438235	0.7963735134815233
R-squared	0.0	-0.02983825492483594

D. Model Diagnostics

The poor performance of the regression models in almost all metrics prompts the selection of model to be diagnosed and analyzed further.

1) Rainbow Test: This test was conducted to assess the linearity assumption of the model

TABLE V. P VALUE OF RAINBOW TEST

Statistic	Value
Rainbow Test Statistic	1.458935322605957
P-value	0.000208766014707

The Rainbow Test, a statistical test for assessing linearity, was applied to evaluate the appropriateness of a linear model for our real estate data. The resulting test statistic of 1.4589 and the extremely low p-value (0.0002) provide strong evidence to reject the null hypothesis of linearity.

2) Breusch-Pagan Test: This test was used to check for heteroscedasticity (unequal varainces) in the residuals

TABLE VI. P VALUE OF BREUSCH PAGAN TEST

Test	P-value
Breusch-Pagan Test	3.647115712982391e-38

The Breusch-Pagan test, a statistical test for homoscedasticity (constant variance of errors), was applied to assess the validity of the assumption of equal variance in our regression models. The resulting p-value of 3.647e-38 is extremely low, providing strong evidence to reject the null hypothesis of homoscedasticity.

The rejection of the null hypothesis suggests the presence of heteroscedasticity in our data, meaning that the variance of the errors (residuals) is not constant across different levels of the predictor variables [12]. This implies that the accuracy of our model's predictions may vary for different ranges of property values or other features.

E. Polynomial Regression

Polynomial Regression, a generalization of linear regression, was employed to capture potential non-linear relationship between features and property prices.

TABLE VII. PERFORMANCE OF POLYNOMIAL REGRESSION

Metric	Test set values
Mean Squared Error (MSE)	0.8190692049756423
Mean Absolute Error (MAE)	0.7090280570256656
Rsquared	0.0

The slight improvement of the polynomial regression suggests that it captures some non-linear relationships, but it might not be fully accounting for all the complexities present in the data.

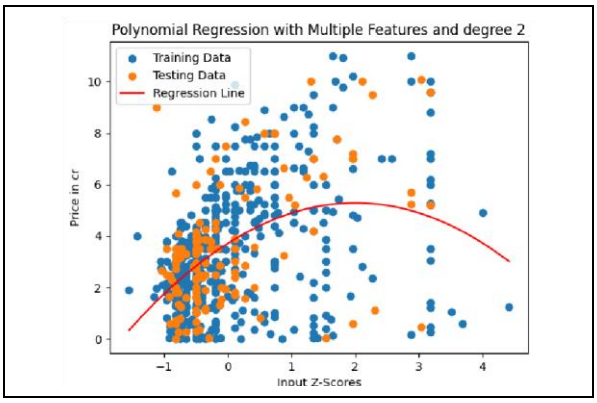


Fig. 7. Plot of Polynomial Regression model of degree 2

VI. Conclusion

Overall, the results highlight the limitations of Linear regression model in capturing the complexities of the Nepalese real estate market. Subsequent exploration of Ridge, Lasso coupled with techniques like gradient descent led to the selection of Ridge regression as the most suitable Linear model for property valuation.

The results of the Rainbow test and Breusch-Pagan test confirms the presence of non-linearity and heteroscedasticity in the relationship between property features and prices. This indicates that the traditional regression models might not be the most suitable for capturing the complexities of the Nepalese real estate market. To properly capture the market dynamics, we might need to employ neural networks or classification and regression tree (CART) models [11].

The results of the polynomial regression model performance were only slightly better than the Ridge regression which might suggest the dataset might have limitation, such as insufficient sample size or missing relevant features, that hinder the performance of all models.

This study of the model has also generated an insight that in the context of Nepalese real estate market, latitude and longitude showed limited sensitivity in predicting property prices. This suggest that other location-based factors, such as proximity to amenities or specific neighborhoods might play a more significant role in determining property values [14].

The data analysis of longitude vs latitude plot showed that the property data of popular destinations like tourist areas which tend to have higher property values compared to nearby areas introduced variance and potentially skew to the model's predictions.

References

[1] C. Dhakal, "Economic Impacts of Real Estate in Nepal," *Interdisciplinary Journal of Management and Social Sciences,* vol. 5, pp. 30-37, 2024.

[2] Statista, "Market Insights, Nepal: Statista," July 2024. [Online]. Available: https://www.statista.com/outlook/fmo/realestate/nepal#value. [Accessed 30 July 2024].

[3] V. a. S. M. a. V. J. Sampathkumar, "Forecasting the Land Price Using Statistical and Neural Network Software," *Procedia Computer Science,* vol. 57, pp. 112-121, 2015.

[4] S. G. a. A. T. a. S. Sharma, "Land valuation and management issues in Nepal," in *Proceedings of International workshop on the role of land professionals and SDI in disaster risk reduction : in the context of Post 2015 Nepal Earthquake, 25-27 November 2015, Kathmandu, Nepal*, Kathmandu, 2015.

[5] M. Kubus, "Locally Regularized Linear Regression in the Valuation of Real Estate," *Statistics in Transition. New Series,* vol. 17, pp. 515-524, 2016.

[6] Republica, "myRepublica," 30 December 2022. [Online]. Available: https://myrepublica.nagariknetwork.com/news/nrb-halts-survey-onreal-estate-price-citing-gaps-in-authenticity-of-data/. [Accessed 28 July 2024].

[7] A. a. K. R. Hoerl, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics,* vol. 12, pp. 55-67, 2012.

[8] L. Nepal, "LalPurja Nepal," [Online]. Available: https://lalpurjanepal.com.np/. [Accessed 18 March 2024].

[9] R. Tibshirani, "Regression shrinkage selection via the LASSO," *Journal of the Royal Statistical Society Series B,* vol. 73, pp. 272-282, 2011.

[10] Ş. a. Y. E. a. L. N. Yalpir, "Comparative evaluation of the performance of different regression models in land valuation," *Advanced GIS,* vol. 4, pp. 10-14, 2024.

[11] M. a. A. K. Razi, "A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models," *Expert Systems with Applications,* vol. 29, pp. 65-74, 2005.

[12] J. a. K. E. a. L. K. a. A. E. a. G. A. a. A. A. a. H. A. Jemna, "Improving the Performance of Linear Regression Model: A Residual Analysis approach," *GSJ,* vol. 8, pp. 212-222, 2020.

[13] A. a. E. R. a. T. H. a. M. W. Alfiyatin, "Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization Case Study : Malang, East Java, Indonesia," *International Journal of Advanced Computer Science and Applications,* vol. 8, 2017.

[14] I. a. K. K. a. D. T. Srour, "Accessibility Indices: Connection to Residential Land Prices and Location Choices," *Transportation Research Record,* vol. 1805, 2002.

[15] M. a. U. A. Behnisch, "Knowledge Discovery in Spatial Planning Data: A Concept for Cluster Understanding," 2015, pp. 49-75.