JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Real Estate Prognostication: A Comparative Study of Regression based Machine learning Models for Real Estate Market Prediction

Prabhu Tejas, Syed Shoaib
PG Scholars (Master of Technology), Artificial Intelligence, REVA University, Bengaluru, Karnataka, India

Abstract: The real estate market has experienced notable price volatility, offering both potential gains and hazards for investors. The emergence of technological tools has enabled both individual and corporate investors to estimate property values with fewer data points than traditional methods require.
In this study, a dataset consisting of the Kuala Lumpur Real Estate market was compiled, yet the absence of a sophisticated analytical framework for price evaluation can result in considerable financial setbacks, particularly for solo investors. This research delves into the potential of various machine learning algorithms to refine the process of property valuation when applied to actual market data.

Keywords – Kuala Lumpur Market, Real Estate Price Prediction, Data preprocessing, Exploratory data analysis (EDA), Feature selection Methods, Correlation metrics, OLS, PLS, Ridge Regression, Decision Tree, Linear regression, Property Valuation Accuracy, Model Building, Price Prediction, Machine learning Algorithms.

I. INTRODUCTION

Kuala Lumpur, as a rapidly developing city, greatly relies on its construction sector for growth. Property investment is a key strategy for many in Kuala Lumpur, attracting interest from various groups and individuals in accurately assessing property values. In 2023, Malaysia's real estate sector contributed roughly 23 billion Malaysian ringgit to the country's gross domestic product (GDP), marking an increase of over five billion Malaysian ringgits compared to the year before. Numerous individuals purchase homes not primarily for the security and stability they offer, but rather as investment assets within their portfolios. In developed nations, real estate represents a significant component of a household's wealth.

Consequently, the value of a property can significantly influences a household's financial portfolio [1]. This research delves into the dynamics of the Malaysian real estate sector, with a specific focus on the Kuala Lumpur housing market. Selected for its representative nature, the Kuala Lumpur market is analyzed to understand trends and patterns that might mirror the broader Malaysian property landscape, highlighting its position as one of the nation's most consistent and stable real estate environments. This study explores various methodologies capable of producing accurate property valuations, applying these techniques to a dataset gathered from the Kuala Lumpur market, sourced from a reputable online platform. This dataset provides insights into various aspects of housing, including the specific district and neighborhood where a property is situated, the number of rooms, the size of the living area in square meters, the number of bathrooms, and the property's valuation. Developing a valuation model is approached as a regression challenge within machine learning, where various techniques are employed to address such issues.

In this study, a dataset consisting of the Kuala Lumpur Real Estate market was compiled. The dataset was subjected to cleaning, manipulation, Exploratory Data Analysis, Data Preprocessing and Model Building Techniques. A suite of predictive models, namely Linear Regression, Lasso Regression, Ridge Regression, Elastic Net Regression, Ordinary Least Squares, Partial Least Squares and Decision Tree Regression, were designed and evaluated. The findings suggest that the application of advanced machine learning methods in developing predictive models markedly enhances the precision of predictions for property listing prices.
.

II. THEORETICAL RESEARCH

Fan et al, [2] conducted a study to explore the significance of the link between housing market prices and property features within Malaysia's resale public housing sector, employing the decision tree methodology for their analysis. Certain researchers have applied clustering techniques to aggregate similar properties, facilitating the subsequent estimation of their values [3]. Companies engaged in real estate often have access to an extensive array of features for analysis, yet a significant challenge arises from the

computational burden associated with managing a high volume of features. This complexity can notably impact the efficiency of developing regression models and calculating gradient descent solutions. Semi-parametric and nonparametric regression models offer a promising avenue for predicting housing prices, often outperforming traditional parametric models in terms of accuracy. Nonparametric models offer the flexibility of fitting the data to a broad class of functions, whereas semi-parametric models blend the strengths of various approaches, adapting the function—be it linear, convex, or another form—to yield the most accurate predictions [4]. The widespread integration of machine learning and artificial intelligence within the real estate sector has fundamentally shifted the landscape from one largely influenced by experience and arbitrage opportunities to a more intelligent, data-driven business model [5]. The present analysis of pricing models in the real estate domain primarily draws upon the hedonic pricing theory, introduced by the renowned economist Sherwin Rosen. This methodology is widely considered viable and has been extensively applied in real estate research by scholars. Rosen's theory posits that the price of a property can be described by a utility function that incorporates various influential factors, including structural features, neighborhood attributes, and environmental conditions [6]. Following this framework, real estate pricing models are typically constructed using a multiple regression approach. This method requires adherence to several foundational assumptions, including the independence of variables, homoscedasticity, and a normal distribution of the residuals. [12] Theoretically, models grounded in multiple regression tend to prioritize statistical inference over prediction because of their inherent characteristics. Consequently, substantial research conducted within this theoretical framework has concentrated on identifying the most influential factor affecting the model. This focus aims to assess the economic value of real estate based on distinct features and to uncover the causal relationships among the variables involved. This approach underscores a deeper investigation into how specific attributes impact property valuations, shifting the emphasis towards understanding the dynamics that drive real estate prices.

III. MULTIPLE LINEAR REGRESSION FOR HOUSING PRICE PREDICTION

Linear Regression: Linear regression is a fundamental statistical technique used to model the relationship between a dependent variable and independent variables. The model assumes that this relationship can be approximated by a linear equation. Mathematically, for a simple linear regression with one independent variable, the model can be represented as:

$$Y=\beta 0+\beta 1X1+\epsilon$$

Ordinary Least Squares (OLS): OLS estimates the linear relationship between a dependent variable (Y) and one or more independent variables (X) by fitting a linear equation of the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon$$

where:

Y: Dependent variable
X_i: Independent variables (i = 1, 2, ..., n)
β_0: Intercept
β_i: Regression coefficients
ε: Error term

The OLS method minimizes the sum of squared residuals (differences between observed and predicted Y values).

Partial Least Squares (PLS): Partial Least Squares (PLS) regression tackles situations where a simple linear relationship between independent variables (X) and the dependent variable (Y) might not fully capture the underlying structure. Unlike OLS, which focuses directly on individual X variables, PLS identifies latent variables – hidden factors within X that explain the most variance in Y.

The actual PLS formula involves matrix operations, but a simplified representation highlights its similarity to OLS:

$$Y = b_1 LV\_X_1 + b_2 LV\_X_2 + ... + b\_n*LV\_Xn + \varepsilon$$

where:
Y: Dependent variable
b_i: Regression coefficients for latent variables (LV) from X
LV_Xi: Latent variables extracted from X.
ε: Error term

Regularized Regression Techniques: These techniques address issues like multicollinearity (correlated independent variables) and overfitting by introducing penalty terms:

Lasso Regression (L_1 Regularization): Lasso adds a penalty term based on the L_1 norm of the coefficient vector (sum of absolute values), encouraging sparsity (setting some coefficients to zero). The equation is:

$$\text{minimize } \Sigma (Y_i - \hat{Y}_i)^2 + \lambda \Sigma |\beta_j|$$

where:

λ : Tuning parameter controlling the strength of regularization.

Ridge Regression (L_2 Regularization): Ridge adds a penalty term based on the L_2 norm of the coefficient vector (sum of squared values), shrinking coefficients towards zero but not necessarily setting them to zero. The equation is:

$$\text{minimize } \Sigma (Y_i - \hat{Y}_i)^2 + \lambda \Sigma \beta_j^2.$$

where:

λ : Tuning parameter controlling the strength of regularization.

Elastic Net Regression: This method combines both L_1 and L_2 penalties, offering flexibility in variable selection and coefficient shrinkage. The equation is:

$$\text{minimize } \Sigma (Y_i - \hat{Y}_i)^2 + \lambda_1 \Sigma |\beta_j| + \lambda_2 \Sigma \beta_j^2.$$

where:

λ_1 and λ_2: Tuning parameters controlling the strengths of L_1 and L_2 regularizations, respectively.

IV. EVALUATION METRICS FOR REGRESSION MODELS:

1. ***Mean Squared Error (MSE):*** This metric measures the average squared difference between predicted and actual values. Lower MSE indicates a better fit.

$$\text{MSE} = (1/n) * \Sigma (Y\_i - \hat{Y}\_i)^2$$

n: Number of data points
Y_i: Actual value for data point i
$\hat{Y}$_i: Predicted value for data point i

2. ***Root Mean Squared Error (RMSE):*** RMSE is the square root of MSE, putting the error in the same units as your data.

$$\text{RMSE} = \sqrt{(\text{MSE})}$$

3. ***Mean Absolute Error (MAE):*** MAE measures the average absolute difference between predicted and actual values. It's less sensitive to outliers compared to MSE.

$$\text{MAE} = (1/n) * \Sigma |Y\_i - \hat{Y}\_i|$$

4. ***R-squared (Coefficient of Determination):*** R-squared represents the proportion of variance in the dependent variable explained by the model. A value closer to 1 indicates a better fit.

$$R^2 = 1 - (\Sigma (Y\_i - \hat{Y}\_i)^2) / (\Sigma (Y\_i - \bar{Y})^2)$$

$\bar{Y}$: Average of the actual values (Y)

v. METHODOLOGY

The process unfolds through a series of structured steps, beginning with the acquisition and initial examination of the dataset to understand its structure and contents. This is followed by a meticulous phase of data cleaning, aimed at rectifying issues like missing values, and preparing the data through encoding categorical variables and refining the set of features for analysis. An exploratory data analysis (EDA) phase then provides insights into the dataset's underlying patterns and relationships. The data undergoes further preprocessing, specifically encoding, to ensure it is in a format suitable for model training. The culmination of this process is the construction of various regression models, including those with regularization and ensemble techniques. These models are rigorously evaluated using a suite of metrics such as the mean absolute error, R2 score, and mean squared error, with cross-validation techniques employed to verify their robustness and predictive power across different data subsets.

About the Dataset - Initially there are 8 features consisting of Location, Price, Rooms, Bathrooms, Car Parks, Property Type, Size and Furnishing containing 50K records.

Initial Data Inspection: The first step involves loading the dataset to get an overview of its structure and content, setting the stage for further actions.

There are 4467 duplicates, 19384 records with Null and junk values.

Data Cleaning: This crucial step focuses on correcting or removing inaccuracies and inconsistencies in the data, including handling missing values and ensuring the data's integrity. Post removal of the duplicates, null and junk values the total record count is 29032.

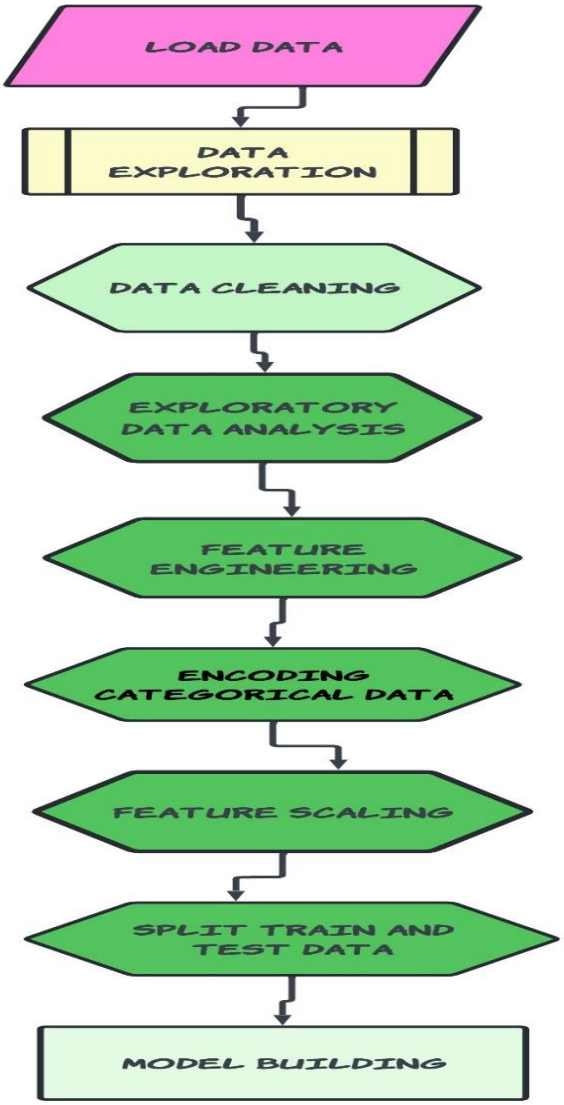
Exploratory Data Analysis (EDA): Through EDA, the dataset is analyzed to uncover patterns, anomalies, trends, and relationships among the variables, informing the subsequent preprocessing steps.
Data Preprocessing for Model Readiness: Key actions here include encoding categorical variables to transform them into a format that can be effectively used by the regression models.

Data Preprocessing:

- Separating rooms into Bedrooms and Additional rooms.
- Creating new features such as Studio and Total Rooms.
- Label encoding and dummy encoding categorical data.

Regression Model Development: This final stage involves building and evaluating various regression models, leveraging techniques such as regularization and ensemble methods to optimize predictive performance. Evaluation metrics and cross-validation are employed to assess and ensure the models' accuracy and generalizability.
.

VI. FLOW-CHART



Workflow

1. DATA Acquisition
2. DATA_CLEANING:
   - Handle missing values (e.g., imputation, deletion).
   - Identify and address outliers.
   - Correct inconsistencies and errors.
3. DATA_PREPROCESSING:
   - Feature scaling or normalization (if necessary).
   - Feature engineering (create new features).
   - Encode categorical variables (if necessary).
4. Exploratory Data Analysis (EDA)
5. Model Building and Training:
   - Split data into training and testing sets.
   - Train the model on the training set.
6. Model Evaluation:
   - Evaluate model performance on the testing set using R-square.

VI. RESULTS AND DISCUSSION

Model	Training R^2 score (%)	Testing R^2 score (%)
Linear Regression	64.59	68.96
Lasso Regression	64.59	68.96
Ridge Regression	64.58	69.44
Elastic Net Regression	48.16	55.71
OLS (Ordinal Least Square)	64.59	64.47
PLS (Partial Least Square)	64.16	68.06
Decision tree regression	99.18	72.00

By analyzing the results of these models (Linear Regression, Lasso Regression, Ridge Regression, Elastic Net Regression), the R^2 scores indicate the proportion of the variance in the dependent variable that is predictable from the independent variables.

- Linear Regression and Lasso Regression have similar R^2 scores for both training and testing sets, indicating they perform similarly in explaining the variance of the target variable.
- Ridge Regression slightly outperforms both Linear and Lasso Regression in terms of R^2 score on the testing set, suggesting it provides a better fit to the test data.
- Elastic Net Regression shows lower R^2 scores compared to other models on both training and testing sets, indicating it may not capture the variance in the data as effectively as the other models.
- Decision Tree Regression model shows the R^2 score for the training set is significantly higher compared to the testing set, suggesting overfitting. The high R^2 score for the training set (close to 1) indicates that the model fits the training data very well. The lower R^2 score for the testing set suggests that the model doesn't generalize as effectively to unseen data, indicating overfitting.

In the comparison between different regression models, Linear Regression and Lasso Regression displayed similar performance, providing decent fits to the data. Ridge Regression exhibited slightly better generalization to unseen data compared to Linear and Lasso Regression, suggesting it might offer better robustness against overfitting. However, Elastic Net Regression appeared less

effective in capturing the variance in the data compared to the other models. Conversely, the Decision Tree Regression model showed signs of overfitting, as evidenced by its significantly higher R^2 score on the training set compared to the testing set. This suggests that while Decision Tree Regression may fit the training data well, it struggles to generalize to new, unseen data. Overall, Ridge Regression seems to strike a balance between model complexity and generalization performance among the evaluated regression techniques, while decision tree model has a higher R^2 value.

The research conducted on regression-based predictive modeling in the Kuala Lumpur real estate market yields several notable conclusions:
Effectiveness of Machine Learning Algorithms: The study demonstrates that advanced machine learning algorithms, including Linear Regression, Lasso Regression, Ridge Regression, Elastic Net Regression, and Decision Tree Regression, significantly enhance the precision of predictions for property listing prices in Kuala Lumpur. Among these, Ridge Regression showed a slightly better performance in terms of generalization to unseen data, indicating its potential as a robust model against overfitting.
Challenges of Overfitting: The Decision Tree Regression model, while showing high accuracy on the training data, indicated a problem of overfitting as it did not generalize as well to the testing set. This highlights the importance of selecting appropriate models that balance fit with the training data and the ability to generalize to new, unseen data to avoid the risk of making overly optimistic predictions.

Role of Data Preprocessing: The research underscores the critical role of data preprocessing, including cleaning, manipulation, and visualization, in developing effective predictive models. Proper handling of missing values, encoding of categorical variables, and feature selection are essential steps to prepare the data for modeling, ensuring the models' performance and accuracy. Potential for Real Estate Valuation: The findings suggest a significant potential for using machine learning models to improve the accuracy and efficiency of real estate valuation. By leveraging a dataset with detailed property characteristics and employing various regression techniques, investors and stakeholders in the Kuala Lumpur real estate market can achieve more reliable and precise property valuations, aiding in investment decision-making processes.
These conclusions highlight the benefits of applying machine learning algorithms to real estate valuation, the need for careful model selection to avoid overfitting, the importance of rigorous data preprocessing, and the overall potential of such models to enhance the accuracy and efficiency of property valuations in the real estate sector.

VII. ACKNOWLEDGMENT

We extend our sincere gratitude to Dr. J.B. Simha for his invaluable guidance and mentorship throughout the entirety of this research endeavors. Dr. Simha's expertise and unwavering support have been instrumental in shaping our methodology and refining our analysis.
Additionally, we would like to express our heartfelt appreciation to Durga, Kishore, Naga Shree, and Ravikumar for their steadfast support and assistance at various stages of this project. Their contributions have been indispensable in facilitating the successful execution of our research objectives.
We are truly grateful for the collective effort and collaboration that has enriched this study and enabled us to achieve our goals.

REFERENCES

[1] Case, B., Clapp, J., Dubin, R., & Rodriguez, M. (2004). Modeling spatial and temporal house price patterns: A comparison of four models. The Journal of Real Estate Finance and Economics, 29, 167–191. Advance online publication. https://doi.org/10.1023/B:REAL.0000035309.60607.53

[2] Fan, G.-Z., Ong, S. E., & Koh, H. C. (2006). Determinants of House Price: A Decision Tree Approach. Urban Studies (Edinburgh, Scotland), 43(12), 2301–2315. https://doi.org/10.1080/00420980600990928

[3] Yang Li, Quan Pan, Tao Yang and Laotian Guee 2016 Reasonable Price Recommendation on Airbnb Using Multi-Scale Clustering Proceedings of the 35th Chinese Control Conference pp27-29.

[4] JooyongShim, OkmyunBin and Changha Hwang 2014 Semi-parametrics partial effects kernel minimum squared error model for predicting housing sales price Neuro computing vol. 124, pp 81-88

[5] J. R. Barr, E. A. Ellis, A. Kassab, C. L. Redfearn, N. N. Srinivasan, and K. B. Voris, “Home price index: a machine learning methodology,” International Journal of Semantic Computing, vol. 11, no. 1, pp. 111–133, 2017.

[6] S. Rosen, “Hedonic prices and implicit markets: product differentiation in pure competition,” Journal of Political Economy, vol. 82, no. 1, pp. 34–55, 1974.

[7] E. Lughofer, B. Trawinski, K. Trawinski, O. Kempa, and T. Lasota, “On employing fuzzy modeling algorithms for the valuation of residential premises,” Information Sciences, vol. 181, no. 23, pp. 5123–5142, 2011.

[8] O. Bin, “A prediction comparison of housing sales prices by parametric versus semi-parametric regressions,” Journal of Housing Economics, vol. 13, no. 1, pp. 68–84, 2004.

[9] Y. Kang, F. Zhang, W. Peng et al., “Understanding house price appreciation using multi-source big geo-data and machine learning,” Land Use Policy, vol. 2020, Article ID 104919, 2020.

[10] J.-G. Liu, X.-L. Zhang, and W.-P. Wu, “Application of fuzzy neural network for real estate prediction,” Advances in Neural Networks - ISNN 2006, vol. 3973, pp. 1187–1191, 2006.

[11] I. V. Lokshina, M. D. Hammerslag, and R. C. Insinga, “Applications of artificial intelligence methods for real estate valuation and decision support,” in Proceedings of the In Hawaii International Conference on Business, Honolulu, Hawaii, USA, January 2003.

[12] V. Kontrimas and A. Verikas, “*e mass appraisal of the real estate by computational intelligence,” Applied Soft Computing, vol. 11, no. 1, pp. 443–448, 2011.

[13] Wong Mei Chin, Nicholas Lee Wen Kit, Jeff Lai Wan Fei, Valuation of Real Estate: A Multiple Regression Approach, ICoMS'19, July 8–10, 2019, Prague, Czech Republic, 2019, Association for Computing Machinery.
ACM ISBN 978-1-4503-7168-1/19/07