Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer
Vol. 24, No. 1, November 2024, pp. 95∼104
ISSN: 2476-9843, accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: 10.30812/matrik.v24i1.4254

Optimizing Hotel Room Occupancy Prediction Using an Enhanced Linear Regression Algorithms

Dewa Ayu Kadek Pramita[1], Ni Wayan Sumartini Saraswati[1], I Putu Dedy Sandana[1], Poria Pirozmand[2], I Kadek Agus Bisena[1]
[1]Institut Bisnis dan Teknologi Indonesia, Denpasar, Indonesia
[2]Holmes Institute Sydney, Australia

Article Info

Article history:

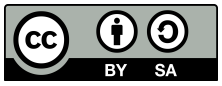
Received July 16, 2024
Revised September 11, 2024
Accepted October 22, 2024

Keywords:

Hotel
Linear regression
Occupancy Prediction
Optimizing Model
Polynomial regression

ABSTRACT

Predicting the correct hotel occupancy rate is important in the tourism industry because it has a major impact on the level of revenue and maintenance of a hotel's reputation. With accurate predictions, hotel performance can be optimized regarding resources, staff, and hotel facilities. The linear regression method has been proven to perform causal predictions well. However, this method has several weaknesses, such as the function of the relationship between dependent variables and independent variables that are not linear, overfitting, or underfitting in building the prediction model. The purpose of this study was to optimize the linear regression model in predicting hotel occupancy rates. The method used in this study was a Linear Regression method optimized with Polynomial Regression and regularization techniques to reduce overfitting using Ridge Regression and Lasso Regression. The results of the model evaluation showed that linear regression, which was optimized with Polynomial Regression and Ridge Regression in the model with the historical data of the Adiwana Unagi occupancy rate, historical data of the hotel occupancy rate in Bali, and the number of tourist visits in Bali, gave the best performance, with a mean absolute error score of 1.0648, root mean square error of 2.1036, and R-squared of 0.9953. The conclusion of this research was optimization using polynomial regression, achieving the best evaluation scores, where the prediction model performance indicates that variable X7 (tourist visit numbers) strongly influences the prediction of the occupancy rate.



Corresponding Author:

Ni Wayan Sumartini Saraswati, +62 819-3302-6640,
Faculty of Technology and Informatics,
Institut Bisnis dan Teknologi Indonesia, Denpasar, Indonesia,
Email: sumartini.saraswati@instiki.ac.id

How to Cite:
D. Pramita, N. W. Saraswati, I. P. Sandana, P. Pirozmand, and I. K. Bisena, "Optimizing Hotel Room Occupancy Prediction Using an Enhanced Linear Regression Algorithms", *MATRIK: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, Vol. 24, No. 1, pp. 95-104, November, 2024.

1. INTRODUCTION

The tourism sector is a significant contributor to foreign exchange for the country [1]. To support the advancement of tourism, providing top-notch service in the hospitality industry is essential. Progress in the hospitality industry will enhance the perception of service quality in this sector, thereby positively impacting the overall image of tourism. The use of artificial intelligence can be beneficial for hotel management. Hotels can increase their competitive intelligence in decision-making through the use of artificial intelligence and machine learning [2]. Competitive intelligence is a forward-looking process used to generate knowledge about a competitive environment to improve organizational performance [3]. One implementation of data science that can be applied for decision support in the hospitality sector is predicting occupancy rates based on previous patterns. The occupancy rate refers to the percentage of rooms booked or occupied at a specific time. Estimating the occupancy rate is a fundamental and critical activity in hotel management because it has a major impact on the level of hotel revenue and the maintenance of the hotel's reputation [4]. Predicting hotel occupancy rates is important in helping hotel management optimize resource management, such as staff, inventory, and facilities. Thus, hotels can reduce waste, increase operational efficiency, and provide better services to guests. In addition, it helps with marketing planning, proper pricing, and strategic decision-making to increase hotel revenue. If the predicted occupancy is high, the hotel can adjust its prices by increasing room prices. Hotels can also optimize services by ensuring that hotel staff are ready to face increased workloads, that services remain optimal, and that hotel supplies and facilities are adequate to respond to surges in demand, including food supplies, cleanliness, and staff availability. Hotels can also carry out additional promotions by encouraging additional sales, such as promotional packages, additional services, or special offers, such as honeymoon packages and family vacation packages. Collaboration with local partners can also be considered to meet additional guest needs or to provide extra services. However, if the predicted occupancy is low, the hotel can adjust prices with price reductions or special offers to increase the attractiveness of the rooms. Hotels can also conduct geo-targeting marketing, focusing on targeted marketing through digital advertising, online promotions, or local partnerships. Hotels can create attractive promotional packages such as discounts for longer bookings or inclusive package offers with added value for service quality. On this occasion, the hotel can also plan maintenance and renovations to remain attractive and meet standards. Hotels can maximize their capacity to obtain the best revenue by maximizing their business strategies based on predicted occupancy rates. Another impact is better hotel service, thus providing a good image for the hotel itself, while a wider impact is a good image for tourism. However, the challenge is to determine an optimal prediction method for hotels.

Linear Regression is a supervised learning method that can be used to make predictions. This method has been proven to be a method with a fairly good success rate and is commonly used to solve prediction problems [5–7]. However, Linear Regression has shortcomings, primarily because it assumes a linear correlationship between independent and dependent variables. If the true correlatioship is non-linear, Linear Regression may not produce accurate results [8–11]. For this reason, the novelty of this research is data trends are analyzed in this study to understand whether the data are closer to a linear or non-linear model and to carry out optimization by adding a polynomial function if the data relationship is non-linear. Linear Regression can also be sensitive to extreme values or outliers, significantly influencing the analysis results. When two or more independent variables have a high correlation (multicollinearity), it is difficult to determine the contribution of each variable to the prediction, and the results may be unstable or ambiguous [12].

There has not been much research that examines hotel occupancy rate prediction models using linear regression. A previous study [13] conducted a literature review on using deep learning methods for this problem. Several deep learning methods used in the literature review, including RNN, have limitations in understanding the relationship between various external variables. Other studies have used time series methods such as ARIMA and ARIMAX [14, 15], one of the methods used is ARIMA with EEMD, which has a higher performance than ARIMA alone, but has limitations in its complexity. Then, there is also research that uses artificial neural networks, deep learning methods [16], and other machine learning methods [17–20]. Machine learning used by one of the studies compared SVM and Linear Regression, where Linear Regression was superior to SVM in predicting hotel occupancy. Only a few studies have performed predictions using Linear Regression [4], but have not optimized this method. This research has not yet examined predictions based on historical occupancy rate data. This is new in this study, considering the holiday season each year; thus, the linear regression model is expected to recognize this recurring pattern well. The next novelty is that this study involves the overall hotel occupancy rate variable in Bali and the number of tourists visits to build a prediction model.

This study aims to optimize the Linear Regression method for predicting hotel occupancy rates by overcoming several weaknesses of this method, as previously mentioned. This study uses Several methods to optimize Linear Regression, including handling nonlinear variables by adding polynomial functions to these variables, which we know as Polynomial Regression. Regularization is performed using Ridge Regression or Lasso Regression technique to overcome multicollinearity. To overcome outliers, identification and handling are carried out using appropriate methods, such as removing them or using robust regression techniques that are less sensitive to extreme values. Future study aims to analyze independent variables that can be used as input for the Linear Regression model thus that it provides the best performance in predicting hotel occupancy rates.

2. RESEARCH METHOD

This research is engineering research, which is conducted by applying science to a design to obtain performance according to the specified requirements. The flow of this study is shown in Figure 1 Data preprocessing aims to eliminate outliers. Mutual information score analysis is used to obtain independent variables that have a strong influence on the dependent or objective variable, while data trend analysis is intended to determine which Linear Regression model is suitable for prediction. In this study, we attempt multiple linear regression and multiple polynomial regression and measure the model's performance. Once the model has been determined, it is adjusted to the training data. The next process of the model that has been trained is regularization, followed by predictions for the hotel occupancy rate for the coming month. The final stage evaluates the performance of the model with evaluation measurement units in the form of the coefficient of determination or R-squared (R2), Mean Absolute Error (MAE), and Root Mean Squared Error (RSME).

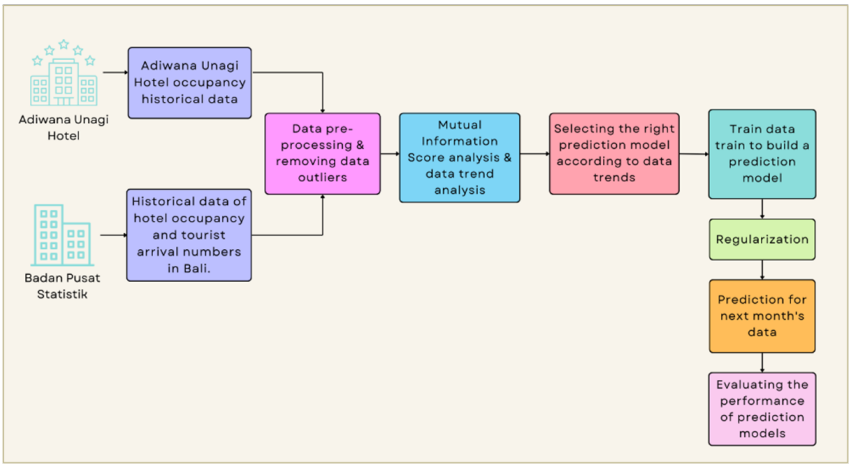


Figure 1. Research flow diagram

2.1. Dataset

We used the Adiwana Unagi Hotel occupancy data from January 2021 until April 2024 as the hotel dataset for prediction. The features used in this study consist of the variables Date, Month, Year, Adiwana Unagi hotel occupancy rate, hotel occupancy rate in Bali, and number of tourist arrivals from the Badan Pusat Statistik (BPS). The occupancy rate for both Adiwana Unagi Hotel and BPS data has a percentage range, namely 0-100%. As 2020 is COVID-19, which will give biased results to the model, the amount of data collected was 40 rows. The sample data are shown in Table 1.

Table 1. Sample Dataset

Date	Bulan	Tahun	Bps	Adiwana	Kunjungan
01/01/2021	Januari	2021	11.15	11.29	282258
01/02/2021	Februari	2021	8.99	5.61	240620
01/03/2021	Maret	2021	10.24	17.17	305582
...	...	...	...	...	...
01/02/2024	Februari	2024	55.27	86.00	1182021
01/03/2024	Maret	2024	52.71	88.47	1081969
01/04/2024	April	2024	57.69	89.52	1627975

2.2. Feature Engineering

The independent variables used in this study are divided into two sources: historical occupancy rate data sourced from the Adiwana Unagi Suites hotel and historical occupancy rate data and tourist visit numbers sourced from BPS. Historical data were then developed to capture seasonal patterns; therefore, this study used seven independent variables, namely X1 is the current month in Adiwana data, X2 is the current month of BPS data, X3 is the +1 month in the previous year for Adiwana data, X4 is +1 month in the previous year for the BPS data, X5 is the current month of the previous year for Adiwana data, X6 is the current month

of the previous year for BPS data, X7 is the number of domestic and foreign tourist visits to Bali in the corresponding month. The dependent variable used was the occupancy rate of +1 month (next month) from the current month in the Adiwana data. The results of determining variables X and Y are in Table 2. After determining the dependent and independent variables, several rows produce missing values for variables that have undergone the shifting process. In this study, missing values have been addressed using statistical methods by replacing the missing values with the mean, median, and mode of the variables Y and X. This is a feature engineering process that improves the quality of the model and removes irrelevant variables [21]. This feature engineering used Mutual Information Score (MSI) analysis. In other words, MSI can help analyze the relevance of each feature, where the higher the MI score value, the more informative and relevant the feature is for use in model predictions.

Table 2. X and Y variables

Date	Bulan	Tahun	Bps	Adiwana	Kunjungan	X1	X2	X3	X4	X5	X6	X7	Y
01/04/2024	April	2024	57.69	89.52	1627975	89.52	57.69	94.32	47.30	93.05	44.31	1627975	NaN
01/03/2024	Maret	2024	52.71	88.47	1081969	88.47	52.71	93.05	44.31	91.64	40.01	1081969	89.52
01/02/2024	Februari	2024	55.27	86.00	1182021	86.00	55.27	91.64	40.01	86.79	41.22	1182021	88.47
01/01/2024	Januari	2024	56.27	85.66	1194566	85.66	56.27	86.79	41.22	82.44	46.16	1194566	86.00
01/12/2023	Desember	2023	62.19	89.07	1686548	89.07	62.19	82.44	46.16	82.95	53.75	1686548	85.66

2.3. Linear Regression

Linear Regression is a statistical technique employed to forecast the value of one variable based on the values of other variables. The variable being predicted variable is known as the dependent variable, while the variables utilized to make these predictions are referred to as independent variables. In this study, Linear Regression was used to predict the occupancy rate for the next month based on its relationship with the independent variables. Linear Regression encompasses Simple Linear Regression, which involves only one independent variable, and Multiple Linear Regression, which involves more than one independent variable. Given that this study involves multiple independent variables, the model developed falls under Multiple Linear Regression. The following illustrates the Simple Linear Regression method shown in Figure 2. The model calculates the slope and intercept of the line of best fit, which illustrates the relationship between the variables. The slope signifies the change in the dependent variable for each unit change in the independent variable. At the same time, the intercept denotes the estimated value of the dependent variable when the independent variable is zero. Linear Regression demonstrates a linear relationship between the independent variable (predictor), represented on the X axis, and the dependent variable (output), represented on the Y axis. Based on the given data points, the model attempts to create a line that best fits these points using Equation 1. In Equation 1, Y_i is th e dependent variable, β_0 is the constant/intercept, β_1 is the slope/intercept, and X_i is the independent variable.

$$Y_i = \beta_0 + \beta_1 X_i \tag{1}$$

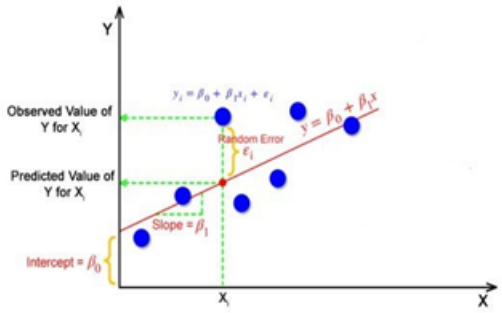


Figure 2. Simple linear regression

The goal of the Linear Regression algorithm was to obtain the best values for β_0 dan β_1. The line of best fit is the line with the minimum error between the predicted and actual values. Next, the development of multiple Linear Regression follows Equation 2. From Equation 2, $\beta_1, \beta_2, \ldots \beta_p$ is the slope coefficient, and $X_1, X_2, \ldots X_p$ are variable predictors, and ε is the error or residual value representing the variability in Y that the model cannot explain. Linear Regression may sometimes struggle to accurately capture certain data points, resulting in an incomplete representation of optimal points in the dataset, as shown in Figure 3. Optimization

can be performed using a Polynomial Regression approach. Figure 3 shows an underfitting condition, namely, a condition when the model is unable to recognize patterns during training, such that the accuracy value of the training data is low, which causes the accuracy value of the test data.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \tag{2}$$

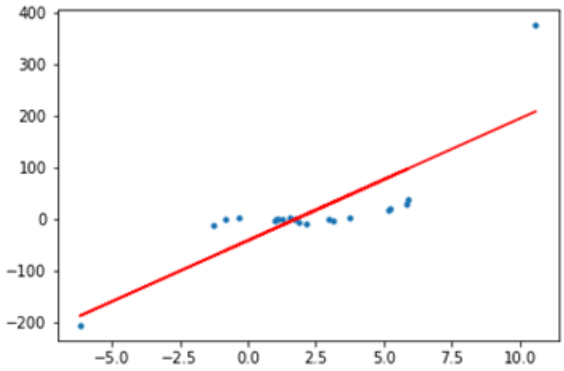


Figure 3. Linear regression is not capable of capturing all points in the data.[22]

2.4. Polynomial Regression

In Polynomial Regression, the relationship between the independent variable X and the dependent variable Y is described using an n-degree polynomial in X. This form of regression captures the appropriate nonlinear relationship between the X values and the conditional mean of Y. The least-squares approach is used to minimize the variance of the coefficients, adhering to the Gauss-Markov theorem. Although Polynomial Regression involves a curvilinear relationship between the dependent and independent variables, it is considered a type of Linear Regression because it models the linear combination of the polynomial terms. The polynomial regression equation is described by the following equation 3.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \ldots + \beta_2 X_1^n + \epsilon \tag{3}$$

While the model maintains linearity regarding the weights assigned to features, including x2 (x square) introduces a quadratic function, thereby adjusting the curve to accommodate quadratic relationships in the data. Based on the illustration in Figure 4 with the same data shown in Figure 3 and with the polynomial model, the data can be better predicted. This study uses a degree parameter of 2, the degree to which the relationship between the dependent and independent variables is described.

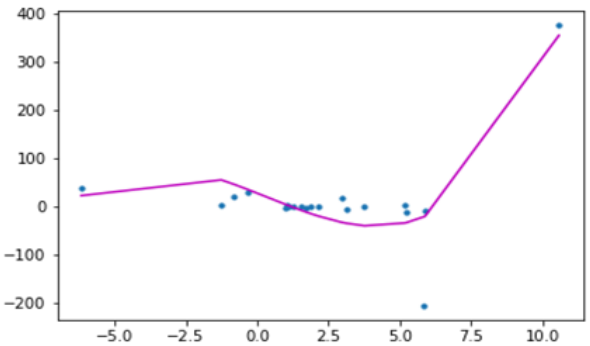


Figure 4. Example of a polynomial regression graph with degree 3 [22]

2.5. Regularization

Another optimization of this method is regularization. Regularization is a method for reducing overfitting in the models [23]. Overfitting is a condition in which the accuracy value of the training data is high and the accuracy value of the testing data is low. Regularization can result in a slight decrease in training accuracy with increased generalization ability. Generalization is the ability of a model to recognize new data. In this case, regularization can improve the model interpretability. There are two types of regularization methods related to Linear Regression: Ridge Regression and Lasso Regression.

Ridge regression is often referred to as L2 regularization, whereas Lasso Regression is known as L1 regularization. Lasso stands for Least Absolute Shrinkage and Selector Operator. This regularization uses a penalty called alpha (α). Alpha can be a value between 0 and infinity. The larger the alpha value, the greater the aggressive penalty. In this study, we used an alpha value of 1. Ridge Regression uses an L2 penalty or the square of the coefficient value, whereas Lasso Regression uses an L1 penalty or the absolute value of the coefficient. Ridge Regression shrinks all coefficients but does not eliminate them, while Lasso reduces some coefficients to zero thus, it can perform feature selection.

2.6. Model Evaluation

R2 denotes the proportion of variance in the dependent variable explained by the independent variables in the model. It ranges from zero to one, with larger values indicating a stronger model fit to the observed data. Mathematically, it can be expressed as shown in equation 4. From Equation 5, the Residual Sum of Squares (RSS) is the total sum of squared residuals, which are the differences between observed values (actual outputs) and predicted values (expected outputs) from the model. Moreover, Total Sum of Squares (TSS) is defined as the sum of the squared differences between each data point and the mean of the response variable. Equation 5 calculates RSS mathematically as the sum of these squared differences. And TSS can be expressed in equation 6.

$$R^2 = 1 - \frac{RSS}{TSS} \tag{4}$$

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{5}$$

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2 \tag{6}$$

In the equation above, y_i is the actual value of the dependent variable at observation i, while y_i is the predicted value of the model at observation i. The symbol $\bar{y}$ is the average of the actual data points. Where n is the number of observations. RMSE, as the error variance's square root, evaluates a model's absolute fit to the data by measuring how closely the predicted values match the actual data points. Mathematically, it can be represented in Equation 7.

$$RMSE = \sqrt{\frac{RSS}{n}} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \tag{7}$$

The formula for RMSE in Equation 7 shows that y_i is the actual value of the dependent variable at observation i, while y_i is the model's predicted value at observation I, as mentioned in Equation 5. The R2 metric is considered superior to the RMSE because the RMSE value is dependent on the units of the variable and is not a normalized measure, causing it to vary with changes in the units of the variable. On the other hand, MAE calculates the average of the absolute differences between the actual and predicted values, described by Equation 8. As mentioned in Equation 5, y_i in MAE is the actual value of the dependent variable at observation i, while y_i di MAE is the predicted value of the model at observation i.

$$MAE = \frac{1}{n} \sum_{i=1}^{1} \mid y_i - \hat{y}_i \mid \tag{8}$$

2.7. Experiment Scenario

The study conducted two experimental scenarios to assess the impact of adding variables to the prediction model. In the first scenario, linear regression was applied using variables X1 to X6. The second scenario used linear regression with an additional variable, X7, which represented tourist visit data. The goal was to determine whether the number of tourist visits could improve the model's predictive performance. The results from both scenarios were then analyzed to evaluate the differences in model performance.

3. RESULT AND ANALYSIS

3.1. Linear Regression and Polynomial Regression Evaluation

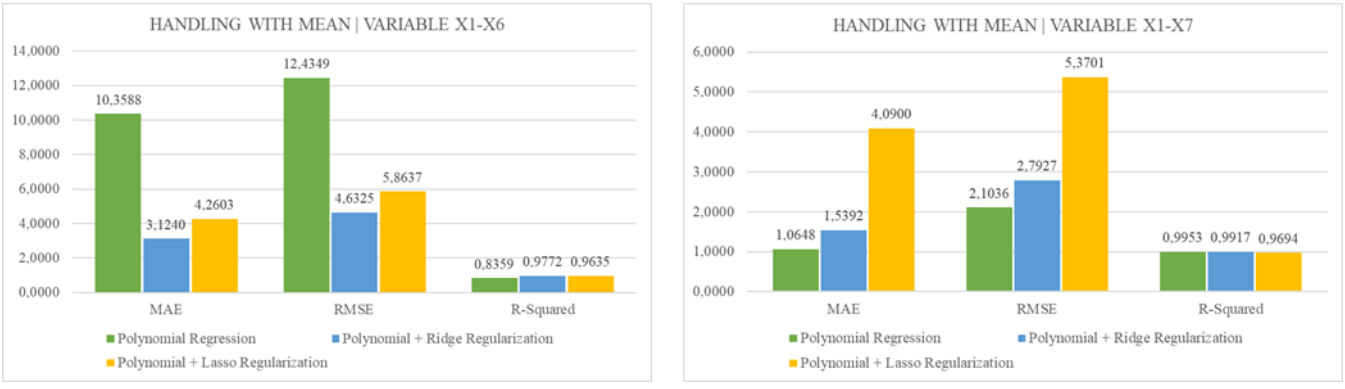
The evaluation results of the Linear and Polynomial Regression models for the two scenarios are listed in Table 3. The evaluation results for Linear and Polynomial Regression when the model uses the independent variables X1-X6 show that Linear Regression is better than Polynomial Regression. The model evaluation results using independent variables X1-X7 show that Polynomial Regression performs better than Linear Regression. Handling missing values with the mean provided the best evaluation values. In both models, Linear and Polynomial Regression show that adding variable X7 produces a better evaluation score. This shows that variable X7, or the number of tourist visits to Bali, relates to the occupancy rate at Adiwana Unagi. An R2 value close to 1 in the model evaluation results indicated that the prediction model was of very good quality.

Table 3. Linear Regression and Polynomial Regression Evaluation Results

Handling Missing Value	Method	MAE		RMSE		R2	
		X1-X6	X1-X7	X1-X6	X1-X7	X1-X6	X1-X7
Mean	Linear Regression	5,5235	5,4894	7,2977	7,1197	0,9435	0,9462
	Polynomial Regression	10,3588	**1,0648**	12,4349	**2,1036**	0,8359	**0,9953**
Median	Linear Regression	5,1251	5,1015	6,6782	6,6758	0,9530	0,9530
	Polynomial Regression	6,5275	1,2738	8,0382	2,6662	0,9319	0,9925
Modus	Linear Regression	8,3580	8,2736	14,2581	12,6687	0,8026	0,8442
	Polynomial Regression	23,6873	2,9092	31,9580	3,7362	0,0085	0,9864

3.2. Evaluation of Polynomial Regression with Regularization

Based on the Linear Regression and Polynomial Regression evaluation, regularization was implemented using Ridge and Lasso to optimize the Polynomial Regression results. Regularization is implemented on missing value data, handled with the mean because the previous evaluation score gave the best score. The prediction model's performance when using regularization is shown in the graph in Figure 5. Referring to Figure 5, the model with independent variables X1-X6 demonstrates that Polynomial Regression optimized with Ridge Regression and Lasso Regression can enhance the performance of Polynomial Regression. The model shows that regularization is successful in optimizing Polynomial Regression with the given independent variable. Based on the evaluation results, Ridge Regression shows a significantly better improvement than Lasso Regression. In the model using independent variables X1-X7, Polynomial Regression optimized with Ridge and Lasso Regularization did not significantly improve the performance of the Polynomial Regression model. The model shows that regularization is unsuccessful in optimizing Polynomial Regression with the X1-X7 independent variable. However, Ridge Regression still performed well in predicting hotel occupancy, with results nearly as close to those of Polynomial Regression. The actual and predicted values in the linear regression and polynomial regression models are compared, and the prediction results shown in the graphs in Figure 6 provide support. These models handle missing values using the mean for variables X1-X7. The prediction line in the Polynomial Regression graph follows the pattern of the actual values better than the Linear Regression graph.



(a) (b)

Figure 5. Implementation and visual results based on DCT, (a) Embedded Lena without DCT, (b)Embedded Peppers without DCT, (c) Embedded Baboon without DCT, (d) Embedded Lena with DCT, (e) Embedded Peppers with DCT, (f) Embedded Baboon with DCT

The improvement in the prediction model's performance, when variable X7 is added, indicates that the number of tourist visits to Bali is a significant factor in predicting the occupancy rate at Adiwana Unagi Hotel. This is because an increase in tourist visits leads to higher demand for hotels, thereby increasing the hotel's occupancy rate. Regarding regularization, Ridge Regression, which exhibits superior evaluation metrics to Lasso Regression, demonstrates its suitability in scenarios involving relevant variables with close interrelationships. Unlike Lasso, which excels in handling high-dimensional variables through automatic feature selection, Ridge regularization maintains all variables but penalizes them proportionally to their squared weights. This study, characterized by a limited number of predictor variables without redundancy and based on MSI analysis indicating close relationships, finds that Lasso regularization may inadvertently remove crucial information, potentially diminishing its performance compared to Ridge regularization.

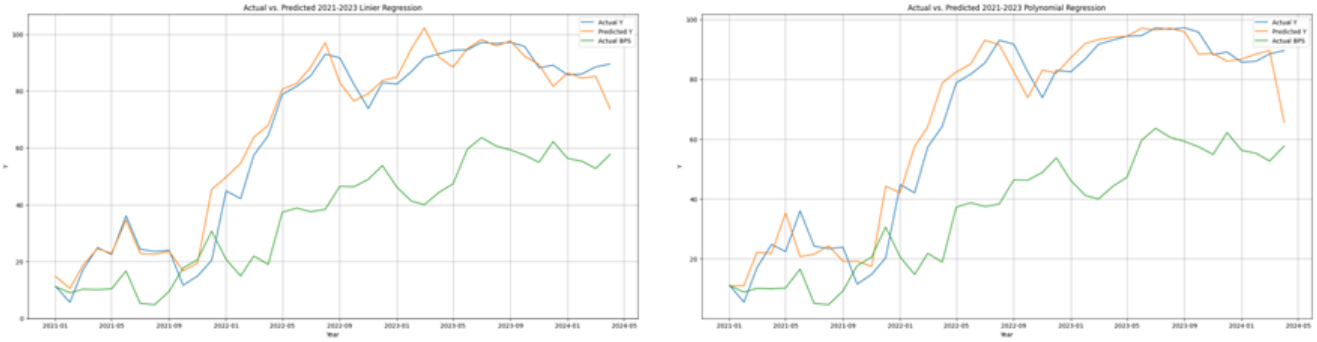


Figure 6. Linear regression and polynomial regression prediction graphs

However, in the model with variables X1-X7, Ridge and Lasso do not perform better than Polynomial Regression, although they still have good performance measures. The choice of alpha value in regularization can be a factor in its success. In addition, this can occur because polynomials are quite good at making predictions without having to use regularization. In addition, Polynomial Regression can capture complex relationships. Thus, regularization can reduce the ability to capture the complexity of relationships between variables. This study found that using polynomial functions can improve the performance of the linear regression prediction model for hotel occupancy rate problems. This has never been done in previous research. Referring to the results of previous studies with different methods, ARIMAX [24] with an RMSE of 4.21%, the polynomial regression model in this study provides better performance with an RMSE of 2.1%. This study found that the variable number of tourist visits to Bali strongly influences the prediction results.

4. CONCLUSION

Evaluation results indicate that Polynomial Regression alone did not significantly improve upon Linear Regression in the model with variables X1-X6. However, a substantial improvement was observed when Polynomial Regression results were optimized with Ridge and Lasso Regression. Ridge Regression achieved the best performance metrics: MAE of 3.1240, RMSE of 4.6325, and R-squared of 0.9772. This shows that the use of regularization to optimize the model has been achieved. Ridge Regression enhanced Polynomial Regression performance in the first scenario by mitigating multicollinearity issues. In contrast, Lasso, suitable for automatic feature selection, yielded slightly inferior results compared to Ridge due to the potential loss of information. In the second scenario, using Polynomial Regression with variables X1-X7, Polynomial Regression itself outperformed Linear Regression. Using the Polynomial function on Linear Regression makes the prediction model achieve better evaluation scores: MAE of 1.0648, RMSE of 2.1036, and R-squared of 0.9953. The prediction model performance indicates that variable X7 (tourist visit numbers) strongly influences the prediction of the occupancy rate at Hotel Adiwana Unagi. Given the non-linear nature of the data, Polynomial Regression effectively overcomes the limitations of Linear Regression by more accurately capturing the non-linear data points. In conclusion, we recommend using a Polynomial Regression model incorporating tourist visit numbers to Bali as the best approach for predicting hotel occupancy rates, providing a solid foundation for stakeholder decision-making.

5. ACKNOWLEDGEMENTS

We thank Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi for their moral and financial support in implementing this research successfully. We also express our deepest gratitude to the MATRIK journal for agreeing to publish the results of our

research. We also thank Adiwana Unagi Suites for providing the data needed for this research. Hopefully, this research will be useful in future research.

6. DECLARATIONS

AUTHOR CONTIBUTION

We thank authors 1 and 2 for formulating the problem and research objectives and for the analysis carried out on this research work. We thank authors 3, 4, 6, and 7 for constructing the linear regression model and all the experiments in this study. We thank author 5 for providing the data needed in this study.

FUNDING STATEMENT

This research is fully funded by the Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi in 2024.

COMPETING INTEREST

There is no competing interest in this research.

REFERENCES

[1] N. W. S. Saraswati, I. K. G. D. Putra, M. Sudarma, and I. M. Sukarsa, "The Image of Tourist Attraction in Bali Based on Big Data Analytics and Sentiment Analysis," in *2023 International Conference on Smart-Green Technology in Electrical and Information Systems (ICSGTEIS)*, 2023, pp. 82–87, https://doi.org/10.1109/ICSGTEIS60500.2023.10424322.

[2] V. Mahalakshmi, N. Kulkarni, K. V. Pradeep Kumar, K. Suresh Kumar, D. Nidhi Sree, and S. Durga, "The Role of implementing Artificial Intelligence and Machine Learning Technologies in the financial services Industry for creating Competitive Intelligence," *Materials Today: Proceedings*, vol. 56, pp. 2252–2255, 2022, https://doi.org/10.1016/j.matpr.2021.11.577.

[3] M. A. Köseoglu, A. Morvillo, M. Altin, M. De Martino, and F. Okumus, "Competitive intelligence in hospitality and tourism: a perspective article," *Tourism Review*, vol. 75, no. 1, pp. 239–242, jan 2020, https://doi.org/10.1108/TR-06-2019-0224.

[4] A. Ampountolas and M. Legg, "Predicting daily hotel occupancy: a practical application for independent hotels," *Journal of Revenue and Pricing Management*, 2023, https://doi.org/10.1057/s41272-023-00445-7.

[5] D. Alita, A. D. Putra, and D. Darwis, "Analysis of classic assumption test and multiple linear regression coefficient test for employee structural office recommendation," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 15, no. 3, p. 295, 2021, https://doi.org/10.22146/ijccs.65586.

[6] K. Le Nguyen, H. Thi Trinh, T. T. Nguyen, and H. D. Nguyen, "Comparative study on the performance of different machine learning techniques to predict the shear strength of RC deep beams: Model selection and industry implications," *Expert Systems with Applications*, vol. 230, p. 120649, 2023, https://doi.org/10.1016/j.eswa.2023.120649.

[7] S. Chakraborty, K. Kalita, R. ˇCep, and S. Chakraborty, "A Comparative Analysis on Prediction Performance of Regression Models during Machining of Composite Materials," *materials*, vol. 14, no. 6689, 2021, https://doi.org/doi:10.3390/ma14216689.

[8] T. Setiyorini and T. Informatika, "Comparison Of Linear Regressions And Neural Networks For Forecasting Electricity Consumption," *Pilar Nusa Mandiri*, vol. 16, no. 2, pp. 135–140, 2020, https://doi.org/10.1016/j.ijepes.2014.12.036.

[9] M. Chakraborty, S. Anirban Mukhopadhyay, and F. Ujjwal Maulik, "A Comparative Analysis of Different Regression Models on Predicting the Spread of Covid-19 in India," in *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*, 2020, pp. 519–524, https://doi.org/10.1109/ICCCA49541.2020.9250748.

[10] Z. Zhou, C. Qiu, and Y. Zhang, "A comparative analysis of linear regression , neural networks and random forest regression for predicting air ozone employing soft sensor models," *Scientific Reports*, pp. 1–23, 2023, https://doi.org/10.1038/s41598-023-49899-0.

[11] W. Kontar, S. Ahn, D. L. Mendoza, M. P. Buchert, J. C. Lin, S. Francisco, B. Area, E. M. Wells, and M. Small, "Bus Travel Time Prediction : A Comparative Study of Linear and Non-Linear Machine Learning Models," in *AICECS 2021 Journal of Physics: Conference Series*, 2022, https://doi.org/10.1088/1742-6596/2161/1/012053.

[12] T. Kyriazos and M. Poga, "Dealing with Multicollinearity in Factor Analysis: The Problem, Detections, and Solutions," *Open Journal of Statistics*, vol. 13, no. 03, pp. 404–424, 2023, https://doi.org/10.4236/ojs.2023.133020.

[13] N. Dowlut and B. Gobin-Rahimbux, "Forecasting resort hotel tourism demand using deep learning techniques – A systematic literature review," *Heliyon*, vol. 9, no. 7, p. e18385, 2023, https://doi.org/10.1016/j.heliyon.2023.e18385.

[14] F. A. Rizalde, S. Mulyani, and N. Bachtiar, "Forecasting Hotel Occupancy Rate in Riau Province Using ARIMA and ARIMAX," in *Proceedings of The International Conference on Data Science and Official Statistics*, vol. 2021, no. 1, 2022, pp. 578–589, https://doi.org/10.34123/icdsos.v2021i1.199.

[15] G. Zhang, J. Wu, B. Pan, J. Li, M. Ma, M. Zhang, and J. Wang, "Improving daily occupancy forecasting accuracy for hotels based on EEMD-ARIMA model," *Tourism Economics*, vol. 23, no. 7, pp. 1496–1514, may 2017, https://doi.org/10.1177/1354816617706852.

[16] Y. M. Chang, C. H. Chen, J. P. Lai, Y. L. Lin, and P. F. Pai, "Forecasting hotel room occupancy using long short-term memory networks with sentiment analysis and scores of customer online reviews," *Applied Sciences (Switzerland)*, vol. 11, no. 21, 2021, https://doi.org/10.3390/app112110291.

[17] M. Y. Anshori, T. Herlambang, V. Asyari, H. Arof, A. A. Firdaus, K. Oktafianto, and B. Suharto, "Optimization of Hotel W Management through Performance Comparison of Support Vector Machine and Linear Regression Algorithm in Forecasting Occupancy," *Nonlinear Dynamics and Systems Theory*, vol. 24, no. 3, pp. 228–235, 2024.

[18] A. S. Akbar and R. H. Kusumodestoni, "Optimization of k value and lag parameter of k-nearest neighbor algorithm on the prediction of hotel occupancy rates," *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 3, pp. 246–254, 2020, https://doi.org/10.14710/jtsiskom.2020.13648.

[19] B. A. Abdelghani, A. A. Mohammad, J. Dari, M. Maleki, and S. Banitaan, "Occupancy Prediction: A Comparative Study of Static and MOTIF Time Series Features Using WiFi Syslog Data," *SSRN Electronic Journal*, 2023, https://doi.org/10.2139/ssrn.4452581.

[20] B. Economics, K. Kozlovskis, Y. Liu, N. Lace, and Y. Meng, "Application Of Machine Learning Algorithms To Predict Hotel Occupancy," *Journal of Business Economics and Management*, vol. 24, no. 3, pp. 594–613, 2023, https://doi.org/10.3846/jbem.2023.19775.

[21] A. Derhab, A. Aldweesh, A. Z. Emam, and F. A. Khan, "Intrusion Detection System for Internet of Things Based on Temporal Convolution Neural Network and Efficient Feature Engineering," *Wireless Communications and Mobile Computing*, vol. 2020, no. 1, p. 6689134, jan 2020, https://doi.org/10.1155/2020/6689134.

[22] P. S, "Understanding Polynomial Regression Model," 2024.

[23] J. Kolluri, V. K. Kotte, M. S. B. Phridviraj, and S. Razia, "Reducing Overfitting Problem in Machine Learning Using Novel L1/4 Regularization Method," in *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*, 2020, pp. 934–938, https://doi.org/10.1109/ICOEI48184.2020.9142992.

[24] F. A. Rizalde, S. Mulyani, and N. Bachtiar, "Forecasting Hotel Occupancy Rate in Riau Province Using ARIMA and ARIMAX," in *The 1'st International Conference on Data Science and Official Statistic*, no. 25, 2021, pp. 578–589, https://doi.org/10.34123/icdsos.v2021i1.199.