Theta: Journal of Statistics, vol. 1, no. 1, pp. 1-7, 2025

Theta: Journal of Statistics

Journal homepage: www.jurnal.untirta.ac.id/index.php/tjs

Comparing Best Subset and Lasso Regression in the Customer Loyalty Prediction in a Restaurant Dataset

Aulia Anggitanniradi[a], Weksi Budiaji[b*], Juwarin Pancawati[c], Sri Mulyati[c]

[a] *Indo Global Transport, Jl. Atang Sanjaya No 21 Tangerang Banten 15125, Indonesia*
[b] *Department of Statistics, Universitas Sultan Ageng Tirtayasa, Jl. Jenderal Sudirman KM 3 Cilegon Banten 42435, Indonesia*
[c] *Department of Agribusiness, Universitas Sultan Ageng Tirtayasa, Jl. Raya Palka KM 3 Serang Banten 42163, Indonesia*
*Corresponding Author: budiaji@untirta.ac.id

INFORMATION

Article information:
Submitted: 03 February 2025
Revised: 17 March 2025
Accepted: 31 March 2025
Available Online: 31 March 2025

Keywords:
Best Subset; Lasso; Regression; Customer loyalty

ABSTRACT

Independent variables such as attributes related to the product, service quality, and purchase satisfaction are often correlated with one another in a customer loyalty research case. For instance, product attributes may overlap with service quality, and both factors jointly influence purchase decisions. To address multicollinearity, models such as best subset and Lasso regression can be employed. These models will be applied to a restaurant customer loyalty dataset. This study was conducted at Warung Tuman Restaurant in South Tangerang, Indonesia, from April to June 2022. We analyze responses from 100 purposively sampled consumers, with loyalty as the dependent variable and X_1 (product attributes), X_2 (service quality), and X_3 (purchase satisfaction) as predictors. Correlation analysis revealed strong positive relationships ($r = 0.44$, $p < 0.00$) among predictors, confirming multicollinearity and justifying the use of best-subset and Lasso. The dataset was split into a 60% training set and a 40% test set, with the training set used to develop predictive models, which were then evaluated for accuracy using the test set. All correlation values demonstrate a significant positive relationship between the independent variables, indicating the suitability of the best subset and Lasso regression applications. The best subset and Lasso regression generate models with two independent predictor variables, i.e. product attributes and purchase satisfaction. The best subset regression exhibits a lower Sum of Squared Errors (SSE), thereby indicating its superior performance compared to the Lasso regression model. To effectively sustain and improve customer loyalty, restaurant managers should prioritize optimizing product attributes and purchase satisfaction factors.

INTRODUCTION

Regression analysis is a fundamental technique in statistics used to model the relationship between a dependent variable and one or more independent variables. It is widely applied in various fields, from economics to social sciences. However, a significant challenge arises when the independent variables are

P-ISSN: 3109-1903/ E-ISSN: 3108-9895. Copyright (c) 2025
DOI: http://dx.doi.org/10.62870/tjs.v1i1.31171

highly correlated with one another, a situation known as multicollinearity. Multicollinearity can lead to inflated standard errors of regression coefficients, making the estimates unstable and less reliable. This issue can severely affect the interpretability and predictive performance of regression models, as it becomes difficult to isolate the individual effect of each predictor [1].

One potential solution to this issue is Best Subset Regression (BSR), which involves evaluating all possible combinations of predictor variables and selecting the subset that minimizes a given model selection criterion, such as the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) [2]. Best Subset Regression can mitigate the effects of multicollinearity by identifying the most relevant predictors, thus providing a more parsimonious model. However, while it is effective in small to moderately sized datasets with low multicollinearity, it becomes computationally expensive and infeasible as the number of predictors increases, especially when dealing with high-dimensional data [3].

Alternatively, Lasso Regression (Least Absolute Shrinkage and Selection Operator) offers another approach to dealing with correlated independent variables. Introduced by Hastie et al. [4], Lasso applies an L1 penalty to the regression coefficients, forcing some coefficients to shrink to zero. This regularization technique not only helps in dealing with multicollinearity but also performs automatic variable selection. Lasso's ability to enforce sparsity makes it particularly useful in high-dimensional settings, where the number of predictors exceeds the number of observations. Recent studies have shown that Lasso tends to outperform stepwise regression, offering better consistency [5].

The multicollinearity is particularly critical in marketing and consumer behavior research, where predictors such as product attributes, service quality, and purchase satisfaction often exhibit strong interdependencies. For example, product attributes might be linked with service quality, while both could influence consumer purchase decisions. This interrelationship can cause challenges in modeling the impact of these variables on consumer loyalty. Therefore, applying methods to handle multicollinearity is crucial for improving the accuracy and interpretability of models used to predict consumer behavior [6].

Despite the growing body of research on the comparison between Best Subset Regression and Lasso in various contexts, their relative merits for customer loyalty prediction remain unclear. It is essential to evaluate how these regression techniques perform when applied to such practical problems. This paper aims to address the better predictive accuracy when modeling loyalty using correlated marketing variables between Best Subset Regression dan Lasso.

RESEARCH METHODS

This study was conducted in Warung Tuman Restaurant, Ciater, Serpong, the city of South Tangerang, Indonesia from April to June 2022, utilizing a dataset to investigate the factors influencing consumer loyalty. The independent variables considered include product attributes (X_1), service quality (X_2), and purchase satisfaction (X_3) [7]. The data was collected from 100 restaurant consumers purposively using purposive sampling. This method was selected to target respondents with direct experience dine in at the restaurant, ensuring relevance to the study's objectives. However, purposive sampling may introduce selection bias, as the sample may not fully represent the broader population of consumers.

Participants were asked about three independent variables: product attribute (7 indicators: product value-price relationship, product quality, product benefits, product distinctive features, product appearance/presentation, product reliability and consistency, menu variety), service quality (4 indicators: warranty/guarantee, delivery communication, complaint handling, problem resolution), and purchase satisfaction (5 indicators: service attitude, communication, ease of access and comfort, restaurant reputation, company competence). On the other hand, the dependent variable is consumer loyalty, which was measured using three indicators. Consumers selected one of seven response options from a Likert scale provided for each indicator, ranging from strongly agree to strongly disagree. The reliability of the Likert scale was assessed using Cronbach's Alpha, with all variables exceeding the threshold of 0.7,

indicating good internal consistency. Additionally, the validity of the measurement scales was confirmed through exploratory factor analysis (EFA), ensuring that the indicators accurately reflected their respective constructs. The seven indicators were chosen due to their good stability and discrimination [8]. Then, the variables analyzed were the sum/total of each indicator within the variable.

Prior to modeling, key regression assumptions were evaluated. Normality of residuals was tested using the Shapiro-Wilk test, while heteroscedasticity was assessed via the Breusch-Pagan test. No significant violations of these assumptions were detected. To assess potential multicollinearity, correlation analysis was performed on these independent variables prior to modeling.

The best subset regression approach was employed to select the most relevant predictors for the model. Best subset regression involves evaluating all possible combinations of independent variables to identify the subset that best predicts the dependent variable. The model is formulated as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \tag{1}$$

where β_0 is the intercept, β_1, β_2, β_3, are the regression coefficients for the independent variables, and ϵ represents the error term. At the core of this statistical task is the best subset problem with a subset size of k, which is defined by the following optimization problem:

$$\min_{\beta} \frac{1}{2} \|y - \boldsymbol{X\beta}\|_2^2 \text{ subject to } \|\beta\|_0 \leq k \tag{2}$$

where the l_0(pseudo)norm of a vector $\boldsymbol{\beta}$ counts the number of nonzeros in $\boldsymbol{\beta}$ and is given by $\|\beta\|_0 = \sum_{i=1}^{3} 1(\beta_i \neq 0)$, where $1(\cdot)$ denotes the indicator function. The algorithms for solving the problem are implemented in the widely used *leaps* statistical package in R [9]. The selection process is based on a selection criterion, i.e BIC, CP, adjusted R^2, which balances model bias and variance [4], [10].

Lasso regression (Least Absolute Shrinkage and Selection Operator) is another technique used in this study for variable selection. Lasso works by imposing an L1 penalty on the regression coefficients, which effectively shrinks some of the coefficients to zero, thus performing variable selection. The lasso model can be expressed as:

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right) \tag{3}$$

where λ is the regularization parameter that controls the strength of the penalty. As λ increases, more coefficients are set to zero, reducing the complexity of the model [4]. An appropriate value for the regularization parameter, lambda, must be selected with caution to ensure the desired level of sparsity. Cross-validation is a widely adopted approach for determining the optimal lambda, a resampling technique in which the training data are partitioned into multiple subsets or folds [11]. The optimal regularization parameter (λ) was determined using 10-fold cross-validation, which partitions the training data into subsets to evaluate model performance across different λ values. The λ yielding the lowest mean squared error (MSE) was selected to balance bias and variance. This method is implemented in the glmnet package in R.

In this study, data was split into a training set and a test set for both the best subset regression and lasso regression models. The training set was used to fit the models, while the test set was used to evaluate their predictive performance. The algorithm for splitting the data was as follows:

1. The dataset was randomized. 60% of the data was allocated to the training set, and the remaining 40% was assigned to the test set.
2. The best subset and lasso models were trained on the training set.
3. Predictive accuracies were evaluated using the test set by calculating the Sum of Squared Error (SSE) for each model. The SSE for each model was computed by summing the squared differences between the observed and predicted values for the test set.

4. The SSE measure allows for the assessment of model accuracy and is particularly useful for comparing models. The model with the lowest SSE was considered to perform better in terms of predictive accuracy.

RESULTS AND DISCUSSION

Correlation Between Independent Variables

The correlation measure used to calculate the correlation between independent variables is Pearson's correlation. Although the correlation values obtained are less than 0.5 (Figure 1), all correlation values indicate a significant positive correlation at the 5% alpha level. Specifically, when product attributes are perceived as higher quality, it is likely that consumers will also rate service quality and purchase satisfaction more positively. This relationship suggests that consumers' overall experience is influenced by the interplay between product quality, the quality of service provided, and their level of satisfaction with the purchase. This positive correlation is consistent with findings from studies conducted at restaurant locations [12] and coffee shops [13].

However, the moderate correlation values also imply that other unmeasured factors (e.g., pricing, ambiance, or brand reputation) may influence consumer loyalty. Future research could expand the model to include these variables for a more comprehensive analysis.

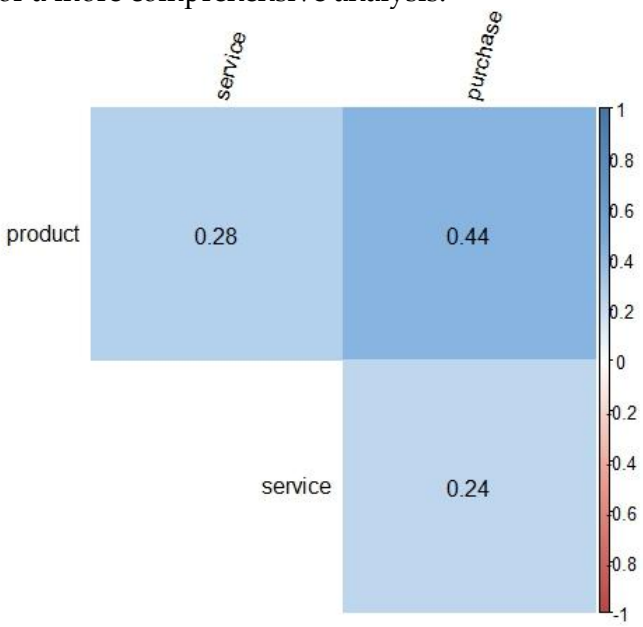

Figure 1. Correlation scores between independent variables (if the color is white, correlation is insignificant

Best Subset Regression

The dataset is divided into two sets: the training set and the test set. The training set consists of 60% of the 100 consumers, meaning 60 consumers were randomly selected for the training set. The results of the best subset regression for the training set, based on the criteria of adjusted R^2, CP Mallow, and BIC (Bayesian Information Criteria), are presented in Table 1. All three criteria show that the best model includes two independent variables.

Table 1. Best subset model evaluation criteria

Criteria	Number of Independent Variable		
	1	2	3
Adjusted R2	0.294	**0.328**	0.323
CP Mallow	4.938	**2.525**	4.000
Bayesian Information Criteria	-16.881	**-17.128**	-13.435

When two variables are used in the best subset regression, the resulting regression equation is

$$y = 5.947 + 0.403\ x_1 + 0.249\ x_3 \quad (4)$$

This equation indicates that consumer loyalty consumer loyalty is primarily driven by product attributes (X_1) and purchase satisfaction (X_3). Several studies have found similar results, suggesting that product attributes [14] and purchase satisfaction [15] influence consumer loyalty. Service quality can be excluded from the consumer loyalty model due to differences in consumer expectations and perceptions. Perceived service quality can vary among consumers, leading to inconsistent consumer loyalty. Consumers with high expectations tend to switch to other brands, even when service quality is perceived as good. From a managerial perspective, these findings underscore the need to prioritize product excellence (e.g., menu diversity, consistency) and post-purchase satisfaction (e.g., staff training, ambiance) over service quality alone.

Lasso Regression

The training set used in the Lasso regression is the same as the training set used in the best subset regression. Cross-validation is applied in Lasso regression to obtain the optimal penalty value (λ). Figure 2 shows that the optimal log(λ) value is -1.387, resulting in an optimal λ of 0.25. Using a λ of 0.25, the regression equation obtained from Lasso is as follows:

$$y = 11.505 + 0.338\ x_1 + 0.166\ x_3 \quad (5)$$

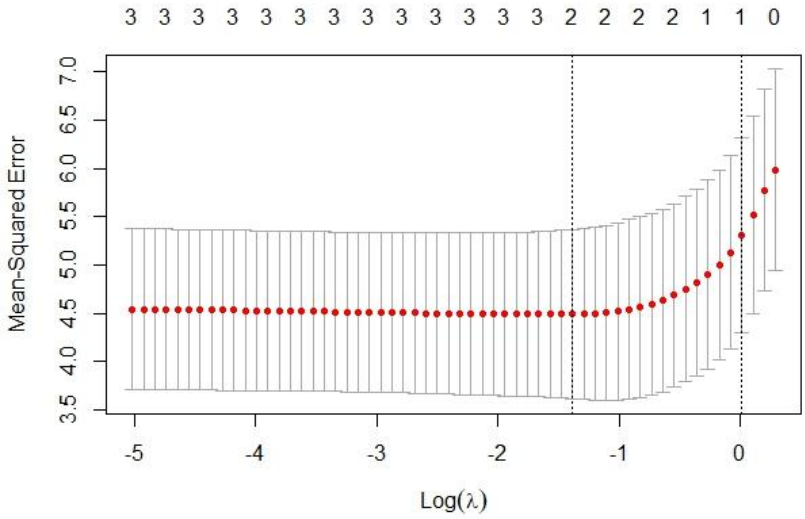

Figure 2. Mean squared Error of various log (λ)

Equivalent to the best subset regression, Lasso regression also excludes service quality as a predictor in the consumer loyalty case for this restaurant data. This consistency across methods strengthens confidence in the model's robustness.

Comparison between best subset and Lasso regression

A comparison of the prediction results between the best subset regression and Lasso regression is applied to the test set. The predicted values from Equations (4) and (5) will be compared with the actual consumer loyalty scores. This calculation results in the Sum of Squared Errors (SSE) for each model, with the model that yields the smallest SSE being considered the best model. Table 2 presents the SSE results for both models.

Table 2. Sum of Squared Errors (SSE) of best subset and lasso models

Model	SSE
Best subset	**115.248**
Lasso	131.515

The best subset regression has a smaller SSE, indicating that it provides better predictions than the Lasso regression. Equation (4) can thus be used as the selected model to predict consumer loyalty.

Limitations and Biases

This study has several limitations that should be acknowledged. First, the use of purposive sampling, while practical, may introduce selection bias by overrepresenting frequent patrons, thereby limiting the generalizability of the findings. Additionally, the model does not account for potentially influential variables such as pricing, location convenience, or cultural preferences, which could further shape consumer loyalty. Finally, reliance on self-reported Likert-scale responses may introduce subjectivity, potentially affecting the accuracy of the measurements.

Management Strategy

To effectively maintain and enhance loyalty, restaurant managers must focus on optimizing product attributes and purchase satisfaction factors. The product attributes that have been found to positively influence loyalty include quality, benefits, distinctiveness, appearance, reliability, menu diversity, and the value-for-money relationship. To strengthen these attributes, restaurant managers should offering a diverse menu to cater to a variety of tastes, and ensuring consistency in the presentation and preparation of dishes. Additionally, positioning the menu as unique or distinctive, possibly by incorporating local flavors or specialty dishes, can make the restaurant stand out in a competitive market.

Purchase satisfaction, defined by service attitude, communication, ease of access, comfort, reputation, and company competence, also plays a critical role in consumer loyalty. Restaurant managers should therefore implement strategies that enhance customer satisfaction. For example, training staff to communicate effectively, and ensure a pleasant atmosphere can significantly improve customer satisfaction.

Strategic Recommendation

While service quality was statistically insignificant in this study, it should not be neglected—improving service may indirectly boost satisfaction (X_3), thereby reinforcing loyalty.

CONCLUSION

Both the best subset regression and Lasso regression generate models with the same number of included predictor variables, namely two variables: product attributes and purchase satisfaction, which are key predictors of consumer loyalty. The comparison between the best subset model and the Lasso model demonstrates that the best subset regression has a smaller Sum of Squared Errors (SSE), thus making it a better model than the Lasso regression model. However, model selection should not rely solely on SSE—factors such as interpretability, computational efficiency, and practical usability in business decision-making should also be considered.

These findings have practical implications for businesses aiming to enhance customer loyalty strategies. By focusing on improving product attributes and ensuring high purchase satisfaction, companies can strengthen consumer retention. Future research could expand on this work by incorporating additional variables (e.g., brand reputation or demographic factors) or testing these models on larger, more diverse datasets to improve generalizability. Further comparative studies could also

evaluate other performance metrics (e.g., AIC, BIC, or out-of-sample validation) to provide a more comprehensive assessment of regression techniques.

REFERENCES

[1] J. Hair, W. Black, B. Babin, and R. Anderson, Multivariate data analysis. Pearson Education, 2009.

[2] M. Hofman, C. Gatu, E. Kontoghiorghes, A. Colubi, and A. Zeileis, "lmSubsets: Exact Variable-Subset Selection in Linear Regression for R," J. Stat. Softw., vol. 93, no. 3, pp. 1–21, 2020, doi: https://doi.org/10.18637/jss.v093.i03.

[3] J. Seedorff and J. Cavanaugh, "Assessing Variable Importance for Best Subset Selection," Entropy, vol. 26, no. 9, p. 801, 2024, doi: https://doi.org/10.3390/e26090801.

[4] T. Hastie, R. Tibshirani, and J. Friedman, The elements of statistical learning: Data mining, inference, and prediction, 2nd ed. Springer, 2009.

[5] D. Zhou, R. Chahal, I. Gotlib, and S. Liu, "Comparison of Lasso and Stepwise Regression in Psychological Data," Methodology, vol. 20, no. 2, pp. 121–143, 2024.

[6] S. Molinillo, R. Aguilar-Illescas, R. Anaya-Sánchez, and E. Carvajal-Trujillo, "The customer retail app experience: Implications for customer loyalty," J. Retail. Consum. Serv., vol. 65, p. 102842, 2022.

[7] Y. Kristiawan, H. Hartoyo, and B. Suharjo, "Customer Satisfaction: Service Quality or Product Quality (Case Study at Fast Food Restaurant in Jabodetabek)", *Binus Business Review*, vol. 12, no. 2, pp. 165–176, Jul. 2021.

[8] W. Budiaji, "Skala Pengukuran dan Jumlah Respon Skala Likert," J. Ilmu Pertan. Dan Perikan., vol. 2, no. 2, pp. 125–131, 2013.

[9] D. Bertsimas, A. King, and R. Mazumder, "Best subset selection via a modern optimization lens," Ann. Stat., vol. 44, no. 2, 2016.

[10] G. Brooks, "Best-Subset Selection Criteria for Multiple Linear Regression," Gen. Linear Model J., vol. 42, no. 2, 2016.

[11] A. Pak, A. Rad, M. Nematollahi, and M. Mahmoudi, "Application of the Lasso regularisation technique in mitigating overfitting in air quality prediction models," Sci. Rep., vol. 15, p. 547, 2025, doi: https://doi.org/10.1038/s41598-024-84342-y.

[12] R. Setiawati and I. Bernarto, "Effects of Service Quality, Food Quality, and Price Fairness Customer Satisfaction at Japanese restaurant 3 Wise Monkeys, Jakarta," Bp. Int. Res. Crit. Inst.-J. BIRCI-J., vol. 5, no. 2, pp. 8921–8934, 2022.

[13] A. Salsabillah, H. Khoirullah, and F. Mustikasari, "Product Quality, Service Quality, Price, and Location Influence Towards Coffee Shop Customer's Satisfaction," Manag. Stud. Entrep. J., vol. 5, no. 1, 2024.

[14] M. Risal and M. Aqsa, "Consumer Loyalty as Impact of Marketing Mix and Customer Satisfaction," MIMBAR, vol. 7, no. 2, pp. 297–304, 2021.

[15] A. Española, A. Janaban, and E. Martir, "Restaurants' Attributes Customers' Satisfaction and Loyalty," Int. J. Sci. Manag. Stud. IJSMS, vol. 7, no. 2, 2024, doi: 10.51386/25815946/ijsms-v7i2p123.