ICFBE 2024

The 8th International Conference on Family Business and Entrepreneurship

PREDICTING BANK LOAN APPROVAL USING LOGISTIC REGRESSION AND FEATURE SELECTION METHOD

Dadang Amir Hamzah [1], Fika Lestauli Sigalingging [2]

[1.]*Actuarial Science Program, President University, dadang.hamzah@president.ac.id*

[2]*Actuarial Science Program, President University, fika.sigalingging@student.president.ac.id*

ABSTRACT

Examining bank loan application is a long process that requires a detail check in every stages. This process is important in the banking industry, as it directly impacts the bank's risk management and profitability. However, due to a long process in making decision the customer wait a long time to get the decision which results to the customers dissatisfaction. Therefore, to improve the examination process and provide a quick decision result, the more effective tools is required. Logistic regression is a machine learning method that able to predict the binary output based on the probability value. This method takes the value from the multiple regression method and convert it into probability value using the activation function called sigmoid function. This paper applies the logistic regression method to predict bank loan approvals based on several features considered as independent variables. This research uses the secondary data taken from www.kaggle.com. The model performance is measured using the confusion matrix that consist of accuracy, precision, recall, and F1 score. This research construct three models based on data type. The first model is constructed using numerical data only, the second model is constructed using categorical data only, and the third model is constructed by combining numerical data type and categorical data type. It is determined that the first model return 87.8% accuracy, 95.94% precision, 87.57% recall, and 91.56% F1 score. The second model return accuracy 69%, precision 96.81%, recall 69.87%, and F1 score 81.17%. Moreover, the return 86.6% accuracy, 96.81% precision, 85.64% recall, and 90.88% F1 score. Based on these results, it is concluded that the best method in processing loan bank application data is to use the third model that is the model that includes both categorical and numerical data type.

Keywords: *Machine Learning, Logistic Regression, Bank Loan Approval Prediction, Artificial Intelligence, Finance.*

1. Introduction

Financial institutions play an important role in business and development. Bank loan is one of the most important financial services that help business growth. Bank encounter a difficult problem in deciding whether a loan application is creditworthy or lead to a high risk. Analysing the loan application require a long and detail evaluation process. This process includes checking important applicant's profile such as work status, credit history, and income. This process results to a long waiting time for making decision and dissatisfaction from the applicants. Nowadays, banks are considering an advanced methods for analysing bank loan application in order to improve the process and reduce the time and risk using data-driven decision-making. This data-driven decision making requires an advance method such as machine learning approach that will help the banking industry to make the loan approval prediction. Using this prediction, Banks can work more effectively in managing their resources, concentrate their marketing efforts, and create financial products that focus on their customers' demands. This study utilizes a dataset consisting of historical loan application data from a financial institution. This dataset contains information on various factors such as the applicant's age, income, employment status, credit score, loan amount, loan purpose, and repayment history. logistic regression is a supervised learning algorithm used for binary classification problems, where the outcome variable has two possible classes, such as 1/0, yes/no, or true/false. Despite its name, it is a classification method, not a regression one, that models the probability of a binary outcome based on one or more predictor variables.

2. **Literature Review**

One of the earliest applications of logistic regression in the financial sector, particularly for loan approvals, is in credit risk modeling. (Altman, 1968) developed the Z-score model, which laid the foundation for using statistical techniques like logistic regression to assess credit risk. Later studies, such as Ohlson (1980), showed how logistic regression could model binary outcomes like loan default or approval based on borrower characteristics, such as income, credit history, and employment status. The logistic model's ability to estimate the probability of a binary outcome made it ideal for predicting whether an applicant would default on a loan or whether their application would be approved.

Many studies have explored the use of logistic regression specifically in predicting loan application approval. For example, Abdou & Pointon (2011) conducted a comprehensive analysis of credit scoring techniques and emphasized that logistic regression is a robust and reliable tool for predicting loan outcomes. In their research, they demonstrated that logistic regression models can efficiently predict loan application status using variables such as loan amount, duration of loan, age of the borrower, and credit score. The authors noted that logistic regression performed well compared to other machine learning algorithms due to its simplicity, interpretability, and performance in smaller datasets.

In the context of bank loan approval, logistic regression's performance highly depends on the feature selection process. Several studies, such as Tobback et. al. (2017), have pointed out that the inclusion of relevant predictors, such as income, employment history, and credit utilization ratio, plays a crucial role in improving the model's predictive accuracy. Their research emphasized that logistic regression's strength lies in identifying the most significant variables influencing loan application outcomes. Properly engineered features—such as credit score, debt-to-income ratio, and loan-to-value ratio—have been proven to significantly improve the performance of logistic regression models in predicting bank loan approval decisions.

Logistic regression is frequently used in the financial sector to assess loan eligibility. Studies such as Arowolo et al. (2022) highlight how logistic regression models help banks predict whether a loan application will be approved or rejected by analyzing customer attributes such as income, employment status, and credit score. The strength of logistic regression lies in its ability to predict probabilities and provide interpretability, allowing decision-makers to understand which factors contribute most to the likelihood of loan approval The results of the research will offer insightful information about how loan applications are evaluated. The logistic regression model developed in this study aims to provide support in decision-making regarding loan approval. For further study and advancement in this field, this will serve as the foundation.

Similarly, Bindal & Chaurasia (2018)compared logistic regression with decision tree-based methods and found that logistic regression offers robust performance in predicting loan approval, especially when combined with techniques to handle class imbalance and regularization.

A significant area of focus in recent research is the impact of feature engineering on logistic regression performance. According to Calibo & Ballera (2019), properly selecting and transforming features such as credit utilization, debt-to-income ratio, and loan-to-value ratio can dramatically improve the accuracy of logistic regression models. Feature selection and data preprocessing techniques, including scaling and handling missing values, play a critical role in the success of these models.

Class imbalance is a common challenge in loan approval datasets, where approved loans typically outnumber rejected ones. Zhou et al. (2023)explored the use of logistic regression in highly imbalanced datasets and demonstrated that resampling techniques, such as Synthetic Minority Over-sampling (SMOTE), improve the model's performance by balancing the classes Recent research trends have also focused on hybrid models that combine logistic regression with other machine learning algorithms. For instance, Xiao et al. (2022) introduced a hybrid model that uses logistic regression as a base classifier, enhanced with imbalanced learning techniques to predict loan defaults more accurately. This approach improved the predictive power of the model without sacrificing interpretability.

Hamzah (2021)applied logistic regression for predicting the travel insurance policy claims. The study uses logistic regression as a classification tool to assess the probability that an insurance claim will be made based on a set of input features, such as demographic information and travel-related factors. The logistic regression model successfully identifies key factors that affect the likelihood of filing a claim. For instance, certain characteristics, such as longer trip duration or travel to riskier destinations, might increase the probability of a claim. The model's performance is evaluated in terms of its accuracy and ability to predict claims correctly.

Kannan et al. (2023)discusses the development and implementation of a loan approval prediction system using various machine learning algorithms. The primary aim of the study is to enhance the accuracy of loan approval predictions, ensuring that loans are granted only to eligible applicants who are likely to repay, thereby reducing the risk of default for financial institutions. The results of the simulation showed improvements in prediction accuracy across several machine learning algorithms. Logistic Regression, Decision Tree, SVM, and Naive Bayes achieved performances of 86%, 74%, 86%, and 86%, respectively, which outperformed results from previous studies.

Singh Sandhu et al. (2023) address the critical role of machine learning (ML) in automating the loan approval process in banks. The researchers emphasize the importance of accurate loan approval predictions to minimize the risk of defaults and ensure that loans are approved for applicants capable of repayment. The study demonstrates that using machine learning models leads to improvements in prediction accuracy. Logistic Regression, Decision Tree, SVM, and Naive Bayes all performed well, with Random Forest achieving the highest accuracy at 86%. This marks an improvement over previous models, which had shown lower predictive performance in similar settings.

Kumar et al. (2019)applies eight different algorithms were utilized to train the models, including Logistic Regression, Random Forest, Decision Trees, Linear Regression, Support Vector Machine (SVM), Naïve Bayes, K-means, and K Nearest Neighbors (KNN) for predicting loan approval. The final results indicated that the models produced varying outcomes. Across both datasets, Logistic Regression achieved 83.24% and 78.13% accuracy, followed by Naïve Bayes with 82.16% and 77.34% accuracy. Random Forest also performed well, highlighting its effectiveness in the loan approval prediction process.

Ekadjaja (2018)focuses on identifying key factors that influence the approval of bank loans for MSMEs. MSMEs play a crucial role in Indonesia's economic growth, but they often face challenges in accessing formal financial sources such as bank loans. This study attempts to address these challenges by investigating the variables that determine bank loan approval for MSMEs operating in Tanah Abang, Jakarta. The study finds that Business Age since Establishment, Total Assets Turnover, Owner's Education Level, Collateral Amount, and Loan Repayment Criteria have a positive and significant effect on bank loan approval. However, Owner's Total Assets, Credit Duration, and Good Relationship with the Bank do not significantly impact loan approval.

3. Research Method

This section is summarized from Peng et al. (2002). The process of applying logistic regression for predicting the bank loan application is conducted throught the following steps.

Step 1 – Data Preprocessing

Loan Application Data sourced from the Kaggle website (J. R. Quinlan, 1987). It consists of 12 independent variables, including Gender, Married, Dependents, Education, Self-employment, Applicant income, Co-applicant income, Loan amount, Loan amount term, Credit history, Property area, and Total income. The dependent variable is Loan Status that consist of "Yes" or "No" answer. The study employs a total of 6000 datasets for analysis.

Cleaning Data

When using secondary data, not all of them provide complete data information. Some data had no values. Missing values are shown as NA (not available). This NA data should be removed because it can interfere with the analysis process and result in inaccurate prediction results. For this reason, the researcher created a function to remove NA values from the data. Next, the researcher generates dummy variables to represent categorical attributes for use in the regression model. Subsequently, the data type is transformed to a binomial format, with values of either 0 or 1. The researcher employs the ifelse function to convert multiple variables into dummy variables.

Data Transformation

Researchers convert characteristic factors with character data types into factors by using the as.factor function. Factor types are used to group data into different categories in no particular order. While integer is used to represent numeric values that have an order or can be counted. The variables that are transformed include gender, married, dependent, education, self-employed, property area, and loan status. This conversion is crucial for accurate data analysis and modeling, as it ensures that categorical variables are properly treated. By converting these variables, researchers can perform more precise statistical analyses and improve the reliability of their results.

Split the data into Training and Testing

The data is divided into two categories for the analysis using the logistic regression method: training data and testing data. Training data is a dataset that has already been gathered and has all known properties, including the target class property. On the other hand, the classification rules that are obtained from the training data are assessed using the testing data. In order to avoid overfitting, this division guarantees that the model is trained on one set of data and verified on a different set. By doing so, researchers can assess the model's performance and generalizability to new, unseen data. In this research we allocate 70% for tracing data and 30% for testing data.

Step 2 – Construct the Logistic Regression Model

Begin with initial estimates for the coefficients $\beta_0, \beta_1, ..., \beta_n$. These values can either be zeros or small random numbers. Next, For each training example$i$, calculate the linear combination z_i that is

$$z_i = \beta_0 x_0 + \beta_1 x_1 + ... + \beta_n x_n \qquad (1)$$

Step 3 – Apply the sigmoid function

To determine the predicted probability $\hat{y}_i$ apply the sigmoid function that is

$$\hat{y}_i = \sigma(z_i) = \frac{1}{1+e^{-z_i}} \qquad (2)$$

This function will convert the value from Step 2 into the value between 0 and 1.

Step 4 – Set the Decision Rule

Once the value is converted the next step is to determine the decision rule. The decision rule for this problem is if $z_i < 0.5$ the predicted value is 0 or “No”. On the other hand if $z_i \geq 0.5$ the predicted value is 1 or “Yes”.

Step 5 – Evaluate the Prediction Value

The prediction value is evaluated using the loss fucntion. The loss function is calucalting the difference between the actual value with the prediction value. The more smaller the loss function then the better the prediction. The loss function for this process is defined as

$$J(\beta) =- \frac{1}{m}\sum_{i=1}^{m}\left[y_i \log log\,(\hat{y}_i) + \left(1 - y_i\right) \log log\,(1 - \hat{y}_i)\right] \qquad (3)$$

where m denotes the number of samples, y_i is the actual value and $\hat{y}_i$ is the predicted value.

Step 6 – Optimize the Model (Gradient Descent)

To improve the model, each β_i is step 2 is updated using the method called the method of Gradient Descent. This method updates the β by reducing it with the gradient of the loss function multiplied with the learning rate α. The gradient of the loss function $J(\beta)$ is determined using the following formula

$$\frac{\partial J(\beta)}{\partial \beta_i} = \frac{1}{m}\sum_{i=1}^{m}\left(y_i - \hat{y}_i\right)x_{ij}$$

Where x_{ij} is the j −th feature of the i −th training example. Then, the formula of updating the β value is as follows

$$\beta_i^{(new)} = \beta_i^{(old)} - \alpha\frac{\partial J(\beta)}{\partial \beta_i} \tag{4}$$

The learning rate α determines the step size for each update during the optimization process. In this research we use $\alpha = 0.01$. This step 1 to step 6 is repeated for n-times or until the desired value of loss function is achieved. In this research we repeat the process until $n = 1000$ times.

Step 7 – Model Testing and Evaluation

Once the lost function is optimized, the next step is to test the model by applying it to the traing data.
Then the model performance is evaluated using the following metrics.

Accuracy is an evaluation metric used in classification problems to measure how well a model has performed. It represents the proportion of correct predictions made by the model out of the total number of predictions. In the context of binary classification, accuracy can be calculated using the confusion matrix, which includes True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). The formula is

$$Accuracy = \frac{TP+TN}{Total\ Data} \tag{5}$$

Precision is a performance metric used in classification problems, especially binary classification. It focuses on the positive predictions made by a model and answers the question: "Of all the positive predictions made by the model, how many are actually correct?". Precision measures the model's accuracy in predicting positive outcomes. A high precision means that when the model predicts a positive class, it is correct most of the time, meaning fewer false positives. Precision is particularly important when false positives are costly or undesirable. The formula of Precision is

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

Recall is a performance metric used in classification problems. It measures the model's ability to correctly identify all relevant instances of the positive class. Specifically, recall answers the question: "Of all the actual positive instances, how many did the model correctly identify?". Recall focuses on identifying all the actual positive cases. A high recall means that the model is good at detecting the positive class and has fewer false negatives. In other words, it tries to minimize the cases where actual positives are missed by the model. The recall formula is

$$Recall = \frac{TP}{TP+FN} \tag{7}$$

F1-Score is a performance metric used in classification problems, particularly when dealing with imbalanced datasets. It is the harmonic mean of **precision** and **recall** and provides a balanced measure that considers both false positives and false negatives. The F1 score is especially useful when you need to find a balance between precision (how many of the predicted positive instances are correct) and recall (how many of the actual positive instances were correctly identified). The formula of F1 score is

$$F1\,Score = \frac{2TP}{2TP+FP+FN} \tag{8}$$

4. Results and Discussion

When using secondary data, not all of them provide complete data information. Some data had no values. Missing values are shown as NA (not available). This NA data should be removed because it can interfere with the analysis process and result in inaccurate prediction results. For this reason, the researcher created a function to remove NA values from the data. Next, the researcher generates dummy variables to represent categorical attributes for use in the regression model. Subsequently, the data type is transformed to a binomial format, with values of either 0 or 1. The researcher employs the ifelse function to convert multiple variables into dummy variables.

Researchers convert characteristic factors with character data types into factors by using the as.factor function. Factor types are used to group data into different categories in no particular order. While integer is used to represent numeric values that have an order or can be counted. The variables that are transformed include gender, married, dependent, education, self-employed, property area, and loan status. This conversion is crucial for accurate data analysis and modeling, as it ensures that categorical variables are properly treated. By converting these variables, researchers can perform more precise statistical analyses and improve the reliability of their results.

The data is divided into two categories for the analysis using the logistic regression method: training data and testing data. Training data is a dataset that has already been gathered and has all known properties, including the target class property. On the other hand, the classification rules that are obtained from the training data are assessed using the testing data. In order to avoid overfitting, this division guarantees that the model is trained on one set of data and verified on a different set. By doing so, researchers can assess the model's performance and generalizability to new, unseen data.

In this analysis, researchers used all independent variables including Gender, Married, Dependents, Education, Self-employment, Applicant income, Co-applicant income, Loan amount, Loan amount term, Credit history, Property area, and Total income. Table 1 displays the prediction values, indicating that 108 records with a rejected loan status were properly expected to be in that position, whereas 14 records with a rejected loan status were mistakenly forecasted to be in that state since they were anticipated to be approved for loan status. Then, 331 records with an authorized loan status were accurately projected to be in such status, whereas 47 records had an inaccurate prediction.

Table 1: Confusion matrix of Case 1

	Actual	
Prediction	0	1
0	108	407
1	14	331

The accuracy rate of 87.8% indicates the effectiveness of using loan status data with the logistic regression model. Additionally, the F1 score stands at 91.56%, reflecting that the classification model performs well in terms of both precision and recall. The precision is 95.94% and the recall is 87.57%.

In this analysis, researchers used only categorical independent variables including Gender, Married, Dependents, Education, Self-employment, and Property area. Table 2 displays the prediction values, indicating that 11 records with a rejected loan status were accurately expected to be in that status, whereas 11 records with an erroneously forecasted state were anticipated to be accepted for loan status. Then, 334 records with an authorized loan status were predicted accurately, whereas 144 records had an inaccurate approved loan status prediction.

Table 2: Confusion Matrix of Case 2

	Actual	
Prediction	0	1
0	11	144
1	11	334

The accuracy rate of 69% indicates the effectiveness of using loan status data with the logistic regression model. Additionally, the F1 score stands at 81.17%, reflecting that the classification model performs well in terms of both precision and recall. The precision is 96.81% and the recall is 69.87%.

In this analysis, researchers used only numerical independent variables including Applicant income, Co-applicant income, Loan amount, Loan amount term, Credit history, and Total income. Table 3 displays the prediction values, indicating that 99 records with a rejected loan status were accurately expected to be in that state, whereas 11 records were mistakenly forecasted to be in that position since they were anticipated to be accepted. Next, 334 records with an authorized loan status were accurately predicted, whereas 56 records had an inaccurate approved loan status prediction.

Table 3: Confusion Matrix of Case 3

	Actual	
Prediction	0	1
0	99	56
1	11	334

The accuracy rate of 86.6% indicates the effectiveness of using loan status data with the logistic regression model. Additionally, the F1 score stands at 90.88%, reflecting that the classification model performs well in terms of both precision and recall. The precision is 96.81% and the recall is 85.64%.

Based on the logistic regression analysis, below is the confusion matrix comparison table between the above 3 cases. Table 4 data shows that the first example has the greatest values for Accuracy, Precision, Recall, and F1 score. This data indicates that the best results are achieved when all dependent variables are processed.

Table 4: Coparison results of Confusion Matrix.

Evaluation	Case 1: Combination of Categorical and Numerical Variables	Case 2: Only Categorical data type	Case 3: Only Numerical data type
Accuracy	87.8%	69%	86.6%
Precision	95.94%	96.81%	98.81%
Recal	85.87%	69.87%	85.64%
F1 Score	91.56	81.17%	90.88%

5. Conclusion and Implications

According to the logistic regression analysis, factors such as Education, Co-Applicant Income, Loan Amount Term, Credit History, and Property Area were found to have a significant impact on loan status. Higher levels of education tend to increase the chances of loan approval, while a higher co-applicant income also positively contributes to approval. Additionally, a shorter loan amount term and a good credit history are factors that support loan approval.

Based on the results of confusion matrix analysis that processes data using combination of categorical and numerical variables, the accuracy value is 87.8%, precision 95.94%, recall 87.57%, and F1 score 91.56%. The results obtained in data processed with categorical variables are accuracy 69%, precision 96.81%, recall 69.87%, and F1 score 81.17%. While the results for numerical variables data are accuracy 86.6%, precision 96.81%, recall 85.64%, and F1 score 90.88%. These data suggest that the best method for processing bank loan application data is to use logistic regression while including the combination of categorical and numerical variables.

In conclusion, the Logistic Regression model exhibits strong performance in accurately predicting loan approval outcomes based on the provided data. Its high precision, recall rates, and F1 score further validate its effectiveness in assessing bank loan approvals. This suggests that the Logistic Regression model can be a valuable tool for processing and analyzing data related to bank loan approvals. The recommendation expected in this research is that credit companies/banks pay close attention to the factors that can influence the risk of debtor default such as financial stability, economic conditions, borrower characteristics, loan specifics, and external factors and carry out a detailed analysis of their performance in managing credit risk. In this way, it is hoped that credit companies/banks can increase the effectiveness of their credit risk mitigation strategies and maintain the health of their credit portfolios.

References

Abdou, H. A., & Pointon, J. (2011). CREDIT SCORING, STATISTICAL TECHNIQUES AND EVALUATION CRITERIA: A REVIEW OF THE LITERATURE. *Intelligent Systems in Accounting, Finance and Management*, *18*(2–3). https://doi.org/10.1002/isaf.325

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, *23*(4). https://doi.org/10.2307/2978933

Amir Hamzah, D. (2021). Predicting travel insurance policy claim using logistic regression. *Applied Quantitative Analysis*, *1*(1). https://doi.org/10.31098/quant.613

Arowolo, M. O., Adeniyi, O. F., & ... (2022). A Prediction Model for Bank Loans Using Agglomerative Hierarchical Clustering with Classification Approach. *Covenant ...*, *10*(2).

Bindal, A., & Chaurasia, S. (2018). Predictive risk analysis for loan repayment of credit card clients. *2018 3rd IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology, RTEICT 2018 - Proceedings*. https://doi.org/10.1109/RTEICT42901.2018.9012366

Calibo, D. I., & Ballera, M. A. (2019). Variable selection for credit risk scoring on loan performance using regression analysis. *2019 IEEE 4th International Conference on Computer and Communication Systems, ICCCS 2019*. https://doi.org/10.1109/CCOMS.2019.8821664

Ekadjaja, M. (2018). FACTORS DETERMINING BANK LOAN APPROVAL AS SOURCE OF FINANCING FOR MICRO, SMALL, AND MEDIUM ENTERPRISES (MSME) IN JAKARTA. *Jurnal Muara Ilmu Ekonomi Dan Bisnis*, *2*(1). https://doi.org/10.24912/jmieb.v2i1.1563

J. R. Quinlan. (1987). Credit Approval [Dataset]. In *UCI Machine Learning Repository*. UCI Machine Learning Repository.

Kannan, M. K. J., Kannan, J., Nithej, A. R., Akhil, P., Gowda, P., & Pareek, P. (2023). ML Based Loan Approval Prediction System A Novel Approach. *International Journal of Innovative Research in Computer and Communication Engineering | An ISO*, *11*(3).

Kumar, R., Jain, V., Sharma, P., Awasthi, S., & Jha, G. (2019). Prediction of loan approval using machine learning. *International Journal of Advanced Science and Technology*, *28*(7). https://doi.org/10.56726/irjmets41180

Ohlson, J. A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, *18*(1). https://doi.org/10.2307/2490395

Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *Journal of Educational Research*, *96*(1). https://doi.org/10.1080/00220670209598786

Singh Sandhu, H., Sharma, V., & Jassi, V. (2023). Loan Approval Prediction Using Machine Learning. In *Emerging Trends In Engineering And Management*. https://doi.org/10.56155/978-81-955020-3-5-01

Tobback, E., Bellotti, T., Moeyersoms, J., Stankova, M., & Martens, D. (2017). Bankruptcy prediction for SMEs using relational data. *Decision Support Systems*, *102*. https://doi.org/10.1016/j.dss.2017.07.004

Xiao, F., Chen, W., Wang, J., & Erten, O. (2022). A Hybrid Logistic Regression: Gene Expression Programming Model and Its Application to Mineral Prospectivity Mapping. *Natural Resources Research*, *31*(4). https://doi.org/10.1007/s11053-021-09918-1

Zhou, Y., Chen, S., Zhong, Y., & Deng, X. (2023). Study on Learning Method of Logistic Regression Classification for Class Imbalance Problem. *Learning and Analytics in Intelligent Systems*, *31*.

https://doi.org/10.1007/978-3-031-29016-9_34