Analysis of the default data set Credit card client data Using the Decision Tree Algorithm and Logistics Regression Algorithm

Muhamad Calvin Syah Putra
Program Sistem Informasi, Universitas Multimedia Nusantara, Tangerang, Indonesia
muhamad.putra@student.umn.ac.id

***Abstract*— A default is a serious credit card status that affects not only your position with that credit card issuer but also your credit status in general and your ability to obtain approval for other credit-based services. This study aims at the case of customer payments in Taiwan and the accuracy of estimates of the probability of default among the six data mining methods. Since the actual probability of default is unknown, this study is to estimate the true probability of default.**

Index Terms - Data Mining, default of credit card clients Data Set, Decision tree, Logistics Regression.

I. INTRODUCTION

A. Background of the study

In times of economic crisis, consumers who are negatively affected (due to job loss or other circumstances) must prioritize how they spend their money. In these situations, credit card payments are often at the bottom of the list. And if you get to the bottom and find there's nothing left to give, the card probably won't get paid. If enough time has passed, you can then enter the credit card's default state.

Other necessities, such as prescriptions or diapers, may also be your top priority. If you are unemployed and looking for work (or planning to work in the future), your list should also include keeping your credit report clean for potential employers who may withdraw credit reports as part of the hiring process.

Whatever your priorities, it's easy to see what's important when listed and why credit card debt seems to come last. It's not that debt isn't important; it's just that other things have to come first. Still, it would be a big mistake not to make a payment if you can pay the minimum. Credit is a powerful tool and having it for you when things get back to normal is smart.

If you don't pay your credit card bill for a long time, your card may go into default status. Default is different from overdue

payments. You are late after 30 days of no payment. Default usually occurs after six consecutive months of not paying the minimum payment due. Your credit score will be negatively affected. When your account is debited, the damage increases. Late payments are not reported immediately, for example if you are not due for several days. This usually happens when you are at least 30 days late.

Credit card defaults occur when you become very delinquent on your credit card payments. A default is a serious credit card status that affects not only your position with that credit card issuer but also your credit status in general and your ability to get approved for other credit-based services.

In the months leading up to default, your (late) payment status will be reported to the three major credit bureaus, and your credit score will be affected by your late payment. If you apply for a new credit card or loan after default, your application is likely to be rejected because the lender thinks you are at risk of defaulting on your new credit obligations. In fact, some lenders won't approve you at all until you settle the default balance (or your credit report goes down).

B. Problem

This study aims at the case of customer payments in Taiwan and the accuracy of the default probability estimates among the six data mining methods. From a risk management perspective, the predictive results of the estimated probability of default will be more valuable than the classification results - credible or non-credible clients.

Since the actual probability of default is unknown, this study is to estimate the true probability of default. With the real probability as the response variable (Y), and the default prediction probability as the independent variable (X), then the results of simple linear regression (Y = A + BX) show that the forecasting model generated by the artificial neural network has the highest coefficient of determination; the regression intercept (A) is close to zero, and the regression coefficient (B) is close to one. Therefore, among the six data mining techniques, the artificial neural network is the only one that can accurately estimate the true default probability Units

II. STUDY LITERATURE

A. Decision tree

Decision tree is a classification method that uses a tree structure, where each node or node represents an attribute and the branch represents a value from the attribute, while the leaves are used to represent a class. The top node of the decision tree algorithm is called the root.

Breiman et al. (1984) stated that the decision tree method is a very popular method

to use because the results of the model formed are easy to understand. Another name for the decision tree is CART (Classification and Regression Tree). This method is a combination of two tree species, a classification tree and a regression tree. And that's why this method is called a decision tree, because the rules formed are similar to the shape of a tree. The tree is formed from a binary recursive sorting process in data groups, so that the value of the response variable in each data group makes the sorting results more homogeneous.

The concept of this decision tree is to convert data into decision trees and decision rules. One of the main benefits resulting from the use of decision trees is its ability to simplify complex decision-making processes so that decision-makers can interpret solutions to problems. There are three types of nodes in the decision tree, namely:

1) Root

   The root is the top node, there is no input to this node and there is no output or it can have more than one output.

2) Internal Node

   This internal node is a branching node, there is only one input and at least two outputs in this node.

3) Leaf

   Leaf is the end node or terminal node, there is only one input and no output (end node) at this node.

The stages of decision tree formation:

1) Construction

   Tree begins with the formation of roots (located at the top). Then the data is broken down using attributes suitable for use as sheets.

2) Tree pruning

   Namely determine and remove branches that are not needed on the tree that has been formed. This is because the decision tree that can be made can be large so it can be simplified by pruning based on the confidence value (confidence level). Tree planting is done in addition to reducing the size of the tree to also reduce prediction errors in new cases of solving results and solutions.

3) Formation of decision rules

   Namely making decision rules from the tree that has been formed. The rule can be in the form of an if-then derived from the decision tree by tracing from root to leaf. For each node and branch, if specified, then the sheet value is entered. After all the rules are created, the rules can be simplified or combined.

Benefits of Decision Tree:

Decision trees are also useful for exploring data, finding hidden relationships between a number of potential input variables and a target variable. Decision trees combine data exploration and modeling which makes

them an excellent first step in the modeling process even when used as a final model for several other techniques.

Advantages and Disadvantages of Decision Tree

Advantages of Descion Trees :

- Easy integration into database systems.
- Has good accuracy.
- Can find unexpected combinations of data.
- Decision areas that were previously complex and highly global can be made simpler and more specific.
- Can eliminate unnecessary calculations. Because with this method, the sample is only tested based on certain criteria or classes.
- With flexible feature selection from different internal nodes, the selected feature distinguishes the criteria from other criteria in the same node.

Disadvantages of Decision Trees:

- Overlap occurs especially when very many classes and criteria are used. It can also lead to longer decision times and memory requirements.
- The accumulation of the number of errors from each level in a large decision tree.
- Difficulty in designing an optimal decision tree.
- The results of the decision quality obtained by the decision tree method are highly dependent on how the tree is designed.

Decision Tree Consists of Three Types of Nodes, namely:

1. Decision nodes – usually represented by boxes
2. Opportunity vertices – usually represented by circles
3. End vertices – usually represented by a triangle

Application of Decision Tree:

A decision tree is a decision support tool that uses a decision model that is shaped like a tree. The decision tree describes the various possible alternatives to solve a problem and there are also potential factors that can affect the alternative along with the final estimate when an alternative is chosen. A decision tree is a method that can be used to display algorithms that contain only conditional control statements.

The use of this decision tree is generally in operations research, especially in decision analysis. The purpose of using a decision tree is to identify the strategies that are most likely to achieve a goal and it is a popular tool in machine learning.

A decision tree is a flowchart-like structure in which each internal node represents a possible attribute, each branch represents the outcome of that probability and each leaf node represents a class name (decisions are made after all attributes have been computed). The path from root to leaf represents the classification rule.

In decision analysis, decision trees and related diagrams are used as visual and analytical decision support tools, where the expected value or utility of the alternatives will be calculated.

B. Logistics Regression

Logistics Regression is a statistical analysis method to describe the relationship between the response variables (dependent variable) which has two categories or more with one or more explanatory variables (independent) variable) categorical or interval scale. Logistics Regression is a regression non-linear, used to explain the relationship between X and Y which are non-linear, non-normal distribution Y, response diversity is not constant that cannot be explained by a linear regression model normal.

$$Y = X1 + X2 + X3 + \ldots + Xn$$

(biner non-metric) (non-metric or metric)

Logistic regression objective:

- The first objective (explanation) has a focus on identifying the independent variables that affect group membership in the dependent variable. The results of this analysis will provide an understanding of the 'reasons' why each observation belongs to one group and not another.

- The second objective is to develop a classification systema logistic-based model that can predict group membership from an observation. Thus the measure of prediction accuracy is classification accuracy (not R2, as in multiple regression).

Example of logistic regression application

Logistics Regression can be used for various applications in various fields such as:

- Predicting the success or failure of a new product.
- Deciding which customer to get credit or not.
- Predicting a startup will be successful or not.
- Predicting visitors will buy or not.
- as well as other classifications or predictions of a binary nature.

Types of Logistic Regression

Binary Logistic Regression: This is a Logistic Regression which only has 2 outputs (classifying into 2 different classes). Example: Positive-Negative, Obesity-Not Obesity.

Multinomial Logistic Regression: Is a Logistic Regression that has 2 or more outputs (classifies into 2 different classes). For example, Sentiment Analysis class positive, negative, and neutral sentences.

Ordinal Logistic Regression: Is a Logistic Regression that has 2 or more outputs with respect to the order. (classifying into 2 different classes by paying attention to

the order). An example is dividing the class of students in the range of Grade Point Average 1.xx, 2.xx, 3.xx, and 4.00.

Logistic Function is a function formed by equating the Y value in Linear Function with Y value in Sigmoid Function. The purpose of the Logistic Function is to represent the data that we have in the form of a Sigmoid function.

We can form a Logistic Function by performing the following steps:

- Perform the Inverse operation on the Sigmoid Function, so that the sigmoid function changes its shape to Y = ln(p/(1-p).

- Equalize it with the Linear function Y = b0+b1*X so that we get the equation ln(p/(1-p) = b0+b1*X.

- Change the equation ln(p/(1-p) = b0+b1*X into logarithmic form so that the equation P = 1/(1+e^-(b0+b1*X))

III. METODOLOGI

A. Object Penelitian

This study aims at the case of customer payme This study focuses on processing customer default payment case data in Taiwan and compares the prediction accuracy of default probability among six data mining methods. From a risk management perspective, the predictive accuracy results from the probability of default estimates will be more valuable than the results of the classification binary - credible or non-credible clients.

The dataset that will be used for the purposes of this study includes the accumulation of default customer payments in Taiwan from a risk management perspective. The characteristics of the dataset are Multivariate, containing 30000 number of instances. and has the characteristics of Integer and Real attributes.

B. Method of collecting data

The data used is secondary data, namely data that was previously collected from other parties. secondary data collected from the UCI Machine Learning Repository website. UCI is a collection of databases, domain theory, and data generators used by the machine learning community to analyze algorithms.

C. Research Method

The research method is the process by which researchers solve the problems posed in research activities. Therefore, it can be said that the research method is the main way used by researchers to achieve goals and obtain answers to the problems being carried out. In this study, the researcher uses a quantitative design methodology that relies on the ability to analyze data so that the managed data can be of added value for the decision-making process of certain parties.

Data processing will be carried out using R.R which is a programming language and also a calculation program used to support statistical and graphic analysis activities. R is accessed using RStudio. RStudio is an integrated development environment (IDE) for R. In addition, before processing the data, the data must first be validated to ensure the accuracy and certainty of the data to be used in research. R will also facilitate this validation function.

IV. DATA ANALYSIS AND DISCUSS

A. Data Visualization

Data *visualization* adalah tampilan berupa grafis atau visual dari informasi dan data. Dengan kata lain, data *visualization* mengubah kumpulan data menjadi hal lebih sederhana untuk ditampilkan. Dengan menggunakan elemen visual tersebut, pembaca akan lebih mudah memahami tren, *outliers*, dan pola dalam suatu data. Dalam bisnis, data *visualization* memungkinkan para pembuat keputusan untuk melihat analitik yang disajikan secara visual. Dengan begitu, mereka dapat memahami konsep yang sulit atau mengidentifikasi pola baru. Hal ini akan membuat pengambilan keputusan menjadi lebih mudah dan tepat.

1. Visualization Data with Boxplot

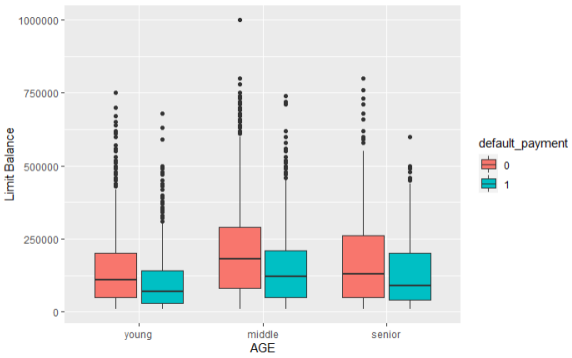

Figure 1.Visualization Data with Boxplot

2. visualization data with barplot

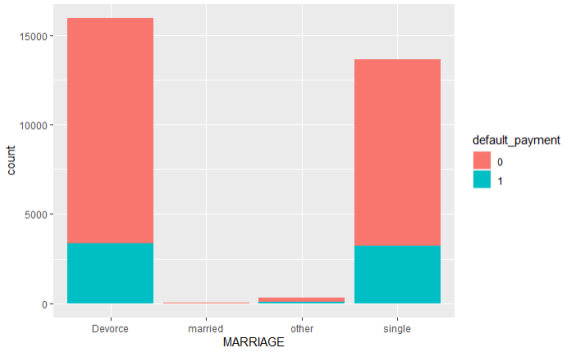

Figure 2.Visualization Data with Boxplot

3. visualization data dengan scale fill

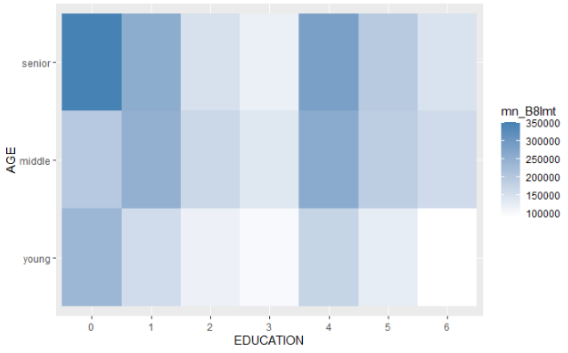

Figure 3.Visualization data dengan scale fill

B. Application of the Decision Tree Algorithm.

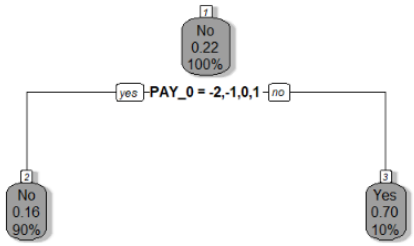

Figure 4. Plot Tree Rpart

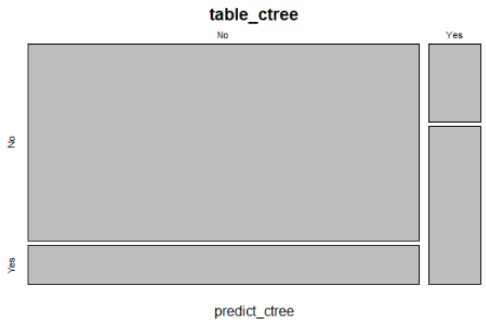

Figure 5. Table_ctree

C. Application of the Logistic Regresion Algorithm.

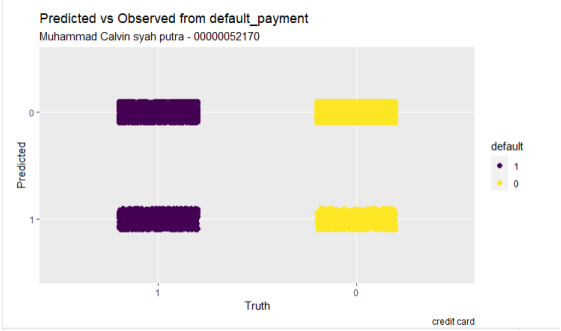

Figure 6. Predicted vs Observed from default_payment

D. Comparison of the two models

An accuracy rate above 50% indicates that the model can be used with a higher degree of accuracy than random guessing. A higher level of accuracy than others will indicate which model is better to use in real-life predictions.

```
y_pred     1     0
     1  1399   654
     0  2549 13398

               Accuracy : 0.8221
                 95% CI : (0.8164, 0.8276)
    No Information Rate : 0.7807
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.372

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.35436
            Specificity : 0.95346
         Pos Pred Value : 0.68144
         Neg Pred Value : 0.84016
             Prevalence : 0.21933
         Detection Rate : 0.07772
   Detection Prevalence : 0.11406
      Balanced Accuracy : 0.65391

       'Positive' Class : 1
```

Figure 7. logistic regresion

```
predict_ctree   No  Yes
          No  8846 1747
          Yes  466  941

               Accuracy : 0.8156
                 95% CI : (0.8085, 0.8225)
    No Information Rate : 0.776
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3613

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9500
            Specificity : 0.3501
         Pos Pred Value : 0.8351
         Neg Pred Value : 0.6688
             Prevalence : 0.7760
         Detection Rate : 0.7372
   Detection Prevalence : 0.8828
      Balanced Accuracy : 0.6500

       'Positive' Class : No
```

Figure 8. Decision tree

E. Conclusion and Suggestions

In general, based on data analysis activities of customer default payment cases in Taiwan using two algorithms that have been described in the results analysis section, it can be concluded that the Logistic Regression algorithm is better used in analyzing this data because of its very high accuracy compared to the Dicision tree algorithm. The following is a description of the conclusions of the Logistic Regression and Dicision tree algorithms.

1. Logistic Regression

Confusion Matrix	Logistic Regression
Accuracy	0.8221
kappa	0.372
95%CI	(0.8164, 0,8276)

2. Dicision Tree

Confusion Matrix	Dicision Tree
Accuracy	0.8156
kappa	0.3613
95%CI	(0.8085, 0.8225)

It can be seen from the confusion matrix, using the Logistics Regression model is more complex than using the decision tree model. First, Logistic Regression has higher accuracy than decision trees. second, the 95% CI value of Logistic Regression is also greater than the 95% CI value of the decision tree, although the kappa value of the decision tree is larger. It can be concluded that the Logistics Regression model is more suitable to be used to compare the default probability prediction accuracy between the six data mining methods. To estimate the true default probability.

REFERENSI

Asfihan, A. (2021, July 28). *Decision Tree Adalah : Jenis, Manfaat, Kelebihan dan Kekurangannya*. Diambil kembali dari Adalah.Co.Id: https://adalah.co.id/decision-tree/

Bucci, S. (2021, Februari 11). *Credit card default: How it happens, what to do about it*. Diambil kembali dari Bankrate: https://www.bankrate.com/finance/credit-cards/credit-card-default/

Dr. Ir. Samuel Tarigan, M. &. (2018). *LOGISTIC REGRESSION*. Diambil kembali dari https://elearning.ithb.ac.id/pluginfile.php/15300/mod_folder/content/0/Logistic%20Regression%20v6%2026%20Januari%202019%20VJ%20ST.pdf?forcedownload=1

IRBY, L. (2021, Juni 29). *What You Can Do About Credit Card Default*. Diambil kembali dari The Balance: https://www.thebalance.com/what-is-credit-card-default-960209

Pradana, A. (2021, Januari 18). *Data Visualization: Cara Tampilkan Data agar Mudah Dipahami*. Diambil kembali dari glints.com: https://glints.com/id/lowongan/data-visualization-adalah/#.YV78EdpBzDd