COSTING: Journal of Economic, Business and Accounting
Volume 7 Nomor 6, Tahun 2024
e-ISSN : 2597-5234

CREDIT CARD DEFAULT PREDICTION USING MACHINE LEARNING

PREDIKSI GAGAL BAYAR KARTU KREDIT MENGGUNAKAN PEMBELAJARAN MESIN

Kevin Naufal Widyadhana
Universitas Pembangunan Nasional "Veteran" Jakarta 60111 Indonesia,
kevinnaufal@upnvj.ac.id

ABSTRACT

The credit card industry had been around for decades and is a product of changing consumer habits also increasing the national income. There has been a significant increase in the number of card issuers, issuing banks to transaction volumes. However, with the increase in credit card transactions, the amount due and the arrears rate of credit card loans are also issue that cannot ignore. This issue is crucial for the successful development of the banking industry in the future. The study focused on modeling and predicting an individual's willingness to repay credit card loans. The methods used in this study are machine learning with random forest approach, artificial neural network, support vector machine, logistic regression, and naïve Bayes. There are 11 variables to be analyzed in this study, and the performance of the five methods will be compared to the evaluation of ROC, and AUC. The result of this research as follows. The random forest method is considered the most appropriate for processing the credit card default dataset with AUC 89%. This model can contribute to the settlement of default probabilities and is of great help to the credit card industry. Based on the PDP, managerially it can be determined that for income and credit card limits the range of 7-50 million is more prone to default
***Keywords**: Credit Card, Classification, Default Prediction, and Imbalance Data*

ABSTRAK

Industri kartu kredit telah ada selama beberapa dekade dan merupakan produk dari perubahan kebiasaan konsumen serta peningkatan pendapatan nasional. Telah terjadi peningkatan yang signifikan dalam jumlah penerbit kartu, bank penerbit hingga volume transaksi. Namun, dengan meningkatnya transaksi kartu kredit, jumlah tagihan dan tingkat tunggakan pinjaman kartu kredit juga menjadi isu yang tidak bisa diabaikan. Hal ini sangat penting untuk keberhasilan pengembangan industri perbankan di masa depan. Penelitian ini berfokus pada pemodelan dan prediksi kemauan seseorang untuk melunasi pinjaman kartu kredit. Metode yang digunakan dalam penelitian ini adalah machine learning dengan pendekatan random forest, jaringan syaraf tiruan, support vector machine, regresi logistik, dan naïve bayes. Terdapat 11 variabel yang akan dianalisis dalam penelitian ini, dan kinerja kelima metode tersebut akan dibandingkan dengan evaluasi ROC, dan AUC. Hasil dari penelitian ini adalah sebagai berikut. Metode random forest dinilai paling tepat untuk mengolah dataset gagal bayar kartu kredit dengan AUC 89%. Model ini dapat memberikan kontribusi dalam penyelesaian probabilitas gagal bayar dan sangat membantu industri kartu kredit. Berdasarkan PDP, secara manajerial dapat ditentukan bahwa untuk pendapatan dan limit kartu kredit kisaran 7-50 juta lebih rentan mengalami gagal bayar.
Kata Kunci: Kartu Kredit, Klasifikasi, Prediksi Gagal Bayar, Dan Data Imbalance

INTRODUCTION

Payment failure (default) on credit cards has become a major problem in various financial institutions in the world. The number of defaults on credit cards is motivated by external factors from customers of financial institutions. The improvement of the system in predicting payment failure is expected to reduce the risk of financial loss for banks. There are two mechanisms to avoid losses due to default, among others: default prevention and default scoring system. Default prevention is a proactive method, which prevents defaults from occurring. On the other hand, a scoring system is needed to predict credit card defaults.

Default credit card deals with the illegal use of credit card information for purchases. Credit card transactions can be done both physically and digitally. In

physical transactions, credit cards are involved during the transaction. In digital transactions, this can happen over the phone or the internet. The cardholder usually provides the card number, expiration date, and card verification number by telephone or website.

The emergence of e-commerce in the last decade has also triggered an increase in the volume of credit card transactions in Indonesia. The volume of credit card transactions in 2013 in Indonesia was around 239 million transactions and increased in 2024 to 349 million transactions [1]. Along with the increase in the number of credit card usage, the number of default cases continues to increase. Although many authorization techniques have been applied the case of credit card defaults has not been effectively deterred. The rise of credit card default cases has had a major impact on the financial industry. The disadvantages of credit card defaults affect the merchandisers, where they cover all costs, including card issuer fees, fees, and administration fees. Because the marchant has to bear the loss, some items are priced higher, or discounts and incentives are lowered. Therefore, it is very important to reduce losses, and an effective default prediction system to reduce or eliminate default cases is important. from various studies on the prediction of credit card defaults. Here are some machine learning methods used, artificial neural networks (ANN), naïve bayes, Logistic Regression, support vector machines (SVM) dan random Forest.

LITERATURE REVIEW

Default credit card

Default Credit card occurs when the debtor is in arrears on credit card payments. This is a serious credit card status that affects not only the debtor's position with the credit card issuer, but also the debtor's credit status in general and your ability to get approved for credit cards, loans, and other credit-based services. Data regarding credit card transactions has information about the type of card, account number, location and time of the transaction, number of transactions, balance, credit limit, and others. This is information that is used as the basis of research to determine or differentiate. default from the user or detect data noise or outliers.

Credit Risk Scoring

Credit Risk Scoring Method is a standard tool in measuring individual risk levels, which is made based on statistical methods through an assessment of historical data that includes parameters or criteria that are estimated to have a significant influence on customer failure to repay loans (default). The key factors that influence the failure of credit payments consist of two categories, namely the financial category and the non-financial category, and each of these categories consists of several criteria and sub-criteria used to determine credit risk ratings. The purpose of the feasibility assessment with this method is to measure credit risk individually which is expected to be fulfilled [2].

Support Vector Machine (SVM)

Support Vector Machine (SVM) was developed by Boser, Guyon, Vapnik, and was first presented in 1992 at the “Annual Workshop on Computational Learning Theory”. The basic concept of SVM is actually a harmonious combination of computational theories that have existed decades before, such as the hyperplane margin. The kernel was introduced by Aronszajn in 1950, and so were the other supporting concepts. However, until 1992, there had never been an attempt to

assemble these components.

Naïve Bayes

Naïve Bayes algorithm is a form of data classification using probability and statistical methods. This method was first introduced by British scientist Thomas Bayes, which is used to predict future opportunities based on past experience, so it is known as Bayes' theorem. The Bayes theorem method is then combined with naivety which is assumed to be with conditions between independent attributes. The Naive Bayes algorithm can be interpreted as a method that has no rules, Naive Bayes uses a branch of mathematics known as probability theory to find the greatest opportunity from possible classifications, by looking at the frequency of each classification in the training data. Naive Bayes is also a very popular classification method and is included in the ten best algorithms in data mining, this algorithm is also known as Idiot's Bayes, Simple Bayes, and Independence Bayes. Bayesian classification has similar classification capabilities to decision trees and neural networks. Naive Bayes classification is a statistical classification that can be used to predict the probability of membership of a class. Bayes rule is used to calculate the probability of a class. The Naive Bayes algorithm provides a way of combining previous probabilities with possible conditions into a formula that can be used to calculate the probability of each possibility that occurs.

Logistic Regression

Logistics Regression (sometimes called logistic model or logit model) is one part of regression analysis that is used to predict the probability of an event occurring, by matching the data to the logit function of the logistic curve. This method is a general linear model used for binomial regression. Like regression analysis in general, this method uses several independent variables, both numerical and categorical. Logistics Regression does not require assumptions of normality, heteroscedasticity, and autocorrelation, because the dependent variable contained in Logistics Regression is a dummy variable (0 and 1), so the residuals do not require the three tests. For the assumption of multicollinearity, because it only involves independent variables, it still needs to be tested. For this multicollinearity test, the goodness of fit test can be used, which is then followed by hypothesis testing to see which independent variables are significant, so that they can still be used in research. Furthermore, among the significant independent variables, a correlation matrix can be formed, and if there are no independent variables that have a high correlation with each other, it can be concluded that there is no multicollinearity disorder in the research model [3].

Random Forest

Random forest is a bagging method, which is a method that generates a number of trees from sample data where the creation of one tree during training does not depend on the previous tree then the decision is taken based on the most votes [4]. Two concepts that form the basis of random forest are building an ensemble of trees via bagging with replacement and random selection of features for each tree that is built. The first thing means that each sample taken from the dataset for the training tree can be used again for another training tree, while the second means that the features used during training for each are a subset of the features owned by the dataset [5].

Artificial Neural Network (ANN)

Neural Networks or better known as Artificial Neural Networks is an information processing using performance characteristics similar to the process of delivering human nerve impulses. [5].

Neural Network developed with a mathematical equation model using the following assumptions:

i. Information processing occurs in the simplest elements called neurons (nodes).
ii. Between neurons with each other are interconnected and have connections.
iii. Each connection connects one node with another node and has a certain weighting value.
iv. Each node has an activation function (usually non-linear) as an input and also to determine the output result.

Another characteristic of a neural network is the existence of an architecture which is a connection between nodes. It takes an activation function which is a method of determining the weighting between connections and is commonly referred to as training or learning.

METHODS

Flowchart Diagram

This research will go through six stages, Pre-Processing data (data cleansing) , Descriptive analysis (for all data variables) , correlation testing (Perform a correlation test between predictor variables with a correlation plot to detect pairs of closely related predictor variables, indicated by a correlation value of more than 0.70. If there is a pair of variables that have a value of more than 0.70, a feature extraction will be carried out, with PCA), divide the data into training data and testing data (It is known that the proportion of data between default and non-default is 1.7% and 98.3%. Then the proportion of data will be divided into 70 : 30 with stratified random sampling method), classifying data using predetermined machine learning methods (support vector machine, naïve bayes, logistic regression, random forest, artificial neural network) , Evaluation result (AUC, PDP) , comparing result and implement the result into managerial nalysis.

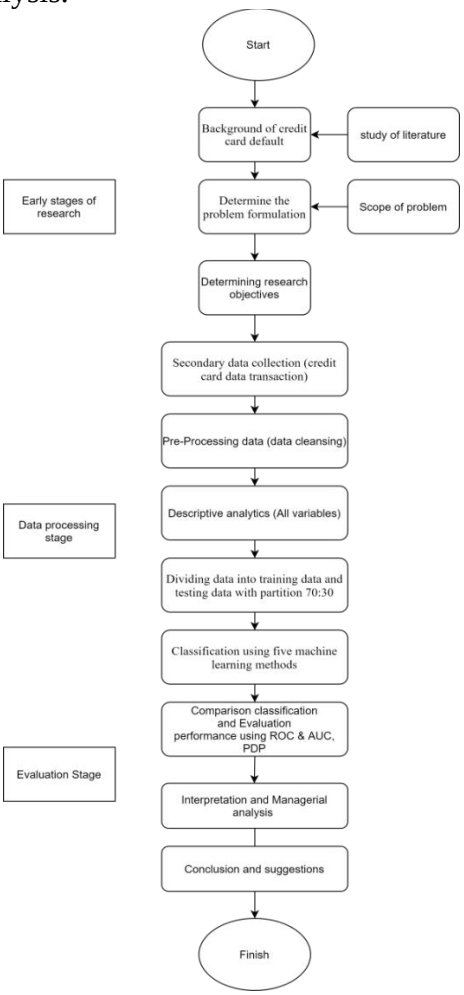


Figure 1. Research methodology

Data source

Original collection of credit card data from financial institutions at PT. XYZ is used in this research experiment. This is based on cardholders from the territory of Indonesia in April 2024. Does not collect full transaction data and is limited to 1000 transactions recorded, with 17 transactions classified as Default. The features used in the experiment are given in Table

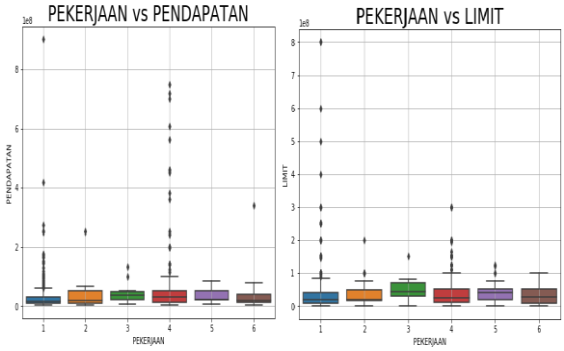
Table 2. Research data features

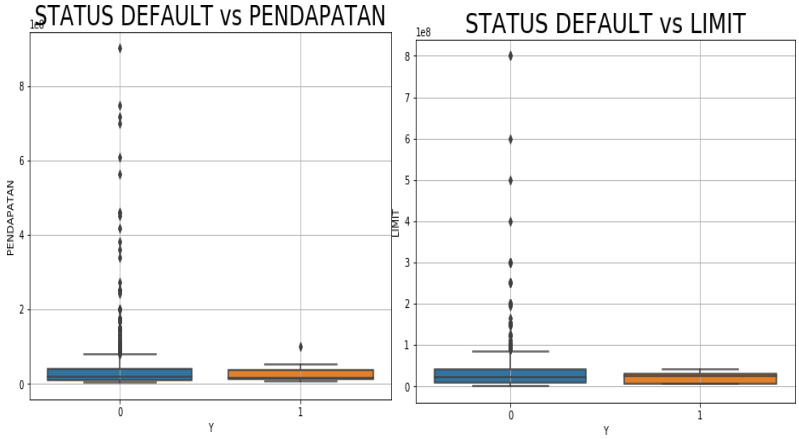
Variable	Information	Measurement Scale
X1	*Amount*	Numerical
X2	*Limit*	Numerical
X3	Age	Ordinal/Category
X4	Income	Numerical
X5	Province	Nominal/Category
X4	Collectability	Ordinal/Category
X5	Profession	Nominal/Category
X4	Credit Card Type	Nominal/Category
X5	Marital status	Nominal/Category
X4	*Gender*	Binary
Y	*Default*	Binary

RESULTS

Empirical analysis

Exploratory Data Analysis (EDA) played an integral part in understanding the credit card dataset. It was vital to get familiar with different relationships within the data through different types of plots before moving towards classification. Analyzing these relationships helped us with interpreting the outcomes of the models. Asking questions about these relationships provided us with additional knowledge about relationships that we may not have known existed. This section will further investigate data distribution and ask specific questions about the data lying within the dataset. Credit card has nine features of transaction. Then, there will be some correspondence between variables to determine the level of correlation formed between variables that are indicated to have a correlation.





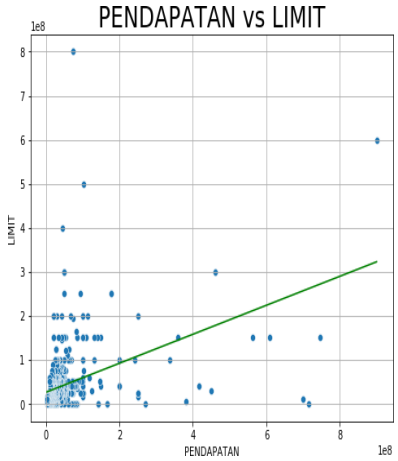


Figure 2. Graph of the relationship between several variables

Based on the graph of the correspondence between variables, it can be concluded that there is no relationship between variables for each comparison of variables, so there is an unsignificant effect between variables. after that a correlation check will be carried out based on the correlation matrix for each variable likely, payment status, income, credit card limit, and employment. The relationship between income and limit is not formed due to a large number of outliers (can be seen in the plot of many points that are spread evenly, but there is one gathering point). The outlier cannot be removed because it is a characteristic of the data itself (because there are quite a lot of outliers).

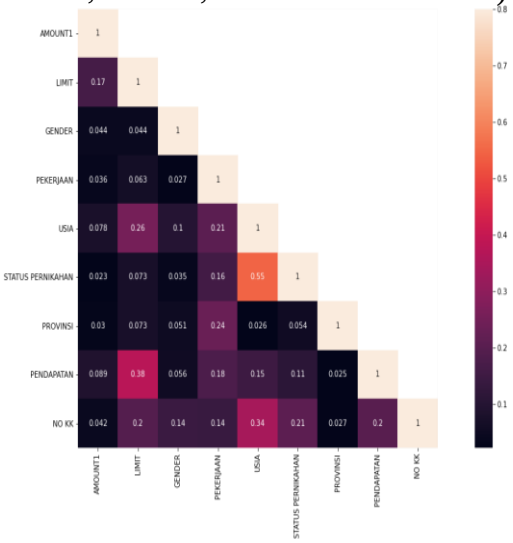


Figure 3. Correlation matrix of several variables

If there is one pair of predictor variables that has a correlation value greater than 0.7. Then the feature extraction will be carried out using the PCA (Principal Component Analysis) method. Based on figure 3, none of the pairs of variables have a correlation value > 0.7. So, there is no need for feature extraction. With no single highly correlated variable, classification can be continued. In this paper, we used five machine learning algorithms, the random forest, support vector machine, artificial neural network, logistic regression, and naïve bayes to work out a model for credit card default prediction. The results of the model are shown below with their classification report to get a better understanding of the accuracy and other scores of the five models.

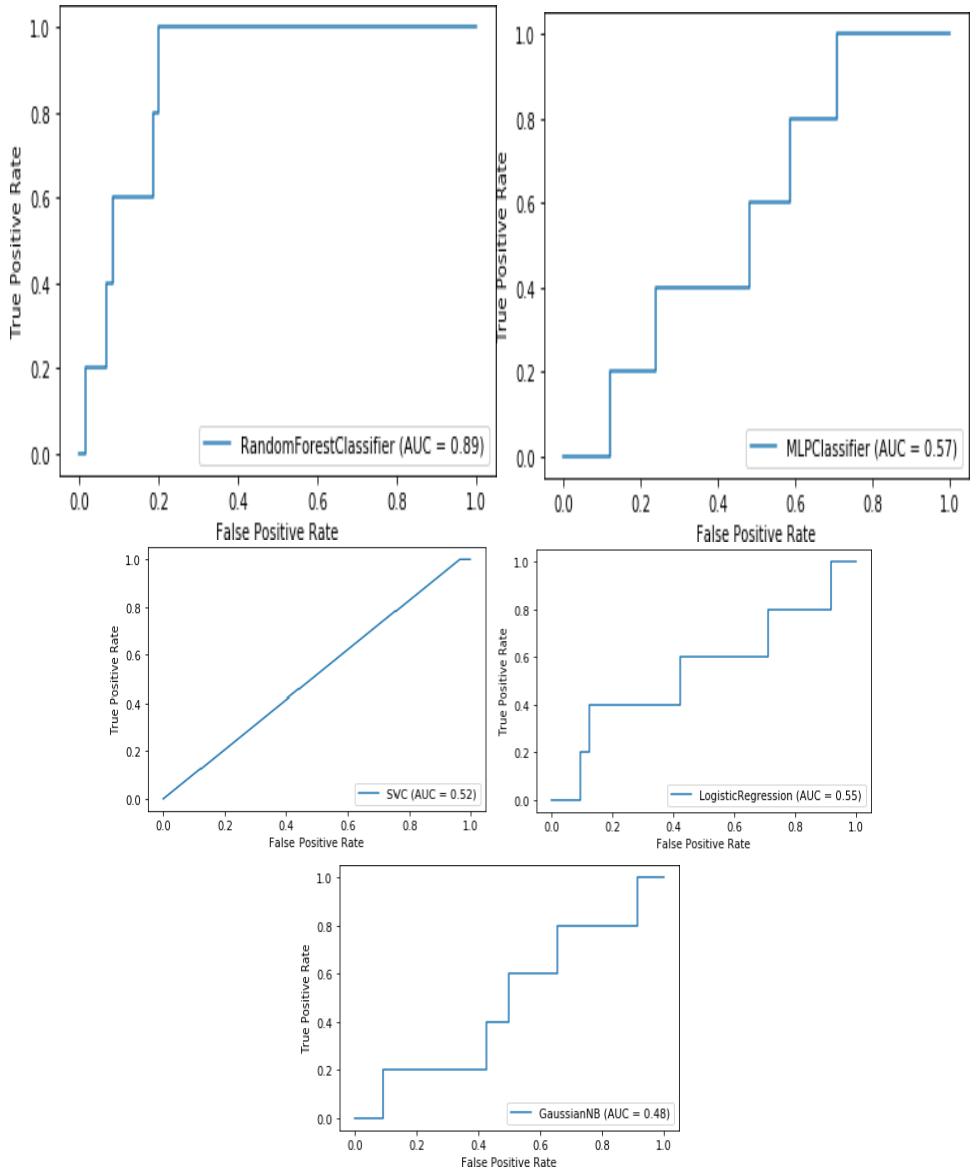


Figure 4. ROC Plot

Based on the ROC plot for each variable, it is known that the steeper the resulting graph, the better the classification method is. Based on Figure 4, it is known that the random forest method produces a fairly steep graph, which means that the random forest method can classify better than the other three methods. The Naïve Bayes method produces the worst classification. After that, the AUC score can be seen as follows.

Table 3. Summary of AUC score from several classification methods

Classification Methods	AUC
Random Forest	0,89
Artificial Neural Network (*ANN*)	0,57
Regresi Logistik	0,55
Support Vector Machine (*SVM*)	0,52
Naïve Bayes	0,48

The results obtained in table 2 are the best results from each classification method (based on several comparison coefficients). Based on table 2, it is known that the Random Forest classification produces the highest AUC score, which is 0.89. The Naïve Bayes

method produces the lowest AUC score. The AUC score obtained corresponds to the resulting ROC curve.

After knowing the best classification method in this classification based on the AUC value, a partial dependency plot will be shown for several variables that have the most influence on the default, namely limits, age, and income. The following is a Partial Dependence Plot of the three variables.

***Partial Dependence Plot* (PDP)**

Partial Dependence Plots can show whether the relationship between targets and features is linear, monotonous, or more complex. For example, when applied to a linear regression model, the partial dependency plot always shows a linear relationship. For classifications where the Machine Learning model outputs probabilities, the partial dependency plot displays the probabilities for a given class assigned different values for the features in. An easy way to handle multiple classes is to draw one line or plot per class. In this study, PDP uses income, age, and limit variables. The following is the PDP of the three variables.

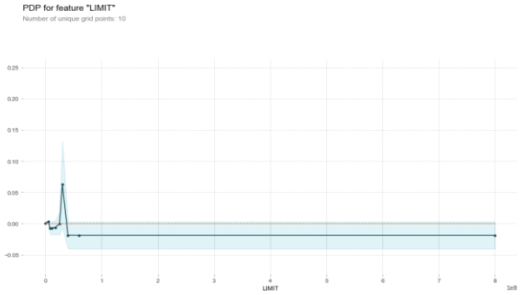


Figure 5. PDP of the limit variable

Based on Figure 5, it is known that there was an increase in plots of around 7 to 30 million, and after that, there was no plot movement. It can be concluded that around the credit limit of 7 to 30 million, there is an increase in the probability of default. So, to apply for credit with a limit of 7 to 30 million, you have to tighten the requirements because in that range, defaults are the most common. The following is the PDP of the age variable.

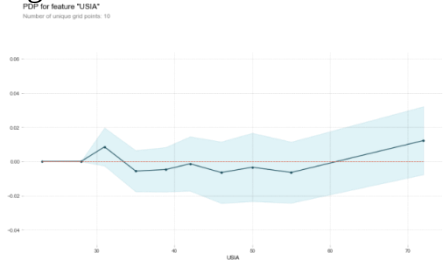


Figure 6. PDP of the Age variable

Based on Figure 6, it is known that there is an increase in the probability around the age of 27 to 32, and after the age of 60 years. It was concluded that in the age range of 27 to 32 years, and after the age of 60 years there were many defaults. So for credit card service providers, around the age of 27 - 32 and above 60 years, are given a more stringent selection in credit card considerations. After that, the PDP will be obtained from the income variable.

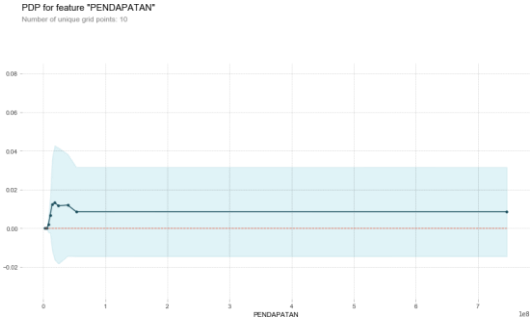


Figure 7. PDP of the income variable

Based on Figure 7, it is known that there is an increase in the probability in the range of 7 – 50 million, and after 50 million it has a positive but stable probability. It is concluded that in the income range of 7 – 50 million rupiah, there is an increase in the probability of default, and above 50 million rupiah has the same default probability. So, all income levels have a probability of default, but the lowest is in the income range of 7 – 50 million Rupiah. So, for credit card service providers, there is no need for a selection distinction based on income except in the income range of 7-

50 million rupiah because the probability generated is higher. However, all income levels have a probability of default.

In PDP based on income and credit card limits, the movement is relatively the same, around 7-50 million. Based on several theories regarding the structure of company workers with the salary they receive for each structure, it is known that for the range 7-50 million are workers with the structure as managers (managers). Meanwhile, below 7 million are operators (production), and above 50 million are policymakers (strategic). Based on the PDP, it can be explained that for the level of manager-class workers who have the highest probability of default.

CONCLUSIONS AND SUGGESTION

This paper aimed to explore, analyse, and build a machine learning algorithm to correctly identify whether a person, given certain attributes, has a high probability to default on a credit card. This type of model could be used by credit card transaction to identify certain traits of future borrowers that could have the potential to default and not pay back their loan by the designated time.

1) Based on 5 classification methods used in this study (Logistics Regression, Naïve Bayes, Support Vector Machine, Random Forest, and Artificial Neural Network). The resulting accuracy rate is based on the highest AUC score obtained by the random forest method, with an AUC score of 0.89. So, for the classification of default credit can use random forest.
2) Based on the PDP of several variables, it is concluded that
   a. For the Limit variable, there is an increase in the default probability at the limit of around 7 – 30 million Rupiah. So that requests for credit cards with a limit of 7 – 30 million Rupiah require a more stringent selection, because the probability of default is quite large.
   b. For the Age variable, there is an increase in the probability of default at the limit around the age of 27 – 32 Years, and more than 60 years. So that requests for credit cards with the applicant's age around 27-32 years, and above 60 years require a more stringent selection, because the probability of default has increased.
   c. For the Income variable, there is an increase in the default probability at a limit of around 7 – 50 million Rupiah. However, for the entire income range it yields a positive probability of default.
3) Based on the PDP of several variables, it is concluded that based on the results of the PDP on income and credit card limits, it is known that managers class workers have a higher probability of default than other working classes. So that based on PDP on income and credit card limits, it can be used as a reference for selection for credit card applicants.

   So, any income there is a default probability. So, there is no difference in selection based on income, because each level of income has a default probability that tends to be the same. Since, the algorithm puts some of the non-defaulters in the default class, we might want to look further into this issue to help the model accurately predict capable credit card users. For the management, it can take into account the conversion of positions that can be universally allocated, which means equalizing a job into a level of position. With the universal job conversion, data quality, especially job data, becomes more

valid and accurate for selecting credit card applicants whether to be accepted or rejected.

REFERENCES

2020. *bi.go.id.* 12 4. Accessed 12 4, 2020.

PT. Bank Rakyat Indonesia. 1999. *Buku Pedoman Pelaksanaan Kredit Unit Retail Banking (PPK URB).* Jakarta: PT. Bank Rakyat Indonesia.

Hosmer, D. W., and S. Lemeshow. 2000. *Applied Logistic Regression.* USA: John Wiley & Sons.

Law, A., and M. Wiener. 2002. "Classification and Regression by Random Forest."

Sazona, V. 2015. *Implementation and Evaluation of a Random Forest Machine Learning Algorithm.*

Fausett, L. V. 1994. *Fundamentals of neural networks (1st ed.), Englewood CLiffs.* New Jersey: Prentice Hall.