CHAIR OF APPLIED STATISTICS
SCHOOL OF THE FREE UNIVERSITY OF BERLIN

Bachelor Thesis

A comparison of logistic regression and classification tree using credit card client data

Sangho Kim

Supervisor:	Prof. Dr. Timo Schmid
Semester:	Summer Semester 2020
Author:	Sangho Kim
Matric. No.:	5123651
Address:	Triftstrasse 67, 13353 Berlin
Email:	ttyy21@naver.com
Phone:	+4915771504359
Subject:	Bachelor Economics

Submitted: 17. July 2020

Abstract

Die vorliegende Arbeit versucht, den binären Datensatz mithilfe der logistischen Regression und des Klassifizierungsbaums zu klassifizieren. Da die logistische Regression und der Klassifizierungsbaum ihre Vor- und Nachteile bei der Analyse von Binärdaten haben, zeigen sie unterschiedliche Ergebnisse. Die Ergebnisse der beiden Methoden werden miteinander verglichen. Zur Entwicklung dieser beiden Methoden werden die Variablenauswahl für die logistische Regression und das Pruning für den Klassifizierungsbaum eingeführt. Zusätzlich wird das Random Forest angewendet, um das Klassifizierungsbaummodell zu kompensieren. Anschließend werden die Ergebnisse der einzelnen Methoden erläutert. Das logistische Regressionsmodell und das Klassifikationsbaummodell zeigen in dieser Studie ihre Stärken nicht vollständig. Es gibt keinen besonderen Unterschied zwischen den logistischen Regressionsmodellen und dem Klassifizierungsbaum zeigt eine Form eines unvollständigen Baummodells. Die Ergebnisse zeigen jedoch, dass einige Variablen für das Modell wirksam sind. Daher sollten diese Methoden basierend auf dem Zweck der Studie ausgewählt werden, und andere Methoden können in weiteren Studien berücksichtigt werden.

Abstract

This study tries to classify the binary data set using logistic regression and classification tree. Since the logistic regression and the classification tree have their advantages and disadvantages of analysis of binary data, they show different results. The results of the two methods are compared with each other. In order to develop these two methods, variable selection for the logistic regression and pruning for the classification tree are introduced. In addition, the random forest method is applied to compensate the classification tree model. Afterwards, the results of the each method are explained. The logistic regression model and the classification tree model do not completely show their strengths in this study. There is no particular difference among the logistic regression models and the classification tree shows a form of imperfect tree model. However, the results show that a few variables are effective on the model. Therefore, these methods should be selected based on the purpose of study and other methods can be considered in further study.

Contents

List of Figures

List of Tables

1. Introduction

Between 1998 and 2005, economy of many Asian countries was highly growing (for Economic Co-operation & Development 2017, P. 12). With this growth, the consumer market of credit card was growing. Since the regulation of credit card issuing was weaker than today, many firms of credit card recklessly issued credit cards to the customers. The customers also imprudently paid on their credit card. Around 2005, the economic growth in the Asian countries was slowing down. Many credit card holders could not pay back their bill. In the end, they become a bad credit. The bad credit crisis had a riffle effect on the whole community. To amend the situation, governments decided to strengthen further the regulation of credit card issuing. In addition, the government and credit card firms started to analyze this problem. Among the analyses, they focused on whether customers in this crisis were defaulted or not. From this analysis about the classification of the customers, the government could get information that was used for the base of the regulation. The firms of credit card also could weed out future customers that will be at a higher risk of injudicious use of credit card. Therefore, the research on the classification of the customers who are defaulter or non defaulter is important (on Financial Services. Subcommittee on Oversight et al. 2002, P. 10).

For the classification of credit card holders, government and credit card firms might use various methods. Popular methods among them are logistic regression and classification tree. The two methods are famous for the binary classification. This study uses the logistic regression and the classification tree. The aim of the study is to classify the customers whether they are defaulter or non defaulter. In order to achieve this goal, this study mainly compares the results of the logistic regression and the classification tree. In chapter 2, the data set is simply explained. Binary logistic regression and classification tree are explained with some mathematical expressions in chapter 3.1 and 3.2 respectively. Then, a comparison between the logistic regression and classification tree is described in chapter 3.3. In chapter 3.4, measurements of model performance are introduced. Before the explained classification methods are applied, the data is explored in chapter 4.1. Logistic regression and classification tree are applied to the data in chapter 4.2 and 4.3. For the better performance of classification using trees, random forest method is used in chapter 4.4, After

that, the used methods are compared in chapter 4.5. In chapter 5, results of application are summarized. In addition, weakness of methods and possibility of extensive research questions are discussed.

2. Data Set

The data set which is used in this study is consisted of a few variables. It should be explained as the preparing the modelling. In this chapter, details of the data set are checked. The meaning of each variable's name and classes is explained.

2.1. Data Information

The data set (Yeh 2009) is about the information of customer's default payment of credit card in Taiwan 2005 (Yeh & Lien 2009, P. 2473 ff.). The number of customers is 30,000 and the number of attributes is 25. The attributes consist of personal information and kinds of customers' payment records from April to September.

2.2. Variable Information

LIMIT_BAL is the amount of given access line of each customer which is measured in new Taiwan dollar (NT dollar). It includes not only the individual consumer credit card but also his or her family credit. PAY_# indicates whether the customers have repaid their bill in each month from April to September or not. # can be replaced by one of six numbers (0, 2, 3, 4, 5 and 6), which means there are six different variables from PAY_0 to PAY_6. Each number means month (April = 6, May = 5, June = 4, July = 3, August = 2, September = 0). For example, PAY_0 is the monthly payment record in September. PAY_3 = -2 means paying duly for two months in May and PAY_2 = 1 means that the payment has been delayed for one month in August. BILL_AMT# is the amount of bill statement and PAY_AMT# is the amount of previous payment in month. Here, # can be also replaced by one of six numbers (1, 2, 3, 4, 5 and 6) and each number corresponds to a month (April = 6, May = 5, June = 4, July = 3, August = 2, September = 1). For example, BILL_AMT1 is total amount of bill statement in September. Lastly, the name of variable 'default.payment.next.month' is too long to write. It is cut to 'default' for further application.

Table 2.1.: Variable description in the data set of customer's default payment

Variable	Description
ID	Unique identification number assigned to each customer
LIMIT_BAL	Amount of given credit access line (NT dollar)
SEX	Gender (1 = male, 2 = female)
EDUCATION	Highest degree obtained (0 = no education, 1 = graduate school, 2 = university, 3 = high school, 4 = others, 5 = unknown, 6 = unknown)
MARRIAGE	Marital status (0 = unknown, 1 = married, 2 = single, 3 = others)
AGE	Age in years
PAY_#	Monthly payment record (NT dollar), # can be replaced by one of six numbers (April = 6, May = 5, June = 4, July = 3, August = 2, September = 0)
BILL_AMT#	Total amount owed (NT dollar), # can be replaced by one of six numbers (April = 6, May = 5, June = 4, July = 3, August = 2, September = 1)
PAY_AMT#	Amount of previous payment (NT dollar), # can be replaced by one of six numbers (April = 6, May = 5, June = 4, July = 3, August = 2, September = 1)
default.payment.next.month	Default payment (1 = yes, 0 = no)

3. Methods

In chapter 2, the details and variables of the data set are checked. In chapter 3, classification methods for binary variable are explained. For the binary classification, there are many methods such as logistic regression (Cox 1958), classification tree (Quinlan 1986), K-nearest-neighbor algorithm (Fix 1951), support-vector machines (Vapnik & Chervonenkis 1974) and others. Among these methods, this study focuses on logistic regression and classification trees, since they are mostly used for classification of binary data. Most part of explanation about the logistic regression and the classification tree is based on Casella et al. (2013).

3.1. Binary Logistic Regression

Assume that Y_i is a binary variable whose units only can have two possible states so that these two states can be labeled as 1 and 0:

$$Y_i = \begin{cases} 1 & \text{for state 1;} \\ 0 & \text{for the other state.} \end{cases} \tag{3.1}$$

In order to classify a binary variable, regression models might be used, since the probability of a certain state is modeled by using the regression. In this case, it is important which model will be used. One can naively use a linear regression model which is very simple. However, there are several problems of the applying linear regression model to explain a binary response variable, because it only has two expressions. First, the estimated target variable Y by using linear regression model is implausible. Suppose that the following linear model is used to explain binary Y:

$$Y = x\beta + \epsilon.$$

x is design matrix and ϵ is error term with $\epsilon \sim N(0, \sigma^2)$. Using this model, the expected value of $E(Y)$ can be estimated. Since Y is binary, it is binomial distributed. In other words, $E(Y)$ a probability of state 1 that only can take values in interval [0, 1]. However, some of the estimates of binary Y from the linear regression model can be outside of the probability interval, [0, 1] (Casella et al. 2013, P. 131). This is showed in figure 3.1. The figure shows

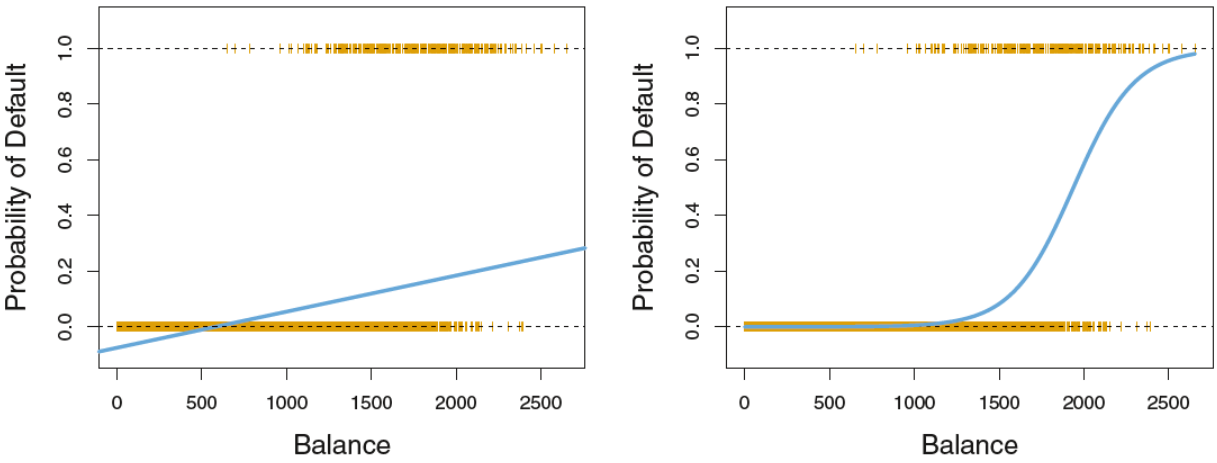

Figure 3.1.: Example of classification binary data using linear regression and logistic regression (Casella et al. 2013)

the estimated probability by linear regression model (left) and logistic regression model (right). In the left figure, the blue line is outside of [0, 1], which shows that the simple linear regression model is not suitable for a binary response variable. Therefore, another suitable regression model should be applied and the logistic regression model (right) is suitable for the binary response variable. The logistic regression model is also one of the most popular regression model for the binary response variable.

3.1.1. Logistic Regression Model

There are a binary response variable Y and p independent variables $x_1, ..., x_p$. Y is binomial distributed with $E(Y_i) = P(Y_i = 1) = \pi_i$ and with $Var(Y) = \pi_i \cdot (1 - \pi_i)$. The aim of a logistic regression model is to model the effect of independent variables on the probability π_i. The probability is modeled by using a standard logistic distribution function $Logist(\eta_i)$, since the values of $Logist$ takes values in interval, [0, 1]:

$$\pi_i = P(Y_i = 1|x_i'\beta) = Logist(e^{\eta_i}) = \frac{e^{\eta_i}}{1 + e^{\eta_i}},$$

where η_i is a linear predictor:

$$\eta_i = x_i'\beta = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik},$$

or logit link function:

$$g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}.$$

Regression parameters $(\beta_0, ..., \beta_k)$ can be estimated by maximum likelihood estimation. The maximum likelihood estimation (Fisher 1922) is a method that estimates the parameters of probability distribution (Rossi 2018, P. 227). For the interpretation of the estimated beta, a concept of odds is needed. Odds of an event are the ratio of the probability that the event will happen $(P(Y_i = 1))$ to the probability of which the event will not happen $(P(Y_i = 0))$ (Daly & Bourke 2008, P. 154).

$$Odds(\pi_i) = \frac{P(Y_i = 1)}{P(Y_i = 0)} = \frac{\pi_i}{1-\pi_i}.$$

When π_i is replaced by $\frac{e^{\eta_i}}{1+e^{\eta_i}}$:

$$\frac{\pi_i}{1-\pi_i} = \frac{\frac{e^{\eta_i}}{1+e^{\eta_i}}}{\frac{1}{1+e^{\eta_i}}} = e^{\eta_i}.$$

When it takes logarithm:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i.$$

This shows:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}.$$

However, the estimated parameter beta cannot be directly interpreted. Instead, odds ratio is used for the interpretation. The odds ratio is a ratio of two odds for two different events:

$$OR(\pi_1, \pi_2) = \frac{Odd(\pi_1)}{Odd(\pi_2)}.$$

When odds are at multivariable model:

$$\frac{P(Y = 1)}{P(Y = 0)} = \frac{\pi_i}{1-\pi_i} = e^{\eta_i} = e^{\beta_0} \cdot e^{\beta_1 x_{i1}} \cdot ... \cdot e^{\beta_k x_{ik}}$$

For example, if x_{i1} increased by 1 to $x_{i1} + 1$, and it applies to the odds:

$$Odds(\pi_{1,x+1}) = \frac{P(Y = 1|x_{i1} + 1)}{P(Y = 0|x_{i1} + 1)} = e^{\beta_0} \cdot e^{\beta_1(x_{i1}+1)} \cdot ... \cdot e^{\beta_k x_{ik}},$$

and

$$Odds(\pi_{2,x}) = \frac{P(Y=1|x_{i1})}{P(Y=0|x_{i1})} = e^{\beta_0} \cdot e^{\beta_1 x_{i1}} \cdot ... \cdot e^{\beta_k x_{ik}}.$$

and for the odds ratio (OR):

$$OR = \frac{Odds(\pi_{1,x+1})}{Odds(\pi_{2,x})} = e^{\beta_1}.$$

Generally, the estimated parameter β_1 is interpreted by whether β_1 is positive, zero or negative. When β_1 is positive, it means that the chance $\frac{P(Y=1)}{P(Y=0)}$ will increase. The chance $\frac{P(Y=1)}{P(Y=0)}$ remains the same at $\beta_1 = 0$. At the negative value of β_1, the chance $\frac{P(Y=1)}{P(Y=0)}$ will decrease.

Since $\hat{\beta}_i$ can be estimated by the maximum likelihood method, $\hat{\eta}_i$ can also be estimated:

$$\hat{\eta}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_k x_{ik}.$$

The probability is estimated by the *Logist* with $\hat{\eta}_i$:

$$\hat{\pi}_i = Logist(e^{\hat{\eta}_i}) = \frac{e^{\hat{\eta}_i}}{1 + e^{\hat{\eta}_i}},$$

For classification of output of logistic regression, a cut-off probability is required. Based on the estimated probability, cut-off probability of the classification can be used to determine two groups and then all cases are classified into the defined groups. The logistic regression sets the cut-off probability at 0.5 as a default. It can be chosen depending on a situation as well (Liu et al. 2011, P. 555 f.).

3.2. Classification Tree

Classification tree is a type of decision tree (Belson 1959). Both classification tree and regression tree are type of decision tree. When a response variable of data set is continuous, the regression tree is used. The classification tree is used when the target variable is categorical. In this study, the response variable is binary. Thus, this study uses the classification tree method. In this chapter, the classification tree method is explained.

A general classification tree is a tree-like model and it has a flowchart-like structure. The figure 3.2 shows name of each part of tree. Root node is the topmost node in a tree. Internal node is any node of a tree and it has child node. The child node means a node of steps. Terminal node, also called as leaf, is the node in the final step. It has no child node (Ahmad 2020, P. 54 ff.).

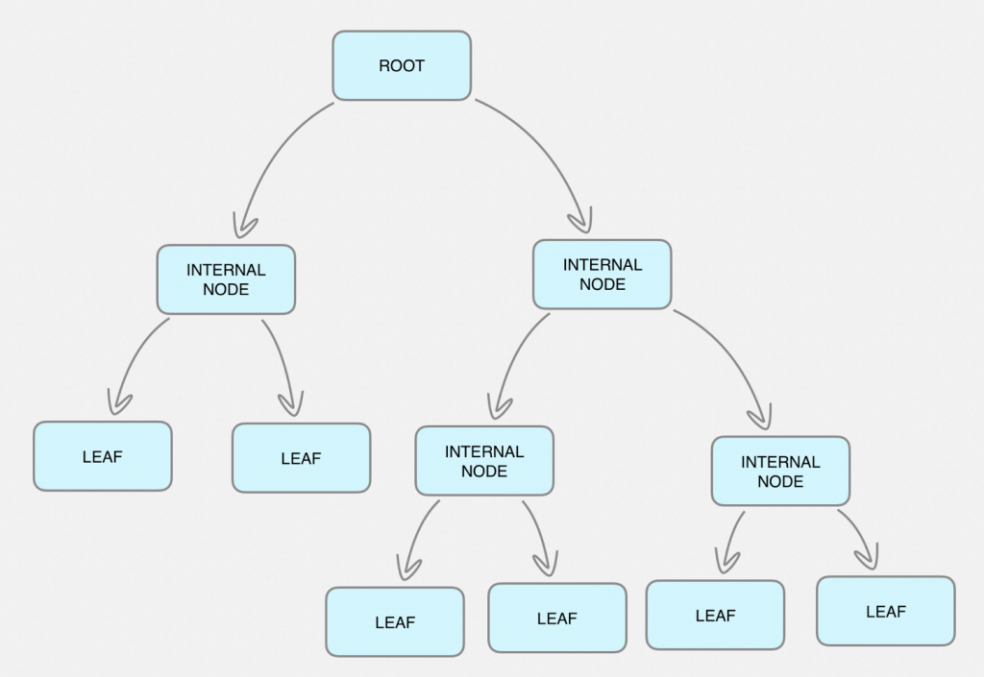

Figure 3.2.: Structure of tree (Ahmad 2020, P. 54 ff.)

A tree grows with the splitting of nodes in two branches from a root to internal nodes. During the splitting of the nodes, best variables should be chosen (Fox 2000, P. 54). In order to choose the best variables, a measurement is required. There are three representative measurements such as: classification error rate (Breiman et al. 1984), Gini-index (Breiman et al. 1984) and entropy (Shannon 1948).

First, classification error rate (CER) is a fraction where an observation does not belong to the most commonly occurring class. It indicates how good the model is in predicting the outcome of new observations. The classification error rate E is:

$$E = 1 - \max_k \hat{p}_{mk}.$$

$\hat{p}_{mk}$ is a proportion of training observations of m-th region and n-th class.

Gini-index is one of the popular measures of a node impurity that is homogeneity of labels at each node (Bruce et al. 2020) and measures the total variation across the k classes. The Gini-index G is (Casella et al. 2013, P. 312):

$$G = \sum_{k=1}^{K} \hat{p}_{mk} \cdot (1 - \hat{p}_{mk}).$$

It is going to be a small value when $\hat{p}_{mk}$ is close to 0 or 1. In addition, its small value means

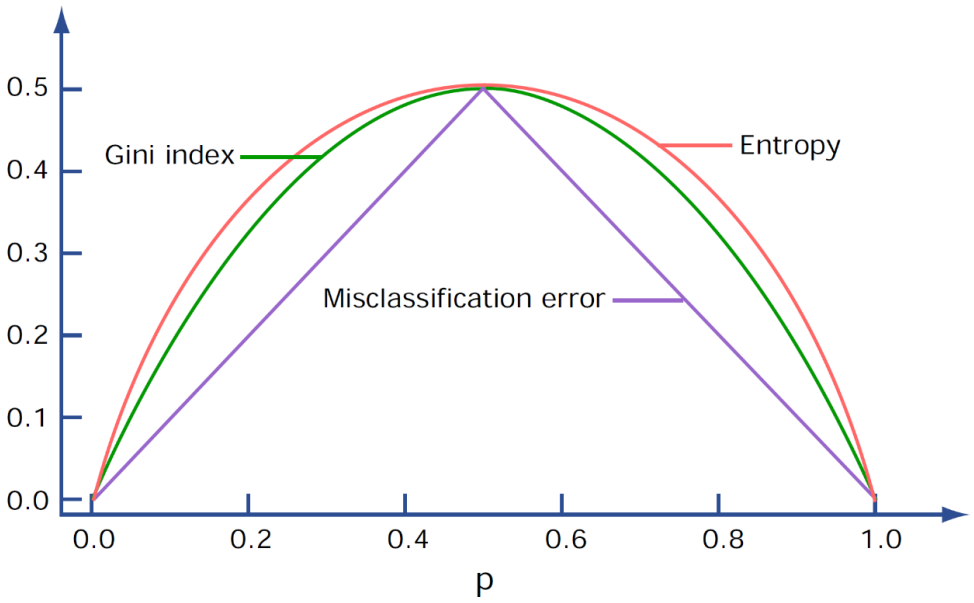

Figure 3.3.: Sensitivity of Gini-index, entropy and classification error rate (Rutkowski et al. 2020, P. 65)

that a node contains predominant observations from a single class.

Entropy is also another famous measurement of impurity and it is similar with the Gini-index. The entropy D is:

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \cdot \log \hat{p}_{mk}.$$

If the $\hat{p}_{mk}$ is near 0 or 1, the entropy will take a value near 0. The process of the entropy will be stopped if the decrease in impurity is smaller than a threshold. In addition, lower entropy values are preferred for making decision splitting (Casella et al. 2013, P. 312).

Gini-index and entropy are pretty much the same. They are required to build classification tree and used to measure impurity level that evaluates the quality of a particular split. In figure 3.2, the classification error rate (purple) is less sensitive to the node purity than the Gini-index (green) and entropy (red).

However, overfitting can be occurred in the tree making process. The overfitting means that a model fits on a particular set of data and the model may fail to fit to another data set (Burnham & Anderson 2002, P. 45). To avoid it, the tree model should be pruned. Pruning the tree reduces not only the size of decision trees but also the complexity, and hence improves predictive accuracy (Casella et al. 2013, P. 22). Before the pruning, an ideal size of the tree should be found. According to K-fold cross-validation experiment, the ideal

size can be found. It is a validation technique: the complete data set is randomly split into k data sets. $k-1$ sets are train set and the k-th set is validation set (Olson & Delen 2008, P. 141 ff.). The classification model is trained and tested k times. With the size of the result of the experiment, the tree is pruned. During the pruning, misclassification rate is used as a criterion (Ripley 2019a, P. 2), and then the tree is finally made.

3.2.1. Ensemble Method: Random Forest

Classification tree method has an advantage: it is easy to explain results to people who have no fundamental knowledge in statistic. However, this method generally does not have a better level of predictive accuracy than other methods. In the view of accuracy, other regression and classification approaches are preferred. In general, a large number of terminal nodes will tend to decrease the bias, whereas it can increase a variance. The high variance causes low accuracy level of prediction (Casella et al. 2013, P.315 f). Additionally, this method is not robust. Trees can be easily changed by a small change of data. Because of its disadvantages, ensemble methods are often used. The ensemble methods combine several techniques into one predictive model in order to decrease variance and to improve the predictions. In other words, the ensemble method gives better accuracy, avoids overfitting and reduces the variance. There are different ensemble methods, however, in this study, random forest (Breiman 2001) for classification tree is implemented.

To explain random forest, explanation of bagging (Breiman 1996) should be done first. Bagging is a compound word of bootstrap and aggregating. Here, the bootstrapping (Efron 1979) is a resampling process of sample data to make an inference about a population from sample data and helps to develop robust models (Barros et al. 2018). The bagging is a general purpose procedure for reducing the variance of statistical learning method by averaging a set of observations. The way to reduce variance is to take many training sets, build prediction models using each training set and average the prediction. But it is not practical, since there is only a single training set. Instead, according to bootstrap, repeated samples are made from the single training set. In case of classification tree, the method is trained on the b-th bootstrapped training set in order to record the class predicted by each of the B classification trees. Finally, the decision is settled by majority vote (Casella et al. 2013, P 316 ff.).

It will be better to use random forest, if there is a strong predictor in a data set. The strong predictor is used for first split in many trees of bags. Consequently, all the bagged trees will look like similar to each other. It leads to correlated bagged trees and there are no substantial gains of accuracy that are caused by averaging their predictions. Therefore,

the over bagged trees are improved by the random forest method according to a random small tweak that decorrelates the trees(Casella et al. 2013, P. 319 ff.).

The procedure of the random forest is: the number of trees on bootstrapped training samples is built. Next, a random sample of m predictors is chosen as split candidates from p predictors in every split point of a tree. In case of the classification tree, only one of the m predictors ($m = \sqrt{p}$) is allowed to use by the split (Casella et al. 2013, P. 321).

3.3. Comparison of Logistic Regression Model versus Classification Tree Model

In regard to the shape of the model, logistic regression model is predefined, while classification tree is not. Since the logistic regression models probability that is related to the binary response variable, the shape of the logistic function is S-curve. However, the shape of the classification tree is not predefined, though the general shape of classification trees is top-down. The classification tree model fits in best possible classification based on the data (Dangeti 2017, P. 134).

Logistic regression is parametric, whereas classification tree method is non-parametric. An assumption of parametric statistics is that a distribution of probabilities about a fixed set of parameters models a population of sample data. In other words, if a model predicts a probability of a new observation, the observation will have to come from the same distribution. Even though parametric methods involve fewer assumptions of structure and distributional form than non-parametric methods, they contain strong assumptions about independence. On the contrary, the classification tree is non-parametric. Non-parametric statistics are based on parametrized families of probability distribution. It is also either on being distribution-free or on having a specified distribution. Non-parametric estimates do not need assumptions about the distribution of probabilities (Dangeti 2017, P. 134).

For the better prediction, independent variables in logistic regression should be continuous in nature. Because the logistic function is continuous at any independent variable, continuous independent variables are better for predicting of probability. However, in classification tree, it will provide the better results if most of the variables are categorical in nature. Since internal nodes of trees split into two separated nodes, categorical variables have an advantage to get good results (Gupta 2015, P. 34).

Effect of outliers on models is different in logistic regression and classification tree. The outliers can seriously deteriorate a performance of logistic regression. In the logistic regression, mean-variance relationship means that a scaling factor for vertical displacement

is a continuous function of the fitted S-curve. Farther out in tails of the curve, the mean is closer to either 0 or 1. Leading to smaller variance can have more substantial impacts on estimates and inferences. Leading to smaller variance which makes seemingly small perturbations can have more substantial impacts on estimates and inference. Conversely, the outliers are dealt with grace in the classification tree. The probable outliers in the data will have a negligible effect, because the nodes are determined based on the sample proportions in each split region (Ayyadevara 2018, P. 69 ff.).

Effects of independent variables can be checked in logistic regression, but not in classification tree. Although an exact value of the effect of the independent variables on the probability of y could not be measured, trends or directions (negative or positive) can be checked by the sign of their coefficient. However, this check is impossible in the classification tree, since the tree has only some nodes that only show the independent variables with the split value (Menard 2010, P. 205).

As for the determination of variable significance, logistic regression model uses Wald test (Wald 1939) and likelihood ratio test (Neyman & Pearson 1928). Both tests are classical approach to hypothesis testing. The Wald test can be used to determine the distribution of a suitable test statistic while the null hypothesis is valid. The likelihood ratio test is calculated on the basis of the likelihood ratio (O'Connell 2006, P. 16). But the classification tree uses root node to determine variable significance. There is the most significant independent variable and the predicted probabilities of the independent variables can be explained through serial interactions (Below n.d., P. 424).

Lastly, interpretation of the logistic regression is more complex, relative to the classification tree. For the interpretation of the logistic regression, some knowledge is necessary. Nonetheless, if a reader has background information about statistic, he can get more and deeper information than the classification tree's (Brezina 2018, P. 118). In contrast, the interpretation of the classification tree is explicit and easily interpreted, even by a non expert (Williams 2011, P. 205).

3.4. Measurement of Model Performance

Logistic regression and classification tree are explained and compared with each other. In this chapter, some methods which can measure and compare performances of two models are explained.

Table 3.1.: Example of confusion matrix

	actual class		
predicted class		cat	not cat
	cat	5 ($n_{1,1}$)	2 ($n_{2,1}$)
	not cat	3 ($n_{1,2}$)	7 ($n_{2,2}$)

3.4.1. Confusion Matrix

Confusion matrix (Miller & Nicely 1955) summarizes correct and incorrect classification that a classifier produced. The classification is summarized for the validation data. For example, the table 3.1 is an example of confusion matrix. The table shows the results of actual classified class and predicted classified class. A sum of diagonal cells ($n_{1,1} + n_{2,2}$) gives the number of correct classification and a sum of off-diagonal cells ($n_{1,2} + n_{2,1}$) gives the counts of misclassification when the models tried to classify animals as 'cat' or 'not cat'. Confusion matrix gives estimates of the true classification and misclassification rate. Estimated misclassification rate (ERR) is (Matignon 2005, P. 36):

$$ERR = \frac{n_{1,2} + n_{2,1}}{n_{1,1} + n_{1,2} + n_{2,1} + n_{2,2}},$$

while estimated accuracy is (Matignon 2005, P. 36):

$$Accuracy = 1 - ERR = \frac{n_{1,1} + n_{2,2}}{n_{1,1} + n_{1,2} + n_{2,1} + n_{2,2}}.$$

In the table, the estimated misclassification rate is $\frac{3+2}{5+2+3+7} = \frac{5}{17} \approx 0.2941 = 29.41\%$ and the estimated accuracy is $1 - 0.2941 \approx 0.7058 = 70.85\%$.

3.4.2. Receiver Operating Characteristic Curve and Area Under the Curve

Another measurement is receiver operating characteristic (ROC) curve (Woodward 1953). The curve visualizes a diagnostic ability of a classification system. The table 3.2 shows a two class prediction problem. P is one class or state of the binary variable and N is the

Table 3.2.: Two class prediction

	actual class		
predicted class		Positive (P)	Negative (N)
	Positive (P)	True Positive (TP)	False Negative (FN)
	Negative (N)	False Positive (FP)	True Negative (TN)

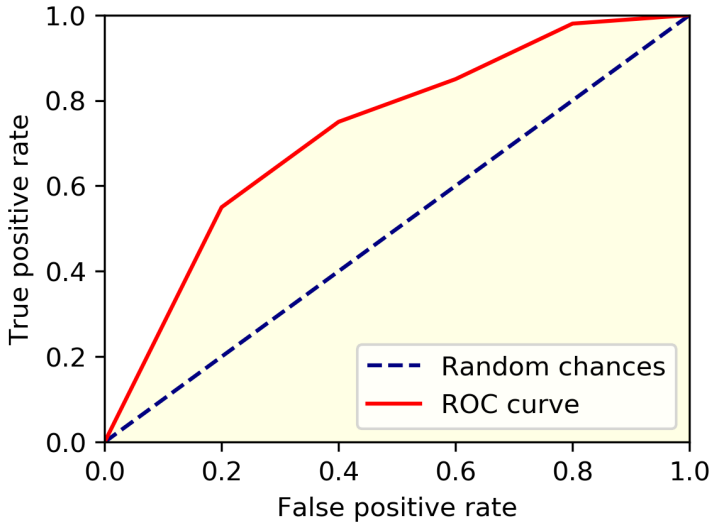

Figure 3.4.: Example of ROC curve (Daldrup 2007, P. 207)

other state of the y. TP is is the true positive classified outcome if the outcome from a prediction is P and the actual value is also P. With the same logic, FN is the outcome when the predicted outcome is P but the actual outcome is N.

ROC curve is created by plotting true positive rate against false positive rate at various threshold settings. The curve displays the two types of error for all possible thresholds. Because the curve takes into account all possible thresholds, it is useful to compare different classifiers (Casella et al. 2013, P. 146). For clarifying the ROC curve, terms should be defined. True positive rate (TPR) is called as *sensitivity* and it measures true positive rate of the model based on the formula:

$$TPR = Sensitivity = \frac{TP}{TP + FN}.$$

False negative rate (FNR) is also named as $specificity$:

$$FNR = Specificity = \frac{TN}{TN + FP}.$$

False positive rate (FPR) is $1 - TPR$:

$$FPR = 1 - Specificity = \frac{FP}{TN + FP}.$$

The ROC curve is defined by FPR (x-axis) and TPR (y-axis) in figure 3.3. The figure

depicts relative trade-offs between true positive and false positive. Each prediction result or instance of a confusion matrix represents one point in the ROC space. A ROC curve of the best performance will hug the top left corner. On the other hand, the worst performance ROC curve is $y = x$ (Casella et al. 2013, P. 147).
Area under the curve (AUC) (Woodward 1953) gives the overall performance of classifier. This is often used to summarize the sequence of repeated of measures on any individual (Fitzmaurice et al. 2012, P. 83). Because the AUC is related to the ROC curve, an AUC of the best performance is 0.95 and the worst is 0.5. The general rules of the AUC is: $0.5 < AUC < 0.6$ means poor classification; $0.6 \leq AUC < 0.7$ means fair classification; $0.7 \leq AUC < 0.8$ means acceptable classification; $0.8 \leq AUC < 0.9$ means excellent classification; $AUC \geq 0.9$ means outstanding classification (Satapathy et al. 2016, P. 192).

4. Application and Discussion

In this chapter, the data set will be investigated. The data set is computed by using the open source software R. Firstly, exploratory data analysis explains data. Next, methods and measurements which are already mentioned in chapter 3 are used to analyze the data set: logistic regression, classification tree and random forest. Finally, the results from the logistic regression and the classification tree are compared with each other.

4.1. Exploratory Data Analysis

Exploratory data analysis (EDA) (Tukey 1977) is an approach of analysis of data sets. It proceeds initial investigations on data to discover patterns or to check assumptions (Hartwig & Dearing 1979, P. 5). Before exploratory data analysis, ID variable should be omitted, since it is random generated numbers to identify all customers. Above all, missing data should be checked. The missing data means that no data value is stored for the variable in an observation (McKnight et al. 2007, P. 2). There is no missing data.

Correlation of the response variable with other variables should be checked and it is showed in figure A.1. There are two types of correlation: between independent and dependent variable, or between independent variables. The correlation between PAY_# and default variable belongs to the first type. The PAY_# variables are also more correlated with the default variable if the month is close to due date (October). As the correlation between the independent variables, both PAY_# and BILL_AMT# variables are correlated within each other. Two squares that are made of blue points in the figure show the correlation of the variables. If there is a correlation, it means that one variable will be associated with the other variable, then the significant variable will be treated as insignificant (Davies & Logan 2014, P. 25). Therefore, PAY_# and BILL_AMT# variables should be carefully considered at the model building.

A figure A.2 shows: defaulter (blue) in LIMIT_BAL have the slightly lower amount of credit line than non defaulter. Next, the number of female customer (blue) is bigger than the male customer, but the difference of default ratio of the female (20.77%) and the male customer (24.16%) is 3.39%. At the EDUCATION variable, most of the customers belong to three classes (1, 2 and 3). The default ratio between these classes are respectively 19.23%,

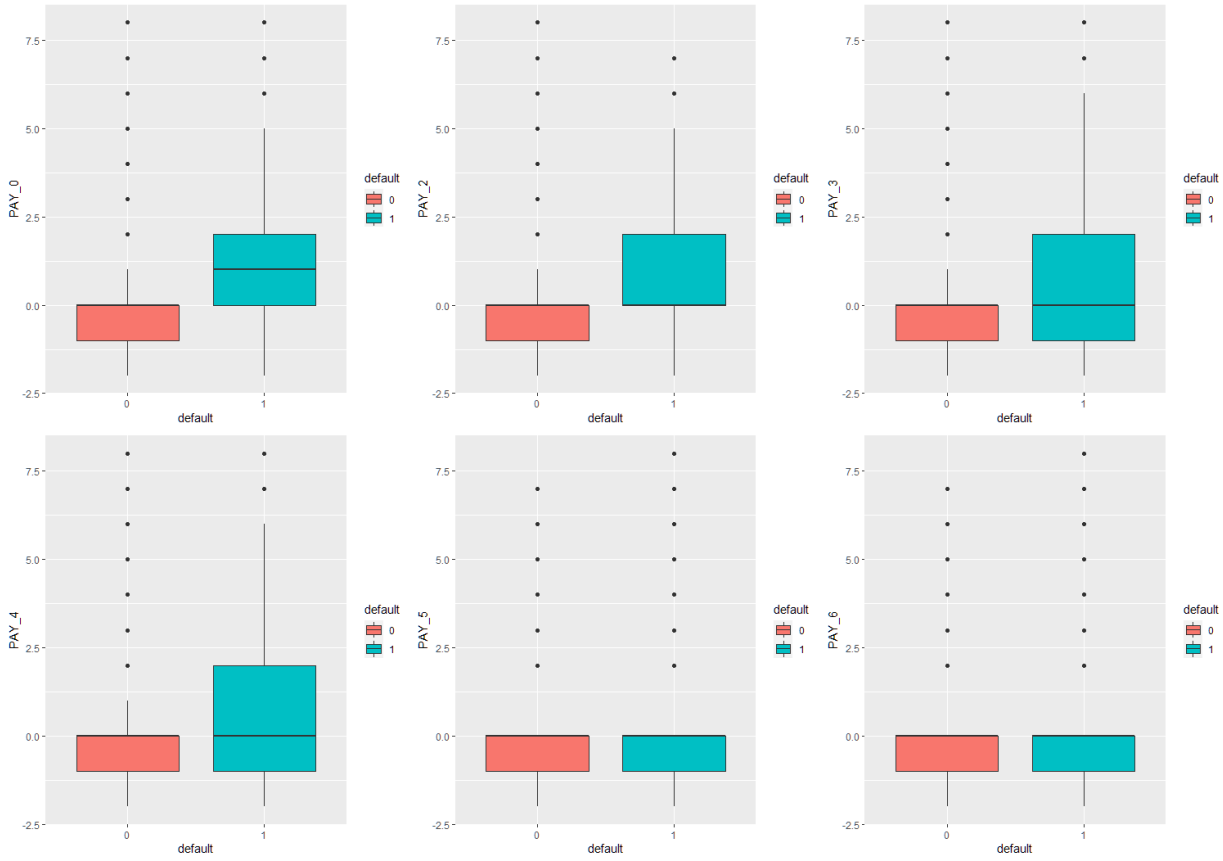

Figure 4.1.: Plots of variable: PAY_#

23.73% and 25.15%. There are also small differences between the classes. The histograms about the MARRIAGE variable show that the most customers are married (class=1) or single (class=2). The number of single customer is slightly higher, but the percentage of default ratio between married (23.47%) and not married customers (20.92%) is about 2.5%. AGE variable in figure A.3 also has a no difference between classes. The two plots, blue and red, seem almost same. Overall, these variables have no big characteristic that may affect the predicting.

In figure 4.1, from PAY_0 to PAY_4, defaulter (blue) have a higher value than the non defaulter's (red) and the difference of PAY_0 and PAY_2 is bigger than of PAY_3 and PAY_4: the higher value means that the payment has been delayed for longer time. In PAY_5 and PAY_6, there is almost no difference between the classes. This means that the more close to the due date (October), the bigger the difference. However, in figure A.4 (BILL_AMT) and A.5 (PAY_AMT), plots of all variables seem similar. There is no difference between defaulter and non defaulter.

4.2. Application of Logistic Regression Model

4.2.1. Preparing for Analysis

The variables SEX, EDUCATION and MARRIAGE are discrete. The discrete variables in the data set should be transformed as categorical variables (Horton & Kleinman 2010, P. 94) in R (LeDell et al. 2020). Then, the data set should split into train and test data set to get a better accuracy of the model. A reason for splitting is: although a model has a good result in the certain data set, this model can have a bad result in another data. Therefore, the model can test its accuracy by the test data that have never been introduced in the model building (Hua et al. 2017). Generally, the ratio between train and test data is 70:30 or 80:20 (Abbasi 2017, P.291). In this study, the ratio is 70:30. There are total 30,000 different cases. The training data set contains 21,000 and the test set contains 9,000 cases.

4.2.2. Logistic Regression Modelling

Table 4.2 shows a summary table from results of a logistic regression model with all variables. The table has five columns, from left to right: variable, coefficient (estimated beta) (Bravais 1844), estimated standard error of beta (Pearson 1893), z-value (test statistic of standard t-test) (Altman 1968) and p-value (Fisher 1925). The p-value is the probability of obtaining results as extreme as the observed results of a statistical hypothesis test if the null hypothesis is actually true (Cumming 2014, P. 129 ff.). As a base value of p-value, 0.01 or 0.05 is used (Butler 2009, P. 481). Among them, 0.05 is widely accepted as the base value (Chin & Lee 2008, P. 130). When the p-value of a variable is lower than the base value, this variable is significant. Conversely, the variable is insignificant when its p-value is higher than the base value.

If the estimated parameter is positive, then the chance will increase. The chance will decrease if the estimated parameter is negative. Besides, if it is zero, the chance will remain the same. In table 4.1, LIMIT_BAL, SEX, PAY_6, BILL_AMT1, BILL_AMT4, BILL_AMT6 and all PAY_AMT variables have negative estimated beta. Furthermore, the size of the estimated beta is also the important factor to the model. It means that the larger the estimated beta is, the bigger effect on the model. The estimated beta of EDUCATION variable is the largest among the variables. This means that it has the biggest effect on the model. In contrast, the estimated beta of LIMIT_BAL, all BILL_AMT# and all PAY_AMT# variables is zero. It means that these variables do not affect on the model.

p-value shows whether the variable is significant or insignificant. For this, the p-value should be compared. To compare them, a base value is required. In this study, 0.05 is used

Table 4.1.: Table from logistic regression with all variables

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-13.4759	86.2000	-0.16	0.8758
LIMIT_BAL	-0.0000	0.0000	-4.62	0.0000
SEX2	-0.1141	0.0367	-3.11	0.0019
EDUCATION1	10.9057	86.1973	0.13	0.8993
EDUCATION2	10.8467	86.1973	0.13	0.8999
EDUCATION3	10.8005	86.1973	0.13	0.9003
EDUCATION4	9.6380	86.1985	0.11	0.9110
EDUCATION5	9.6152	86.1978	0.11	0.9112
EDUCATION6	10.5173	86.1985	0.12	0.9029
MARRIAGE1	1.5467	0.6777	2.28	0.0225
MARRIAGE2	1.3720	0.6777	2.02	0.0429
MARRIAGE3	1.4770	0.6956	2.12	0.0337
AGE	0.0064	0.0022	2.87	0.0041
PAY_0	0.5743	0.0211	27.25	0.0000
PAY_2	0.0835	0.0241	3.47	0.0005
PAY_3	0.0726	0.0270	2.68	0.0073
PAY_4	0.0383	0.0298	1.28	0.1992
PAY_5	0.0303	0.0319	0.95	0.3422
PAY_6	-0.0157	0.0264	-0.60	0.5518
BILL_AMT1	-0.0000	0.0000	-3.95	0.0001
BILL_AMT2	0.0000	0.0000	1.48	0.1400
BILL_AMT3	0.0000	0.0000	0.89	0.3715
BILL_AMT4	-0.0000	0.0000	-0.23	0.8195
BILL_AMT5	0.0000	0.0000	0.68	0.4987
BILL_AMT6	-0.0000	0.0000	-0.66	0.5091
PAY_AMT1	-0.0000	0.0000	-4.92	0.0000
PAY_AMT2	-0.0000	0.0000	-3.24	0.0012
PAY_AMT3	-0.0000	0.0000	-0.68	0.4954
PAY_AMT4	-0.0000	0.0000	-0.72	0.4744
PAY_AMT5	-0.0000	0.0000	-0.92	0.3571
PAY_AMT6	-0.0000	0.0000	-2.02	0.0434

as the base value of the p-value, since it is widely used. In table 4.1, LIMIT_BAL, SEX, MARRIAGE, AGE, PAY_0, PAY_2, PAY_3, BILL_AMT1, PAY_AMT1, PAY_AMT2 and PAY_AMT6 have the p-value less than 0.05. They are significant in this model.

There is a pattern that the variables are significant when they are close to the due date (October). For example, the p-value from PAY_4 to PAY_6 is greater than 0.05 and from PAY_0 to PAY_3 is less than 0.05. The variables of three months (April, May and June) which are relative far from the due date (October) are less significant than the variables of July, August and September. However, this pattern does not correspond to BILL_AMT# and PAY_AMT#: only BILL_AMT1 (amount of bill statement in September) is significant among BILL_AMT#. Also PAY_AMT6 is significant in the model, though it is the amount of previous payment in April.

One remarkable thing is that some variables are significant, but they have no effect on the model: although LIMIT_BAL, BILL_AMT1, PAY_AMT1, PAY_2, and PAY_AMT6 have the p-value less than 0.05, they have zero estimated beta. It means that they are significant variables, but no effect on the model. On the contrary, EDUCATION variable has the highest estimated beta, but it has the p-value much greater than 0.05. Despite of its effect on the model, it is insignificant. For these reasons, these variable can have a possibility that significant variables will be treated as insignificant if they are correlated each other. Therefore, relevant variables should be selected.

4.2.3. Variable Selection

In this study, four models are compared with some criteria such as: first one is the model with all variable (full model), second one is the forward selection model, third one is backward selection model and fourth one is the model with selected variables that have clear difference at the exploratory data analysis. The criteria that are used for drawing comparisons are AIC (Sakamoto et al. 1986), Bayesian information criterion (BIC) (Schwarz et al. 1978) and AUC. The AIC measures a relative goodness of the fitting of a statistical model. The BIC is also used to measure the relative goodness of models. A model which

Table 4.2.: Measurement index of each model: AIC, BIC and AUC

Model	AIC	BIC	AUC
All variables	19,620	19,712	0.7200
Forward selection	27,881	28,097	0.7224
Backward selection	19,615	19,789	0.7199
Model 3	19,835	19,874	0.7068

has the smallest AIC or BIC is preferred (Efroymson 1960). The stepwise function in R selects a formula-based model by Akaike information criterion (AIC) (Sakamoto et al. 1986). The AUC is explained with the ROC curve in chapter 3.4.2.

The first model is already explained in chapter 4.2.2. The second model uses the forward variable selection. This selection uses stepwise regression. The stepwise regression is an automatic procedure that chooses the variables in the regression model. During the stepwise procedure, each variable is considered whether it is added or subtracted in each step. The forward selection starts with no variables in the model. A variable is added when this addition significantly improves a fitting of the model. These procedures last until no more improvement of the model exists. The third model is made by the backward forward variable selection. It is also one of the stepwise regression. But the backward selection starts with all candidate variables and a variable is deleted for improvement of model fitting. This process also last until there is no more improvement of the model. Lastly, the model with selected variables which have clear difference at the exploratory data analysis in chapter 4.1 has only four variables from PAY_0 to PAY_4. In the plot of LIMIT_BAL, SEX, EDUCATION, MARRIAGE, AGE, PAY_5 and PAY_6 in figure A.2, A.3 and A.4, there is no difference between classes. Also, all variables of BILL_AMT# and PAY_AMT# show no difference in figure A.5. Thus, the variables from PAY_0 to PAY_4 are selected even if the variables from PAY_0 to PAY_4 are correlated with each other and with the response variable. If they are correlated, then the significant variable will be treated as insignificant. Nonetheless, the reason for the selection is that only they show the difference. This model is named as model 3 in this study.

Table 4.3 shows a few measurement indexes of the relative goodness for a statistical model: a forward selection model has much higher AIC and BIC value than the other models. Hence, the forward selection model is excluded in this comparing. The backward selection model has the lowest AIC, but the model with all variables has the lowest BIC. The difference of the BIC between the model with all variables and the backward selection model is bigger than the difference of the AIC between these two models. The model with all variables has a next larger AUC, but this is only 0.001 larger than the backward selection model's. In this comparing, model 3 has no advantages from the variable selection. To sum up, in the view of the measurement indexes of the relative goodness, the model with all variables is best model in this study. However, the gaps of the indexes between the models are very narrow.

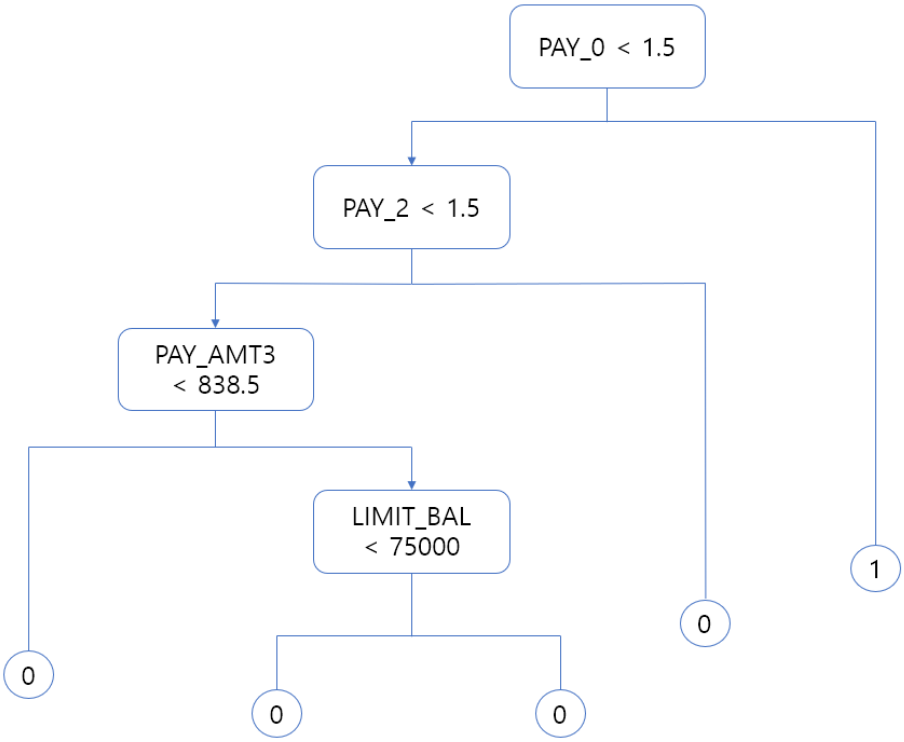

Figure 4.2.: Pruned classification tree

4.3. Application of Classification Tree Model

In this chapter, the classification tree method is used to analyze the data set. To find the best attribute for each split step, criterion or measurement is needed. In chapter 3, three different measures are introduced. However, for the data appliacation, Gini-index is used because it is more sensitive to node purity than classification error rate's sensitivity (Hand 1997, P. 134). A tree before pruning has 16 terminal nodes and its misclassification error rate 18%. In order to prune the tree, it is better to find an ideal size of the tree. A result of K-fold cross-validation shows in figure A.7 that the ideal size of the tree is five. Figure 4.1 shows the pruned tree and has five terminal nodes and its misclassification error rate is also 18%.

In terms of the accuracy, the model accuracy measured by correct classification rate from the pruned tree (81.86%) shows that the classification tree can perform well in classifying the customers. In the tree, four variables are used. These variables are PAY_0, PAY_2, PAY_AMT3, and LIMIT_BAL. Four terminal nodes are 0 and one terminal node is 1. The root node splits data depending on whether the payment of credit in September is delayed less than 1.5 month or not. When the payment is delayed exactly 1.5 month or

more, this customer is classified as defaulter. When the payment is delayed less than 1.5 month, the data is split again by first internal node PAY_2. The first internal node PAY_2 has also a split value 1.5. When the payment is delayed less than 1.5 month, the data is split again by first internal node PAY_AMT3. But, a customer is classified as non defaulter when the payment is delayed 1.5 month or more. The next internal node is PAY_AMT3. When the amount of previous payment is exactly 838.5 or more, then the data is split again by last internal node LIMIT_BAL. But two splits from LIMIT_BAL $<$ 75,000 yield two terminal nodes that have the same predicted value. Even though this split does not reduce the classification error, it leads to increased node purity (Casella et al. 2013, P. 314). For this reason, four terminal nodes from the left are classified as non defaulter. However, a problem of this classification tree is that the structure is monotonous and only PAY_0 variable affects the final result. Even if there are 23 variables in the data set, only PAY_0 is effective. It means that the PAY_0 is an important variable. The reason is that the PAY_0 has recorded the past payment from April to September, since it is the history of past payment and September is only one month before the due date (October). But the problem in practice is that government institution and firms of credit card cannot actually prevent that customers get a credit card who will have a bad credit. Because one month from September to October is too short to make a regulation or classify the credit card holders. Therefore, another approach should be required.

4.4. Application of Random Forest Model

Again, the random forest is applied to the train data set. As explained, random predictor m should be defined. Since there are 23 possible predictors p in data set, m should be defined as $\sqrt{p}$. Next, for the better modelling of random forest, it would be generally preferred to get trees as many as possible. Moreover, larger trees are preferred than smaller trees, because the larger trees with enough depth convey information by splitting each node. However, making many and larger trees costs a lot and it does not significantly decrease the error rate, over a certain number of trees. Thus, the number of tree is limited to 500 at the random forest modelling.

Figure 4.3 visualizes the importance scores of variables from the random forest. First, SEX, EDUCATION and MARRIAGE have relative low level of the importance. The variable PAY_0 has the highest variable importance score. The gap of the importance value between PAY_0 and the other variable is huge. It is common in classification tree method and random forest method that PAY_0 variable is important. However, PAY_2

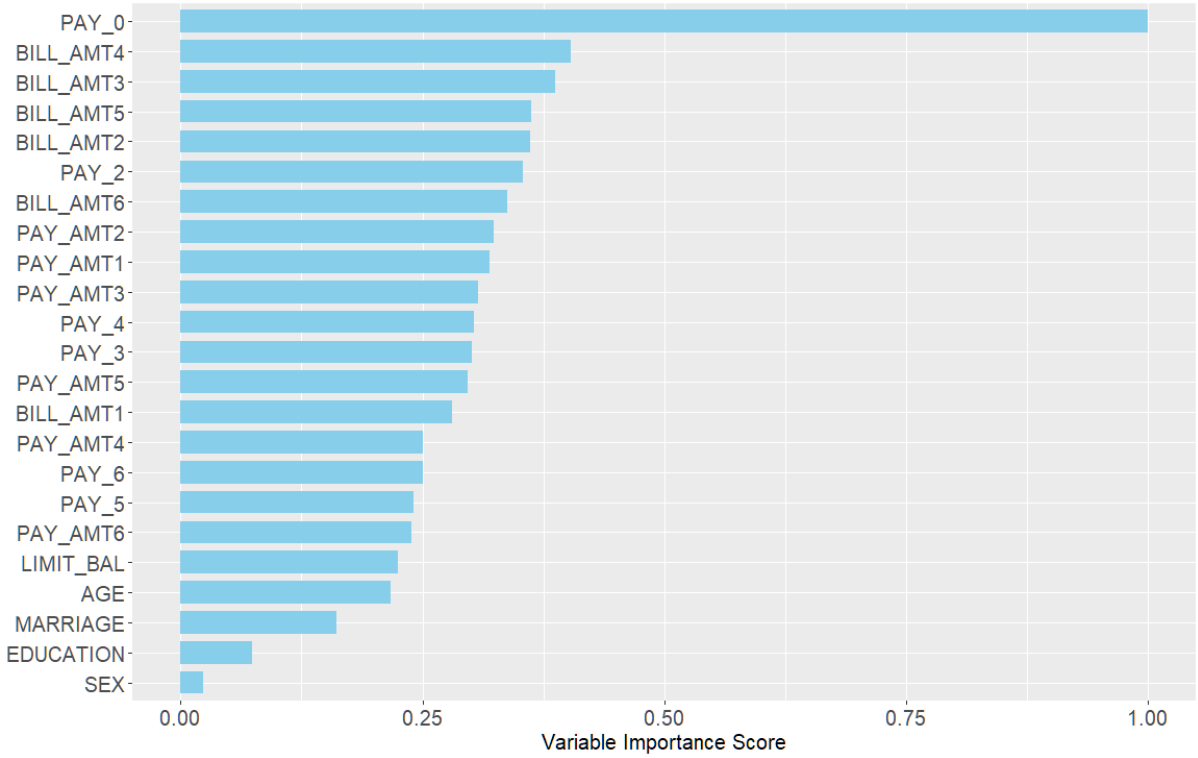

Figure 4.3.: Variable importance score

that locates in the first internal nodes of the classification tree is less important in the random forest model. LIMIT_BALL is also ranked very low in the figure. It is the last internal node of the classification tree. On the contrary, variables from Bill_AMT2 to BILL_AMT5 which are not even showed in the classification tree have more importance in random forest. In conclusion, there are differences between using a single classification tree and using a random forest.

4.5. Comparison of Logistic Regression Model and Classification Tree Model

Logistic regression and classification tree are the popular method of analysis for the binary data. But some points of the result in this study between these two methods are different. First, several variables that seem unimportant in logistic regression, are shown in classification tree as important. For example, LIMIT_BAL and PAY_AMT3 can be considered as unimportant variables in the logistic regression, since their coefficients are very close to zero. However, these variables are located in the third and fourth internal

node of classification tree. The discrete variables, SEX and AGE, are significant in logistic regression. Their p-values are low enough and their coefficients are not zero. But, they are not shown in the classification tree as important and their importance score of the random forest is low.

But, the logistic regression and the classification tree have a similar model accuracy. The logistic regression model with all variables (full model) has 80.94% of accuracy by confusion matrix. The optimal model by the backward selection and the model 3 have 80.87% and 81.01% of accuracy respectively. In the classification tree, the accuracies of trees before pruning and after pruning are same (81.68%). Overall, both the logistic regression model and the classification tree model have the accuracy over 80%. This value means that the model performs well.

If there is no assumption and the problem of variable correlation is considered, then the classification tree will be better for the classifying the model. In figure A.1, some variables suffer the correlation problem. It can affect the significance of variables. In the view of the correlation, the classification tree with no assumptions is better choice for the classifying. However, the classification tree could not cover all variables. In this study, the mentioned variables in the root node and the internal nodes are only four, but there are entirely 23 variables. Therefore, since there are advantages and disadvantages of these two methods, no one can say which model is better than the others. The method should be chosen depending on what the researcher focuses on.

5. Conclusion

This study aims to classify customers who had credit cards into defaulter and non defaulter. In order to achieve the goal, two classification methods are implemented. The result is: the logistic regression model that uses all variables is the best model than the other logistic regression model with selected variables. The classification tree model gives an insufficient model to classify the customer. The random forest model which compensates the tree method shows that PAY_0 variable has the biggest importance score. Among these results, it should be noted that PAY_# variables are the highly influential in the data set. Among them, PAY_0 is regarded as most important.

Although the logistic regression method and the classification tree method have strengthens, these strengthens are not showed well in this study. For example, the logistic regression model especially has a strength to give a measure of the relevancy of a predictor and its direction of association (Massaron & Boschetti 2016, P. 112). But, since some variables (LIMIT_BAL, BILL_AMT# and PAY_AMT#) have zero coefficients, they can not exactly give the measures. Next, the classification tree model does not require normalization of data (Rathore et al. 2018, P. 568) and missing values in the data also do not affect the process of building classification tree (Dangeti 2017, P. 134). However, this data set does not require the data-normalization and there are no missing values in the data set.

Even though logistic regression and classification tree are widely used for the classification problems, it is not enough to classify the data. Therefore, it can be better to consider other methods and there is a few possible improvements. For example: Bayesian network (Pearl 1985), support vector machine (Boser et al. 1992), neural network (Rosenblatt 1958), boosting (Kearns 1988) and others. Because there are many different types of methods, it is possible to find other optimal methods that are suitable for the data set. In the classification tree, another way is to try a different index on pruning. The Gini-index is used for the tree modelling and the pruning. In further study, for example, the entropy can be used. In addition, other tree packages in R can be applied to model the tree. In this study, only *tree* package (Ripley 2019b) was considered for the tree modelling. However, there are two more packages for tree modelling: *ctree* (Hothorn et al. 2006) and *rpart* (Therneau & Atkinson 2019). The *ctree* uses the unbiased recursive partitioning based on permutation tests and

rapart is based on the methodology of classification and regression trees. If *rapart* or *ctree* is used, it is possible that they will give another result.

A. Appendix

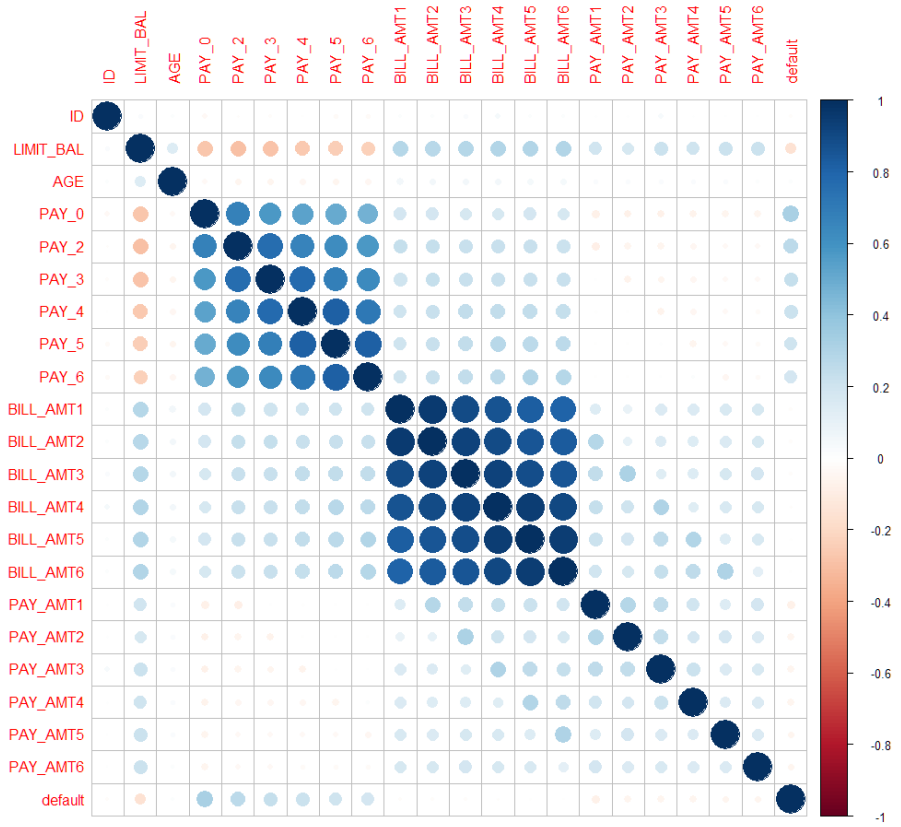

Figure A.1.: Correlation of variables check

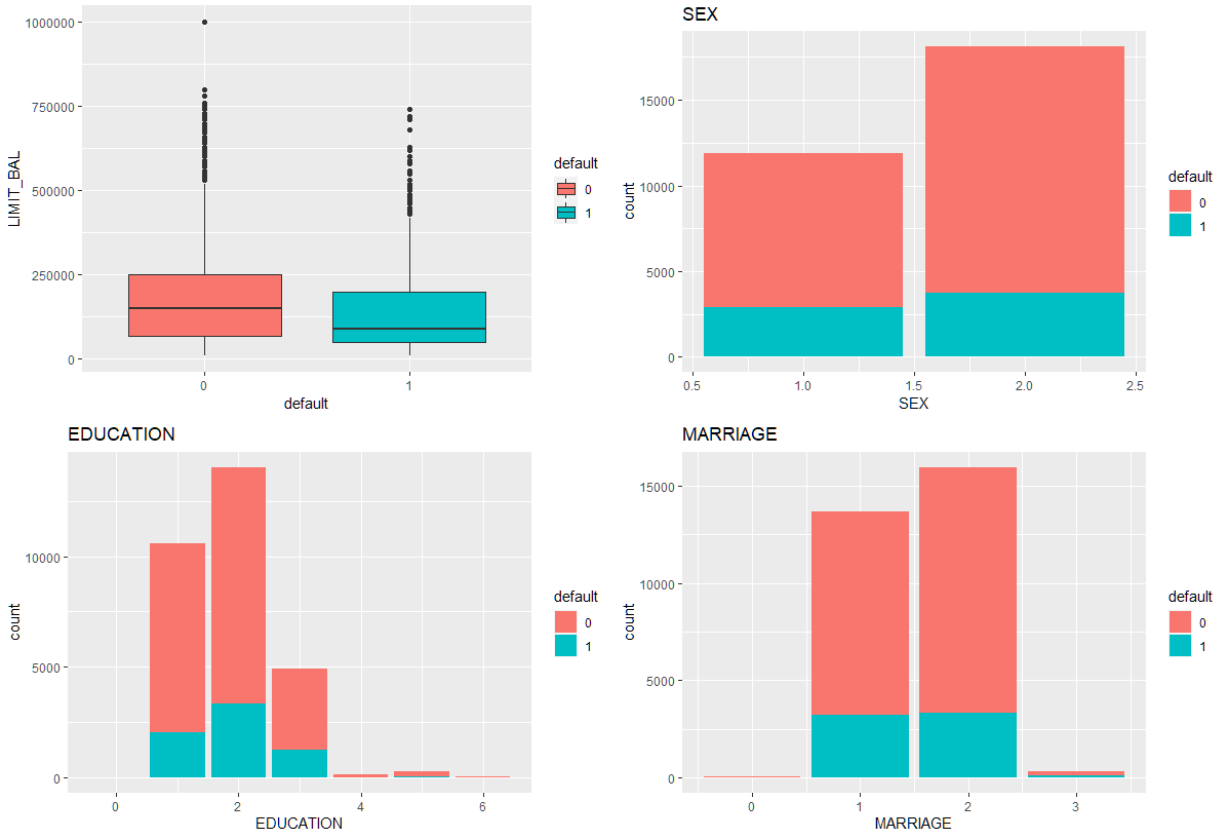

Figure A.2.: Plots of variable: LIMIT_BAL, SEX, EDUCATION and MARRIAGE

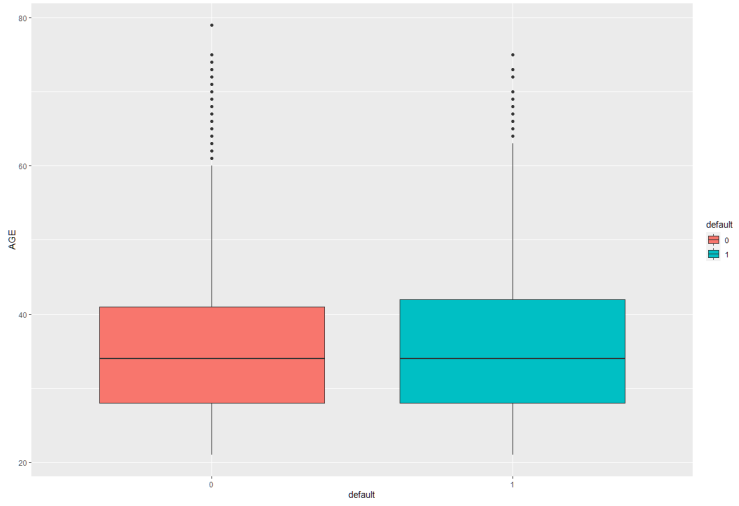

Figure A.3.: Plots of variable: AGE

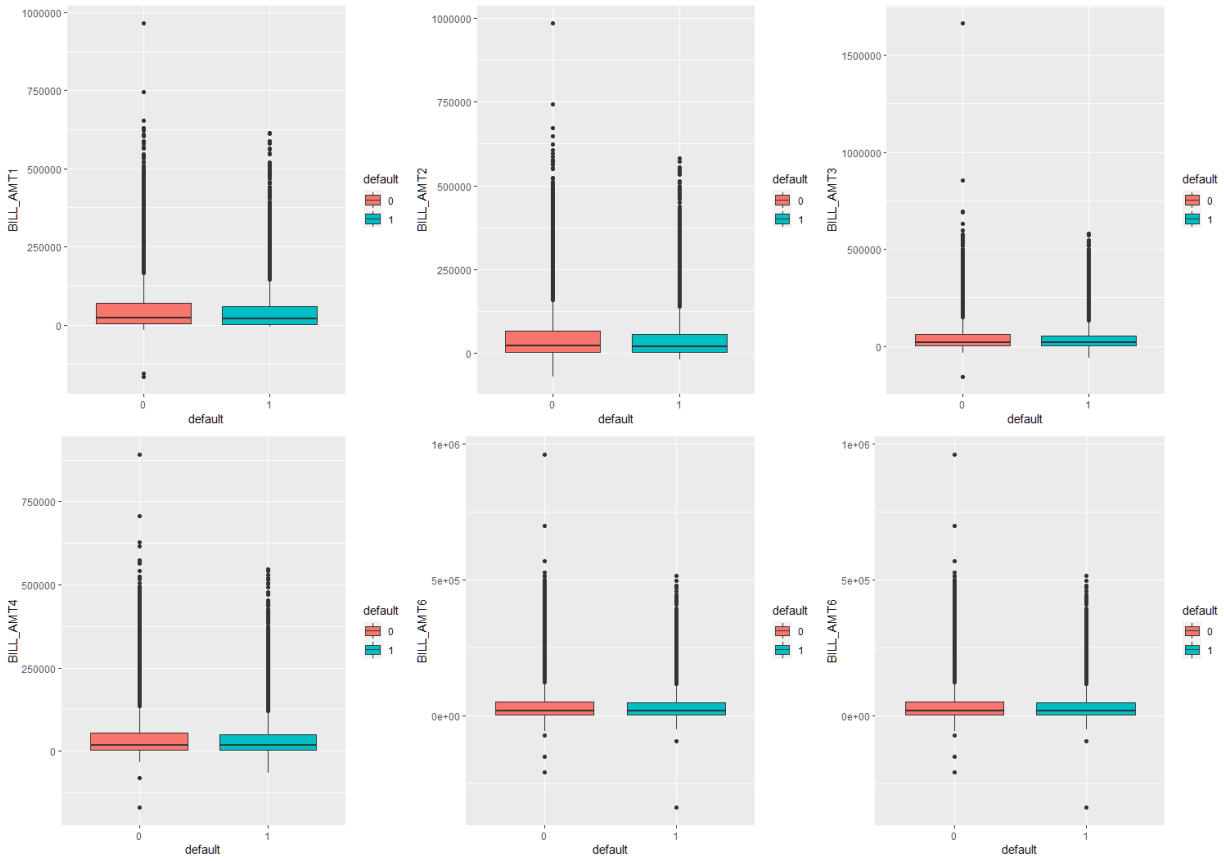

Figure A.4.: Plots of variable: BILL_AMT#

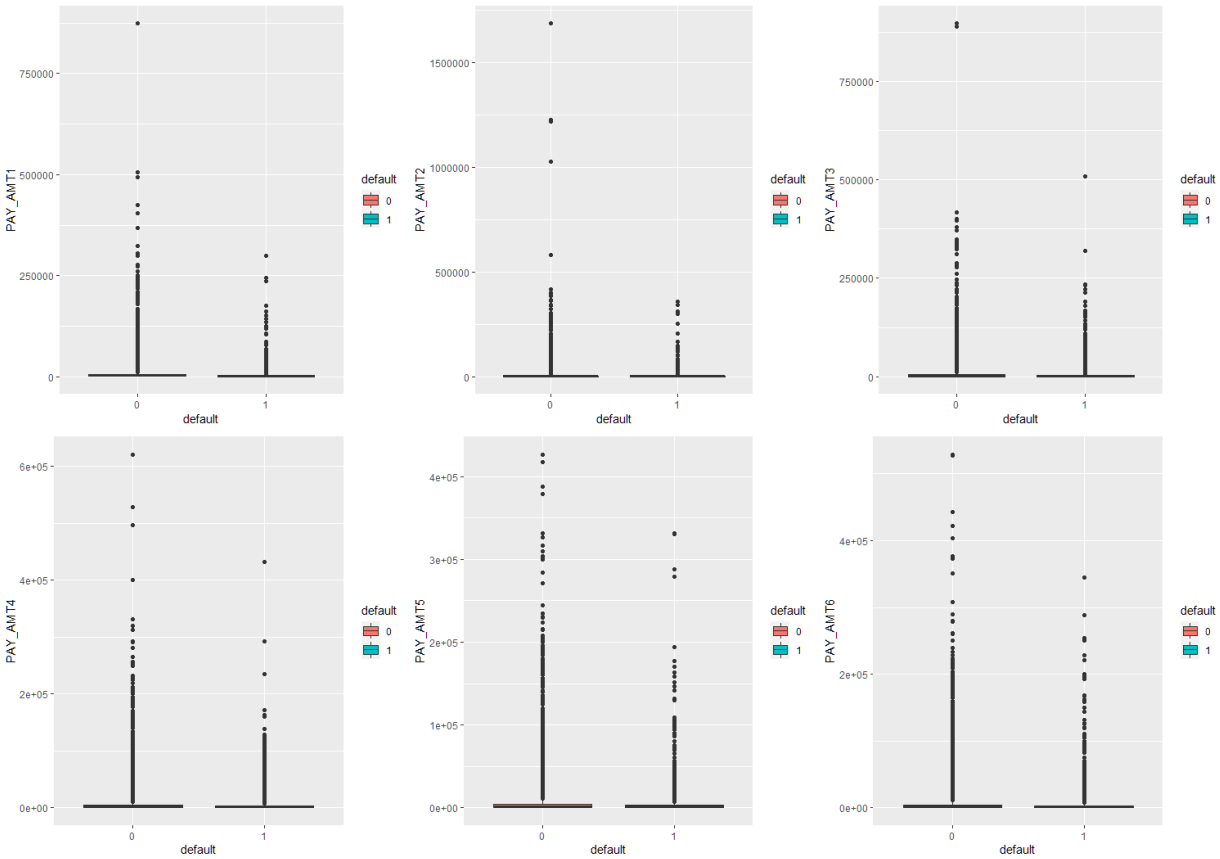

Figure A.5.: Plots of variable: PAY_AMT#

Table A.1.: Table with variables of forward selection from the model with all variables

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-13.1178	82.4391	-0.16	0.8736
PAY_0	0.5793	0.0177	32.79	0.0000
LIMIT_BAL	-0.0000	0.0000	-4.42	0.0000
PAY_3	0.0803	0.0204	3.95	0.0001
PAY_AMT1	-0.0000	0.0000	-6.02	0.0000
BILL_AMT1	-0.0000	0.0000	-4.91	0.0000
MARRIAGE1	1.3180	0.5162	2.55	0.0107
MARRIAGE2	1.1290	0.5164	2.19	0.0288
MARRIAGE3	1.2388	0.5331	2.32	0.0201
EDUCATION1	10.8107	82.4375	0.13	0.8957
EDUCATION2	10.7257	82.4375	0.13	0.8965
EDUCATION3	10.7032	82.4375	0.13	0.8967
EDUCATION4	9.6549	82.4384	0.12	0.9068
EDUCATION5	9.4412	82.4378	0.11	0.9088
EDUCATION6	10.4998	82.4384	0.13	0.8987
PAY_AMT2	-0.0000	0.0000	-4.67	0.0000
BILL_AMT3	0.0000	0.0000	1.98	0.0481
PAY_2	0.0815	0.0202	4.04	0.0001
SEX2	-0.1118	0.0307	-3.64	0.0003
PAY_5	0.0544	0.0177	3.07	0.0021
AGE	0.0054	0.0019	2.89	0.0038
PAY_AMT4	-0.0000	0.0000	-2.14	0.0324
PAY_AMT5	-0.0000	0.0000	-2.12	0.0338
PAY_AMT6	-0.0000	0.0000	-1.71	0.0876
PAY_AMT3	-0.0000	0.0000	-1.60	0.1096
BILL_AMT2	0.0000	0.0000	1.63	0.1039

Table A.2.: Table with variables of backward selection from the model with all variables

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-13.4811	86.2938	-0.16	0.8759
LIMIT_BAL	-0.0000	0.0000	-4.82	0.0000
SEX2	-0.1129	0.0367	-3.08	0.0021
EDUCATION1	10.9046	86.2911	0.13	0.8994
EDUCATION2	10.8449	86.2911	0.13	0.9000
EDUCATION3	10.7985	86.2911	0.13	0.9004
EDUCATION4	9.6375	86.2923	0.11	0.9111
EDUCATION5	9.6204	86.2916	0.11	0.9112
EDUCATION6	10.5233	86.2923	0.12	0.9029
MARRIAGE1	1.5494	0.6772	2.29	0.0221
MARRIAGE2	1.3746	0.6772	2.03	0.0424
MARRIAGE3	1.4758	0.6951	2.12	0.0337
AGE	0.0064	0.0022	2.88	0.0039
PAY_0	0.5759	0.0210	27.39	0.0000
PAY_2	0.0837	0.0239	3.50	0.0005
PAY_3	0.0714	0.0269	2.65	0.0080
PAY_4	0.0547	0.0237	2.31	0.0211
BILL_AMT1	-0.0000	0.0000	-4.08	0.0000
BILL_AMT2	0.0000	0.0000	2.83	0.0047
PAY_AMT1	-0.0000	0.0000	-5.20	0.0000
PAY_AMT2	-0.0000	0.0000	-3.21	0.0013
PAY_AMT5	-0.0000	0.0000	-1.59	0.1113
PAY_AMT6	-0.0000	0.0000	-2.01	0.0449

Table A.3.: Table with variables from model 3

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.3775	0.0188	-73.09	0.0000
PAY_0	0.5982	0.0212	28.26	0.0000
PAY_2	0.0886	0.0235	3.76	0.0002
PAY_3	0.1059	0.0265	4.00	0.0001
PAY_4	0.0395	0.0235	1.68	0.0921

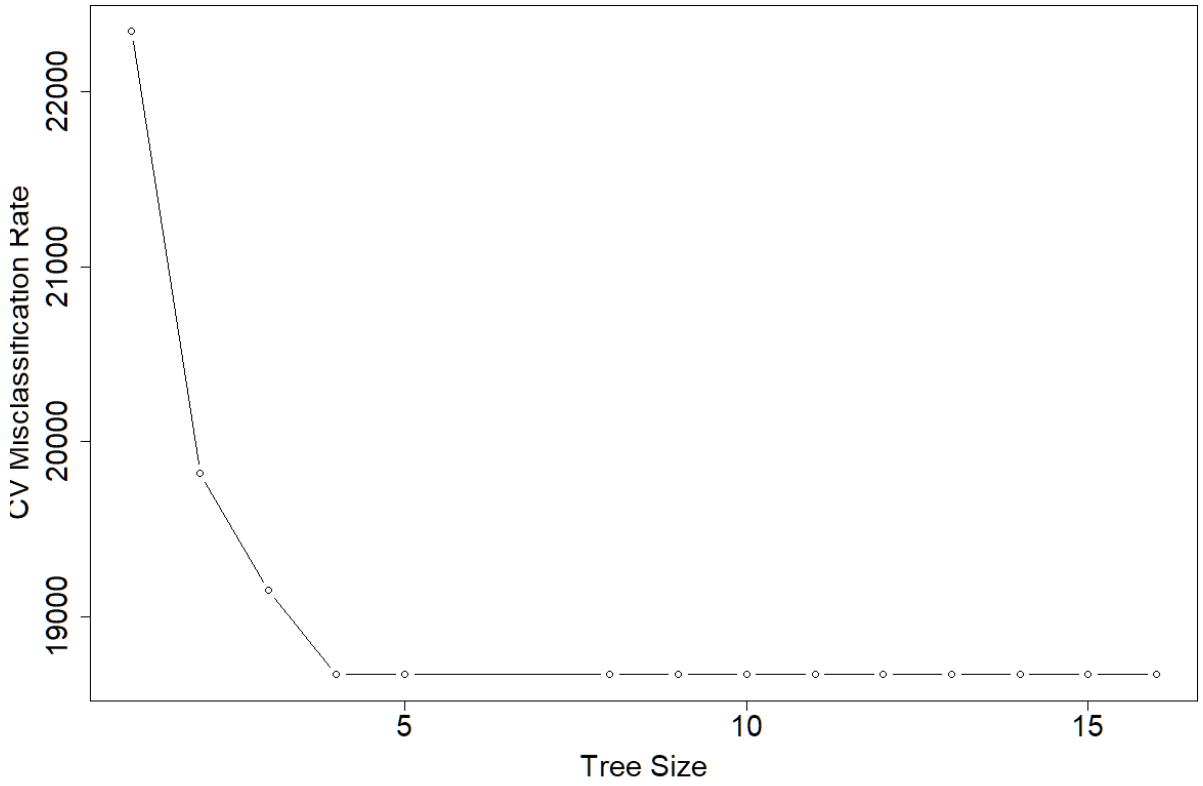

Figure A.6.: K-fold cross-validation from the unpruned tree

References

Abbasi, M. A. (2017). *Learning apache spark 2*. Packt Publishing Ltd.

Ahmad, I. (2020). *40 algorithms every programmer should know - hone your problem-solving skills by learning different algorithms and their implementation in python*. Birmingham: Packt Publishing Ltd.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, *23*(4), 589–609.

Ayyadevara, V. (2018). Pro machine learning algorithms. *Apress, New York*.

Barros, A., van Gulijk, C., Haugen, S., Erik Vinnem, J., & Kongsvik, T. (2018). Safety and reliability-safe societies in a changing world.

Below, M. C. A. L. (n.d.). database marketing analyzing and managing customers.

Belson, W. A. (1959). Matching and prediction on the principle of biological classification. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *8*(2), 65–75.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on computational learning theory* (pp. 144–152).

Bravais, A. (1844). *Analyse mathématique sur les probabilités des erreurs de situation d'un point*. Impr. Royale.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press.

Bruce, P., Bruce, A., & Gedeck, P. (2020). *Practical statistics for data scientists: 50+ essential concepts using r and python*. O'Reilly Media.

Burnham, K. P., & Anderson, D. R. (2002). A practical information-theoretic approach. *Model selection and multimodel inference, 2nd ed. Springer, New York*, *2*.

Butler, J. M. (2009). *Fundamentals of forensic dna typing*. Academic press.

Casella, G., Fienberg, S., & Olkin, I. (2013). Springer texts in statistics.

Chin, R., & Lee, B. Y. (2008). *Principles and practice of clinical trial medicine*. Elsevier.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, *20*(2), 215–232.

Cumming, G. (2014). The new statistics: Why and how. *Psychological science*, *25*(1), 7–29.

Daldrup, A. (2007). *Konzeption eines integrierten iv-systems zur ratingbasierten quantifizierung des regulatorischen und ökonomischen eigenkapitals im unternehmenskreditgeschäft unter berücksichtigung von basel ii* (Vol. 56). Cuvillier Verlag.

Daly, L., & Bourke, G. J. (2008). *Interpretation and uses of medical statistics*. John Wiley & Sons.

Dangeti, P. (2017). *Statistics for machine learning*. Packt Publishing Ltd.

Davies, B., & Logan, J. (2014). *Reading research, fifth canadian edition - e-book - a user-friendly guide for health professionals*. Edinburgh, New York: Elsevier Health Sciences.

Efron, B. (1979). Computers and the theory of statistics: thinking the unthinkable. *SIAM review*, *21*(4), 460–480.

Efroymson, M. (1960). Multiple regression analysis. *Mathematical methods for digital computers*, 191–203.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *222*(594-604), 309–368.

Fisher, R. A. (1925). Theory of statistical estimation. In *Mathematical proceedings of the cambridge philosophical society* (Vol. 22, pp. 700–725).

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2012). *Applied longitudinal analysis* (Vol. 998). John Wiley & Sons.

Fix, E. (1951). *Discriminatory analysis: nonparametric discrimination, consistency properties*. USAF School of Aviation Medicine.

for Economic Co-operation, O., & Development. (2017). *A decade of social protection development in selected asian countries*. OECD.

Fox, J. (2000). *Multiple and generalized nonparametric regression* (Vol. 7). Sage.

Gupta, A. (2015). *Learning apache mahout classification*. Packt Publishing Ltd.

Hand, D. J. (1997). *Construction and assessment of classification rules*. Wiley.

Hartwig, F., & Dearing, B. E. (1979). *Exploratory data analysis* (No. 16). Sage.

Horton, N. J., & Kleinman, K. (2010). *Using r for data management, statistical analysis, and graphics*. CRC Press.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674.

Hua, Q., Azeem, S. U., & Ahmed, S. (2017). *Machine learning with tensorflow 1. x: Second generation machine learning with google's brainchild-tensorflow 1. x*. Packt Publishing Ltd.

Kearns, M. (1988). Thoughts on hypothesis boosting. *Unpublished manuscript*, *45*, 105.

LeDell, E., Gill, N., Aiello, S., Fu, A., Candel, A., Click, C., ... Malohlava, M. (2020). h2o: R interface for the 'h2o' scalable machine learning platform [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=h2o` (R package version 3.30.0.1)

Liu, Y. Y., Yang, M., Ramsay, M., Li, X. S., & Coid, J. W. (2011). A comparison of logistic regression, classification and regression tree, and neural networks models in predicting violent re-offending. *Journal of Quantitative Criminology*, *27*(4), 547–573.

Massaron, L., & Boschetti, A. (2016). *Regression analysis with python*. Packt Publishing Ltd.

Matignon, R. (2005). *Neural network modeling using sas enterprise miner*. AuthorHouse.

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction*. Guilford Press.

Menard, S. (2010). *Logistic regression: From introductory to advanced concepts and applications*. Sage.

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some english consonants. *The Journal of the Acoustical Society of America*, *27*(2), 338–352.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part i. *Biometrika*, 175–240.

O'Connell, A. A. (2006). *Logistic regression models for ordinal response variables* (Vol. 146). Sage.

Olson, D. L., & Delen, D. (2008). *Advanced data mining techniques*. Springer Science & Business Media.

on Financial Services. Subcommittee on Oversight, U. S. C. H. C., Investigations, on Financial Services Subcommittee on Oversight, U. C. H. C., & Staff, I. (2002). *Patriot act oversight: investigating patterns of terrorist financing: hearing before the subcommittee on oversight and investigations of the committee on financial services, us house of representatives, one hundred seventh congress, second session, february 12, 2002* (Vol. 4). US Government Printing Office.

Pearl, J. (1985). Bayesian netwcrks: A model cf self-activated memory for evidential reasoning. In *Proceedings of the 7th conference of the cognitive science society, university of california, irvine, ca, usa* (pp. 15–17).

Pearson, K. (1893). Contributions to the mathematical theory of evolution. *Journal of the Royal Statistical Society*, *56*(4), 675–679.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*(1), 81–106.

Rathore, V. S., Worring, M., Mishra, D. K., Joshi, A., & Maheshwari, S. (2018). *Emerging trends in expert applications and security: Proceedings of iceteas 2018* (Vol. 841). Springer.

Ripley, B. (2019a). tree: Classification and regression trees [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=tree` (R package version 1.0-40)

Ripley, B. (2019b). tree: Classification and regression trees [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=tree` (R package version 1.0-40)

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Rossi, R. J. (2018). *Mathematical statistics: an introduction to likelihood based inference*. John Wiley & Sons.

Rutkowski, L., Jaworski, M., & Duda, P. (2020). *Stream data mining: Algorithms and their probabilistic properties*. Springer.

Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). Akaike information criterion statistics. *Dordrecht, The Netherlands: D. Reidel*, *81*.

Satapathy, S. C., Mandal, J. K., Udgata, S. K., & Bhateja, V. (2016). Information systems design and intelligent applications. *Advances in intelligent System and Computing*, *2*(1), 219–223.

Schwarz, G., et al. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461–464.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, *27*(3), 379–423.

Therneau, T., & Atkinson, B. (2019). rpart: Recursive partitioning and regression trees [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=rpart` (R package version 4.1-15)

Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2). Reading, MA.

Vapnik, V., & Chervonenkis, A. (1974). *Theory of pattern recognition.* Nauka, Moscow.

Wald, A. (1939). Contributions to the theory of statistical estimation and testing hypotheses. *The Annals of Mathematical Statistics*, *10*(4), 299–326.

Williams, G. (2011). *Data mining with rattle and r: The art of excavating data for knowledge discovery*. Springer Science & Business Media.

Woodward, P. (1953). *Probability and information theory, with applications to radar. new york: Mcraw-hill book co. inc.* London: Pergamon Press Ltd. First published.

Yeh, I.-C. (2009). Default of credit card clients data set [Computer software manual]. Retrieved from `https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients#` (Accessed: 23.04.2020)

Yeh, I.-C., & Lien, C.-h. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, *36*(2), 2473–2480.

Declaration of Authorship

I hereby affirm that I have written this thesis independently and to the best of my knowledge and belief. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I declare that all information sources and literature used are indicated in the thesis.

Berlin, 17. July 2020

.....

(Sign of Author)