Classification Consumer Credit for Missing Value Dataset

I Noviandi* and I D Sumitra
Universitas Komputer Indonesia (UNIKOM), Bandung, Indonesia

*noviiandiqbal@gmail.com

Abstract. The objective of the study is to find the best method to construct a model that could predict the future failure as a function of variables obtained from the customer profile. Decision Tree and Logistic Regression are classification algorithm. One of Decision Tree algorithm is Classification and Regression Tree (CART). It can used to analyze numeric and categorical data. Logistic Regression is more accurate than Decision Tree. In fact, there is some missing value in datasets. Amelia II is the best method to estimate missing value for numeric and categorical data. This study combines Amelia II to estimate missing value, Decision Tree to screening and re-categorization variable and Logistic Regression to classifying debtor into ‘good’ and ‘bad’ risk classes. We found that the accuracy of this combined method constant until 40% missing value. The Correct Classification Rate (CCR) value for 10% - 40% same as the CCR value for dataset without missing value. Otherwise, the accuracy decreased for missing value above 40%. This method is effective if missing value of the dataset below 40%. We recommend the bank to apply this method for classify risk of debtor if the missing value is below 40%.

1. Introduction
Credit decisioning is an important financial problem that requires many decision variables in a continually changing market. The objective is to construct a model that could predict the future failure as a function of variables from the customer profile. Machine Learning techniques have been proposed for this problem. However, the outcomes are difficult to understand. They are difficult to apply in heavily regulated industry [1]. In fact, there is some missing value in datasets. We need to solve this problem as well.

Classification is one of data analysis task. It is used to predict future data trends by generate a model. Logistic regression and decision tree are the two most popular algorithms of classification used by previous research. They are perform well [2]. This study combines Amelia II to estimate missing value, Decision Tree to screening and re-categorization variable and Logistic Regression to classifying debtor into ‘good’ and ‘bad’ risk classes.

Decision Trees are easy to understand. User easy to interpret the outcome [3]. Logistic regression is a regression model where the dependent variable is categorical. It is computationally inexpensive, easy to explain, useful in knowledge representation, and easy to interpret [4]. Blankers et al [5] did a research to compare 4 imputation programs. There are 4 methods to be compared NORM, MICE, Amelia II, and SPSS MI. Based on the research, Amelia II is the most accurate than the others. The objective of the study is to find the best method to construct a model that could predict the future failure as a function of variables obtained from the customer profile.
2. Experimental Method

The data for this analysis is German Credit Dataset as shown in Table 1. University of California Irvine (UCI) repository of machine learning provided it. There is 1000 customer in the dataset. This dataset contained 20 numerical and categorical independent variable. The dependent variable consists of two categories that classifying debtor into ‘good’ and ‘bad’ risk. There is 700 ‘good’ risk debtor (70%) and 300 ‘bad’ risk debtor (30%). This dataset used by the scientist to research about logistic regression and decision tree several times.

| No  | Variable                                           | Type      |
|-----|----------------------------------------------------|-----------|
| 1   | Status of existing checking account                | Categoric |
| 2   | Duration in Month                                  | Numeric   |
| 3   | Credit History                                     | Categoric |
| 4   | Purpose                                            | Categoric |
| 5   | Credit amount                                      | Numeric   |
| 6   | Savings account/bonds                              | Categoric |
| 7   | Present employment since                           | Categoric |
| 8   | Installment rate in percentage of disposable income| Numeric   |
| 9   | Personal status and sex                            | Categoric |
| 10  | Other debtors / guarantors                         | Categoric |
| 11  | Present residence since                            | Numeric   |
| 12  | Property                                           | Categoric |
| 13  | Age in years                                       | Numeric   |
| 14  | Other installment plans                            | Categoric |
| 15  | Housing                                            | Categoric |
| 16  | Number of existing credits at this bank            | Numeric   |
| 17  | Job                                                | Categoric |
| 18  | Number of people being liable to provide maintenance for | Numeric |
| 19  | Telephone                                          | Categoric |
| 20  | foreign worker                                     | Categoric |

The Research stage shown in Figure 1.

- Split the dataset into 90% data training to build the model and 10% data testing for accuracy measurement.
- Generate missing value from 10% to 90% from data training. Therefore, we have 10 datasets.
- Impute missing value using Amelia II for datasets with missing value. Therefore, we have 10 complete datasets.
- Screening and Recategorization the variables of this 10 datasets using CART.
- Use Logistic Regression to generate a model from the datasets that the variables have been screening and recategorization by CART in previous step.
- Measure the accuracy using CCR.

3. Results and Discussion

3.1. Imputing Missing Value Using Amelia II
Missing values is vital to handle in order to manage data successfully. If the missing values are not handled correctly then the result of analyze can be inaccurate. Due to improper handling, the result obtained will differ from ones where the missing values are present. We used Amelia II to imputing missing value for a dataset with 10% - 90% missing value. For dataset with 80% missing value there is 10 observation that all of the variables are missing, and for dataset with 90% missing value there is 103 observation that all of the variables are missing. In this case, Amelia cannot impute the missing value. After this step, we have 10 complete datasets.

3.2. Screening and Recategorization Variable using CART
The next step is screening and recategorizing variable using CART. Leo Breiman et al developed Classification and Regression Tree (CART) in 1982. CART is capable to analyze numerical and categorical data. The parameter for this method is Gini Index. Gini Index is a parameter to separate independent variable into targeted class. It always separate independent variable into two group. It provides a hierarchy binary decision tree [7]. The result of screening variable as shown in Table 2.

Table 2. Screening Variable using CART.

| Dataset            | Variable Used in Tree Construction                                      |
|--------------------|-------------------------------------------------------------------------|
| Full Dataset       | checking_status; credit_amount; credit_history; duration; purpose       |
|                    | savings_status                                                          |
| Missing Value 10%  | checking_status; credit_amount; credit_history; duration                |
|                    | employment; property_magnitude; purpose; savings_status                 |
|                    | age; checking_status; credit_amount                                     |
| Missing Value 20%  | credit_history; duration; employment; installment_commitment; other_payment_plans; purpose |
| Missing Value 30%  | age; checking_status; credit_history; duration; employment; job; purpose |
|                    | residence_since; savings_status                                         |
| Missing Value 40%  | age; checking_status; duration                                           |
|                    | employment; installment_commitment; personal_status; property_magnitude; purpose |
| Missing Value 50%  | checking_status; credit_amount; duration; num_dependents; other_payment_plans; residence_since |
| Missing Value 60%  | age; checking_status; credit_amount; credit_history; duration; existing_credits; property_magnitude; purpose |
|                    | residence_since; savings_status                                         |
| Missing Value 70%  | age; checking_status; credit_amount; employment; housing; property_magnitude; purpose; savings_status |
| Missing Value 80%  | age; credit_amount; credit_history; duration; employment; housing       |
|                    | installment_commitment; num_dependents; personal_status; purpose; savings_status |
| Missing Value 90%  | age; checking_status; credit_amount; duration                           |
|                    | employment; job; other_payment_plans; own_telephone; personal_status; purpose |
Table 2 shows the result of screening variables using CART. The number of variables that affect different for each dataset. Only effected variable used in logistic regression. Variables that have no effect will be excluded. Each effected variable was recategorization into two categories. Splitting category rules based on decision tree result from the CART. The example of the decision tree as shown in Figure 2.

**Figure 2.** Decision Tree for Full Dataset.

**Figure 2** shows the Decision Tree for Full Dataset using CART. In that decision tree, checking status was recategorized into two categories, BC, and AD. Previously this variable were 4 categories. So does the others categorical variable. Duration, one of numerical was categorized into two categories, <22 and >=22. The recategorize variables that used in decision tree will be analyzed in logistic regression.

### 3.3. Generate Model Using Logistic Regression

Logistic regression is computationally inexpensive, easy to explain, right in knowledge representation, and easy to interpret. Logistic regression could build a model that can predict the risk of each observation. It proven as a robust algorithm [8]. Logistic regression has been used in many research applications such as social research, medical research, prediction of bankruptcy, market segmentation and customer behaviour [9, 10, 11, 12].

The accuracy of Logistic Regression model could be measured by Correct Classification Rate (CCR). CCR is how much the model can predict correctly if data testing was inputted to the model. The CCR of the model shown in **Figure 3**. According to this figure, we can see that the CCR constant in between 74% - 75% for up to 40% missing value. This value same as dataset without missing value. Otherwise, the CCR decrease to 63% - 66% for 50%-66% missing value.
4. Conclusions
This study aims to analyze the performance of combine methods Amelia II, CART and Logistic Regression. Amelia II for imputing missing value, CART for screening and recategorization variable and Logistic Regression for modeling. We found that the accuracy of this combined method in between 74% - 75% for up to 40% missing value. The CCR value for missing value below 40% same as the CCR value for dataset without missing value. Otherwise, the accuracy decreased for missing value above 40%. This method is effective for dataset with missing value below 40%.

Acknowledgements
We acknowledged Universitas Komputer Indonesia (UNIKOM) for supporting this study and UCI Repository of Machine Learning for provide the dataset.

References
[1] Khandani A E, Kim A J and Lo A W 2010. Consumer credit-risk models via machine-learning algorithms Journal of Banking and Finance 34(11), 2767-2787
[2] Hadden J, Tiwari A, Roy R and Ruta D 2007 Computer assisted customer churn management: State-of-the-art and future trends Computers and Operations Research 34(10) 2902-2917
[3] Singh D 2014 Analysis of Data Mining Classification with Decision tree Technique (Doctoral dissertation, MPUAT, Udaipur)
[4] Harrington P 2012 Machine learning in action 5 pp 12-20 Greenwich, CT: Manning
[5] Blankers M, Koeter M W and Schippers G M 2010 Missing data approaches in eHealth research: simulation study and a tutorial for nonmathematically inclined researchers Journal of medical Internet research 12(5) pp 120-140
[6] Bache K and Lichman M 2013 UCI machine learning repository
[7] Venkatesan E and Velmurugan T 2015 Performance analysis of decision tree algorithms for breast cancer classification Indian Journal of Science and Technology 8(29) 1-8
[8] Lee T S, Chiu C C, Chou Y C, and Lu C J 2006 Mining the customer credit using classification and regression tree and multivariate adaptive regression splines Computational Statistics and Data Analysis 50(4) 1113-1130
[9] Flagg J C, Giroux G A and Wiggins C E 1991 Predicting corporate bankruptcy using failing firms Review of financial Economics 1(1), 67
[10] Olsen W K, Warde A and Martens L 2000 Social differentiation and the market for eating out in the UK International Journal of Hospitality Management 19(2) 173-190
[11] Suh E H, Noh K C and Suh C K 1999 Customer list segmentation using the combined response model Expert systems with Applications 17(2) 89-97
[12] Laitinen E K and Laitinen T 2000 Bankruptcy prediction: Application of the Taylor's expansion in logistic regression International review of financial analysis 9(4) 327-349