# THE MOST IMPORTANT VARIABLES FOR CREDIT RISK MODEL

*E.I. Gubin Ph.D., Associate Professor*
*T. Phana, student 8PM9I*
*Tomsk Polytechnic University*
E-mail: phana@tpu.ru

## Introduction

Banks and finance organizations very often use credit risk models for evaluation potential customers when providing credit. Traditional credit risk uses a method, which weights various factors including credit history, length of credit history, types of credit used and recent credit inquiries and so on. The important part to create the credit risk model is the variable reduction by using statistical methods SAS. Among the variety of statistical methods that are employed to analyze data set very popular regression methods. There are widely used in examining the relationship between target variable and other predictor variables. For class of regression methods, logistic regression is well suited for studying categorical variables. Logistics regression also helpful choosing the most important variables of the predictive model to make the model more accurate.

## Research methods

Firstly, obtain the data set of customer's information of the bank and have the target variable in category type (Good/Bad, 0/1, or etc.). We have variables "GB" – target variable. If "GB"=0, then "Good customer" and if "GB"=1 then "Bad customer"

Before to analyze create credit risk model we need to do four steps: 1) To choose target variable; 2) To cleaning data sets. 3) To choose most important variables. 4) And finally create the credit risk model (scorecard). For choosing most important variables we will use logistic regression SAS. For data partition, we will divide data set to training data and test data. After that, we will evaluate the training and validation data with logistic regression**.**

## Results

In the Table 1 show about description of all the variables in this dataset and type of the variable The dataset contains all the basic information of the customer that want to loan money from the bank. In this dataset has 2993 customers (rows) and 26 variables (attributes). In the Table 1 show about description of all the variables in this dataset and type of the variable. This dataset already passed the cleaning data method [1].

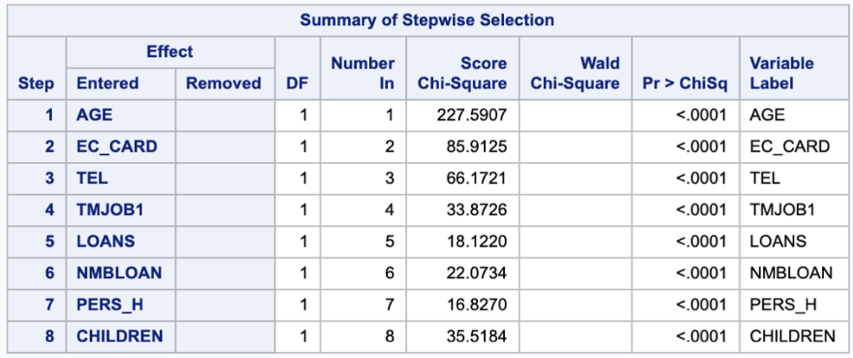
Table 1. Dataset variables

| *Variable* | *Explanation* | *Format* |
|---|---|---|
| *AGE* | *Age* | *Numeric* |
| *BUREAU* | *Credit Bureau Risk Class* | *Numeric* |
| *CAR* | *Type of Vehicle* | *Char* |
| *CARDS* | *Credit Cards* | *Char* |
| *CASH* | *Requested cash* | *Numeric* |
| *CHILDREN* | *Num of Children* | *Numeric* |
| *EC_CARD* | *EC_card holders* | *Numeric* |
| *FINLOAN* | *Num finished Loans* | *Numeric* |
| *GB* | *Good/Bad* | *Numeric* |
| *INC* | *Salary* | *Numeric* |
| *INC1* | *Salary+ec_card* | *Numeric* |
| *INCOME* | *Income* | *Numeric* |
| *LOANS* | *Num of running loans* | *Numeric* |
| *LOCATION* | *Location of Credit Bureau* | *Numeric* |
| *NAT* | *Nationality* | *Char* |
| *NMBLOAN* | *Num Mybank Loans* | *Numeric* |
| *PERS_H* | *Num in Household* | *Numeric* |
| *PRODUCT* | *Type of Business* | *Char* |

| | | |
|---|---|---|
| *PROF* | *Profession* | *Char* |
| *REGN* | *Region* | *Numeric* |
| *RESID* | *Residence Type* | *Char* |
| *STATUS* | *Status* | *Char* |
| *TEL* | *Telephone* | *Numeric* |
| *TITLE* | *Title* | *Char* |
| *TMADD* | *Time at Address* | *Numeric* |
| *TMJOB1* | *Time at Job* | *Numeric* |

Implementation of logistic regression models in the SAS System in PROC LOGISTIC is very similar to how OLS regression models are implemented in PROCs REG and GLM. If you are already familiar with how to perform OLS regression in PROC REG, then learning how to use PROC LOGISTIC for binary outcome modeling is a straightforward task. As with PROCs REG and GLM, separate output data sets containing predicted values and parameter estimates can be created for subsequent analysis or “scoring.”

For data analysis, we are used logistic regression PROC LOGISTIC (SAS) [2-5] and splitted the data into two datasets 70% into training data and 30% into validation data.



**Summary of Stepwise Selection**

| Step | Effect Entered | Effect Removed | DF | Number In | Score Chi-Square | Wald Chi-Square | Pr > ChiSq | Variable Label |
|---|---|---|---|---|---|---|---|---|
| 1 | AGE | | 1 | 1 | 227.5907 | | <.0001 | AGE |
| 2 | EC_CARD | | 1 | 2 | 85.9125 | | <.0001 | EC_CARD |
| 3 | TEL | | 1 | 3 | 66.1721 | | <.0001 | TEL |
| 4 | TMJOB1 | | 1 | 4 | 33.8726 | | <.0001 | TMJOB1 |
| 5 | LOANS | | 1 | 5 | 18.1220 | | <.0001 | LOANS |
| 6 | NMBLOAN | | 1 | 6 | 22.0734 | | <.0001 | NMBLOAN |
| 7 | PERS_H | | 1 | 7 | 16.8270 | | <.0001 | PERS_H |
| 8 | CHILDREN | | 1 | 8 | 35.5184 | | <.0001 | CHILDREN |

Finally, as we can see, from 26 initial variables algorithm selected only 8 most important variables: AGE, EC_CARD, TEL, CHILDREN, TMJOB1, NMBLOAN, LOANS, and PERS_H. Another 18 variables is not so important because they influence on target variable “GB” not significant.
We tested the probability (PR>ChiSq) of observing a Chi-Square statistic as extreme as, or more so, than the observed one under the null hypothesis; the null hypothesis is that all of the regression coefficients in the model are equal to zero. Typically, PR>ChiSq is compared to a specified alpha level, our willingness to accept a type I error, which is often set at 0.05 or 0.01. The small p-value from the all three tests would lead us to conclude that at least one of the regression coefficients in the model is not equal to zero.

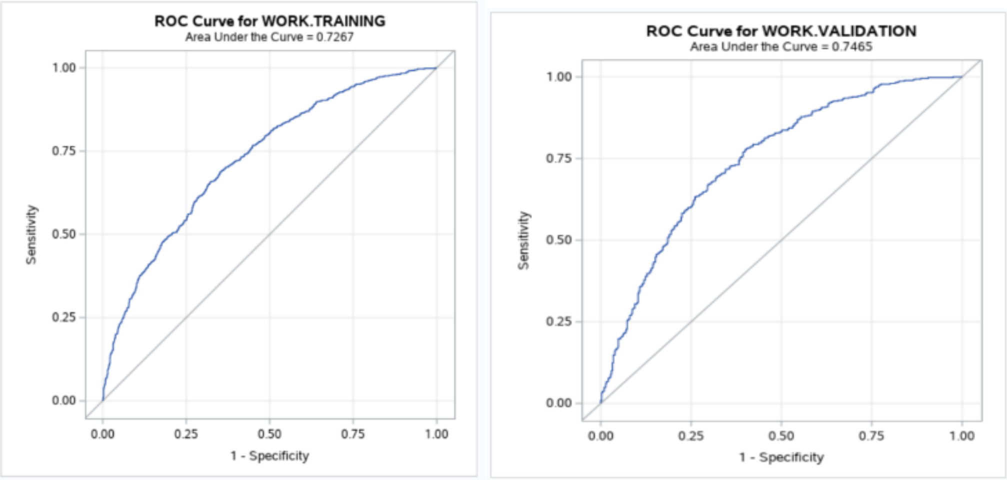


Fig. 1. ROC curve to show possibility of TRAINING Dataset and VALIDATIN Dataset

In Figure 1, we can see ROC curve is to the upper left corner, the higher the overall accuracy of the test. Therefore, the ROC curve on the TRAINING data shows that it has 72.67% accuracy, and for the VALIDATION, data has 74.65% accuracy. This result is quite well!

## Conclusion

After analysis initial data, using Logistic regression model, our predict model show good result – ROC curve for validation data more than 74%. We can see that the variable selection algorithm decided that the model would include CHILDREN, PERS_H, AGE, TMJOB1, TEL, NMBLOAN, EC_CARD and LOANS variables. In addition, our results are 72.67% accuracy for the TRAINING dataset and 74.65% accuracy for the VALIDATION dataset.

## References

1. Huang Shan, Gubin E**.** Data cleaning for data analysis // Молодежь и современные информационные технологии: Труды XVI Междунар. научно - практической конференции студентов, аспирантов и молодых ученых. Томск, 2018г. - С. 387-389.
2. Kallunki J[1].-P., Broussard J., Boehmer E. Using SAS in Financial Research (2002)(en)(166s)
3. Allison, P. (2012). Logistic Regression using SAS (Theory & Applications), 2nd edition. SAS Institute, Cary, NC
4. Baesens B., Rosch D., & Schenle H. (2017). Credit Risk Analytics. Wiley & SAS Institute
5. Agresti, A. (2013). Categorical Data Analysis, 3rd edition. John Wiley & Sons, Inc., Hoboken, NJ