Northwestern MSDS-498 Artificial

Model #101: Credit Card Default Model

Model Development Guide

Andrew Stevens

September 1, 2022

# Introduction

.

# The Data

University of California Irvine hosts a Machine Learning Repository (Dua and Graff 2019) which includes the default of credit card clients in Taiwan prepared to compare the predictive abilities of selected data mining methods (Dua and Graff 2019). The response variable of the dataset is a binary indicator for whether a customer defaulted on their credit card debt. Delinquency is defined as missing a single payment due date, while default is not making a specific number of consecutive payments (Cagan 2020). Entering default involves collections actions and likely losses for the creditor, so a company would seek customers unlikely to default.

The predictor variables included in the dataset can be divided into two categories about the customer: demographic attributes and billing/payment history. The demographic attributes are comprised of SEX, EDUCATION, MARRIAGE, and AGE. The billing/payment history variables comprise six months of history including repayment status, billing amount, and payment amount.

Before the data can be engineered into features consumable by different modeling methods, each datatype and feature class must be reviewed for sufficient and consistent data quality. The dataset was first checked for empty values, and zero nullity was reported across all fields. The dataset was focused to only explanatory and target variables.

ccd\_focus <- subset(credit\_card\_default, select=ID:DEFAULT)

Then the dlookr () packages describe function was used to generate descriptive statistics, and the field of interest `na` was searched for values greater than 0 (indicating any empty records).

desc <- describe(ccd\_focus)

any(desc$na > 0)

>> [1] FALSE

The data dictionary must next be used to review whether invalid values exist and must be cleaned. Appendix A provides the complete dictionary with each field explicitly defined.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 1: Data Dictionary - Abridged** | | |  |
|  | |  |  |
| Fields | Variable | Valid  Values |  |
|  | |  |  |
| X1 | LIMIT\_BAL | > 0 |  |
| X2 | SEX | (1,2) |  |
| X3 | EDUCATION | (1:4) |  |
| X4 | MARRIAGE | (1:3) |  |
| X5 | AGE | Int, >0,<120 |  |
| X6-X11 | PAY\_# | (-1,1:9) |  |
| X12-X17 | BILL\_AMT# | numeric |  |
| X18-X23 | PAY\_AMT# | numeric |  |
| Z | DEFAULT | binary |  |
|  | |  |  |

# Feature Engineering

In the practice of credit risk modeling, features are usually engineered by aggregating customer transactional data to determine behavioral patterns (Bahnsen et al 2016). We will also consider and test approaches for binning and combining demographic attributes of the customer, dependent on each specific model’s needs.

The AGE attribute is received as integers indicating years of age for each customer. Because age is a discrete variable with high cardinality, discretization can bring it closer to a knowledge-level representation (Peng et al 2009) and is essential for models such as trees/forests. Age has been initially separated by decade, and testing will be performed on more evenly distributed bins or perhaps other approaches.

|  |  |
| --- | --- |
| **Table 2: Resulting distribution of Age Binning** | |
|  | |
| Age\_Group | Freq |
|  | |
| 1-10 | 0 |
| 11-20 | 0 |
| 21-30 | 11,013 |
| 31-40 | 10,713 |
| 41-50 | 6,005 |
| 51-60 | 1,997 |
| 61-70 | 257 |
| 71-80 | 15 |
|  | |

Weight of evidence binning was also tested, which divided the AGE attribute into four classes and the ‘separation’ of response results indicates that it will be a more effective means than based only on decade.

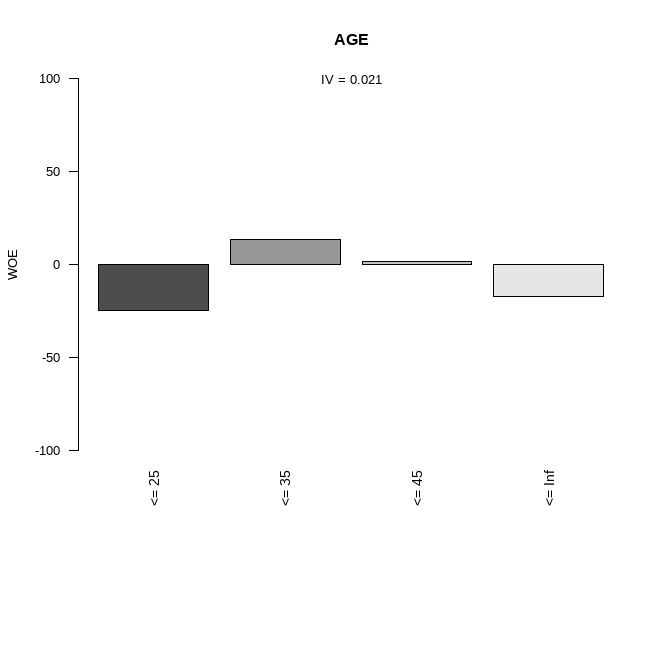


Figure 1: Weight of Evidence Binning Result

In order to produce variables useful and potentially meaningful to statistical models, all transactional data will be replaced with aggregated and computed statistics.

* List out max, avgs…. Generated
* (eng\_credit\_card\_default\_data.R)

# Exploratory Data Analysis

I wasted time attempting several R EDA pacakages that i coudln’t get working (e.g. eda\_web\_report). Didn’t realize till late that One Rule Classification module was for EDA. I’ve got laptop problems right now & don’t seem to have sufficient resources for OneR

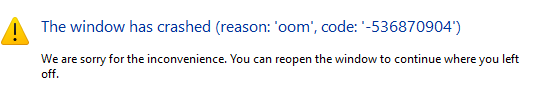


Figure 2: FAIL

Or even exporting to csv….

Error: cannot allocate vector of size 234 Kb

Error during wrapup: cannot allocate vector of size 4.6 Mb

Error: no more error handlers available (recursive errors?); invoking 'abort' restart

The target variable, DEFAULT, is imbalanced, though not severely. This will require adjustment to accuracy measure and potentially modeling choices.



Figure 3: Histogram of Target Variable

The method used to calculate some of the engineered variables will require further adjustment. NAs result in 0 in division

|  |  |
| --- | --- |
| [pay\_ratio1](file:///C:\Users\steve\gits\nw-msds498\report.html#pp_var_1998515190767848884) has 2468 (8.2%) missing values | **Missing** |
| [pay\_ratio2](file:///C:\Users\steve\gits\nw-msds498\report.html#pp_var_6312482923373883349) has 2814 (9.4%) missing values | **Missing** |
| [pay\_ratio3](file:///C:\Users\steve\gits\nw-msds498\report.html#pp_var_-544506946630210523) has 3150 (10.5%) missing values | **Missing** |
| [pay\_ratio4](file:///C:\Users\steve\gits\nw-msds498\report.html#pp_var_8793987982588993322) has 3449 (11.5%) missing values | **Missing** |
| [pay\_ratio5](file:///C:\Users\steve\gits\nw-msds498\report.html#pp_var_483185311441523152) has 3959 (13.2%) missing values | **Missing** |
| [ratio\_avg](file:///C:\Users\steve\gits\nw-msds498\report.html#pp_var_-3479988427685877770) has 5791 (19.3%) missing values | **Missing** |
|  |  |

Eda….

# Predictive Modeling: Methods and Results

methods.

# Comparison of Results

results

# Conclusions

conclusion.

# Bibliography

1. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
2. Yeh, I. C., and 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.
3. Cagan, Michele. Debt 101: From Interest Rates and Credit Scores to Student Loans and Debt Payoff Strategies, an Essential Primer on Managing Debt. First Adams Media hardcover edition, Adams Media, 2020.
4. Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
5. Correa Bahnsen, Alejandro, et al. “Feature Engineering Strategies for Credit Card Fraud Detection.” *Expert Systems with Applications*, vol. 51, June 2016, pp. 134–42. *DOI.org (Crossref)*, https://doi.org/10.1016/j.eswa.2015.12.030.
6. L. Peng, W. Qing and G. Yujia, "Study on Comparison of Discretization Methods," 2009 International Conference on Artificial Intelligence and Computational Intelligence, 2009, pp. 380-384, doi: 10.1109/AICI.2009.385.
7. Liu, and

# Appendix A: Data Dictionary

|  |  |  |
| --- | --- | --- |
| X1 | LIMIT\_BAL | Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit. |
| X2 | SEX | Gender (1 = male; 2 = female). |
| X3 | EDUCATION | Education (1 = graduate school; 2 = university; 3 = high school; 4 = others). |
| X4 | MARRIAGE | Marital status (1 = married; 2 = single; 3 = others). |
| X5 | AGE | Age (year). |
| History of monthly past payment. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above. | | |
| X6 | PAY\_1 | repayment status in September, 2005 |
| X7 | PAY\_2 | repayment status in August, 2005 |
| X8 | PAY\_3 | repayment status in July, 2005. |
| X9 | PAY\_4 | repayment status in June, 2005. |
| X10 | PAY\_5 | repayment status in May, 2005. |
| X11 | PAY\_6 | repayment status in April, 2005. |
| Amount of bill statement (NT dollar) | | |
| X12 | BILL\_AMT1 | bill statement amount in September, 2005 |
| X13 | BILL\_AMT2 | bill statement amount in August, 2005 |
| X14 | BILL\_AMT3 | bill statement amount in July, 2005. |
| X15 | BILL\_AMT4 | bill statement amount in June, 2005. |
| X16 | BILL\_AMT5 | bill statement amount in May, 2005. |
| X17 | BILL\_AMT6 | bill statement amount in April, 2005. |
| Amount of previous payment (NT dollar) | | |
| X18 | PAY\_AMT1 | amount paid in September, 2005 |
| X19 | PAY\_AMT2 | amount paid in August, 2005 |
| X20 | PAY\_AMT3 | amount paid in July, 2005. |
| X21 | PAY\_AMT4 | amount paid in June, 2005. |
| X22 | PAY\_AMT5 | amount paid in May, 2005. |
| X23 | PAY\_AMT6 | amount paid in April, 2005. |