Default of Credit Card Clients

- Factors Affecting Credit Card Default Risks

A final report prepared by:

Albert Fang, Ting Gao

Machine Learning--Data Mining

Wesleyan University

May 2016

Abstract

This project was designed to predict the default of credit card users. The hypothesis is mainly based on the fact that credit card default rate can be predicted by the behaviors of credit card users in previous months. In order to test the hypothesis, we used dataset contains payment data of their customers from a major bank in Taiwan. The idea was to introduce and compare different methods of machine learning by applying those methods to the dataset. This research will provide us with valuable insights regarding how banks determine whom to issue credit cards.

1. Research Question

Nowadays, credit card market is gaining more market share than before. Business Insider estimated that the credit card industry processed $4 trillion in the U.S. in 2014 (Business Insider, Dec.5, 2014). Seven of the largest card issuers -- American Express, JP Morgan, Capital One, Bank of America, Citigroup, Discover and U.S. Bancorp -- reported more than $490 billion in total credit card payments made in the fourth quarter of 2014 alone (Trefis Q4 2014 Bank Review: Credit Card payment Volumes).

In order to reduce the credit default rate of credit card users, and thus reduce the risk of the banks, a number of measurements (for example, monthly income, occupation, and age) were considered by the commercial banks before issuing individual credit cards and deciding their credit lines. In the United States, Social Security Number is an important record to track an individual's credit history because it keeps tracks of an individual's income, health insurance, and credit card history. In addition, some independent organizations such as FICO also created their own credit scores as a credit rating indicator of an individual's credit history.

The major purpose of risk prediction is to use financial information, such as business financial statement, customer transaction and repayment records, etc., to predict business performance or customers' credit risk and to reduce the damage and uncertainty (Yeh, I. C., & Lien, C. H., 2009).

Believing that past credit history can predict an individual's future action, the main research question driving this study is, "What characteristics of an individual are most influential to their predicted credit card default risk?" More specifically, we wish to use various machine-learning mechanisms to predict an individual's credit default risk, based on their credit related record.

This study uses different machine learning models to estimate individual's credit default risk, and comparing the relative accuracy rate of different approaches. Through this project, we wish to apply what we have learned in class, compare their pros and cons on credit default prediction, and try to find links or general correlations among different methods.

2. Methods

2.1 Sample

We obtained our dataset, the "Default of Credit Card Clients Data Set", from the UCI Machine Learning Repository website (http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients). The raw dataset contains payment data for 30,000 individual observations in October 2005 from a major bank in Taiwan.

There are 25 explanatory variables, including Amount of the given credit, Gender, Education, Marital status, Age, History of past payment, Amount of bill statement, Amount of

previous payment. We believe all these variables are significant indicators/predictors of an individual's credit default rate. Our final data set contains 23,150 observations with 34 variables after wrangling.

2.2 Measures

Below is a summary of selected variables of interest and their operational definitions.

Default payment – dependent variable, defines the credit status of the individual predicted by the bank, with 1 being default, and 0 being no default. Of the 23,150 observations in our final data set, 5362 (23.2%) of the credit card holders are expected to have default payment next month.

Education – measures the highest grade achieved by a subject in the data set. Variable is categorical, with 7 levels in the original dataset, and was recoded into 4 levels (1 = graduate school; 2 = university; 3 = high school; 4 = others)

Marital Status – because the amount of the given credit also includes an individual's family credit, marital status can be considered as a proxy for family support. It is a categorical variable with 3 levels (1 = married; 2 = single; 3 = others).

Amount of the given credit – includes both the individual consumer credit and his/her family (supplementary) credit.

History of past payment – tracks the past monthly payment records in the last 6 months (from April to September, 2005). Variable is categorical, with 10 levels (-1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; ,,, 9 = payment delay for nine months and above).

Amount of bill statement – tracks the amount of bill statement history in the last 6 months.

Amount of previous payment – tracks the past 6 month's amount of previous payments.

In addition, in order to have a better understanding of an individual's payment history, we also created some additional variables.

Credit Card Balance – The amount of charges owed to the bank. It's calculated by subtracting amount of previous payment from amount of bill statement. A negative credit card balance means the individual pays more than its statement, while a positive credit card balance indicates the individual still owes some money to the bank. A follow-up binary variable *Balance Cleared* is also created to see the overall balance of the individual over the six months, with 1 meaning a negative overall credit card balance and 0 meaning a positive overall credit card balance.

Credit Card Utilization Rate – One of the most important factors credit scoring models use to calculate credit score, it's calculated by dividing total credit card balances by total credit card limits. In our study, we used the average credit balance over the half year to calculate the utilization rate. Usually lower credit card utilization rate is associated with higher credit score.

2.3 Analyses

2.3.1 Bivariate Analysis

Education v.s. Default Payment

The probability table (Figure 1) shows the relationship between education level and credit card default risk. We can see that excluding education level = 4, which indicates an unknown education level, higher education is associated with lower probability of default risk. Pearson's Chi-squared test also confirms that the relationship is significant (Figure 2).

```
> prop.table(table(credit1$default.payment.next.month,credit$EDUCATION),2)

           1         2         3         4
0 0.80446854 0.75284313 0.74142678 0.95833333
1 0.19553146 0.24715687 0.25857322 0.04166667
```

Figure 1: Two-way table between default payment and education level

```
> chisq.test(credit$default.payment.next.month,credit$EDUCATION)

        Pearson's Chi-squared test

data:  credit$default.payment.next.month and credit$EDUCATION
X-squared = 101.88, df = 3, p-value < 2.2e-16

> chisq.test(credit$default.payment.next.month,credit$EDUCATION)$observed
   credit$EDUCATION
      1    2    3    4
0 6085 8672 2962   69
1 1479 2847 1033    3
> chisq.test(credit$default.payment.next.month,credit$EDUCATION)$expected
   credit$EDUCATION
         1        2         3        4
0 5812.027 8850.971 3069.6786 55.32337
1 1751.973 2668.029  925.3214 16.67663
> chisq.test(credit$default.payment.next.month,credit$EDUCATION)$residuals
   credit$EDUCATION
          1         2         3         4
0  3.580597 -1.902333 -1.943493  1.838759
1 -6.521621  3.464868  3.539836 -3.349076
```

Figure 2: Chi-Squared test and the follow-up post-hoc test between default payment and education level

Credit Card Utilization Rate v.s. Default Payment

Since credit card utilization rate is calculated by dividing total credit card balances by total credit card limits, a lower credit card utilization rate (i.e. a relatively low total credit card balance compared to the total credit card limit) is predicted to have a lower risk of credit card default risk. Logistic regression shows a significant positive correlation between credit card utilization rate and credit card default risk, which confirms our prediction.

```
> fit <- glm(default.payment.next.month ~ UTILIZATION_RATE, data=credit, family = "binomial")
> summary(fit)

Call:
glm(formula = default.payment.next.month ~ UTILIZATION_RATE,
    family = "binomial", data = credit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5694  -0.7709  -0.6421  -0.6013   2.1946

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.61370    0.02695  -59.87   <2e-16 ***
UTILIZATION_RATE  0.93727    0.04663   20.10   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25059  on 23149  degrees of freedom
Residual deviance: 24646  on 23148  degrees of freedom
AIC: 24650

Number of Fisher Scoring iterations: 4
```

Figure 3: Logistic regression between default payment and credit card utilization rate

2.3.2 Machine Learning Techniques

We randomly subset the entire dataset into two groups, a training set (80%) and the remaining testing set (20%). The machine learning procedure predicted the *Default Payment* for the test set based on information learned from the training set, using all the variables in the dataset.

K-nearest neighbor classifiers (KNN)

The K-nearest neighbor classifiers are based on learning by analogy (Yeh, I. C., & Lien, C. H., 2009), where the class label (in this case, *default payment*) is predicted by using a majority vote of its k nearest neighbors. We tuned the model using different k values, and the accuracy rate was maximized with k = 155, at 76.97%. Normalization also improves the performance of KNN. With normalized data set, the accuracy rate improved to 84.43% when k = 155, and

maximized when k = 2, at 89.78%. In order to get a more complete picture of the relative performance of different values, we plot the ROC curves for different values of k, with and without normalization, and compared their AUC. The output summary is shown in Table 1 below. As we can see from Figure 4, which graphs the ROC curves for different values of k, we can see that the kNN classifier performs the best when k = 155, using normalized data. Although the accuracy rate was maximized when k = 2, the AUC was actually the lowest among all four, thus it won't be considered as the optimum model for kNN classifier.

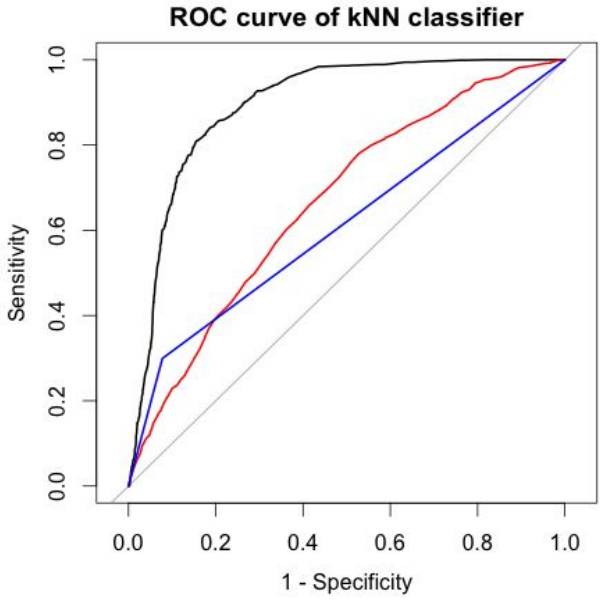

Figure 4: ROC curve of kNN classifiers

Methods	Tuning Parameters	Accuracy	AUC
KNN	K = 155	0.7697127	0.6663
	K = 193	0.7684165	0.6663
KNN w/ normalization	K = 155	0.8107583	0.8939
	K = 2	0.8978181	0.6108

Table 1: Output of KNN classifiers

Naive Bayes

Naive Bayes Classification are based on the Bayesian Theorem under the assumption that all of the features in the dataset are equally important and independent, and normally distributed for continuous features. The tuning method for this classifier is laplace smoothing, which solves the zero count problem. We run the model using no laplace smoothing, and laplace smoothing with different values, and the output is provided in Table 2. We observed that while the accuracy rate was maximized with a large Laplace estimator, the AUC for those models were actually pretty low. The AUC was maximized when Laplace estimator = 1 (the red curve shown on Figure 5).

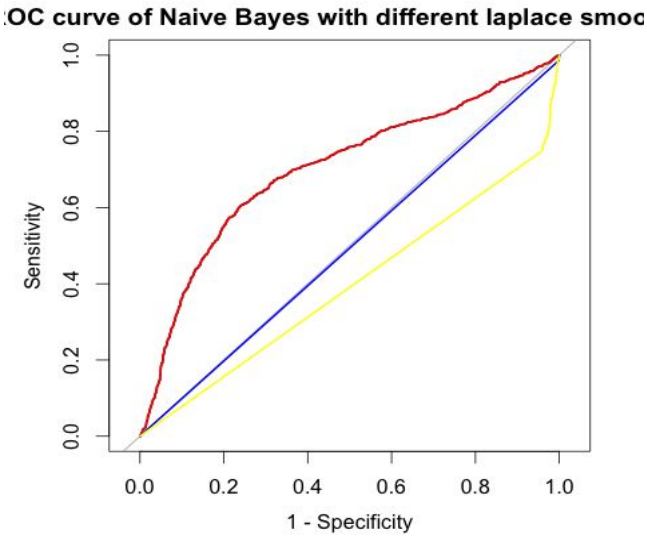

Figure 5: ROC curves of Naive Bayes classifiers

Methods	Tuning Parameters	Accuracy	AUC
Naive Bayes		0.6865414	0.702
Naive Bayes w/ laplace smoothing	Laplace = 1	0.6882696	0.7021
	Laplace = 500	0.7887233	0.4935

	Laplace = 1000	0.7887233	0.3943

Table 2: Output of KNN classifiers

Random Forest

Random Forest models grow trees out as far as possible. The tuning parameters for random forest include *mtry* and *ntree*. We tune the model using different combinations of the two tuning parameters, and the output is shown in Table 3 below. A table showing what variables were important is also shown in Figure 6. The left column (accuracy) tests how worse the model performs without each variable, and the right column (Gini) measures how pure the nodes are at the end of the tree. We can see that among the 30 variables, *PAY_0*, which stands for the repayment status in September, 2005, i.e. current month, is the at the top both measures.

According to Table 3, we see that as we vary the tuning variables, the accuracy rate and AUC does not change much, and when mtry = 5 and ntree =100, the model yields the highest accuracy rate (82.3288%) and AUC (0.7927). The ROC curves also concludes that different parameters have pretty similar performance.

Methods	**Tuning Parameters**		**Accuracy**	**AUC**
Random Forest	Mtry = 5		0.8224238	0.7924
	Mtry = 14		0.8209116	0.7886
	Mtry = 5	Ntree = 1000	0.823288	0.7927
	Mtry = 14	Ntree = 1000	0.8222078	0.7886

Table 3: Output of Random Forest

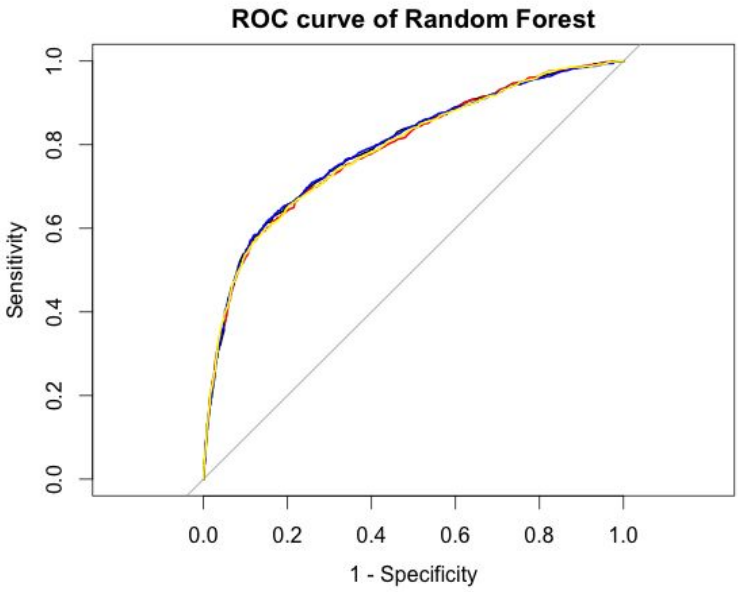

Figure 7: ROC curves of Random Forest classifiers

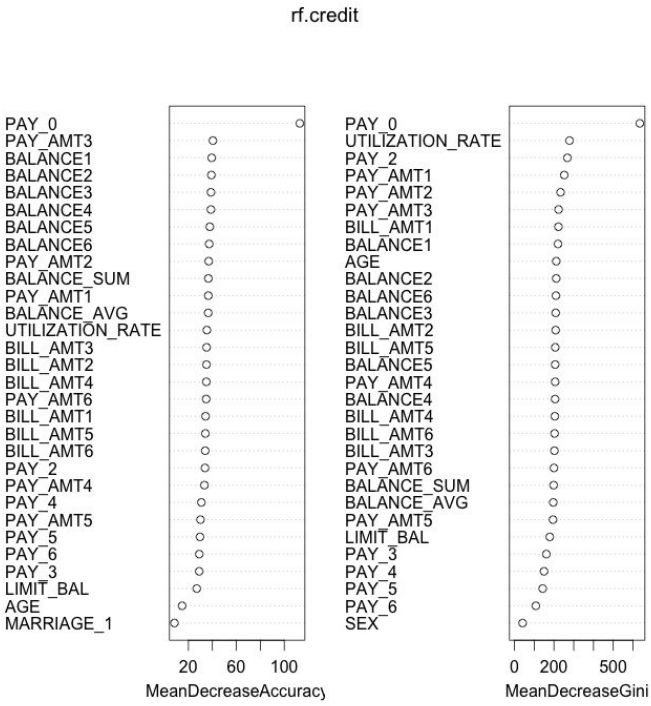

Figure 6: Variables importance as measured by a Random Forest

Logistic Regression

Logistic regression is used when one or more independent variables determine an outcome and the outcome has to be binary. We first fit a binomial model which we will be performing over the entire dataset. We observed that since all p-values are less than 0.05, this regression model is statistically significant. It indicates that these independent variables do have an effect on the dependent variable and they all make intuitive sense. For example, education and default of credit card payment are negatively correlated(lower level means higher education) which implies that when education level goes down, people tend to be less likely to default. Marriage is negatively correlated with default of credit card which suggests that people who are single tend to have a lower default rate because they have a more stable life.

The negative correlation between education and default risk and the negative correlation between marital status and default risk do not seem to follow our common understanding. Therefore, in order to learn about the differences across each education and marital status, we run a second logistic regression using binary (dummy) variables rather than categorical variables for education and marriage. The output shows that compared to the base level, which is graduate school, it's only statistically different between education level = 4, which stands for others, while difference among graduate school, undergraduate, and high school are not statistically significant. As a result, we can conclude that education level is in fact not statistically correlated with default risk, according to the knowledge given in the dataset. It is the education level that's categorized as other "confounded" the negative correlation. For marital status, the output shows

that compared to a married individual, being single has a negative and significant relationship with credit card default risk. One explanation of that might be that being single actually has less household expenses compared to a married individual, which may in term result in a lower default risk. Other variables such as credit line, gender, age, average balance, and utilization rate are all significantly related to default risk for both regressions, and their directions are the same as what we predicted.

```
Call:
glm(formula = default.payment.next.month ~ LIMIT_BAL + SEX +
    EDUCATION + MARRIAGE + AGE + BALANCE_AVG + UTILIZATION_RATE,
    family = "binomial", data = credit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4212  -0.7982  -0.6404  -0.3495   2.8259

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -5.652e-01  1.162e-01  -4.863 1.15e-06 ***
LIMIT_BAL       -4.322e-06  2.465e-07 -17.535  < 2e-16 ***
SEX              1.603e-01  3.250e-02   4.931 8.17e-07 ***
EDUCATION       -3.085e-02  2.459e-02  -1.254 0.209719
MARRIAGE        -2.271e-01  3.391e-02  -6.698 2.11e-11 ***
AGE              2.314e-03  1.880e-03   1.231 0.218495
BALANCE_AVG      2.216e-06  4.628e-07   4.788 1.68e-06 ***
UTILIZATION_RATE 2.729e-01  7.607e-02   3.587 0.000334 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25059  on 23149  degrees of freedom
Residual deviance: 24044  on 23142  degrees of freedom
AIC: 24060

Number of Fisher Scoring iterations: 4
```

```
Call:
glm(formula = default.payment.next.month ~ LIMIT_BAL + SEX +
    EDUCATION_2 + EDUCATION_3 + EDUCATION_4 + MARRIAGE_2 + MARRIAGE_
    AGE + BALANCE_AVG + UTILIZATION_RATE, family = "binomial",
    data = credit2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4145  -0.7983  -0.6399  -0.3436   2.8059

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -7.933e-01  9.421e-02  -8.421  < 2e-16 ***
LIMIT_BAL       -4.270e-06  2.468e-07 -17.299  < 2e-16 ***
SEX1             1.656e-01  3.255e-02   5.086 3.66e-07 ***
EDUCATION_21     8.077e-04  3.847e-02   0.021  0.98325
EDUCATION_31    -4.567e-02  5.067e-02  -0.901  0.36742
EDUCATION_41    -1.755e+00  5.935e-01  -2.958  0.00310 **
MARRIAGE_21     -2.566e-01  3.678e-02  -6.977 3.02e-12 ***
MARRIAGE_31     -1.372e-01  1.406e-01  -0.976  0.32887
AGE              1.082e-03  1.961e-03   0.552  0.58094
BALANCE_AVG      2.218e-06  4.625e-07   4.796 1.62e-06 ***
UTILIZATION_RATE 2.712e-01  7.615e-02   3.561  0.00037 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25059  on 23149  degrees of freedom
Residual deviance: 24024  on 23139  degrees of freedom
AIC: 24046

Number of Fisher Scoring iterations: 5
```

Figure 7: Outputs for logistic regressions

3. Results and Interpretation

Based on the comparison of different models, we found out that kNN is the best model with an accuracy of 0.81and AUC of nearly 0.90. In general, an accuracy of 0.81 is quite good and accurate, which means that banks do take these indicators into account when predicting

individual's credit card default risk. The logistic regression also shows significant correlations between different variables and predicted credit card default risk.

4. Conclusions / Limitations

After conducting this study, we concluded that personal financial statement including all the independent variables mentioned do have a correlation with the default rate of credit cards. We also believe that banks look at these significant indicators to determine an individual's credit line which can be proven using various machine learning method; but correlation does not imply causation and by that we can not simply look at only these indicators. Our dataset came from a bank in Taiwan which can be bias in the sense that certain variables could only be used in Taiwan and we have no way to find out. Even though our project does seem to have some limitation, it gave us a good sense of how banks determine who to issue credit card and determining credit card lines based on the estimation their default credit card risks.