**OPIM 5604 – PREDICTIVE MODELING**

**GROUP PROJECT 2 - DEFAULT OF CREDIT CARD**

**Submitted by – Team 5**

****

**Agrawal, Shivani**

**Kalidindi, Peeyush Varma**

**Krishnan, Nandita**

**Sinha, Ayush**

**Xia, Liyun**

**This file documents all the processes related to pre-processing and modeling techniques for default of credit card**

**Table of Contents**

1. Executive Summary
2. Problem Statement
3. Methodology
   1. Sample
   2. Exploration
   3. Modify
   4. Model
   5. Assessment
4. Results
5. Conclusions & Recommendations
6. References
7. Appendix
8. **Executive Summary:**

This project is about forecasting the defaults of upcoming credit card payments by a customer. Through this model, credit card companies would understand the customers who are more likely to default in their payments and make strategies to reduce/encounter those defaults. Moreover, banks and financial institutions can use this model to target their ideal credit card customer by analyzing the potential customer profiles.

In this data set, we have 24 attributes which include - personal information (like gender, education, marital status & age) and payment information of customers (like limit balance, number of delays in months and bill amount & previous payments from August to September).

Firstly, the data is cleaned by performing basic data treatment procedures like checks for outliers and missing values, etc. Further, new variables were created by feature engineering method which reduces the complexity from 13 variables (limit balance, bill amount & previous payments for 6 months) to 2 variables (credit utilization & remaining payment percentage) and they retain the same knowledge for the model. The bivariate analysis between independent variables and target variable also helped in identifying some interesting relations.

Finally, all different modeling techniques are performed on the cleaned data. After all the analysis and comparisons, Decision tree model was finalized which gave the best output with the maximum accuracy of 1 and lowest misclassification rate on the test dataset.

Through this model we found out that the customer who are using more than the certain percentage of limit are more likely to default and level of education is also inversely proportional to the default in the payment.

Based on the findings we recommended few suggestions to the credit card company which they can consider while encountering these issues.

1. **Problem Statement:**

This research is aimed at the case of customers default payments in Taiwan. Given the last six months of payment history, the goal of the project is to forecast the probability of default credit card payment for the following month. From the perspective of risk management, the result of predictive accuracy of “defaults” will be more valuable than the total accuracy of the model.

1. **Methodology:**

We used the SEMMA (Sample, Explore, Modify, Model, Asses) process to explore and clean the data, and to evaluate the best fit model for predicting the “defaults” by customers.

**3.1 Sample:**

* The data was taken from: <https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients>
* It has 30,000 instances and 24 attributes
* This sample contains information about default payment in Taiwan and we’ve used this information to predict the people that will default on their credit card payments the following month

**3.2 Explore:**

* Missing Values - There are no missing values in this data.
* Outliers – We used tail quantile of 0.1 and Q value of 3 to count the number of outliers present in the data. Count: 2,836 outliers present in BILL\_AMT & PAY\_AMT variables.

Apart from missing values and outliers, we also explored the relationship between the predictor variables and the target variable.

Refer to the **appendix7.1** for screenshots.

Here are some of the important observations gathered from the data explored: -

1. 22% of the sample population of 30,000 people have defaulted on their credit card payments for the following month
2. A relatively lower limit balance median was observed for people who have defaulted indicating that there’s a higher tendency for them to default
3. The data showed that men tend to default more than women
4. The level of education is inversely proportional to the number of people defaulting
5. Married people default more than the single population
6. People who have a history of paying on time or in advance tend to default less

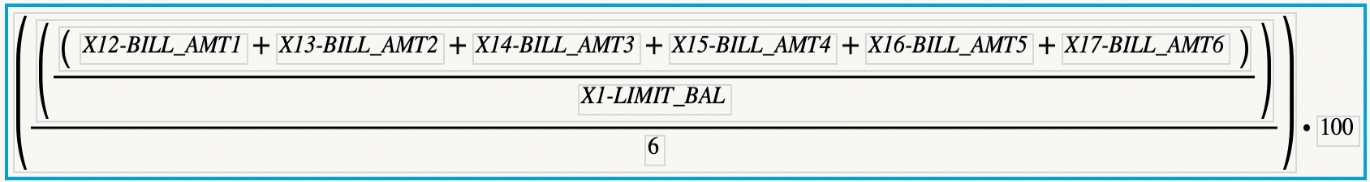
**3.3 Modify:**

* Feature Engineering

Two new columns were created using the information present in the data

1 – Credit Utilization

Credit utilization is the percentage of the limit utilised by the individual. Without including X1-LIMIT\_BAL and X12-BILL\_AMT1 to X17 BILL\_AMT6 (amount of bill statements from the month of September 2005 to April 2005) we can get all the information into one variable. So, we are reducing the complexity of the model by a great degree, while still retaining the same information.

Credit Utilization =

2 – Remaining payment percentage

Remaining payment percentage is the percentage of the average bill amounts that is left to be paid back by the individual. Without including X18- PAY\_AMT1 to X23- PAY\_AMT6 (amount of previous payment from the month of September 2005 to April 2005) we can get all this information into one variable.

Remaining payment % = Graphical user interface, text

Description automatically generated

We’ve removed 13 columns and replaced them with 2 columns which has reduced the complexity by a lot.

To confirm that these two columns had all the information from the 13 columns we ran two different sets of models.

We first ran all the models including all the 13 columns and then ran all the models removing those 13 columns and including the 2 new columns.

All models almost had a similar confusion matrix and misclassification rate and hence we decided to use the credit utilization and remaining payment percentage attributes instead of the 13 attributes (Limit, Bill Amounts and Pay Amounts).

* Fixing Outlier

We found that there were 2916 outliers across BILL\_AMT, PAY\_AMT, credit utilization & remaining payment percentage variables. On comparing the outlier sample with the rest of the data we observed a similar distribution of “number of defaults” and hence we concluded that keeping the outliers won’t add any value to the model. Based on this conclusion we removed the outliers.

**3.4 MODEL:**

* **General Configuration** –
  1. Predictor Variables (X) – X1 – Limit\_Bal, X2 – Sex, X3 – Education, X4 – Marriage, X5 – AGE, X6 – Pay\_0 to X11 – Pay\_6, Cred\_Utilization, Payment %
  2. Target Variable (Y) – “Y Default payment next month"
* **Logistic Regression** –
  1. Configuration:
     1. Personality – Nominal Logistic
  2. Variable Reduction: Removed Age & Remaining payment percentage because the p-value was greater than 0.05
* **Decision Tree** *–*
  1. Number of splits: 3

We chose 3 number of splits because that gives us the simplest model with the highest validation assessment.

* **Bootstrap Forest** *–*
  1. General Configuration
* **Boosted Tree** *–* 
  1. General Configuration
* **Neural Network** –
  1. Configuration:
     1. Hidden Layer Structure: We tried multiple hidden layer structures based on the total number of predictor variables. The following gave us the best results in terms of total accuracy and accuracy of 1:
        + First Layer = 3 TanH
        + Second Layer = 3 TanH
     2. Boosting and fitting options are default
* **Naïve Bayes** *–*
  1. General Configuration
* **K-nearest neighbors** –

1. Best k : 9

* **Ensemble** *–*
  1. Model Averaging: After removing Naïve Bayes and KNN (Owing to relatively lower total accuracy compared to the other models)

**Note:** *On addition to the above models, we also tried stratification to remove the imbalance of the non-defaults in the data. However, when we ran models to this stratified data the accuracy of 1’s reduced instead of showing an improvement. Hence, we decided to use the complete data instead of the stratified data*

(For confusion matrix and results refer to **appendix 7.2 to 7.8**)

**3.5 ASSESS**

* We used the following metrics in JMP to assess our model:
  + **Misclassification rate:** It tells us the percentage of classifications that were incorrect. The misclassification rate for our best model which is decision tree is 18.81%.
  + **Confusion matrix:** It gives us the number of 0s and 1’s the model predicted correctly or incorrectly

Table

Description automatically generated

* + **Accuracy of 1** (output 1 being customer defaulting on credit card payment): Using the confusion matrix we can calculate the percentage of 1’s the model predicted accurately. The accuracy of 1’s for our best model which is decision tress is 68.93%
  + **ROC:** The ROC does not depend on the class distribution. So, it is useful for predicting rare events. It gives us the area under the curve. The larger the area the better the model.

Chart, line chart

Description automatically generated

* + **Lift Curve:** If we take the top 10% of our data then we will get a lift of 3 over our baseline accuracy. It means that we are 3 times more likely to predict a customer who defaults accurately.

Chart, histogram

Description automatically generated

1. **Results**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ***MODEL*** | | ***MISCLASS. RATE*** | ***ACCURACY*** | ***ACCURACY of 1*** |
| ***TRAINING*** | *Logistic Regression* | | **18.39%** | **81.61%** | **70.40%** |
| ***Decision Tree*** | | **18.78%** | **81.22%** | **70.97%** |
| *Bootstrap Forest* | | **17.16%** | **82.84%** | **74.43%** |
| *Boosted Tree* | | **16.78%** | **83.22%** | **76.39%** |
| *Neural Network* | | **18.11%** | **81.89%** | **71.29%** |
| *Naïve Bayes* | | **20.48%** | **79.52%** | **58.02%** |
| *KNN* | | **20.15%** | **79.85%** | **63.6%** |
| *Ensemble\** | | **17.71%** | **82.29%** | **73.02%** |
| ***VALIDATION*** | *Logistic Regression* | **18.44%** | | **81.56%** | **72.06%** |
| ***Decision Tree*** | **18.90%** | | **81.11%** | **67.86%** |
| *Bootstrap Forest* | **18.37%** | | **81.63%** | **67.7%** |
| *Boosted Tree* | **18.59%** | | **81.41%** | **66.34%** |
| *Neural Network* | **18.50%** | | **81.50%** | **66.94%** |
| *Naïve Bayes* | **20.25%** | | **79.75%** | **57.14%** |
| *KNN* | **19.58%** | | **80.42%** | **63.5%** |
| *Ensemble\** | **18.33%** | | **81.67%** | **67.95%** |
| ***TEST*** | *Logistic Regression* | **18.79%** | | **81.21%** | **66.93%** |
| ***Decision Tree*** | **18.81%** | | **81.19%** | **68.93%** |
| *Bootstrap Forest* | **18.85%** | | **81.15%** | **66.3%** |
| *Boosted Tree* | **19.38%** | | **80.62%** | **64.35%** |
| *Neural Network* | **18.88%** | | **81.12%** | **66.06%** |
| *Naïve Bayes* | **21.19%** | | **78.81%** | **54.48%** |
| *KNN* | **20.36%** | | **79.64%** | **58.9%** |
| *Ensemble\** | **18.81%** | | **81.19%** | **66.82%** |

*\*Ensemble excludes KNN and Naïve Bayes*

* **All models performed similarly** with Logistic Regression, Decision Tree, Bootstrap Forest, Boosted Tree, Neural Network, and Ensemble performing at par with each other
* With **Decision Tree, we see the highest accuracy of 1** for the validation and test dataset
* If you look at the table given above, you’ll notice logistic regression has the best total accuracy among all the models however we chose to go with decision tree as our final model as it displays a better accuracy of 1 and performed only slightly worse than the logistic regression model in the total accuracy metric

1. **Conclusions and Recommendations**

**Final Conclusions:**

* The newly created variable credit utilization is the limit\_bal utilized by the individual. Upon analysis it was noticed that a higher credit utilization was directly related to a customer defaulting on their credit card payment for the following month
* During initial data exploration phase, we found out there was high correlation between the education level and the chances of customer defaulting on their credit card. Customers with a lower education level are more likely to default on their credit card payment than those with higher education level
* Remaining payment percentage is the percentage of the average bill amounts that is left to be paid back by the individual. If the remaining payment percentage is high, they are more likely to default on the credit card payment for the following month

**Recommendations:**

* In the case of existing customers, the following steps can be taken:
  + If the credit utilization of the customer is relatively high then the customer’s spending activities should be monitored for a certain period, and they should be warned against repercussions
  + The customer’s limit\_bal should be reduced if there is a trend of high credit utilization every month
* From the perspective of acquiring new customers, the below recommendation can be made:
  + The personal background of a potential customer like their education level can be evaluated before coming to a decision on the issuance of their credit card

1. **References:**

AN EMPIRICAL INVESTIGATION OF CREDIT CARD DEFAULT:

<https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.197.8564&rep=rep1&type=pdf>

https://ptmoney.com/credit-card-default/

1. **Appendix:**

**7.1 Relationship between the predictor variable and target variable**

Chart

Description automatically generatedTable

Description automatically generated

**7.2 Regression – Nominal Logistic**

**Text

Description automatically generatedTable

Description automatically generated**

**7.3 Bootstrap ForestTable

Description automatically generated**

**7.4 Boosted Tree**

**Table

Description automatically generated**

**7.5 Neural Network – 3TanH1 3TanH2**

**Text

Description automatically generated with medium confidence**

**7.6 Naïve Bayes**

**Graphical user interface, application

Description automatically generated**

**7.7 K-nearest neighbors**

**Chart, line chart

Description automatically generated**

**Graphical user interface, table

Description automatically generated**

**7.8 Ensemble – Screenshots**

**Table

Description automatically generated**