Credit Card Default Identification Report

by

Qing Wu

Credit One company

October20th, 2018

**Background and Business goal:**

Since there was an increase in the number of customers who have defaulted on loans in Credit One company over the past year or so, our data scientist team need to build an effective and accurate model to help them to gain insight into what factors affect default rate and predict the customers who would default on loans by using historical data.

**Descriptions of data sources:**

There are the total 30,000 observations in our credit card dataset. The dataset used binary variable – default payment next month (Yes = 1, No = 0), as the response variable. There are 6636 observations (22.12%) are customers with default payment. This dataset used 23 variables as explanatory variables. These attributes are listed below.

|  |  |  |
| --- | --- | --- |
| Variable name | Attribute Information | Description |
| LIMIT\_BAL | Amount of the given credit (NT dollar) | It includes both the individual consumer credit and his/her family (supplementary) credit. |
| SEX | Gender | 1 = male; 2 = female |
| EDUCATION | Education | 1 = graduate school; 2 = university; 3 = high school; 0, 4, 5, 6 = others |
| MARRIAGE | Marital status | 1 = married; 2 = single; 3 = divorce; 0=others |
| AGE | Age (year) |  |
| PAY\_0 | the repayment status in September, 2005 | The measurement scale for the repayment status is: -2: No consumption; -1: Paid in full; 0: The use of revolving credit; 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. |
| PAY\_2 | the repayment status in August,2005 | Same as above. |
| PAY\_3 | the repayment status in July,  2005 | Same as above. |
| PAY\_4 | the repayment status in June,  2005 | Same as above. |
| PAY\_5 | the repayment status in May,  2005 | Same as above. |
| PAY\_6 | the repayment status in April,  2005 | Same as above. |
| BILL\_AMT1 | Amount of bill statement (NT dollar) in September, 2005 |  |
| BILL\_AMT2 | Amount of bill statement in August, 2005 |  |
| BILL\_AMT3 | Amount of bill statement in July, 2005 |  |
| BILL\_AMT4 | Amount of bill statement in June, 2005 |  |
| BILL\_AMT5 | Amount of bill statement in May, 2005 |  |
| BILL\_AMT6 | Amount of bill statement in April, 2005 |  |
| PAY\_AMT1 | Amount paid in September, 2005 |  |
| PAY\_AMT2 | Amount paid in August, 2005 |  |
| PAY\_AMT3 | Amount paid in July, 2005 |  |
| PAY\_AMT4 | Amount paid in June, 2005 |  |
| PAY\_AMT5 | Amount paid in May, 2005 |  |
| PAY\_AMT6 | Amount paid in April, 2005 |  |
| default payment next month | Client's behavior | Y=0 then not default, Y=1 then default" |

**Exploratory data analysis:**

We did exploratory data analysis on the original dataset. We have learned some potential business values from this analysis.

(1) There are the total 30,000 observations in our credit card dataset. There are 6636 observations (22.12%) are defaulted credit card holders. There are 23364 observations (77.88%) are not defaulted credit card holders.

(2) In the defaulted credit card holder group, 43.29% are male, while 56.71% are female. In the not\_ defaulted group, 38.585% are male, while 61.415% are female. It showed proportion of male in the defaulted group is higher than not defaulted group.

(3)The mean of age in the defaulted credit card holder group is 35.75 and 75% of this group is under 42. And mean of age in the not defaulted credit card holder group is 35.41 and 75% of this group is under 41. There are no significant differences in age distributions between these two groups.

(4) The mean of Amount of the given credit in the defaulted credit card holder group is 130109.66, and 75% of this group is under200000. The mean of Amount of the given credit in the not defaulted credit card holder group is 178099.73, and 75% of this group is under 250000.The results show that Amount of the given credit in the defaulted group is lower than not defaulted group.

(5) In the defaulted credit card holder group, 30.68% of this group is graduate level, 50.18% is university level and 18.64% are high school. In the not defaulted group, 36.59% of this group is graduate, 45.80% is university level and 15.75% are high school. The not defaulted group has a larger proportion of higher education level.

(6) In the defaulted credit card holder group, 48.31% of customers are singles, 50.35% of customers are married, and 1.2% of customers are divorce. In the not defaulted group, 46.74% of customers are singles, 54.03% of customers are married, and 1% of customers are divorce. The not defaulted group has a slight larger proportion of married customers.

Exploratory data analysis is a very effective tool to get deep insight into our data through visualization methods. We can get a lot of useful information about our dataset before building up our model. Through EDA, we can explore our data from different angles, visualize our results and get new ideas to analyze our model. Through EDA, we can easily communicate and present our results to stakeholders without technical background.

This analysis indicates gender, amount of the given credit, education and marriage status have some effect on default rate. Age is not a significant factor to affect default rate. Before building up our model, we gain some insight into

factors that affect default rate, which can help us to solve high default rate problem.

**Data Preprocessing and Feature Engineering:**

We use this credit card database for classification to predict Client's payment next month. Data preprocessing was performed. We dropped the column ID, for it can not provide any useful information to build the model. All the continuous features were standardized, while all the categorical features were transformed into dummy variables. The dataset contains no missing values in all attributes. We used recursive feature elimination to perform feature engineering. After dummy variable creation, we totally had 82 predictor variable. After feature selection, we kept 41 predictor variables to build model. We divided this credit card dataset into training and test datasets by using 70/30% split.

**Model building, tuning and model selection:**

Firstly, Logistic Regression, k-nearest neighbors, Decision Tree, Gaussian Naive Bayes, Support Vector Machine are selected to train the model. 10 folds cross-validation was applied to these models, and the accuracies of validation dataset of these five algorithm were quite good and listed below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Logistic  Regression | KNN | CART | NB | SVM |
| Accuracy | 0.820 | 0.796 | 0.801 | 0.794 | 0.818 |

Secondly, we improved the performance of models by using ensemble method. Stochastic Gradient Boosting, Random Forest, Extra Trees Classification, Bagged Decision Trees for Classification are selected to develop these models. 10 folds cross-validation was applied to these models, the accuracies of validation dataset of these four algorithm were quite good and listed below. But these ensemble methods did not improve the accuracy significantly.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algorithm | GBM | Random Forest | Extra Trees | Bagged Decision  Trees |
| Accuracy | 0.821 | 0.801 | 0.799 | 0.801 |

Thirdly, hyper parameters of Logistic Regression, Stochastic Gradient Boosting, KNN, Decision Trees, Support Vector Machine and Random Forest models are tuned and optimized to improve performances.

For Stochastic Gradient Boosting classifier, 10-folds cross-validation is applied and two important hypermeters- Learning rate and n\_estimators are tuned. Learning rate shrinks the contribution of each tree by learning\_rate. Learning\_rate is set to 0.001, 0.01, 0.1. N\_estimators means the number of boosting stages to perform. N\_estimators is set to 50,100,150,200,250,300,350,400. After cross validation, the best parameters are selected. The best model with highest accuracy is that learning\_rate is set to 0.1 and n\_estimators is set to 100.

For logistic regression classifier, 10-folds cross-validation is applied and two Hypermeters-C and Penaltyare tuned. C describes the inverse of regularization strength. Like in support vector machines, smaller values specify stronger regularization. C is set to 15 numbers spaced evenly on a log scale, the starting point is 1e-05, the stop point is 100000000.Penalty is used to specify the norm used in the penalization. Penalty is set to “l1”, “l2”. After cross validation, the best parameters are selected. The best model with highest accuracy is that C is set to 1389495.5 and Penaltyis set to l2.

For Decision Tree classifier, 10-folds cross-validation is applied and four hypermeters- "max\_depth", "max\_features","min\_samples\_leaf", "criterion" are tuned. "max\_depth" means the maximum depth of the tree. "max\_depth" is set to 3,5,7. "max\_features" means the number of features to consider when looking for the best split."max\_features" is set to 3,5,7,9."criterion" means the function to measure the quality of a split. Supported criteria are “gini” for the Gini impurity and “entropy” for the information gain. "criterion" is set to "gini", "entropy". "min\_samples\_leaf" means the minimum number of samples required to be at a leaf node. "min\_samples\_leaf" is set to 30,50,100. After cross validation, the best parameters are selected. The best model with highest accuracy is that "min\_samples\_leaf" is 30 and "max\_features"is 9, "max\_depth" is 7 and "criterion" is "entropy".

For k-nearest neighbors, cross-validation is applied and hypermeter n\_neighbors are tuned. Hyper parameter n\_neighbors means the number of neighbors. n\_neighbors is set to a list of 1,3,5,7,9,11,13,15,17,19,21. After cross validation, the best parameters are selected. The best model with highest accuracy is that n\_neighbors is set to 15.

For support vector machine, cross-validation is applied and hypermeters C and gamma are tuned. C is the parameter for the soft margin cost function. For large values of C, the optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly. Conversely, a very small value of C will cause the optimizer to look for a larger-margin separating hyperplane, even if that hyperplane misclassifies more points. C is set to a list of 1, 10, 100. Gamma is the parameter of a Gaussian Kernel (to handle non-linear classification). The gamma parameter defines how far the influence of a single training example reaches, with low values meaning ‘far’ and high values meaning ‘close’. In other words, with low gamma, points far away from plausible separation line are considered in calculation for the separation line. Where as high gamma means the points close to plausible line are considered in calculation. So a small gamma will give you low bias and high variance while a large gamma will give you higher bias and low variance. Gamma is set to a list of 0.1, 0.01. After cross validation, the best parameters are selected. The best model with highest accuracy is that C is set to 10, and Gamma is set to 0.1.

For Random Forest, cross-validation is applied and two important hypermeters- n\_estimators and max\_features are tuned. 'n\_estimators' that means number of trees in the forest is set to a list of 50,100,150,200,250,300,350,400, and 'max\_features' that means max number of features considered for splitting a node is set to a list of 3,5,7,9,10. After cross validation, the best parameters are selected. The best model with highest accuracy is that 'n\_estimators' is set to 200, and 'max\_features' is set to 7.

**Model evaluation：**

Models evaluations are applied to these models; accuracy is selected to evaluate performance of Cross validation. Metrics of accuracy, recall, precision and F1 score are selected to evaluate the performance of test set, which can help to determine how well the model generalize to the unseen data.

The table below shows accuracy of cross-validation of these tuned models.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Algorithm | Stochastic  Gradient  Boosting | Logistic  Regression | CART | KNN | SVM | Random  Forest |
| Accuracy | 0.821 | 0.821 | 0.808 | 0.815 | 0.821 | 0.804 |

The table below shows the models evaluations results of test set.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algorithm | Accuracy | precision | recall | F1-score |
| Stochastic  Gradient  Boosting | 0.82011 | 0.80 | 0.82 | 0.80 |
| Logistic  Regression | 0.8187 | 0.80 | 0.82 | 0.80 |
| KNN | 0.8163 | 0.80 | 0.82 | 0.79 |
| SVM | 0.8189 | 0.80 | 0.82 | 0.80 |
| Random  Forest | 0.7983 | 0.78 | 0.80 | 0.78 |

**Recommendations:**

Stochastic Gradient Boosting model is the best model we

suggest Credit One Company to use, because this model

has the highest accuracy and F1-score. The accuracy of this model is 0.82011, and the F1-score of this model is 0.80, which is quite good. If we want to future improve accuracy, we can collect more data later and update our model. With more new data, we may build up more accurate model.