Personal Loan Default Risk Project

Yen-Hsiu Chang, Yule Jin, Zhen Li

The work is divided up among team members as follows. Yen-Hsiu Chang leads the project, and he is responsible for cleaning up data, training data using Naïve Bayes classfier, logistic regression, Generalized Additive Model and tuning graphs. Yule Jin applied Random Forest to train data and created slides. Zhen Li’s duty is writing up reports.

**Executive Summary**

Loan interest has been an important source of income for commercial banks. However, bad loans can be hazardous for banks. We, as Data Analytics team of Bank of America, were trying to provide a pre-screening guideline for issuing loans. This project is aimed at predicting the chance of delinquency over 90 days of the borrower on the loan in two years based on applicants’ financial status and past credit activities. This project will both save BoA humongous amount of effort and increase accuracy in determining applicants’ capability of loan repayment.

We applied GBM, GAM, Random Forest, logistic regression, and Naïve Bayes classifier in our model. Based on our analysis, we will maximize the accuracy in identifying the loan becoming delinquency.

**Introduction**

The most common process of approving a personal loan is pre-screening the applicant’s credit history, followed by careful determining the amount and rate of the loan. This process have been greatly relaxed before 2008 because of the popularity of the Mortgage Backed Security. After the subprime debt crisis, banks become more and more cautious with issuing loans to people who could have unhealthy payment habits. Our goal of the project is to create a pre-screening model for banks to quickly screen out applicants who are likely to default in two years without misclassifying applicants who should qualify for loans.

The rest of the report is organized as following. In Section 3, we describe the data and its source, and we present you with a summary of our findings. Section 4 discusses methods deployed and interpret the results into plain English. A conclusion and possible improvement is given out in Section 5.

**Data**

Our data is downloaded from Kaggle.com. The data is consisted of 150,000 anonymous borrowers and whether they have payments 90 days past due is provided. The list of features for each applicant is described in the table below.

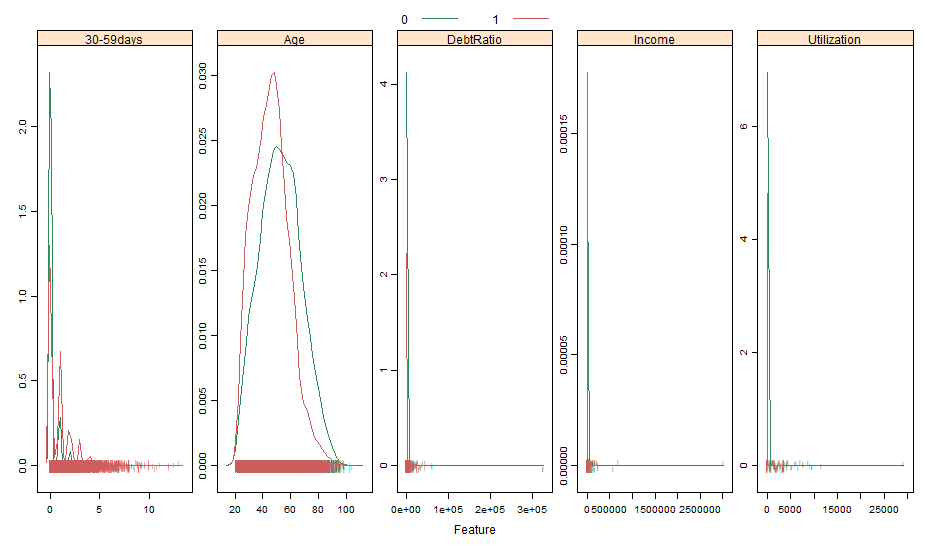
|  |  |  |  |
| --- | --- | --- | --- |
| Variable Name | Name used in R | Description | Type |
| SeriousDlqin2yrs (RESPONSE) | Dlqin2yrs | Person experienced 90 days past due delinquency or worse | 1, 0 |
| RevolvingUtilizationOfUnsecuredLines | Utilization | Total balance on credit cards and personal lines of credit except term loans divided by total credit limits | Percent |
| Age | Age | Age of borrowers in years | Integer |
| NumberOfTime30-59DaysPastDueNotWorse | X30\_59days | Number of times borrower has been 30-59 days past due but no worse in the last 2 years. | Integer |
| DebtRatio | DebtRatio | Monthly debt payments, alimony, living  costs divided by monthly gross income | Percent |
| MonthlyIncome | Income | Monthly income | Real |
| NumberOfOpenCreditLinesAndLoans | OpenCredit | Number of open loans (installments like car loan or mortgage) and Lines of credit (e.g. credit cards) | Integer |
| NumberOfTimes90DaysLate | X90days | Number of times borrower has been 90 days or more past due | Integer |
| NumberRealEstateLoansOrLines | RealEstate | Number of mortgages and real estate loans including home equity lines of credit | Integer |
| NumberOfTime60-89DaysPastDueNotWorse | X60\_89days | Number of times borrower has been 60-89 days past due but not worse in the last 2 years | Integer |
| NumberOfDependents | Dependents | Number of dependents in the family excluding themselves (spouse, children etc.) | Integer |

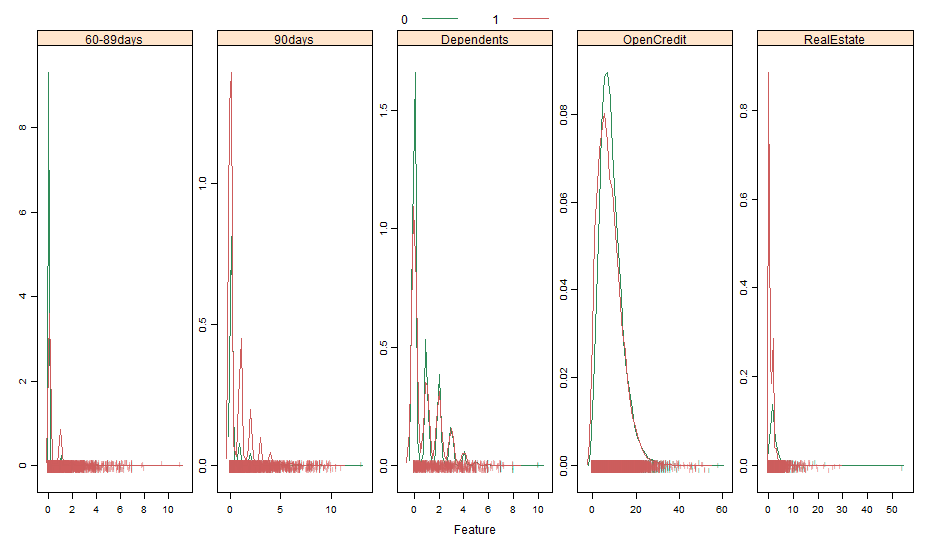
A brief summary of all the data are given in the following table.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dlqin2yrs | Utilization | Age | X30\_59days | DebtRatio | OpenCredit |
| Min. :0.00000 | Min. : 0.00 | Min. : 0.0 | Min. : 0.000 | Min. : 0.0 | Min. : 0.000 |
| 1st Qu.:0.00000 | 1st Qu.: 0.03 | 1st Qu.: 41.0 | 1st Qu.: 0.000 | 1st Qu.: 0.2 | 1st Qu.: 5.000 |
| Median :0.00000 | Median : 0.15 | Median : 52.0 | Median : 0.000 | Median : 0.4 | Median : 8.000 |
| Mean :0.06684 | Mean : 6.05 | Mean : 52.3 | Mean : 0.421 | Mean : 353.0 | Mean : 8.453 |
| 3rd Qu.:0.00000 | 3rd Qu.: 0.56 | 3rd Qu.: 63.0 | 3rd Qu.: 0.000 | 3rd Qu.: 0.9 | 3rd Qu.:11.00 |
| Max. :1.00000 | Max. :50708.00 | Max. :109.0 | Max. :98.000 | Max. :329664.0 | Max. :58.000 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Income | X90days | RealEstate | X60\_89days | Dependents |
| Min. : 0 | Min. : 0.000 | Min. : 0.000 | Min. : 0.0000 | Min. : 0.000 |
| 1st Qu.: 3400 | 1st Qu.: 0.000 | 1st Qu.: 0.000 | 1st Qu.: 0.0000 | 1st Qu.: 0.000 |
| Median : 5400 | Median : 0.000 | Median : 1.000 | Median : 0.0000 | Median : 0.000 |
| Mean : 6670 | Mean : 0.266 | Mean : 1.018 | Mean : 0.2404 | Mean : 0.757 |
| 3rd Qu.: 8249 | 3rd Qu.: 0.000 | 3rd Qu.: 2.000 | 3rd Qu.: 0.0000 | 3rd Qu.: 1.000 |
| Max. :3008750 | Max. :98.000 | Max. :54.000 | Max. :98.0000 | Max. :20.000 |
| NA's :29731 |  |  |  | NA's :3924 |

The histogram of each variable is shown in the graph below.





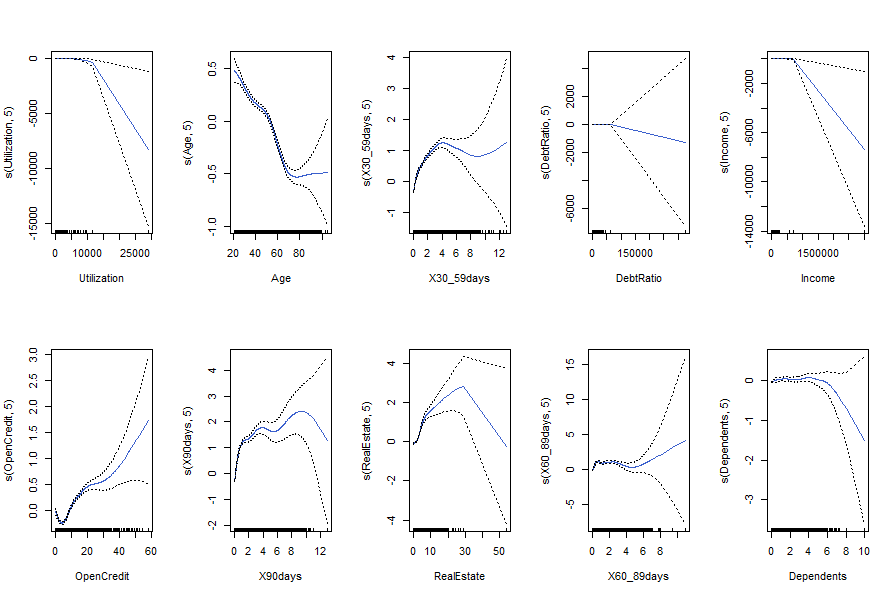
One problem with the dataset is that it has some missing values that needs to be cleaned up. We replaced the missing values of NumberofDependents by 0. All other missing values are imputed with K-Nearest Neighbor Method. Another problem is the data is heavily imbalanced. Out of the 150,000 observations, a roughly 6.7% (10,026 cases) of total observations failed to make payment 90 days past due (response of 1). The imbalance of the dataset could lead to inaccuracy in predicting data. In this case, we deployed Synthetic Minority Over-sampling Technique (SMOTE), and down sampled the majority class and up sampled the minority class to balance out the data.

We divided up the data using a 2-1 ratio. 100,000 data is used for training purpose, and the rest is used for validation.

**Analysis**

In this project, we would like to predict whether an applicant would fail to pay within 90 days past due date. The nature of this project requires a fairly high correct detective rate in the minority class and allows for a small error rate in the majority class to achieve this. To improve accuracy of our model, we applied Receiver Operating Characteristic Curve (ROC) which is a standard technique for summarizing classifier performance over a range of tradeoffs between true positive and false positive error rates. The error rate is given by Area under the Curve (AUC). We compared the AUC for validation set from different models and achieved the optimal model.

GAM is used to fit the data. The graph below shows the nonlinearity in many of the features, such as Age, Open Credit. The graph also indicated that there are outliers in our data. GAM returns relatively sound result.



We also ran Random Forest and tuned the number of feature m from 1 to 6. The results for different parameters are listed in the table below. When m=1, the AUC for the validation set is the highest.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | m=1 | m=2 | m=3 | m=4 | m=5 | m=6 |
| cv\_err\_500 | 0.8883013 | 0.9305431 | 0.9519827 | 0.9558456 | 0.9560081 | 0.9554955 |
| eval\_err\_500 | 0.16022 | 0.15856 | 0.14156 | 0.13884 | 0.1389 | 0.13868 |
| cv\_auc\_500 | 0.904480 | 0.9489756 | 0.9664063 | 0.9691184 | 0.9688117 | 0.9684215 |
| eval\_auc\_500 | 0.8557716 | 0.8543223 | 0.8485134 | 0.8431783 | 0.8413054 | 0.8391322 |

GBM played a big part in our project. We tuned the number of trees among eight different values, 3000 to 10,000, incremented by 1000 each time. We are able to get the largest area under curve using 10,000 number of trees. Were we given more time, we will be able to find the optimal amount.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| n.trees | cv\_err | cv\_auc | test\_err | test\_auc |
| 3000 | 0.1430142 | 0.9346061 | 0.14298 | 0.8574468 |
| 4000 | 0.1321881 | 0.9426279 | 0.13082 | 0.8574171 |
| 5000 | 0.1242373 | 0.9481631 | 0.1214 | 0.8578133 |
| 6000 | 0.1183492 | 0.952135 | 0.11592 | 0.8583111 |
| 7000 | 0.1134863 | 0.9549726 | 0.11128 | 0.8588709 |
| 8000 | 0.1100484 | 0.9570110 | 0.10742 | 0.8588132 |
| 9000 | 0.1064856 | 0.9588786 | 0.10506 | 0.8591457 |
| 10000 | 0.1043104 | 0.9601312 | 0.10212 | 0.8592644 |

GBM outperforms Random Forest from the comparison below.

Random%20Forest%20ROC%20Curve%20AUC%20(ntree=500).pdfGBM%20ROC%20Curve%20AUC.pdf

After applying all five classifiers with the best parameter tuned as stated above, we can see from the table below, the model that produced the best result is GBM. GAM and Random Forest models come second and third respectively. The result is consistent with our expectation. We made a prediction that neither the Naïve Bayes nor logistic regression would work very well in this case. The Naïve Bayes classifier is vastly impacted by outliers, besides, one important assumption of Naïve Bayes is the independence among the variables, which is certainly not the case in this project, so the result of its training is not optimal. For logistic regression, it fails to describe the nonlinearity of some of the most important features.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | AUC | Threshold | Specificity | Sensitivity |
| GBM | 0.858871 | 0.302197 | 0.784391 | 0.776786 |
| GAM | 0.856959 | 0.510335 | 0.804824 | 0.754167 |
| Random Forest | 0.853903 | 0.369 | 0.787929 | 0.776488 |
| Logistic Regression | 0.821404 | 0.437249 | 0.768439 | 0.727083 |
| Naive Bayes | 0.794349 | 0.005914 | 0.748521 | 0.703274 |

../../../ROC%20Curve%20Summary.pdfEvaluation%20set%20ROC%20Curve%20AUC.pdf

**Conclusion**

Our training and testing of the dataset indicated that the Gradient Boosting Model delivers the best result in predicting customers’ default on payment 90 days after due date in two years. We get the variable importance ranking by running *filterVarImp* in R. As can be seen from the table below, utilization, number of credit lines and number of mortgages are the top three factors contributing to the future payment failure.

/Users/Yule/Downloads/Relative Importance From GBM.pdf

If we were given more time to work on the project, we would be able to get a more accurate model by trying out different combinations of shrinkage and number of trees in GBM models. Moreover, for this project, we only used SMOTE to balance the data. If time permits, we could try out more different methods in sampling imbalanced data.

**Reference**

(1) Wang H, Xu Q, Zhou L (2015) Large Unbalanced Credit Scoring Using Lasso-Logistic Regression Ensemble. PLoS ONE 10(2): e0117844.

(2) Liang, Junjie. "Predicting Borrowers Chance of Defaulting on Credit Loans." CS 229 Machine Learning Final Projects, Autumn 2011. Stanford University. Web. 6 Nov. 2015.

(3) Huang, Cheng-Lung, Mu-Chen Chen, and Chieh-Jen Wang. "Credit Scoring With A Data Mining Approach Based On Support Vector Machines." Expert Systems With Applications 33.(2007): 847-856. ScienceDirect. Web. 6 Nov. 2015.

(4) Galar M, Fernndez A, Barrenechea E, Bustince H, Herrera F (2012) A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 42: 463 484.

(5) C. Sei\_ert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, Building useful models from imbalanced data with sampling and boosting, in Proc. 21st Int. FLAIRS Conf., May 2008, pp. 306311.