ORIE 4740 Project

Predicting the Probability of Default Payments

Peng Yuan Chen (pc364), Charlotte Wang (xw476)

Abstract

We are interested in assessing loan loss for financial agencies in Taiwan; by understanding the probability of default, agencies could investigate client's performance and provide better services. This dataset presents several opportunities for interesting analysis. By applying various of techniques, our primary goal is to predict the probability of default payment vary by categories of different demographic and payment variables. The results of our study should help credit card companies identify credit card consumers in stress condition, and take proactive steps to lower default rate.

I. Introduction

Since 2008-2009 financial crisis, credit risk becomes an essential topic among the financial industry, not only within the mortgage markets, but also in the credit card companies. Credit risk is defined as the risk of default on a debt that may arise from a borrower failing to make required payments. This kind of customers is generally regarded as a credit defaulter. As the Federal Reserve Bank continues raising the interest rate these years, the other emerging markets will follow as well, such as Taiwan. As interest rates increase, defaults of credit card payments are likely to happen more frequently. Thus, it is important to apply the statistical data analysis on fitting the best models of the dataset to help the financial industry on predicting the probability of default on credit card payments.

This dataset includes 30,000 observations and different demographic variables including age, gender, due payments, marital status, initial borrowed amount, etc. Since this dataset has a decent amount of data, we applied a variety of techniques to find the best fitted model, including logistic regression with best subset selection and regularization, KNN and various tree-based models. We selected the final best optimal model based on accuracy, AUC, KS, and GINI index, subjected to model interpretability, in order to predict the default payments on customers. We found that performance of Lasso and KNN is higher than pure logistic regression, and they have high interpretability; the predictive power of Random Forest outperforms other methods; however, it is not easy to interpret.

II. Data Cleaning and Exploratory Data Analysis

The default of credit Card Clients Dataset we obtained came from the Kaggle website. Our original dataset consists of 25 predictors with 30,000 observations. Each observation corresponding to each consumer. We are interested in predicting whether or not the consumer will default on his/her credit card the next month, which is a binary classification problem, and we considered

Default.Payment.Next.Month as our response variable. To clean the dataset, we dropped "ID" column, removed all the dataset that showed the value of "N/A" in the response variable Default.Payment.Next.Month, removed all the 0's in the predictor variables SEX, EDUCATION and MARRIAGE. In addition, we replaced the value under the predictor variable EDUCATION to 5 when the value is 6, since the value 5 and value 6 are both unknown under the predictor EDUCATION. We lost less than 1% of observations after cleaning the data with the total observations came down to 29,932. The key variables that we considered are listed in the *Table 1* below.

Table 1. Description of dataset variables

LIMIT_BAL	Amount of given credit in NT dollars
SEX	Gender (1=male, 2=female)
EDUCATION	Education background (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
MARRIAGE	Marital status (1=married, 2=single, 3=others)
AGE	Age in years
PAY_1 to PAY_6	History of past payment in September 2005 to April 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, etc.)
BILL_AMT1 to BILL_AMT6	Amount of bill statement in September 2005 to April 2005
PAY_AMT1 to PAY_AMT6	Amount of previous payment in September, 2005 to April, 2005
Default.Payment.Next.Month	Default status (1=yes, 0 = no)

a. Descriptive Statistics

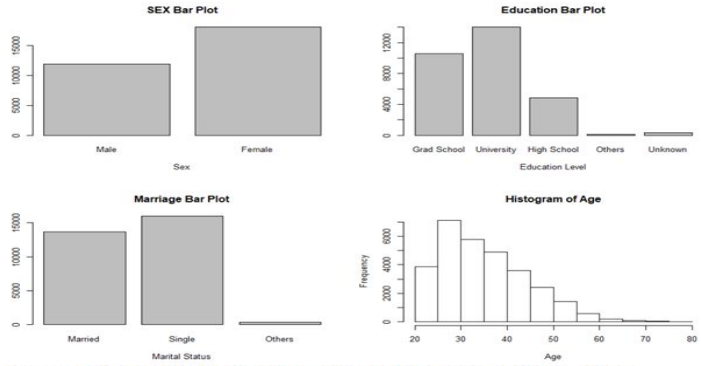


Figure 1. Summary Statistics of Different Demographic Variables

Let's first explore the demographic variables. The distribution of ages is largely one would expect: The average age of the data set is 35.48 years. In *Figure 1,* we can see there are more female than male in the

dataset; the university has the highest number, compared to the graduate school and high school; there are more single than married. From the histogram of the age, we can see that the distribution of the age of the dataset is mainly in the range from 25 to 40, which is the prime ages for the employment population. These key variables are essentially to determine the best fitted model.

Next, *Figure 2* unsurprisingly shows that married couple has higher balance limit than the Single has. In addition, *Figure 3* also shows that as the education background increases, the balance limit increases as well.

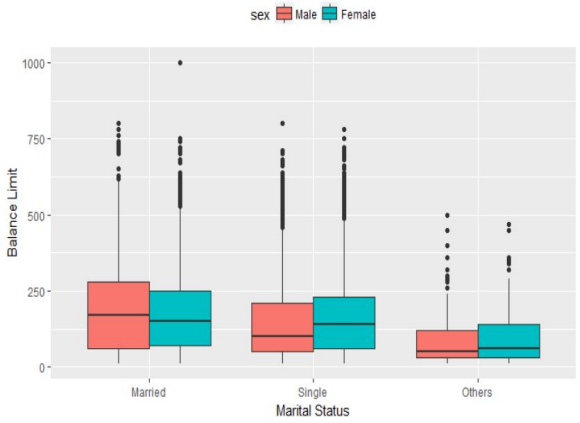


Figure 2. Balance Limit for Different Marital Status

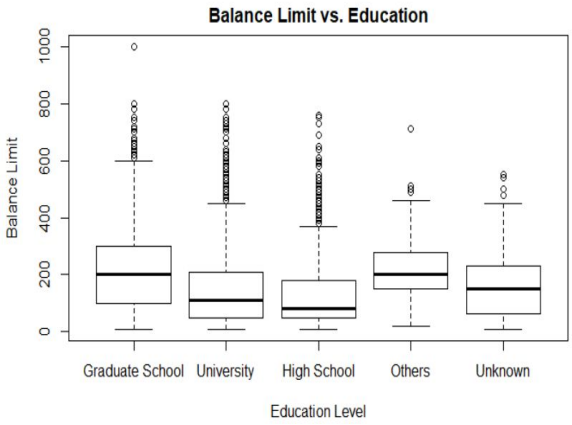


Figure 3. Balance Limit for Different Education Background

After looking at the general trend of the demographic variables, we can look at how each demographic variables associated with the default status. As shown in the *Table 2* below, we can see the default percentage in various demographic variables. One should notice that the dataset is imbalanced, skewed to no defaults.

Table 2. Proportion of Defaults by Various Demographic Variables

	Male	Female	Married	Single	Graduate School	University	High School
No Default%	75.80%	79.20%	76.50%	79.10%	80.80%	76.30%	74.70%
Default%	24.20%	20.80%	23.50%	20.90%	19.20%	23.70%	25.30%

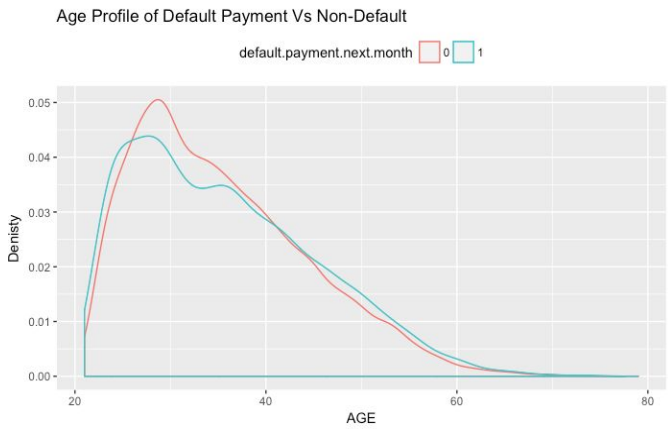
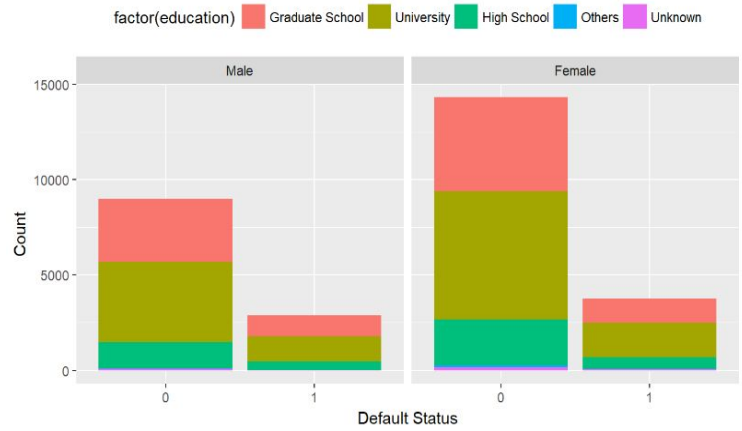


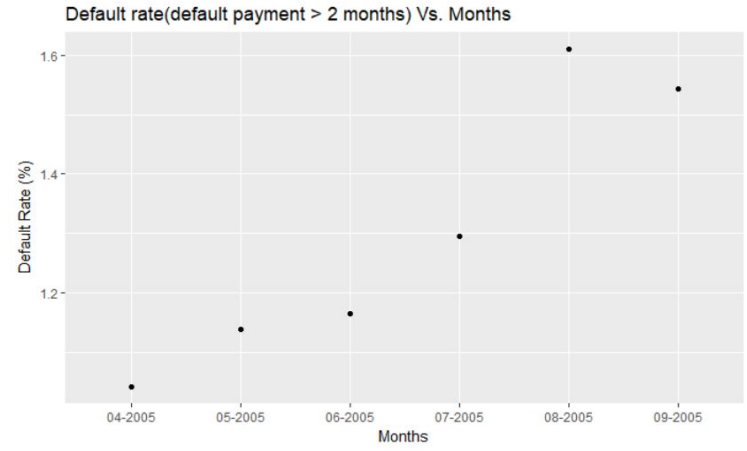
Figure 4. Distribution of Credit Card Default by Age

As age is a continuous variable, we drew a density diagram *(Figure 4)* to visualize the distribution of age in different default status.



We continued uncovering data patterns through conducting feature engineering. We looked into correlations between the response variable and interaction of sex and education. *(Figure 5)*

Figure 5. Proportion of Defaults by Education Background and Gender



From *Figure 6*, we can see that the default rate is linearly increasing from April 2005 to August 2005, then slightly decrease in September 2005, which means our response variable is likely to be correlated with PAY_1 to PAY_6.

Figure 6. Default Rate vs. History of Past Payment

One of the most important takeaways from these graphs and statistics is that there is significant variability in our data set, which gives us considerable freedom to try complicated models.

b. Correlation

We further wanted to examine if high correlation exist among predictor variables.

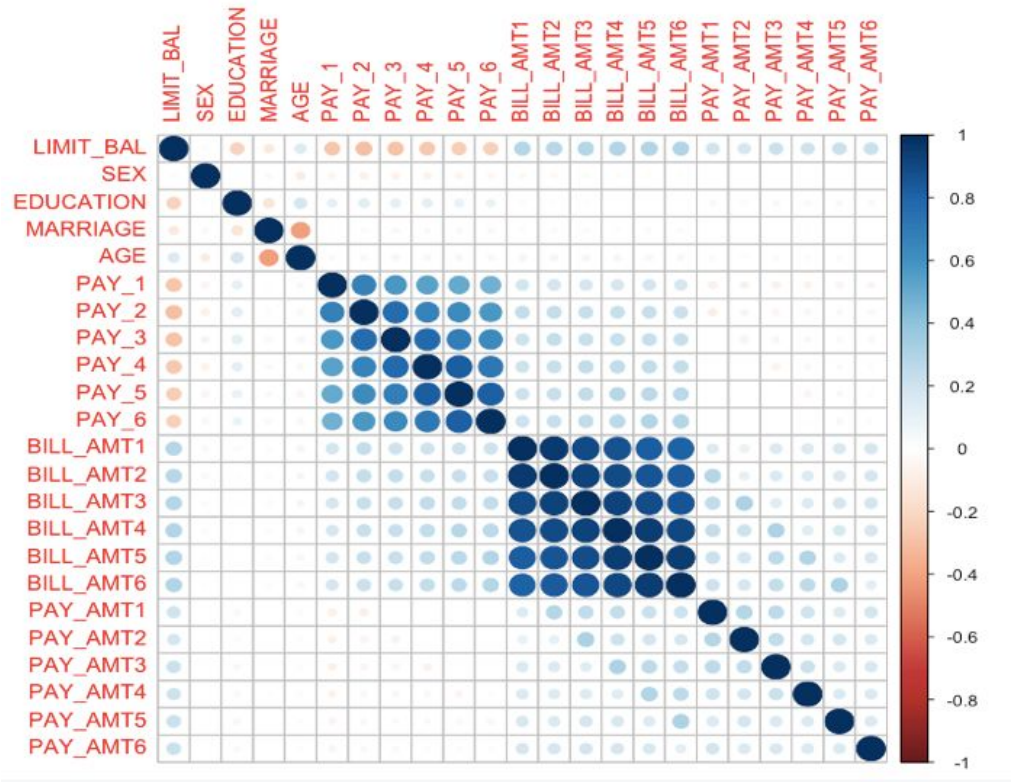


Figure 7. Correlations between Predictor Variables

Figure 7 illustrates that there is a high correlation among the BILL_AMT1 to BILL_AMT6, and among the PAY_1 to PAY_6. This makes sense that payment history or the previous months billing amount has the effects on the next month payment history or the billing amount. However, it indicates that multicollinearity may exist if including all predictors in the model.

III. Model Selection and Evaluation

c. Analytical Setup

Prior to selecting our models for analysis, we split our data into a training set and a test test. We built the model with the training set, 75% of the whole dataset, and used the remaining 25% to test the model validity.

We chose to evaluate model performance via prediction accuracy of target variable on test set. We also generated a ROC curve to measure the diagnostic ability of a binary classifier as its threshold is varied. Ultimately, we were concerned about the area under the curve, AUC. As we achieve high AUC, we consider the model performs well in discriminating between the two categories which comprise the target variable. Additionally, we considered GINI Index and Kolmogorov-Smirnov statistics (KS) as model performance evaluation metrics, as they are the most frequently used in industry. GINI index is calculated based on Lorenz Curve and demonstrates how good is the model compared the a random model. Similar to Gini, KS also captures the discriminatory power of the model in separating "Good" from "Bad". It is the highest separation between the Cumulative Good Rate and Cumulative Bad Rate.

d. Benchmark Analysis

We first considered a naive approach of blindly guessing the credit card default decisions, as we were hoping this may serve as a baseline to judge the success of our various modeling methods. If we blindly guessed that each individual does not default on credit card. 77.8% of consumers don't have default records as we can see from the summary statistics, we have 77.8% confidence in this approach. This will provide us with a baseline performance metric to which we can compare our other models.

e. Logistic Regression Analysis

i. Pure Logistic Regression Model

As we know this is a binary classification problem, as a result, we ran a simple logistic regression. We first fitted a model with all 23 predictor variables included on the training set and found that some variables are not significant if present in the full model. We then tested on the testing set and received an accuracy of 76.5% at the optimal threshold (at which best trade-off between model sensitivity and specificity can be achieved), AUC of 72.2%, KS of 37.8%, and GINI of 44.4%; however, it is not very interpretable, as it includes 12 insignificant predictor variables.

ii. Logistic Regression Model with Best Subset Selection

To improve the model accuracy and interpretability, we performed best subset selection. As what we saw earlier in the correlation plot, some variables are highly correlated, such multicollinearity increases the standard errors of the coefficients, which may in turn make those independent variables insignificant. We tried to find a smaller set of highly-predictive variables that contribute to model success using best subset selection approach. Out of several goodness-of-fit criteria, we chose to use the Bayesian Information Criterion (BIC), which usually results in more parsimonious model than other criteria. As we saw in *Figure 8*, using BIC criterion, we can select 9 variables out of total of 23 variables, which would significantly improve model interpretability.

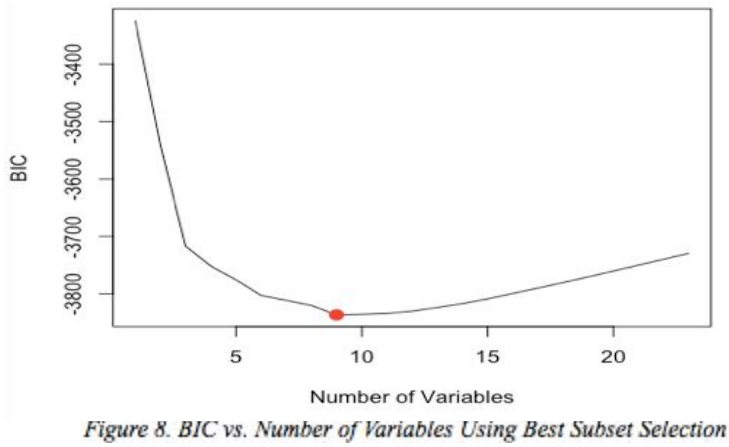


Figure 8. BIC vs. Number of Variables Using Best Subset Selection

We fitted a model with all selected 9 predictor variables included on the training set and found that all variables are significant. We then tested on the testing set and received an accuracy of 77.6% at its optimal threshold, AUC of 71.9%, KS of 38.4%, and GINI of 43.8%. Therefore, this model is decent enough to improve the model accuracy compared with benchmark, and slightly better than pure logistic regression model.

iii. Ridge Regression

Interested to see how regularization would affect our model, we chose to fit ridge regression to our data set. Ridge regression seeks to minimize the same logistic loss function plus the l2 regularizer, with a tuning parameter, lambda. With the addition of regularizer to the objective function, we can stabilize our estimates and produce a unique solution. Specifically, l2 regularizer adds a small penalty for large coefficients and therefore shrinks the coefficients towards 0; however, such solution doesn't help restrict number of variables included in the model. Hence, there is no improvement on model

interpretability. Like most forms of regularization, this will introduce some bias into our model, but it should decrease model variance and possibility of overfitting and hopefully improve accuracy.

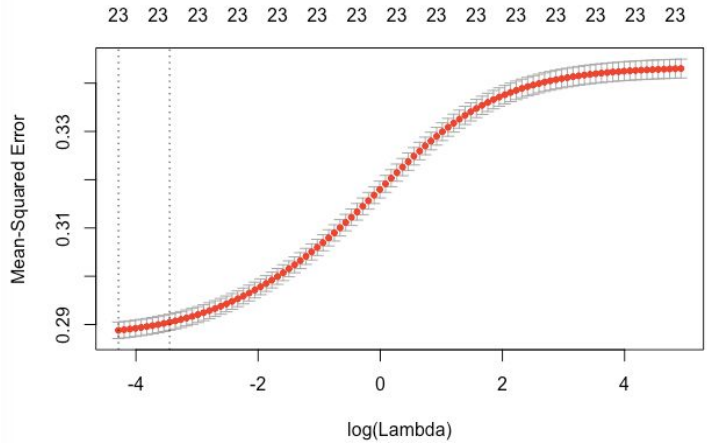


Figure 9. Cross Validation Error at Different Lambda

We performed 10-fold cross validation to select the optimal tuning parameter. This graph shows the corresponding cross validation error given different parameters. As we can see from *Figure 9*, the best lambda is 0.014 at which we achieved the lowest cross validation error.

We fit the final ridge model with the obtained tuning parameter and received an accuracy of 79.4% at its optimal threshold, AUC of 72.1%, KS of 37.6%, and GINI of 44.3%. One may notice that ridge regression significantly improve the accuracy of unregularized logistic regression.

iv. Lasso

Next, we applied a l1-regularized logistic regression to our training set, in order to select the most important feature and further improve model interpretability. By applying this method, the weights for the less important features are restricted to exactly 0. One critical parameter in this regularization method is the tuning parameter, which adjusts the relative importance of the regularizer versus the loss function. It controls the model flexibility, as lambda increases, less variables will be included in the model, hence the model is more interpretable. As with any regularization method that focuses on the size of the coefficients, we first ensured that all of data was appropriately scaled in order to avoid unnecessary bias. Next we performed 10-fold cross validation to select the optimal tuning parameter. This graph shows the corresponding cross validation error and AUC given different parameter lambda.

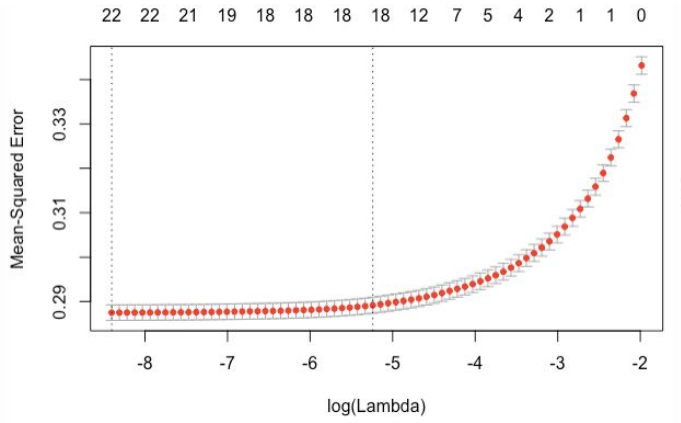


Figure 10. Cross Validation Error at Different Lambda

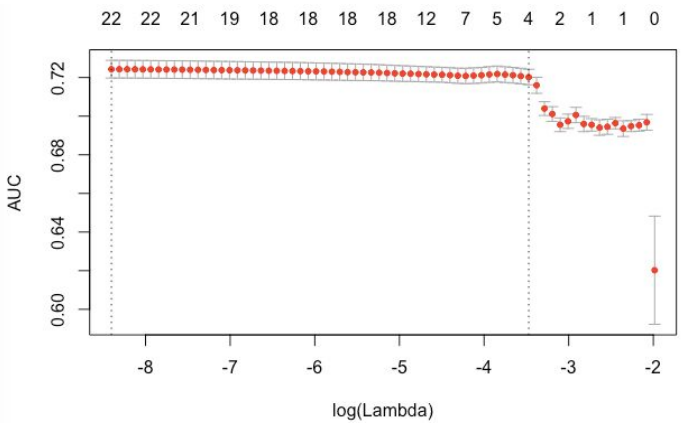


Figure 11. AUC at Different Lambda

We found that under the optimal tuning parameter at which we achieved the lowest cross validation error, only one variable was eliminated *(Figure 10)*. However, looking at *Figure 11*, we could make a trade-off to select 4 variables and still achieve high AUC. Although such model may decrease the model strength, we believe that it significantly improve model interpretability, which would be valuable in commercial use.

We fit the final lasso model with the obtained tuning parameter and received an accuracy of 79.4% at its optimal threshold, AUC of 71.1%, KS of 37.6%, and GINI of 42.2%. One should notice that lasso outperforms unregularized logistic regression in terms of accuracy; additionally, we consider it a decent model as it only includes the most important predictors while achieving high accuracy.

f. Tree-Based Analysis

i. Single Classification Tree

We decided to pursue tree based methods as we anticipated that they will perform very well on this data. We first tried single classification tree, which can be graphically displayed and easy to interpret. It can automatically complete feature selection; the top few nodes on which the tree is split are essentially the most important variables within the dataset.

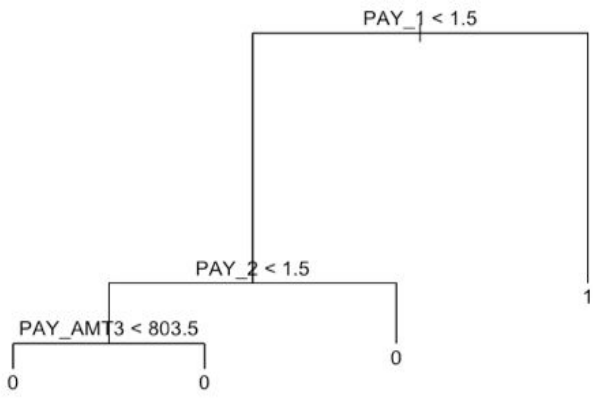


Figure 12. Single Classification Tree

Only three variables are actually used in the tree construction: PAY_1, PAY_2, PAY_AMT3. The most important indicator of default decision appears to be the payment history in September 2005 (PAY_1), since it is the first branch differentiates "good" and "bad" account. The classification tree achieved a high accuracy of 81.9% on the test set.

Next we considered whether pruning the tree might lead to improved results. We applied cross validation to determine the optimal level of tree complexity. Cost-complexity pruning was used to select a sequence of best subtrees.

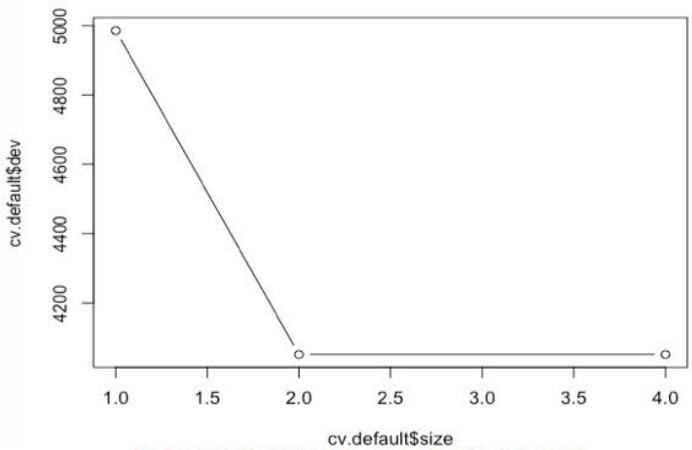


Figure 13. Cross Validation Error vs. Tree Size

Figure 13 shows that as we cut the tree to 2 terminal nodes, the corresponding classification error rate is the lowest. However, this only leaves us with one criterion -- PAY_1 which may be too simple for building a model; thus, we chose not to prune the tree. We then plotted the ROC to visualize the prediction performance at different threshold. AUC is 73.2%, KS is 37.1%, GINI is 46.4%, which further verify that the single tree model is well-performed.

ii. Bagging Classification Tree

We have known that single tree is not very robust, as small changes in the data can cause a large change in the final estimated tree. We applied bagging to reduce variance in decision trees. To do so, we generated different bootstrapped training sets and then averaged the predictions obtained on each bootstrapped set. This bagged model yields a high accuracy of 81.7%. One may notice it is slightly lower

than single tree; however, as mentioned before, the bagged model is more robust with a lower variance. Compared with single tree, bagging is hard to interpret, thus we used GINI index to compare the importance of each predictor.

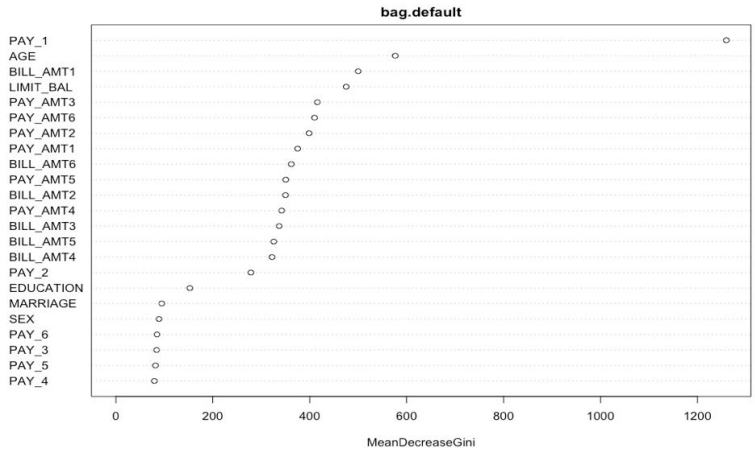


Figure 14. Mean Decrease in GINI for Various Predictor Variables

We can determine the three most important variables by mean decrease in GINI index from this plot, PAY_1, AGE, BILL_AMT1. The larger the mean decrease in GINI, the more importance the variable. We plotted the ROC to visualize the prediction performance at different threshold. AUC is 76.0%, KS is 39.7%, GINI (for model) is 52.0%, which further verify that the bagging model is well-performed.

iii. Random Forest

The next tree-based model we considered is the random forest. Random forest provides an improvement over bagged trees by way of a small tweak that decorrelates the trees. If all of the bagged trees look similar to each other, averaging highly correlated quantities does not lead to a substantial reduction in variance over a single tree. Random forest overcomes this issue by forcing each split to consider only a subset of predictors, thus makes the resulting trees less variable. For the random forest model we used 200 trees in the forest with forcing each split of the tree to consider 5 predictors. (This number of predictors considered at each split is chosen as the square root of number of total predictors.) The final random forest achieved an accuracy of 81.7% and it further reduced the tree variance.

Just like what we did for bagging, we can attempt to judge the importance of individual variable by looking at the reduction in GINI coefficient. We see that the top three most important features are still PAY_1, BILL_AMT1, and AGE if using the mean decrease in GINI as the criterion.

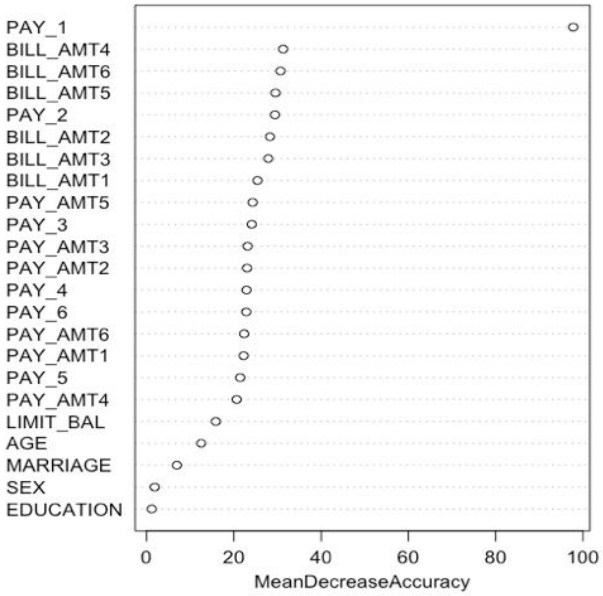


Figure 15. Mean Decrease in Accuracy for Predictor Variables

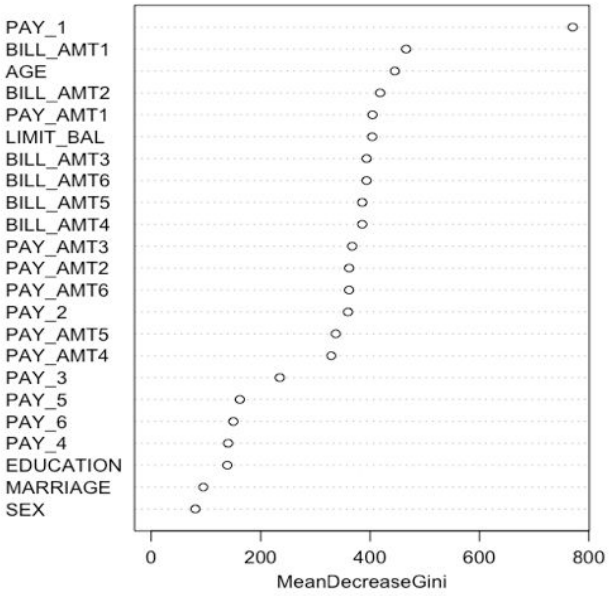


Figure 16. Mean Decrease in GINI for Predictor Variables

We plotted ROC to visualize the prediction performance at different threshold. AUC is 76.4%, KS is 40.3%, GINI (for model) is 52.9%, which further verify that the random forest model performs well.

g. Other Approach -- KNN

We were also interested to see if KNN would be a better classifier for our dataset. We believe it would improve the model performance as KNN is comparatively robust to noisy data and it can thus handle non linearly separable data. We applied cross validation to determine at which k the model can achieve the highest accuracy.

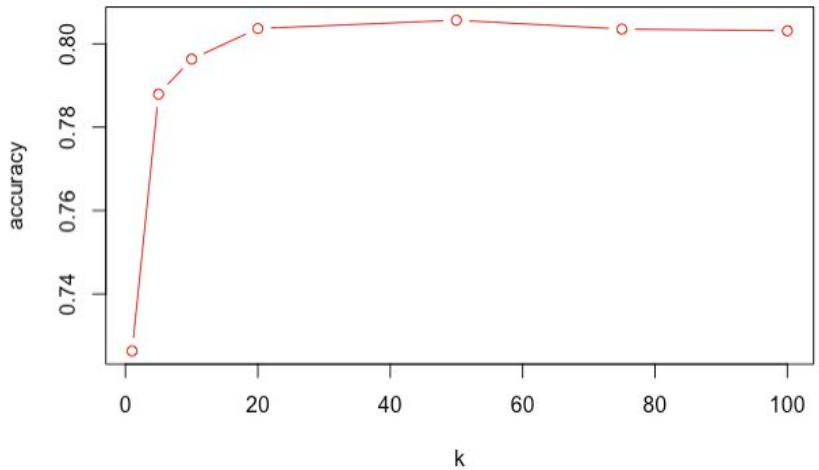


Figure 17. Accuracy vs. k for KNN

As we can see from the plot, the accuracy reaches the maximum when the number of nearest neighbors is 50. The accuracy for this model is 81.3%.

The ROC curve shows the prediction performance at different threshold. AUC is 75.9%, KS is 38.9%, GINI is 51.8%, which further verify that the KNN model is well-performed.

IV. Conclusion

Our goal is to find the model with the highest performance to help us predict the probability of default payment among credit card customers. In total, we applied eight different models to this problem. To evaluate prediction performance, we traced four different metrics: accuracy, area under ROC curve (AUC), KS statistics, and GINI index. The following table summarizes model performance.

Table 3. Performance Metrics for Different Models

Model	Accuracy	AUC	KS	GINI	Interpretability
Logistic Regression	▼ 76.5%	72.2%	37.8%	44.4%	23 Variables
LR with Best Subset	▼ 77.6%	71.9%	38.4%	43.8%	9 Variables
Ridge	▲ 79.4%	72.1%	37.6%	44.3%	23 Variables
Lasso	▲ 79.4%	71.1%	37.6%	42.2%	4 Variables
KNN	▲ 81.3%	75.9%	38.9%	51.8%	50 Nearest Neighbors
Single Tree	▲ 81.9%	73.2%	37.1%	46.4%	3 Split Criteria
Bagging	▲ 81.7%	76.0%	39.7%	52.0%	23 Variables at Each Split
Random Forest	▲ 81.7%	76.4%	40.3%	52.9%	5 Variables at Each Split
Benchmark	— 77.8%	NA	NA	NA	Educated Guess

As we can see, almost all models outperforms benchmark in terms of accuracy except the pure logistic regression. (The difference between best subset selection and benchmark is negligible.) We also noticed that tree-based models and KNN in general performs better than logistic regression models, in terms of any evaluation metrics. We plotted ROC curve to all models and found the area under curve (AUC) are quite high. We included KS stats and GINI index as additional metrics as they are more valuable in the industry. The higher the KS and GINI towards 1, the better the model. We can see that KS and GINI for all models are fairly high.

If we look at the model interpretability, out of all logistic regression models, we would prefer logistic regression with best selection and with l1 regularizer (lasso) as they give the simplest model with

relatively high performance; we would also recommend KNN and random forest as they are considered to be robust and have relatively low variance. This project had given us ample opportunities to explore different model analysis on the dataset. For future analysis, we would like to further improve the robustness of models and apply other learning methods. Some methods we considered but did not have a chance to apply were methods like SVM, XGBoost, which may yield better results.

V. References

Default of Credit Card Clients Dataset. (n.d.). Retrieved May 17, 2018, from https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset.

James, Gareth. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2014.

"Board of Governors of the Federal Reserve System." *Board of Governors of the Federal Reserve System*, Board of Governors of the Federal Reserve System (U.S.), 17 May 2018, www.federalreserve.gov/.