University of Victoria | Mathematics & Statistics

STAT 568 GLM Project

Predictive Models for Credit Card Default

By: *Peijie Xie (V00887527)*

Meixin Liu(V00812082)

Ben Liu (V00863757)

Instructor: Dr. Xuekui Zhang

Last Revised: April 14, 2018

Acknowledgement

This project was the supported by the STAT 568 Generalized Linear Model course. We want to thank our Professor Xuekui Zhang from the department of Mathematics and Statistics at University of Victoria who hold every week meeting with us and provided a lot of professional methodologies and expertise that greatly assisted the project. We also want to thank Dr. Li Xing who provides professional industry experience for this course that greatly improves our interest in the biostatistics field. Again, we would like to show our gratitude to Professor Xuekui Zhang for his hard work, commitment and offering us this great opportunity to collaborate on the project and learn from each other.

Abstract

This report investigated the customer's default status in Taiwan. Principle Component Analysis was applied to reduce dimension and remove collinearity between predictors. Penalized regression methods, such as Lasso, Ridge and Elastic net, were applied to predict the credit card default based on personal bank information. Several classification models, such as Logistic Regression, Naïve Bayes, Linear Discriminant Analysis, K-Nearest Neighbors and Random Forest were also built to improve the predictive results. Accuracy rate, AUC and Accuracy ratio were three main criteria used to evaluate each model performance.

Keywords: Credit Risk, Penalized regression, Classification methods, Performance Evaluation

1. Introduction

Credit risk is the traditional risk of banking industry and managing credit risk is a crucial component in banking business. Currently, to increase their market share, banks tend to over-issued credit cards to many unqualified applicants. Meanwhile, many credit card holders overuse their credit card regardless of their repayment ability. However, it would be too expensive having a close background check for each individual applicants. Therefore, it is a necessity for banks to construct a model to predict whether a client would default his credit card payment or not.

In this report, historical data in Taiwan was analyzed. The rest of this report is organized as follows. In section 2, we introduced our data set and our preliminary data analysis methods. In section 3, we introduced the criteria for comparing different models. Then, in section 4, we introduced parametric models we tried including Penalized Regression (Lasso, Ridge and Elastic Net), Logistic Regression, Naive Bayes and Linear Discriminant Analysis. In section 5, we introduced non-parametric models including K-Nearest Neighbors and Random Forest Methods. In section 6, we compared previous methods and gave our conclusion and discussions. Finally, in section 7, we discussed some potential future work.

2. Preliminary Data Analysis

2.1. Data Description

We used the historical payment data from a bank in Taiwan on October, 2005. The dataset was downloaded from the kaggle website. This dataset has a binary outcome variable to indicate whether the client defaulted or not. Other 23 variables were used as explanatory variables. There were 30000 observations in total and the following table (Table 1) showed the variable names, their categories and meaning.

Table 1: Data Description

Variable Name	Category	Description
Y	Categorical (Nominal)	Default status
$\mathrm{LIMIT}_{\mathrm{BAL}}$	Numerical	Amount of Credit limit
SEX	Categorical (Nominal)	Gender
EDUCATION	Categorical (Ordinal)	Education Background
MARRIAGE	Categorical (Nominal)	Marital status
AGE	Numerical	Age
$\mathrm{PAY}_0 - \mathrm{PAY}_6$	Categorical (Ordinal)	Default time
$\mathrm{BILL}_{\mathrm{AMT1}} - \mathrm{BILL}_{\mathrm{AMT6}}$	Numerical	Bill statement amount
$\mathrm{PAY}_{\mathrm{AMT1}} - \mathrm{PAY}_{\mathrm{AMT6}}$	Numerical	Previous payment amount

2.2. Outlier Detection

Before constructing the predict model, we first used Mahalanobis distance to detect outliers. We chose critical cut-off value. Figure 1 showed the Mahalanobis distance for each observation with the mean value before and after deleting outliers. Finally, 26716 observations were left. Note that many important clients who had large payment or bill amount each month may be regarded as outliers. We believed that those clients would be very crucial for the bank. However, those clients would be required special models. In this project, we only consider a general model.

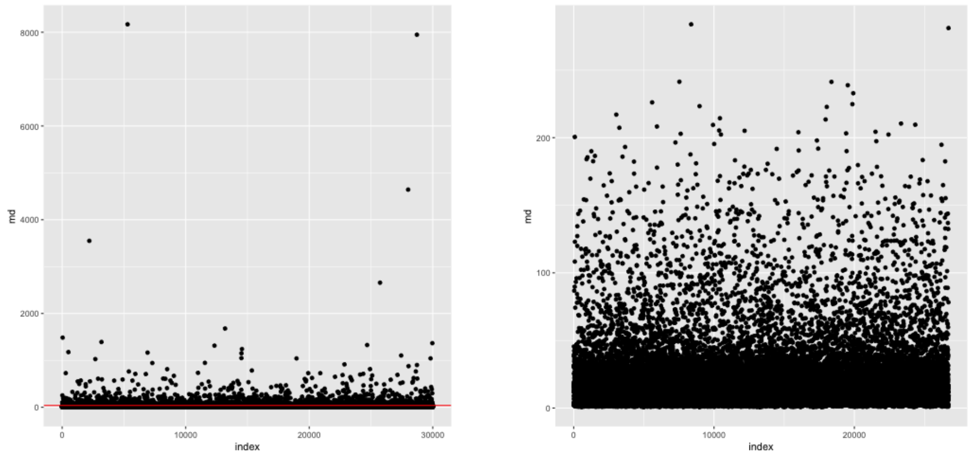


Figure 1: Mahalanobis Distance

2.3. Principal Component Analysis

Principle Component Analysis (PCA) is a multivariate statistics analysis method to perform dimension reduction for correlated variables while retain as much information as possible. It uses linear combinations of original variables called principal components so that each component has the highest possible variance and are orthogonal to each other.

The original dataset contained 23 predict variables, and 18 of which are monthly repayment status, amount of bill statement and amount of payment for different months. It is reasonable to believe that those monthly data shared a strong relationship with each other. Therefore, we first plotted a correlation heatmap (Figure 2) for all variables. The darker the color is in the heatmap, the stronger the linear relationship those two variables share.

From the heatmap, 6 bill payment amount variables share a strong linear relationship, which will lead to the collinearity problem for our future model fitting. Therefore, we perform the PCA on these 6 variables. We plotted the variance explained percentage for each principal component (Figure 3).This plot showed how much information was preserved by each component. From the plot, we can see that the first component has already explained almost

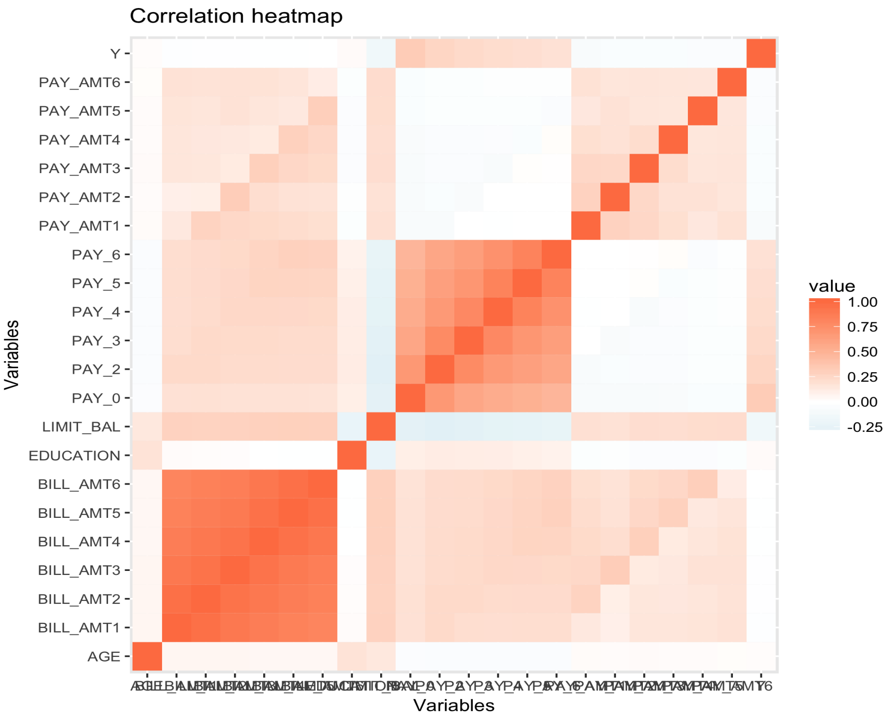


Figure 2: Correlation Heatmap for Predictors

all the variance (information) for the 6 variables. Therefore, we decided to use only the first principal component and the six bill amount variables were reduced to one variable named $BILL_{AMT}$ by linear transformation using the loadings of the first principal component.

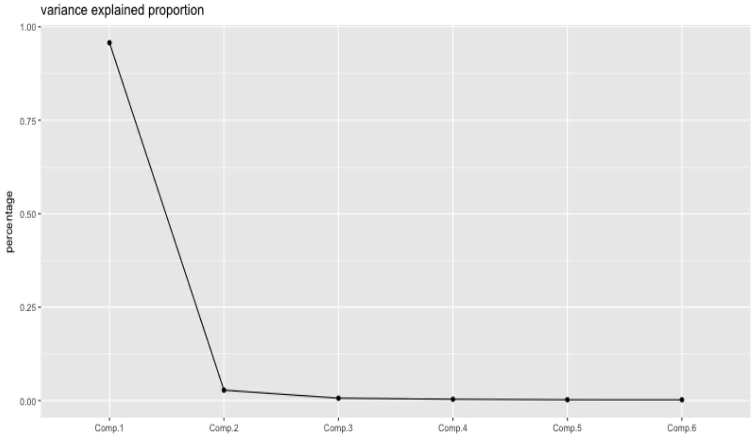


Figure 3: Variance Percentage for Each Component

3. Criteria for Model Comparison

This section will introduce three basic criteria we used for comparing our models which are accuracy rate, ROC curve and lift chart.

3.1. Accuracy Rate

We were dealing with binary classification prediction, that is, there are exactly two classes, positive or negative. In our case, we took 1 as positive and 0 as negative. Accuracy rate is a popular way for measuring accuracy of classification models, and accuracy rate = $\frac{TP+TN}{TP+FN+FP+TN}$, where TP, FN, FP and TN are defined in Table 2.

Table 2: Confusion Table

		Actual Class	
		Positive(1)	Negative (0)
Predicted Class	Positive	True Positive	False Positive
	Negative	False Negative	True Negative
		Sensitivity	1-Specificity
		$TPR = \frac{TP}{TP+FN}$	$FPR = \frac{FP}{FP+TN}$

3.2. Receiver Operating Characteristic Curve

ROC(Receiver Operating Characteristic Curve) is an important way to visualize the tradeoffs between sensitivity and specificity, where sensitivity and specificity is defined in Table 2. Moreover, ROC curve can identify the best threshold for separating positive and negative. According to Vuk and Curk(2006), "Every point on a ROC curve corresponds to a binary classifier, for which we can calculate the classification accuracy and other quality measures."(p.96)

The algorithm for computing the best threshold are as follow. Denoted T as the total number of responses, P as the total number of positive responses and N as the total number of negative responses.

$$\begin{aligned} Accuracy &= \frac{TP + TN}{T} \\ &= \frac{TPR \cdot P}{T} + \frac{(1 - FPR) \cdot N}{T} \\ &= \frac{P}{T} \cdot sensitivity + \frac{N}{T} \cdot specificity \\ &= \frac{P}{T} \cdot y + \frac{N}{T} \cdot x \end{aligned}$$

Thus, to maximize the accuracy rate is to find the point on the ROC curve such that it has tangent line with slope $\frac{N}{P}$. Moreover, we can also get a glance from the area under ROC curve(AUC) of overall accuracy of model. Ideally, "A perfect model would return a squared ROC"(Suite, 2016), that is AUC = 1.

However, both accuracy rate and AUC for ROC can be misleading when the data is highly unbalanced. Knowing that the ROC curve is independent of the ratio between the number of positive class over the number of negative class. Thus, ROC is useful if we do not know the ratio between positive and negative. For our data, the ratio between positive and negative is 22%, which is foregone. For instance, if more than 90% of the data are non-default, and we simply label all clients are will not be default, then our accuracy ratio will be over 90%. In this case, we actually did not make any prediction, but the accuracy rate seemed satisfied. Similarly, for ROC, no matter what the ratio of default and non-default is, the ideal line is always the same. Therefore, those criteria is not suitable when the outcome variable is highly unbalanced.

3.3. Lift Chart

Notice that percentage of positive response is around 22%, which is obvious not a 50% split between positive and negative. Then AUC of ROC is not the best comparison criteria. Therefore, we introduced lift chart as another criteria for our comparison. Unlike the ROC curve, lift curve integrates the percentage of positive on the charts(Suite, 2016) so that the curve depends on ratio of the total number of actual positive class over the total number of actual negative class. The x-axis represented the cumulative percentage of total number of data, and the y-axis showed the cumulative proportion of positive prediction. In the lift chart, 45-degree line is baseline; according to Suite(2016), "any lift going below would be inferior to a random selection". Also, our perfect line represents the prediction without any mis-classification. Here we introduced accuracy ratio to compare lift curves.

$$Accuracy - ratio = \frac{Area\ between\ baseline\ and\ model}{Area\ between\ baseline\ and\ perfect\ line}$$

4. Parametric Models

In this section, we fitted several parametric models, penalized model, logistic regression and classification algorithms. To overcome the overfitting problem, we did 10-fold cross validation for each model. We partitioned data randomly into 10 equal size subsets, and take each single subset exactly once as the testing data and the remaining as training data.

4.1. Lasso, Ridge and Elastic net

First, we chose penalized models for analyzing our high-dimensional data. Here we fit Lasso, Ridge and Elastic net. Notice that high-dimensional data often include many correlated predictor variables. Even though we did principal component analysis to deal with the highly correlated predictor, it was still essential to fit penalized models which can select variables and fit a model simultaneously.

To fit penalized models, we chose Lasso, Ridge and Elastic net methods. In each fold of cross validation, we did a second layer cross validation by cv.glmnet() to choose the tuning parameter λ, which controls the strength of the penalty. Here we chose two values of λ, one minimizes the CV error curve and the other is the largest value of λ such that CV error is within one standard error of the minimum. When using elastic net method, we also did a second layer cross validation to choose α to decide weight between L_1 and L_2 penalty. Since each one of the 10 equal size subsets was retained exactly once as testing data, we combined the 10 times results from the folds which gives the predictions of whole data. Base on the prediction class and actual class of response, we drew ROC curves for each models in Figure 4. It is obvious that each models have similar curves.

Knowing that ROC curve can help identify the best cut-off for separating positive and negative. Thus we chose the best threshold from ROC curve which are shown in Table 3. Accuracy of each model when positive and negative best separated were around 0.82 and notice that Lasso model with 1se lambda had the best accuracy rate with best separating threshold 0.407.

Without choosing specific threshold, we wanted to see the overall performance of each

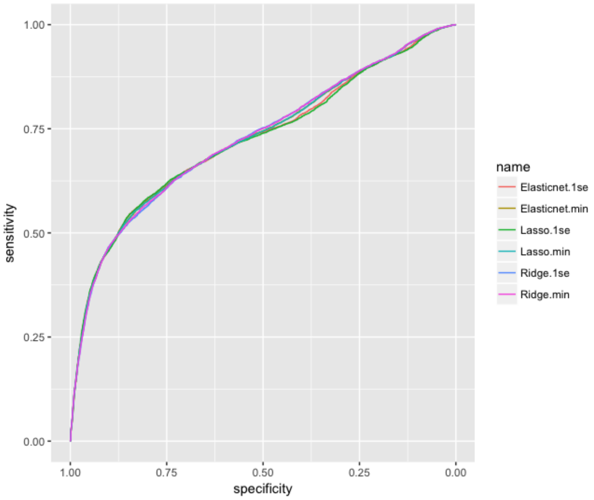


Figure 4: ROC Curves for Penalized Models

models. Since our data is unbalanced, we chose lift chart to analysis fitness of each model. From the lift chart above, it is shown a similarly curve among models. From Table 4, the accuracy-ratio are all around 0.45, and the ridge method with min lambda performed the best with accuracy-ratio 0.4516135.

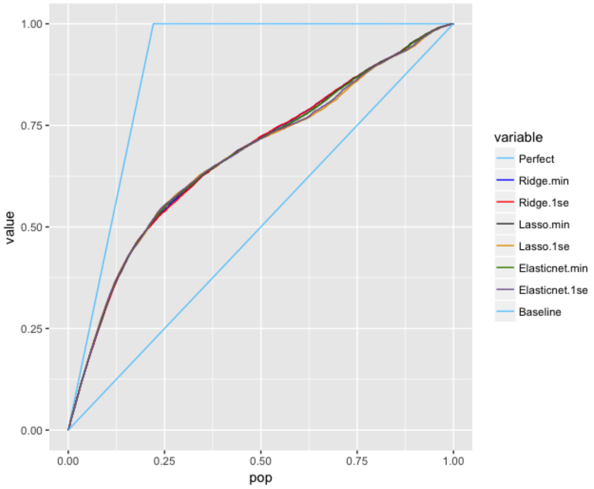


Figure 5: Lift Curves for Penalized Models

Table 3: Accuracy Rate when Positive and Negative are Best Separated

Model	Threshold	Accuracy
Ridge with min Lambda	0.4141284	0.8169262
Ridge with 1se Lambda	0.3969842	0.8161401
Lasso with min Lambda	0.4229622	0.8178245
Lasso with 1se Lambda	0.4069446	0.8192843
Elastic net with min Lambda	0.4257687	0.8180491
ELastic net with 1se Lambda	0.4063384	0.8173379

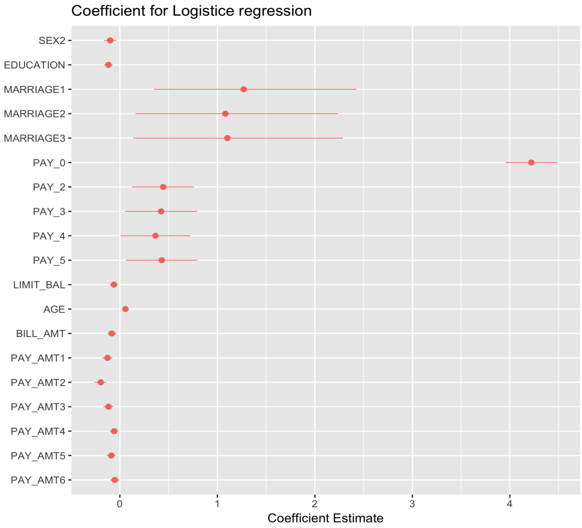
Table 4: Accuracy Ratio for Penalized Models

Model	Accuracy - ratio
Ridge with min Lambda	0.4516135
Ridge with 1se Lambda	0.4501136
Lasso with min Lambda	0.4504514
Lasso with 1se Lambda	0.4434366
Elastic net with min Lambda	0.4507044
ELastic net with 1se Lambda	0.4460900

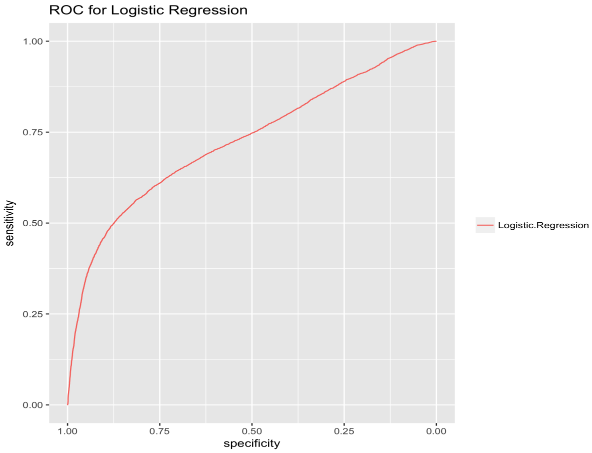
4.2. Logistic Regression

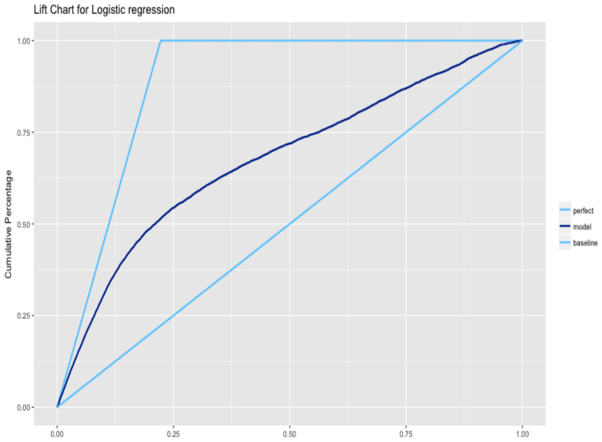
Logistic Regression is a regression model for binary outcomes. In logistic regression, the response variable follows a binomial distribution, and we model the log odds of the event (in our case, default) as a function of predictor variables. We used 10-fold cross validation to avoid the problem of overfitting. After fitting the model, we used stepwise AIC to perform variable selection.

The following plot showed the coefficients for logistic regression. From the plot, we found the PAY_0 variable has the largest value, indicating the the latest payment status might have a large effect for us to predict whether the client default or not.



We used the ROC Curve, the accuracy rate the the accuracy ratio form the lift chart to evaluate the model. The ROC Curve and the lift chart was plotted as follows. We found the Area Under Curve for the ROC Curve is 0.726. Meanwhile, the optimal accuracy rate was 0.817 when the cuf off was chose at 0.4369. The accuracy ration was 0.4521 from the lift chart. We would use these data to compare logistic regression with other models later.





4.3. Linear Discriminant Analysis and Naïve Bayes

We also applied two other classification algorithms to make predictions and compared each model by computing the area under curve. Linear Discriminant Analysis is a classification technique which is very powerful, in addition, able to handle multiple classes. The other one is Naïve Bayes which is built upon Bayes theorem, a theorem pertaining to conditional probabilities. Given some possible set of classes, we use the patterns inside those classes to predict new, unclassified data. The main idea behind Naïve Bayes is each predictor is assumed to be independent of one another. Although, this is a strong assumption, the models tends to perform relatively well in practice after we removed collinearity by PCA. Again, we applied 10-fold cross validation to train the model and compared them by visualizing the ROC.

Figure 6 indicated that the Naïve Bayes outperformed the LDA by having larger AUC.

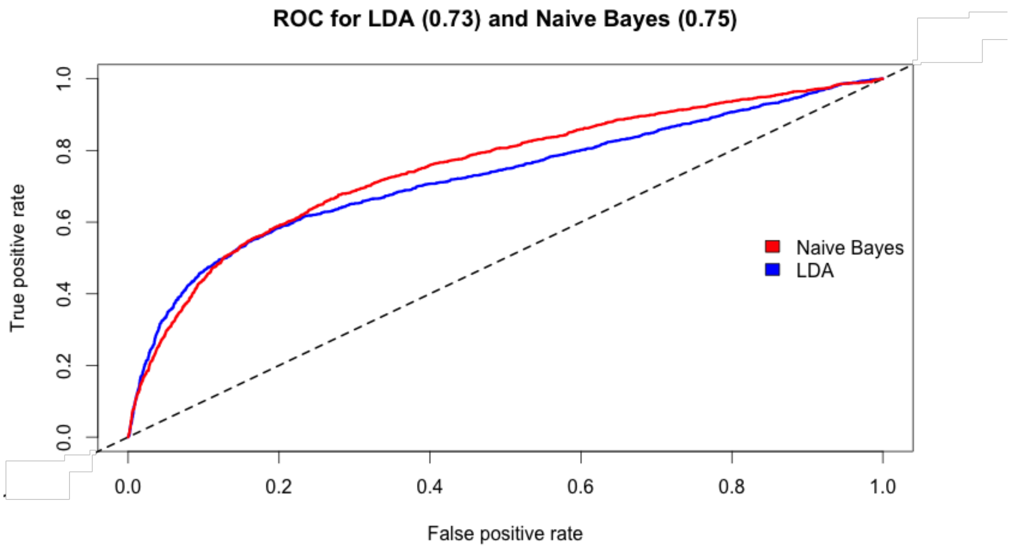


Figure 6: ROC: LDA v.s. Naïve Bayes

5. Non-parametric Models

In this section, we looked at two non-parametric methods, such as K-Nearest Neighbors and Random Forest, to see if model's performance can be further improved.

5.1. K-Nearest Neighbors

KNN is one of the simplest method in massive data mining. First, a positive integer k is specified, along with a new sample. Then, we select k points that are closest to the new sample, and find the most common classification of these points. Finally, this is the classification we give to the new sample. The main thing we needed to consider here was the selection of the tuning parameter k. If k is very small, we are likely to overfit the model, and if k is too large, the model might be underfitted. To solve this trade-off, we tried 100 different numbers of neighbor, and used the 10-fold cross validation for each k.Figure 7 showed that the accuracy rate increased as k increased, but it tended to be stable at around 78%, and the maximum accuracy was achieved at k equals 16, 79.3%. So we fixed this k, and calculated the AUC which was only 0.547.

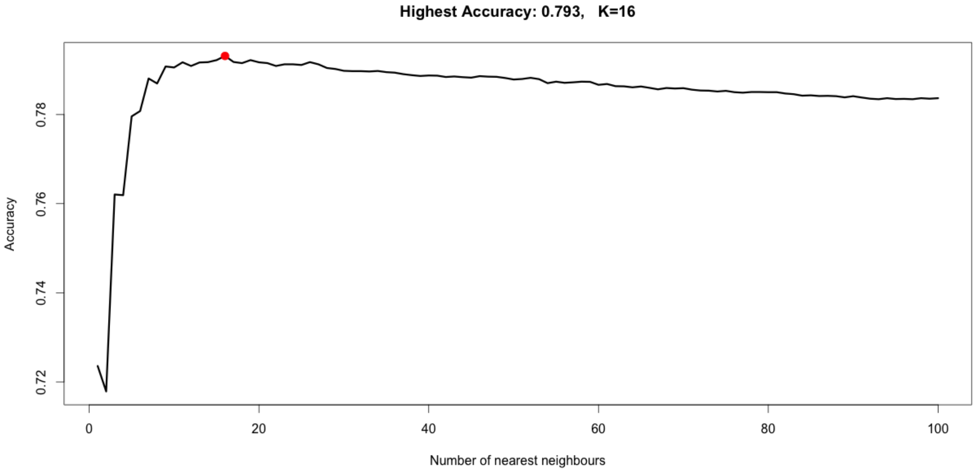


Figure 7: Number of Neighbors

5.2. Decision Tree and Random Forest

Decision tree can accommodate non-linear relations between predictors, and it can be easily visualized and interpreted to stakeholders. We used the CART algorithm to build a decision tree, which repeatedly partitioned the data into small subsets until the final subset is homogeneous in terms of the response variable. The model might also be overfitted in order to achieve the homogeneous split. Hence, it is necessary to prune the tree according to the complex parameter which is used to mention the smallest improvement before further splitting nodes. Figure 8 implied the PAY_0 was an important variable. This was consistent with all our previous results.

Decision tree is very good to give us first hand information about our data, but sometimes it is not stable. In other words, smaller changes in the training set can lead to significant

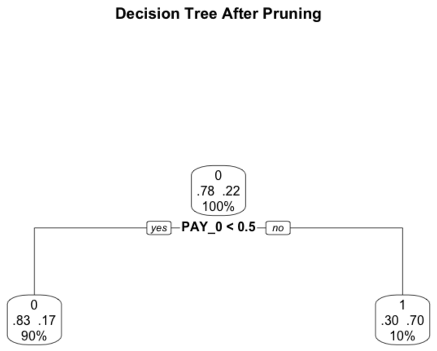


Figure 8: Decision tree after pruning

changes in the tree structures. The stability can be improved using bootstrapped methods. As a result, we applied the random forest algorithm which is more robust and builds a bunch of decision trees on bootstrapped training samples of data. We also checked the importance of variables based on Mean Decrease Gini Index, which can be explained as the larger the value, the more important the variable. Figure 9 indicated that PAY_0, $Bill$, $Amount$, PAY, $Amount$, $Limit$, $Balance$ were important variables, while variables like $Marriage$ and Sex were of less importance. Again, we trained the random forest using 10-fold cross validation and plotted the ROC in Figure 10. The best threshold was given at 0.789, which was defined as the maximum value of the sum of sensitivity and specificity. The AUC was 0.77, larger than all previous models. The lift chart for the random forest was displayed in Figure 11.

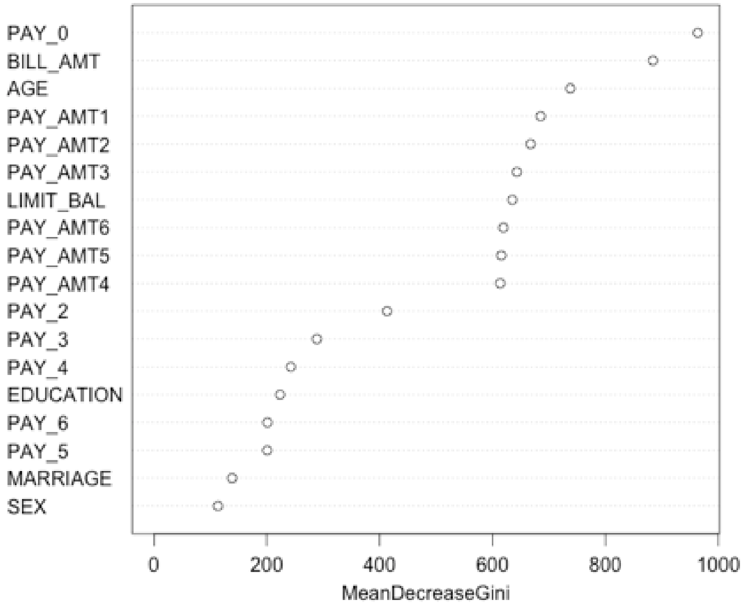


Figure 9: Importance of variables

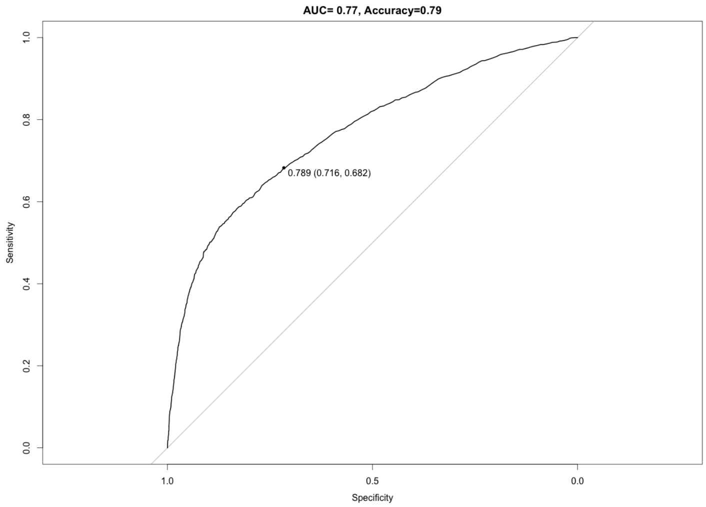


Figure 10: ROC: Random Forest

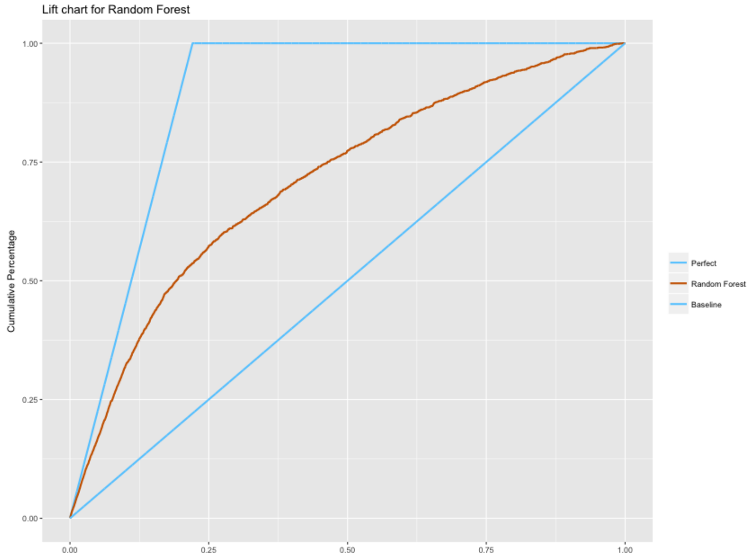


Figure 11: Lift Chart: Random Forest

6. Conclusion

We compared each model performance in Table 5. It was clear that the random forest has the largest AUC than others, while the parametric models had higher accuracy rate in predicting default than non-parametric models. Since the number of 0's accounted for the majority in our data, we further compared these models based on accuracy ratio in Table 6. The random forest won again by having approximately 20% higher accuracy ratio than the penalized regression methods. As a result, the random forest was selected as our best model.

7. Future work

From results we obtained, it shows that for some models have the best performance at specific threshold, but some other models have the best overall performance. For real world problem, the criteria for choosing model are depended. In future analysis, We are also curious if any other models can be considered such that they have better performance

Table 5: Model Comparison

Model	AUC	Accuracy Rate
Random Forest	0.77	0.787
KNN	0.54	0.793
Logistic	0.73	0.816
Ridge (min)	0.73	0.817
Lasso (1se)	0.72	0.819
LDA	0.73	0.812
Naïve Bayes	0.75	0.810

Table 6: Model Comparison

Model	Accuracy Ratio
Random Forest	0.539
Ridge (min)	0.452
Lasso (1se)	0.443
Logistic	0.452

in both accuracy rate and accuracy ratio, such as artificial neural network which can infer unseen relationships between predictors. Outliers removed by Mahalanobis distance could be further explored by fitting above models, and compared with the current methods.

Reference

[1] Baesens B, Van Gestel T, Viaene S, Stepanova M, Suykens J, Vanthienen J. Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the operational research society. 2003 Jun 1;54(6):627-35.

[2] Hrdle W, Simar L. Applied multivariate statistical analysis. Berlin: Springer; 2007 Aug 9.

[3] James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. New York: springer; 2013 Feb 11.

[4] Suite. Big Data and Analytic for Citizen Data Scientists: Lift, ROC, AUC, and Gini [web log comment]. Retrieved from http://www.business-insight.com/blog/2016/07/lift-roc-auc-and-gini/; 2016, July 28.

[5] Vuk, M. & Curk, V. ROC Curve, Lift Chart and Calibration Plot. Metodološki zvezki, Vol.3. Retrieved from https://www.stat-d.si/mz/mz3.1/vuk.pdf; 2003.

[6] Yeh IC, Lien CH. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications. 2009 Mar 1;36(2):2473-80.

[7] Data source https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset