PREDICTING THE PROBABILITY THAT SOMEONE WILL EXPERIENCE FINANCIAL DISTRESS IN THE NEXT TWO YEARS

ADOLPH N. OKINE

STATISTICAL COMPETITION
Department of Mathematics
ILLINOIS STATE UNIVERSITY
2015

TABLE OF CONTENTS

1. Abstract

This paper focuses on improving the state of the art in credit scoring, by predicting the probability that somebody will experience financial distress in the next two years. This was motivated by the shortfall of the credit scoring algorithm, which guesses the probability of default, the methods banks use to determine whether or not a loan should be granted. Classification techniques, namely, the linear discriminant analysis, the quadratic discriminant analysis, logistic regression, random forest and sector vector machine. Principal component analysis is adopted as data reduction technique to analyze its effect on the models. The models were assessed using ROC curve which is a graphical plot that illustrates the performance of the models. The ROC curve was constructed for both training data and test data. The results indicates that the random forest was the best model.

2. Introduction

Banks play a crucial role in market economies. They decide who to fund and who not to and the terms of funding. This is very crucial to the implementation of investment decisions. This is because individuals and firms need access to credit to enable them carry out their investment decisions. Credit scoring algorithms, which guesses the probability of default, are the methods banks use to determine whether or not a loan should be granted. The goal of this paper is to build a model that banks can use to help make the best financial decisions. Predicting the probability that somebody will experience financial distress in the next two years has many benefits that accrue not only to the lenders but also to the borrowers. One main problem with the credit scoring algorithm is the change of patterns over time. The key assumption for any predictive modeling is that the past can predict the future (Berry & Linoff, 2000). In credit scoring, this means that the characteristics of past applicants who are subsequently classified as "good" or "bad" creditors can be used to predict the credit status of new applicants. Sometimes, the tendency for the distribution of the characteristics to change over time is so fast that it requires constant refreshing of the credit scoring model to stay relevant. Therefore predicting the probability that somebody will experience financial distress based on other financial information is the right way to go since this will help banks and financial institutions make the best financial decisions.

I obtained the data from an ended competition on Kaggle website "[Give me some credit challenge] (http://www.kaggle.com/c/GiveMeSomeCredit)". In this competition, the goal is to predict whether a borrower will experience financial distress in the next two years. Therefore, one could treat this as a classification problem with two classes. For the competition, historical data are provided on 146,076 borrowers. The historical data

contain eleven variables with "*Del*" treated as the response variable and all other variables used as predictors. The data description is below:

- *Del*(Y/N): Person experienced 90 days past due delinquency or worse.
- *Util*(percentage): Total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits.
- *Age* (integer): Age of borrower in years.
- *Del3059* (integer): Number of times borrower has been 30-59 days past due but no worse in the last 2 years.
- *Debt_Ratio* (percentage): Monthly debt payments, alimony, living costs divided by monthly gross income.
- *Income* (real): Monthly income. Transformed by cube root.
- *Credit_Lines* (integer): Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards).
- *Del90* (integer): Number of times borrower has been 90 days or more past due.
- *RealEstate* (integer): Number of mortgage and real estate loans including home equity lines of credit.
- *Del6089* (integer): Number of times borrower has been 60-89 days past due but no worse in the last 2 years.
- *Dep* (integer): Number of dependents in family excluding themselves (spouse, children etc.)

3. Data Exploration

3.1. Missing Values

Income was NA 17.2% of the historical data set provided. As a predictor that I considered potentially useful, I decided to compute the *Income* for those customers using a regression technique based on other customer characteristics. The remaining 82.8% of the historical data was divided in training (2/3) and test data (1/3). The regression techniques considered.

a. Reduced Linear Regression Model

After building a linear regression model with all the variables using the training data, I realized some of the variables were not significant *(Del, Util, Del90)*. I decided to run another regression model using the significant variables and from the F-test below the reduced model is equivalent to the full model. Using the test data to make predictions, $MSE = 50.04386$

Null and Alternative hypothesis:

$H_0: \ \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$

$H_1: \ \beta_j \neq 0$, *for at least one value of* j

```
Analysis of Variance Table

Model 1: Income ~ (Del + Util + age + Del3059 + Debt_Ratio + Credit_Lines +
    Del90 + RealEstate + Del6089 + Dep) - Del - Util - Del90
Model 2: Income ~ Del + Util + age + Del3059 + Debt_Ratio + Credit_Lines +
    Del90 + RealEstate + Del6089 + Dep
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1  80261 826581
2  80258 826556  3    24.375 0.7889 0.4998
```

b. Forward and Backward Stepwise Selection

Forward stepwise selection begins with a model containing no predictors, and then adds predictors to the model, one-at-a-time, until all of the predictors are in the model. In particular, at each step the variable that gives the greatest additional improvement to the fit is added to the model. The Backward Stepwise selection backward stepwise selection begins with the full least squares model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time. The forward and backward stepwise algorithm gave the same model selecting all the 10 predictor variables for the model. $MSE = 50.0186$

c. Ridge Regression

The ridge regression minimizes the quantity below, where $\lambda \geq 0$ is a tuning parameter to be determined seperately and the term $\lambda \sum_{j=1}^{p} \beta_j^2$ is called the shrinkage penalty. The graph below shows that the best lambda to use is $\lambda = 0.6067512$. The $MSE = 51.21711$.

$$\sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

d. Lasso Regression

The Lasso regression minimizes the quantity below where, $\lambda \sum_{j=1}^{p} |\beta_j|$ is the penaltly term. One diffrence between the lasso and ridge regression is that for the lasso model, the penalty has the effect of forcing some of the coefficients estimates to be exactly equal to zero. From the graph below, the best lamda to use is $\lambda = 0.04381235$. The

lasso regression forced four coefficients estimates to be exactly zero *(Del, Util, Del3059, and Del90).* $MSE = 50.75989$.

$$\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right|$$

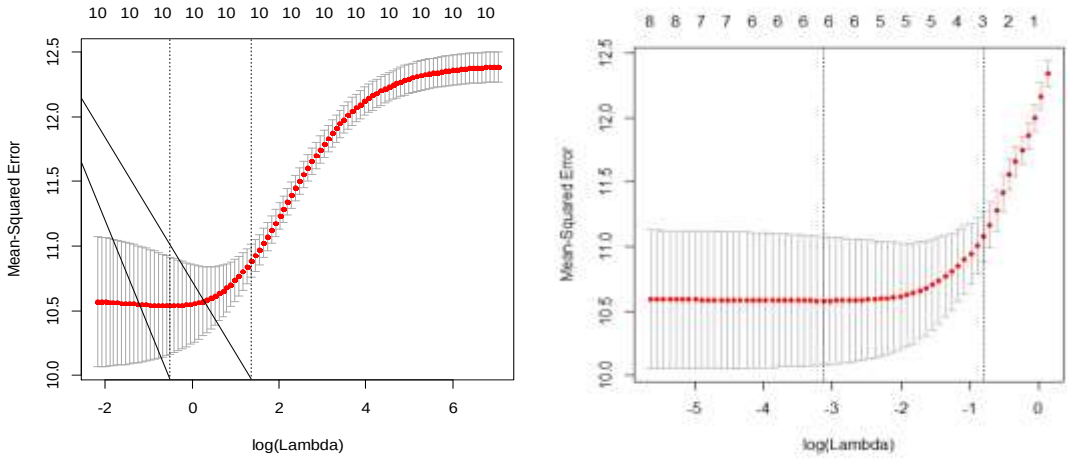


Figure 1: Left: best lambda graph for ridge regression. Right: best lambda graph for lasso regression.

e. Principal Component Regression

Let $Z_1, Z_2, \ldots, Z_M$ represent $M < p$ linear combinations of our original linear combination p predictors. That is, $Z_M = \sum_{j=1}^{p} \phi_{jm} X_j$. The key idea of PCR is that often a small number of principal components suffice to explain most of the variability in the data, as well as the relationship with the response. If the assumption underlying PCR holds, then fitting a least squares model to $Z_1, Z_2, \ldots, Z_M$ will lead to better results than fitting a least squares model to $X_1, X_2, \ldots, X_M$, since most or all of the information in the data that relates to the response is contained in $Z_1, Z_2, \ldots, Z_M$ and by estimating only $M < p$ coefficients we can mitigate over fitting. From the graph below, after the 8th principal component the *MSEP* hits its minimum indicating that the best number of components to use is eight. Using the eight principal components for regression, the $MSE = 50.00511$ on the test data.

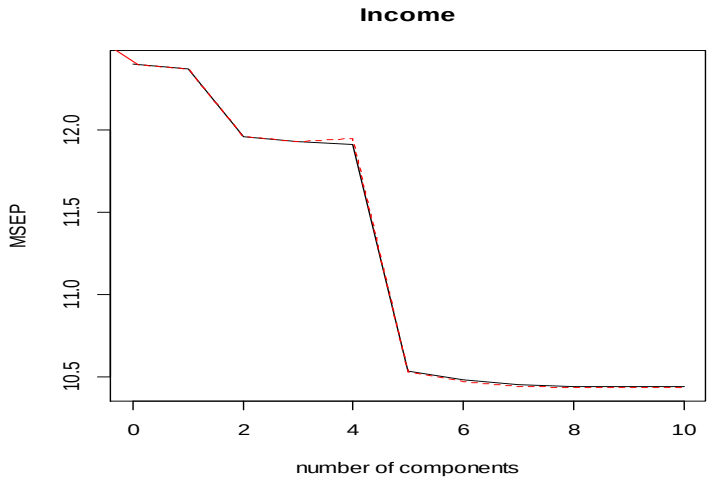


Figure 2: Number of components for PCR

f. Summary

From the table below, the best mean squared error from the test data comes from the principal component regression model. To predict the missing data I used the PCR model.

Models	OLS	Forward/Backward	Ridge	Lasso	PCR
MSE	50.04386	50.0186	51.21711	50.75989	50.00511

3.2. Correlation

The table below shows the correlation values between the predictor variables. Most of the correlation values are not significant with the exception of three, the correlation coefficient between *Del3059, Del90 and Del6089.*

	Util	age	Del3059	Debt_Ratio	Income	Credit_Lines	Del90	RealEstate	Del6089	Dep
Util	1.00	-0.01	0.00	0.00	0.01	-0.01	0.00	0.01	0.00	0.00
age	-0.01	1.00	-0.06	0.03	0.06	0.16	-0.06	0.04	-0.05	-0.21
Del3059	0.00	-0.06	1.00	-0.01	-0.03	-0.05	0.98	-0.03	0.99	0.00
Debt_Rati	0.00	0.03	-0.01	1.00	-0.31	0.05	-0.01	0.12	-0.01	-0.04
Income	0.01	0.06	-0.03	-0.31	1.00	0.20	-0.04	0.21	-0.03	0.21
Credit_Lir	-0.01	0.16	-0.05	0.05	0.20	1.00	-0.08	0.43	-0.07	0.07
Del90	0.00	-0.06	0.98	-0.01	-0.04	-0.08	1.00	-0.04	0.99	-0.01
RealEstate	0.01	0.04	-0.03	0.12	0.21	0.43	-0.04	1.00	-0.04	0.12
Del6089	0.00	-0.05	0.99	-0.01	-0.03	-0.07	0.99	-0.04	1.00	-0.01
Dep	0.00	-0.21	0.00	-0.04	0.21	0.07	-0.01	0.12	-0.01	1.00

Table 1: Correlation between predictor variables.

3.3. Challenges with the data

There are two classes (0- No serious delinquency in two years and 1-Serious delinquency in two years) in the data set. The biggest challenge with the data was that almost 90% of the training data set are 0's which means there is a huge risk that the models are more likely to predict 0's for 1's. To address this problem, I used just 29,542 of the historical data which contains all the 1's. I then randomly selected 19,694 as the training data and 9,847 as the test data.

4. Classification with original data.

Classification is a multivariate technique concerned with separating distinct sets of objects (or observations) and with allocating new objects (observations) to previously defined groups. The immediate goal of classification is to sort objects (observations) into two or more labeled classes. The emphasis is on deriving a rule that can be used to optimally assign new objects to the labeled classes. (Johnson & Wichern, 2014). I used five classification techniques and evaluates which give the best prediction results on the test data.

4.1. Linear Discriminant Analysis (LDA)

The Linear Discriminant Analysis models the distribution of the predictors *X* separately in each of the response classes (i.e. given *Y)*, and then use Bayes' theorem to flip these around into estimates for $\Pr(Y = k | X = x)$. LDA is based on the assumption that the data follows a normal population and the covariance matrices of the two groups are equal. In general, for the LDA assigns an observation $X = x$ to the class to which $\delta_k(x)$ from below is the largest.

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

I used the training data in building the linear discriminant model. Prediction results from the training and test data and defining positive as "Serious delinquency in two years" are summarized below. The accuracy rate and true positive rate increase for the test data which is good.

Data	True Pos Rate	False Pos Rate	True Neg Rate	False Neg Rate	Accuracy Rate	Error Rate
Training	0.76	0.01	0.99	0.24	0.91	0.09
Test	0.83	0.01	0.99	0.17	0.94	0.06

4.2. Quadratic Discriminant Analysis (QDA)

Like LDA, the QDA classifier results from assuming that the observations from each class are drawn from a Gaussian distribution, and plugging estimates for the parameters into Bayes' theorem in order to perform prediction. However, unlike LDA, QDA assumes that each class has its own covariance matrix. Under this assumption, QDA assigns an observation $X = x$ to the class to which $\delta_k(x)$ from below is the largest.

$$\delta_k(x) = -\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\Sigma_k| + \log \pi_k$$

I used the training data in building the quadratic discriminant model. Prediction results from the training and test data and defining positive as "Serious delinquency in two years", are summarized below. The total positive rate for the QDA is low and the error rate is high, which means the QDA is worse than the LDA.

Data	True Pos Rate	False Pos Rate	True Neg Rate	False Neg Rate	Accuracy Rate	Error Rate
Training	0.44	0.01	0.99	0.56	0.81	0.19
Test	0.42	0.00	1.00	0.58	0.80	0.20

4.3. Logistic Regression

In its simplest setting, the response variable in the logistic regression is restricted to two values. Let the response variable be 1 if the observational unit belongs to population 1and 0 if it belongs to population 2. Assign **z** to population 1 if the estimated odds ratio is greater than 1 or

$$\frac{\hat{p}(z)}{1-\hat{p}(z)} = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 z_1 + \ldots + \hat{\beta}_r z_r\right) > 1$$

I used the training data in building the logistic regression model. Prediction results from the training and test data and defining positive as "Serious delinquency in two years" are summarized below. The results from the logistic regression is better than the LDA.

Data	True Pos Rate	False Pos Rate	True Neg Rate	False Neg Rate	Accuracy Rate	Error Rate
Training	0.88	0.01	0.99	0.12	0.95	0.05
Test	0.89	0.00	1.00	0.11	0.96	0.04

4.4. Random Forest

Random forests builds a number forest of decision trees on bootstrapped training samples. But when building these decision trees, each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set of *p* predictors. The split is allowed to use only one of those *m* predictors. A fresh sample of *m* predictors is taken at each split, and typically we choose $m = \sqrt{p}$. From the decision tree and variable importance plot below, *Income* is the most important variable in predicting probability of serious delinquency and when income is less than 21.37 the customer will be delinquent in 2 years.

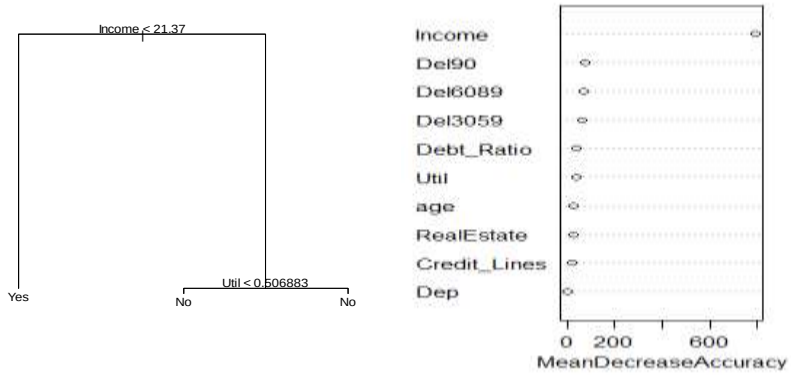


Figure 3: decision tree(left) and variable importance plot (right)

I used the training data in building the random forest model with $m = 5$. Prediction results from the training and test data and defining positive as "Serious delinquency in two years" are summarized below. The random forest has the best result in terms of true positive rates and overall accuracy rate.

Data	True Pos Rate	False Pos Rate	True Neg Rate	False Neg Rate	Accuracy Rate	Error Rate
Training	0.91	0.01	0.99	0.09	0.97	0.03
Test	0.93	0.00	1.00	0.07	0.97	0.03

4.5. Sector Vector Machine.

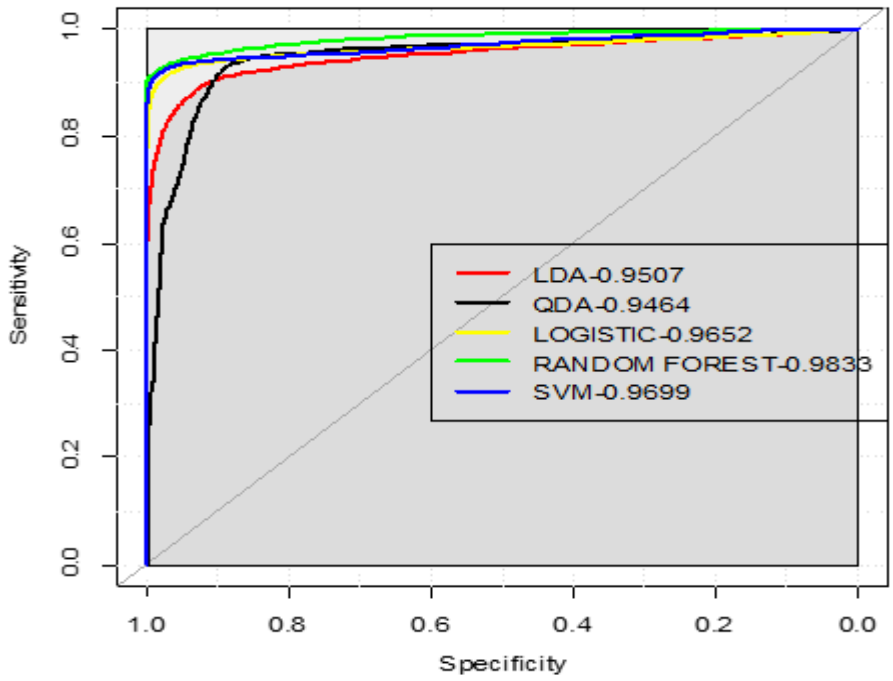
A Support Vector Machine provides a binary classification mechanism based on finding a hyper plane between a set of samples with +ve and -ve outputs. It assumes the data is linearly separable. If the data is not linearly separable due to noise (the majority is still linearly separable), then an error term will be added to penalize the optimization. If the data distribution is fundamentally non-linear, the trick is to transform the data to a higher dimension so the data will be linearly separable or

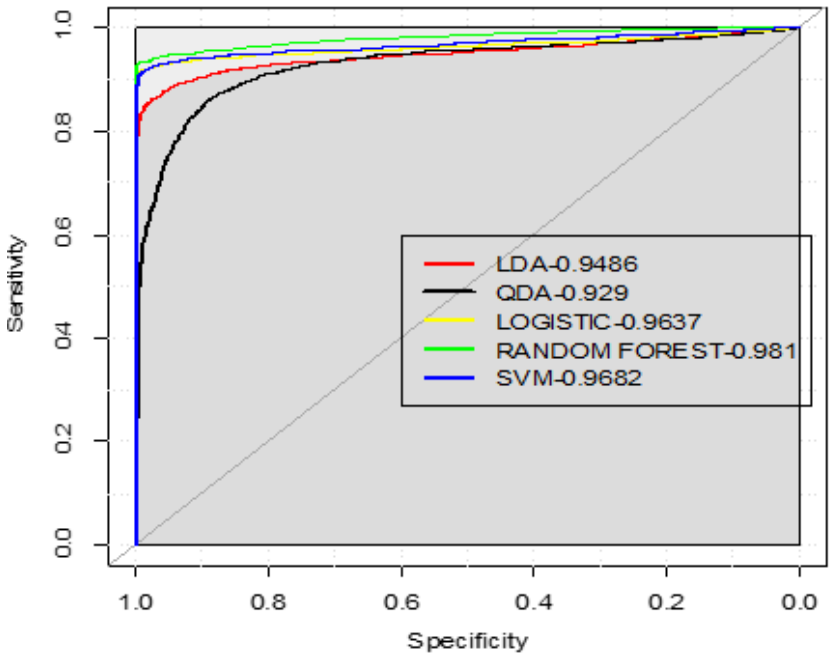
performing a kernel function in the original. SVM predicts the output based on the distance to the dividing hyper plane. This doesn't directly estimate the probability of the prediction. We therefore use the calibration technique to find a logistic regression model between the distance of the hyper plane and the binary output. Using that regression model, we then get our estimation.

The prediction results from the training and test data and defining positive as "Serious delinquency in two years" are summarized below. The results from the SVM is better than the logistic model but worse than the random forest model.

Data	True Pos Rate	False Pos Rate	True Neg Rate	False Neg Rate	Accuracy Rate	Error Rate
Training	0.89	0.00	1.00	0.11	0.96	0.04
Test	0.90	0.00	1.00	0.10	0.96	0.04

4.6. Summary



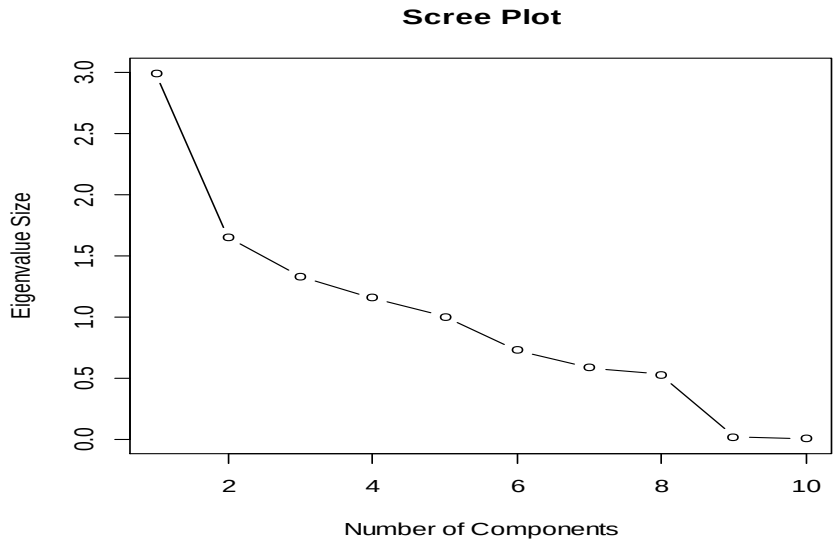


Graph 5: ROC on train data (top) and ROC on test data (bottom)

5. Principal Components

A principal component analysis is concerned with explaining the variance–covariance structure of a set of variables through a few linear combinations of these variables. Its general objectives are data reduction and interpretation. Although p components are required to reproduce the total system variability, often much of this variability can be accounted for by a small number k of the principal components. If so, there is almost as much information in the k components as there is in the original p variables. The k principal components can then replace the initial p variables, and the original data set, consisting of n measurements on p variables, is reduced to a data set consisting of n measurements on k principal components (Johnson & Wichern, 2014).

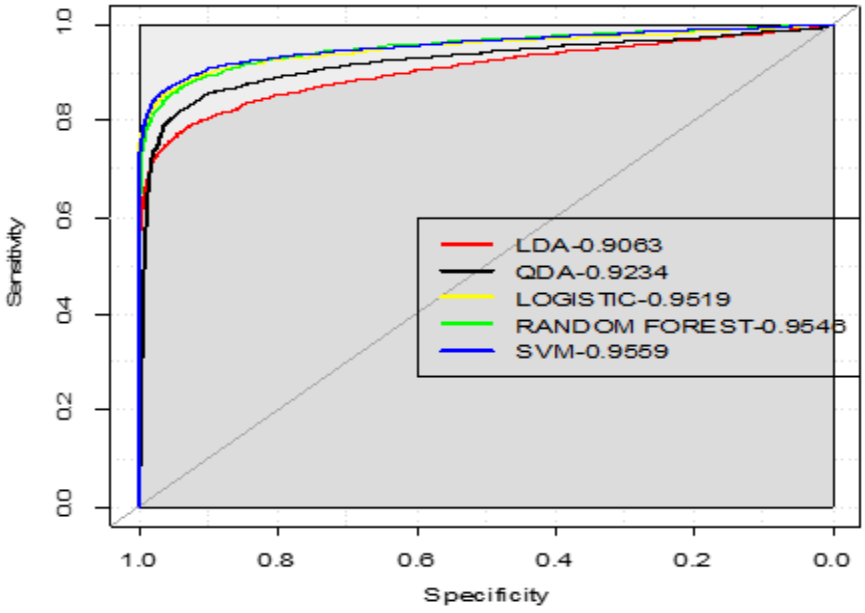
The scree plot below is that of the scaled data since the original values are very different. From the scree plot and the table that contains the importance of principal components below, after the fifth principal component almost 80% of the variability in the data can be explained. Hence, I build the classification models using the 5 principal components.



Graph 6: Scree plot showing the 5 principal components can replace the 10 variables.

5.1. Summary: Test data

Using the 5 principal component scores from the principal component analysis, the results from the models are summarized below. The SVM model has the highest area under curve value of 0.9559 and the Random Forest model result was worse than that on the original model.



Graph 7: ROC on test data using principal components

6. Conclusion

Motivated by the shortfall of the credit scoring algorithm, this paper focused on improving the state of the art in credit scoring, by predicting the probability that somebody will experience financial distress in the next two year. I obtained the data from an ended competition on Kaggle website and classification techniques, namely, the linear discriminant analysis, the quadratic discriminant analysis, logistic regression, SVM and the Random Forest were used to build models. The principal components was used as a data reduction technique which resulted in the lowest MSE in the regression model when predicting missing values for "*Income*" but did not improve the results in the classification. The results indicates that the Random Forest was the best predictive model and should be used in conjunction with the credit scoring algorithms to calculate the probability of default.

References

Berry, M., & Linoff, G. (2000). *Mastering Data Mining: The Art and Science of Customer Relationship Management.* New York: John Wiley & Sons Inc.

Fraley, C., & Raftery, A. E. (2002). Model-Based Clustering, Discrimination Analysis and Density Estimation. *Journal of the American Statistical Association*, 458.

Johnson, R., & Wichern, D. (2014). *Applied Multivariate Statistical Analysis.*

Kaggle . (2014, October 24). Retrieved from Kaggle web site: http://www.kaggle.com