Statistical Learning and Bankruptcy Prediction

MA678 Midterm Project

Zixuan Liu (zliu203)

1. Abstract

The dataset is about bankruptcy prediction of Polish companies. The data was collected from Emerging Markets Information Service, which is a database containing information on emerging markets around the world. The bankrupt companies were analyzed in the period 2000-2012, while the still operating companies were evaluated from 2007 to 2013.

2. Introduction

2.1 Question

Prediction of firm bankruptcies have been extensively studied in the field of accounting to monitor the financial performance by all shareholders. Especially I want to start my own business after graduate. I want to find out the change that I will success to build my own company. The aim of this project is to examine the relationships between these parameters and develop an effective multilevel model to assess how these measurement correlated with the firm bankruptcies.

2.2 Data Source and Description

The dataset I use is called `Polish Companies Bankruptcy Data Set` which is hosted by UCI Machine Learning Repository and collected from EMIS, a database containing information on emerging markets around the world. The bankrupt companies were analyzed in the period 2000-2012, while the still operating companies were evaluated from 2007 to 2013. In this project, I will use partial data called `3year` for bankruptcy prediction. It contains financial rates from 3rd year of the forecasting period and corresponding class label that indicates bankruptcy status after 3 years. The data `3year` contain contains 64 variables and 10503 observations in total. The dependent variable is the class variables with levels 0 or 1, indicating the company bankruptcy or not. Some variables as financial ratio could affect the company be classified as bankruptcy or not. For example, the first variable is “net profit/total assets” which is return on assets (ROA), a financial ratio that shows the percentage of profit a company earns in relation to its overall resources. It is possible that the higher the ROA, the less likely the company will be bankrupt.

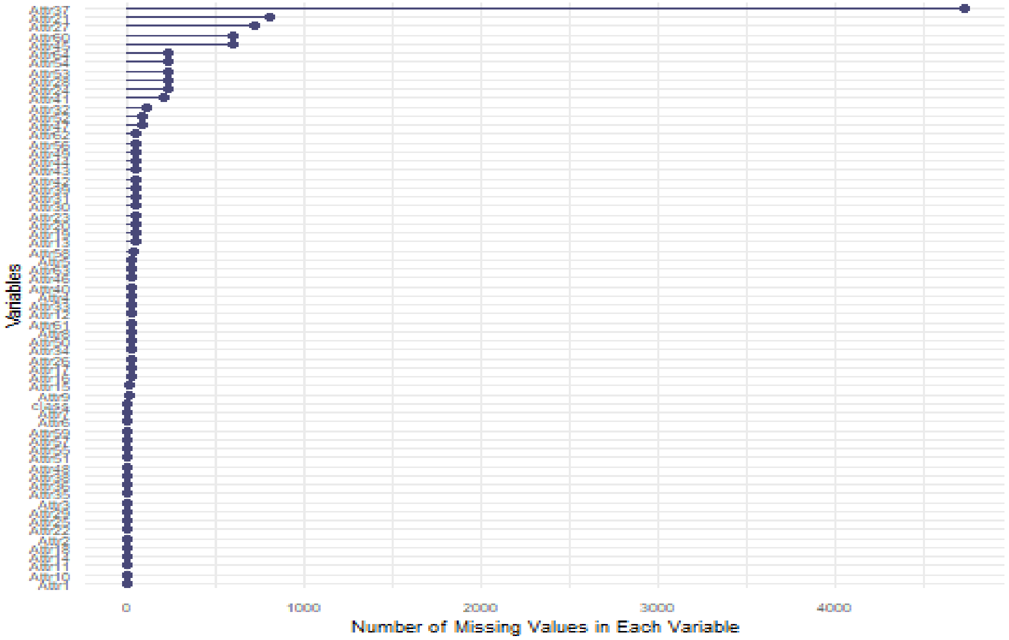
3. Methods

3.1 Missing Values & Data Preprocessing

3.1.1 Missing Values

First I conduct basic data preprocessing. Missing values for dataset are shown in the histogram below.

The plot above shows that attr 37 has the highest missing value. Due to the large number of missing values in each dataset, completely delete missing values will result to a large amount of data loss. Thus, I use variable means to replace missing values. I also drop the first variable `id` and factorize variable `class`.



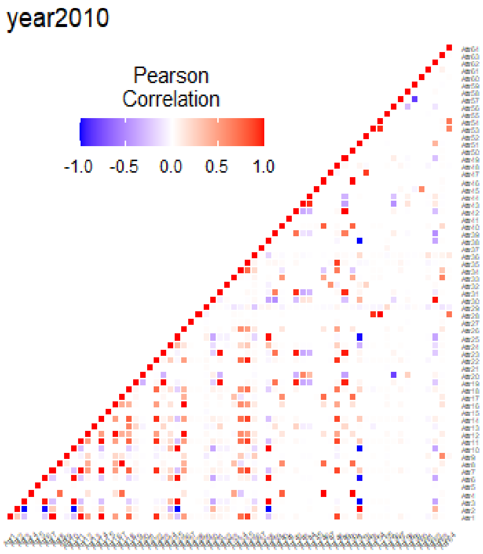
3.1.2 Heatmap

Shown in below is a correlation map for the year 2010 data that decribes the relationship between the different features.

I can find that attr 2 (total liabilities/ total assets) and attr 10 (Equity/ total assets) have highlist negative correlation. Then I noticed that

$$Assets = Liabilities + Owner'sEquity$$

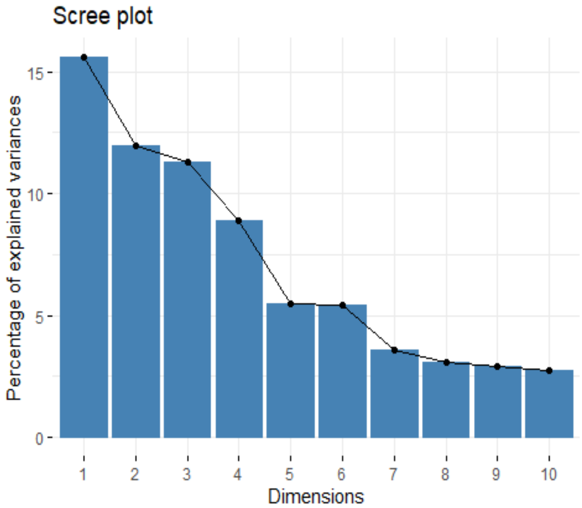
which can explain correlation that if total assets are keep same, when liabilities are increased, equity must be decreased.



3.2 Fitting model

3.2.1 PCA

Principal Component Analysis (PCA) is a useful tool for exploratory data analysis which is a dimensionality reduction or data compression method and can select a subset of variables from a larger set, based on the highest correlations variables. I want to use PCA to find a direction that displays the largest variance from the variables.

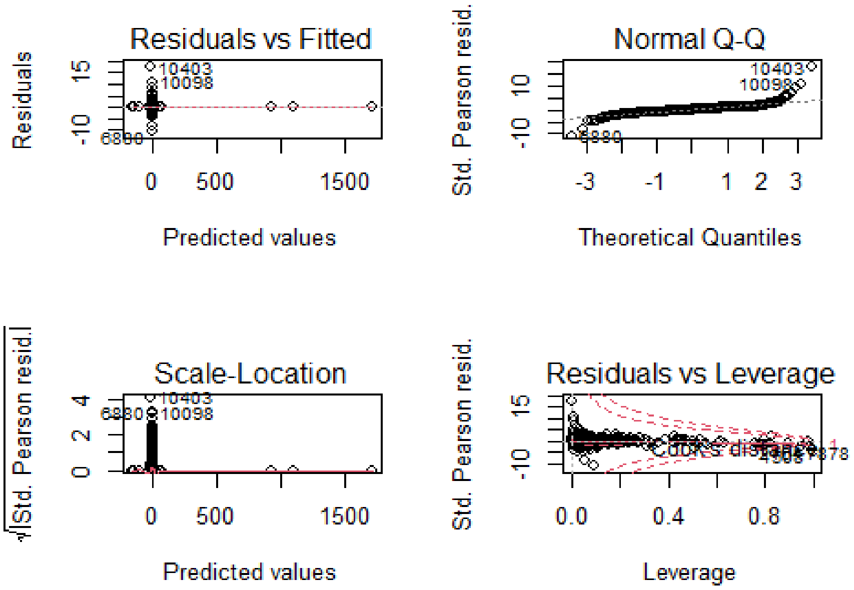


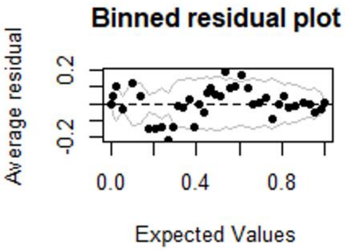
As the plot shown above, although the first three principal components are much more influential than others, PC1 only explains around 25% of the variance which is not as higher as normal dataset. As I know that anything above 30% is a good loading, My data show a lowly correlated and I cannot find The largest variance by using PCA.

3.2.2 Residuals

In my models, ***Class*** is the response variable, 0 means the firm did not bankrupt; 1 means that the firm bankrupt. Then I took logistic mixed-effect model for this data. The model can be write as: fit1 = glm(class~., data = train_1_resample,family = “binomial”)

To check the fitted models, plot the residuals for my model.





4. Results

I want to look at my Diagnostic Plots individually. I don't think the residuals vs fitted plot show that my model is good, because that there are no obvious pattern in this plot. most of the points are around 0 and there are few that's are over 800.

By looking at the normal QQ plot, We can see this plot shows that the residuals are normally distributed, since the residuals are lined well on the straight dashed line.

The scale-location plot shows the residuals are not spread equally alone the ranges of predictors. The red line is not horizontal and it's not randomly spread points.

In residuals vs leverage plot, this plot is far beyond the Cook's distance lines (the other residuals appear clustered on the left because the second plot is scaled to show larger area than the first plot).

Based on the above plots, I don't think my model is appropriate to fit the data since the average residuals does not have an regular pattern. This might due to I averaged out the missing value, or might because that I delete one variable since it has a huge amount of missing value. If I got more time, It will be interesting to figure it out why the model does not seem so good, and I want to apply some prediction for the data to test out whether the firm will bankrupt or not.

In terms of classification method. I included Principal Component Analysis (PCA) for classification. I used PCA to find the most influential components and use them to lower dimensions and hence improve model accuracy. Besides, in terms of tuning parameters, since I only consider parameters in the range of 1 to 10, I also expanded the grid and investigate more on variance-bias trade off issue.

5. Discussion

5.1 Obstacle

The biggest obstacle in the project is cleaning missing values. As listed in the previous section, the dataset include a great number of missing values. Thus, how to deal with missing value is the major task in data reprocessing. I believe there would be substantial loss in data if I simply delete all missing values. In this project, I finally used mean approach to substitute each missing value with the mean of the corresponding variable. There are still other approaches which can be applied to resolve missing values, such as K Nearest Neighbor.

5.2 Future work

In terms of classification methods, there are a lot of other methods can be considered and investigated. For example, I can include Principle Component Analysis (PCA) for classification. I can use PCA to find the most influential components and use them to lower dimensions and hence improve model accuracy. Besides, in terms of tuning parameters, since I only consider parameters in the range of 1 to 10, I can also expand the grid and investigate more on variance-bias trade issue in the future. Meanwhile, considering missing values, as I stated before, there are also other approaches can be applied including K Nearest Neighbors.

5.3 Reference

[1] Sudheer Chava and Robert A. Jarrow, Bankruptcy Prediction with Industry Effects, Review of Finance 8: 537-569, 2004, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.495.4409&rep=rep1&type=pdf

[2] Sunarira Ashraf and Elisabete, Do Traditional Financial Distress Prediction Models Predict the Early Warning Signs of Financial Distress?, Volume 11, Issue 5, June 1994, Pages 545-557, https://doi.org/10.1016/0167-9236(94)90024-8

[3] Sai Surya Teja Maddikonda and Sree Keerthi Matta, Bankruptcy Prediction: Mining the Polish Bankruptcy Data, https://github.com/smaddikonda/Bankruptcy-Prediction/blob/master/Bankruptcy% 20Prediction%20Report.pdf

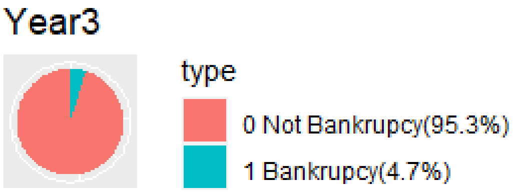
[4] Polish Companies Bankruptcy Data Set, UCI Machine Learning Repository,http://archive.ics.uci.edu/ ml/datasets/Polish+companies+bankruptcy+data

6. Appendix

6.1 Imbalance Data

6.1.1 Pie Chart

I did a pie chart to show the imbalance in response variable



The pie charts above show that the data is imbalanced. It has `0` with above 95.3%. The I used the SMOTE method to oversimple the minority group and achieve a more balanced dataset.

6.1.2 SMOTE Algorithm For Unbalanced Classification

Synthetic Minority Oversampling Technique (SMOTE) is a widely used oversampling technique. I used SMOTE algorithm to creates artificial data based on feature space similarities from minority samples and achieve a more balanced dataset.First I took the difference between the feature vector and its nearest neighbor, than Multiplied this difference by a random number between 0 and 1, and added it to the feature vector under consideration. Finally I chose a random point along the line segment between two specific features to get the new balanced data set: 5445 Bankrupt instances and 7425 Non-bankrupt instances.

Bankruptcy	Frequency
0	10008
1	495

Bankruptcy	Frequency
0	7425
1	5445

6.2 Split data

To build classification models, I splited dataset into training and testing dataset by split ratio = 0.8. I first use training data to fit various classification models, and then use testing data to make predictions and calculate model accuracy. I take Year 2010 dataset as an example.

6.3 Ridge Regression and Cross Validation

In Ridge and Lasso regressions, I will used loss function:

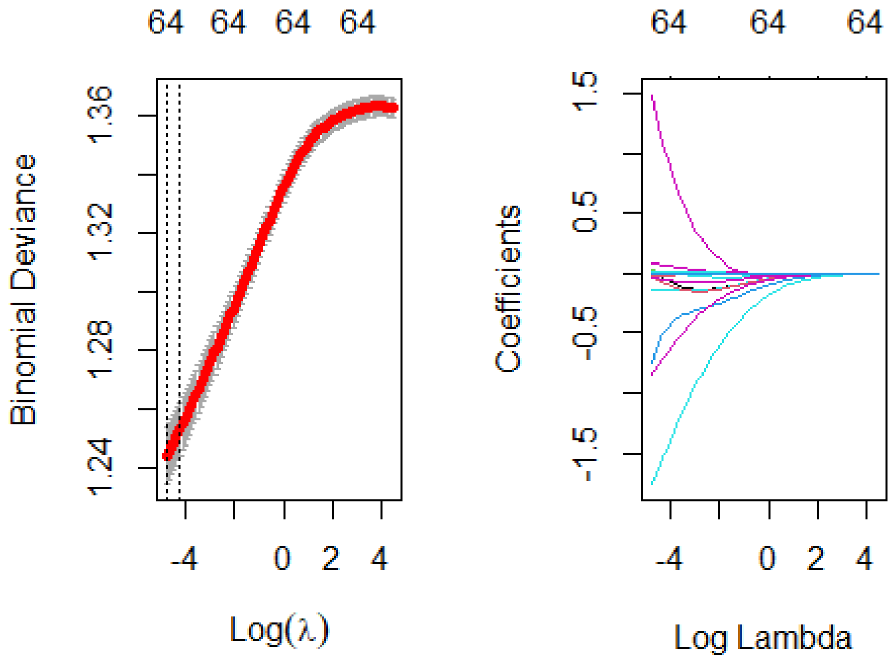
$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j}^{p} \beta_j x_{ij})^2 + (1-\alpha)\lambda \sum_{j}^{p} \beta_j^2 + \alpha\lambda \sum_{j}^{p} |\beta_j|$$

I will use built-in cross-validation function: cv.glmnet() to choose the turning λ. Since the choice of the cross-validation folds is random, I set a random seed at first. And in 'glmnet' function, I will use

$$alpha = 0$$

to determines the regression model:

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



```
## [1] 0.008607136
```

The left plot shows the training mean squared error as the function of λ. The second plot shows the coefficients for different values of λ and it shows that when λ becomes larger, the coefficients are tend towards 0 . I also see that the value of λ which results in the smallest cross-validation error is 0.009847234 I also create a class predictions table based on the predicted probability of bankruptcy. If probability is greater than 0.5, it will show 1, otherwise will show 0.

	Actual 0	Actual 1
Predicted 0	1293	751
Predicted 1	192	338

The calculated error rate is 36.6355866%.

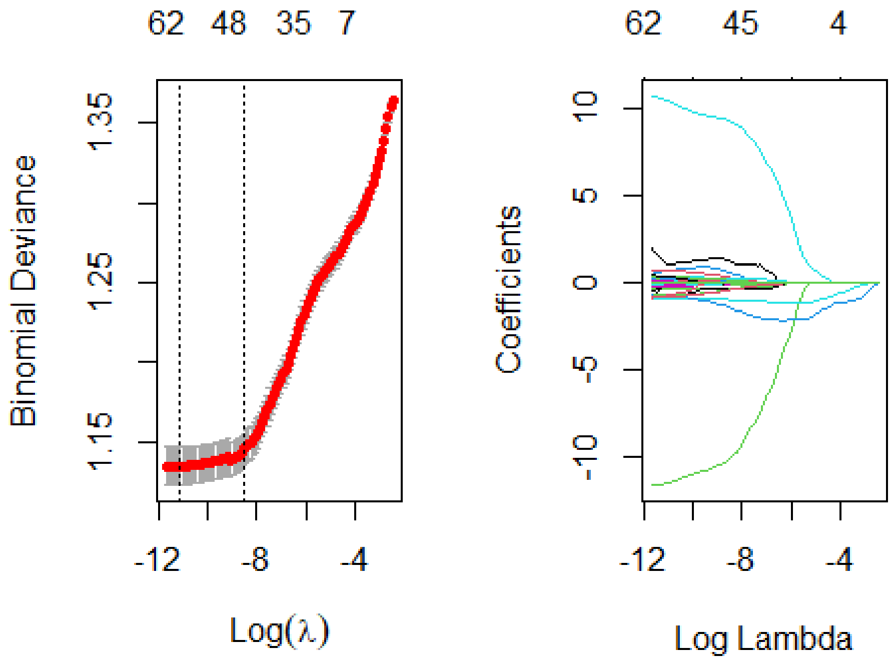
6.4 Lasso Regression and Cross Validation

In lasso regression, I want to test if the lasso can yield either a more accurate or a more interpretable model than ridge regression. I will also use the loss function and cv.glmnet to fit model, and in this time, I will use the argument

$$alpha = 1$$

. I will fit the lasso regression model:

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



```
## [1] 1.504122e-05
```

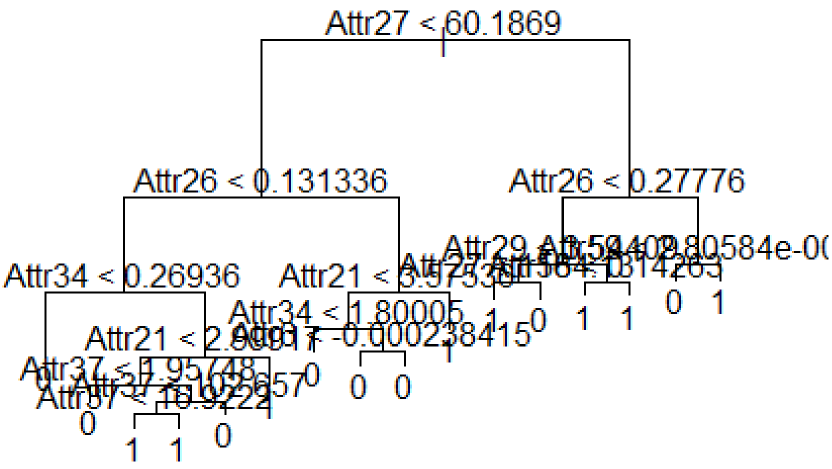
These two plots show how λ changes the mean squared error and the coefficients for different ??. And I can see that 2.274839e-05 is the smallest cross-validation error for ??.

	Actual 0	Actual 1
Predicted 0	1171	434
Predicted 1	314	655

The calculated error rate is 29.0598291%.

6.5 Decision Tree

Decision tree is a non-parametric supervised learning method that recursively partition the feature space into hyper-rectangular subsets, and make prediction on each subset. I created a decision tree model to predict the response `class`: whether the company goes bankrupt or not, using all 64 attributes in the Polish Bankruptcy dataset. The model learns simple decision rules inferred from the 64 attributes. My decision tree model gives equal weights to all 64 attributes and use Gini indices as measure of quality of the splits. The model uses "Attr27" "Attr13" "Attr34" "Attr21" "Attr58" "Attr39" "Attr6" "Attr26" "Attr29" "Attr59" in the actual model built with `Attr27` as the root node, which features profit on operating activities / financial expenses of each company. The daughter nodes of the root node, `Attr21` and `Attr26`, represents sales in current year / sales in the previous year, and (net profit + depreciation) / total liabilities, respectively.



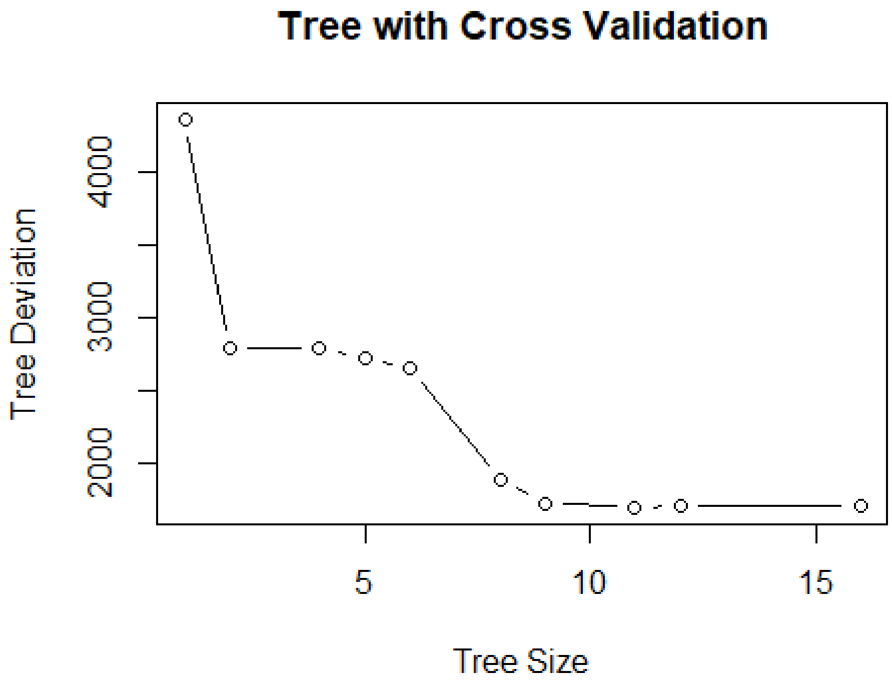
Before pruning, the confusion matrix of my tree model is

	Actual 0	Actual 1
Predicted 0	1287	226
Predicted 1	198	863

The error rate is:

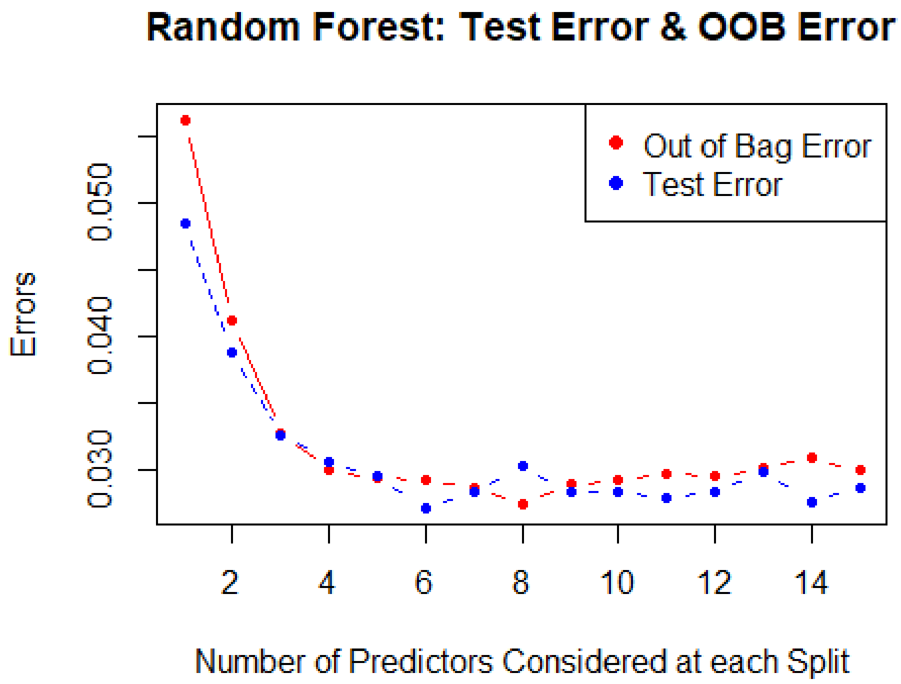
```
## [1] 0.1647242
```

I then consider using cross-validation to prune the tree, and the optimal tree size stays at 14, which gives us the same model I had.



6.6 Random Forests

A random forest is a mega estimator that fits a number of decision tree classifiers. Trees are built with randomly selected subsets of features and the best split within the chosen subset is used for these trees. The randomness in this tree-building method yields larger bias and smaller variance due to averaging, and the decrease in variance would overcompensate the increase in bias. I consider different mtry, number of predictors considered at each split, from 1 to 15, and plot the errors. I find the testing error is minimized when mtry=13.



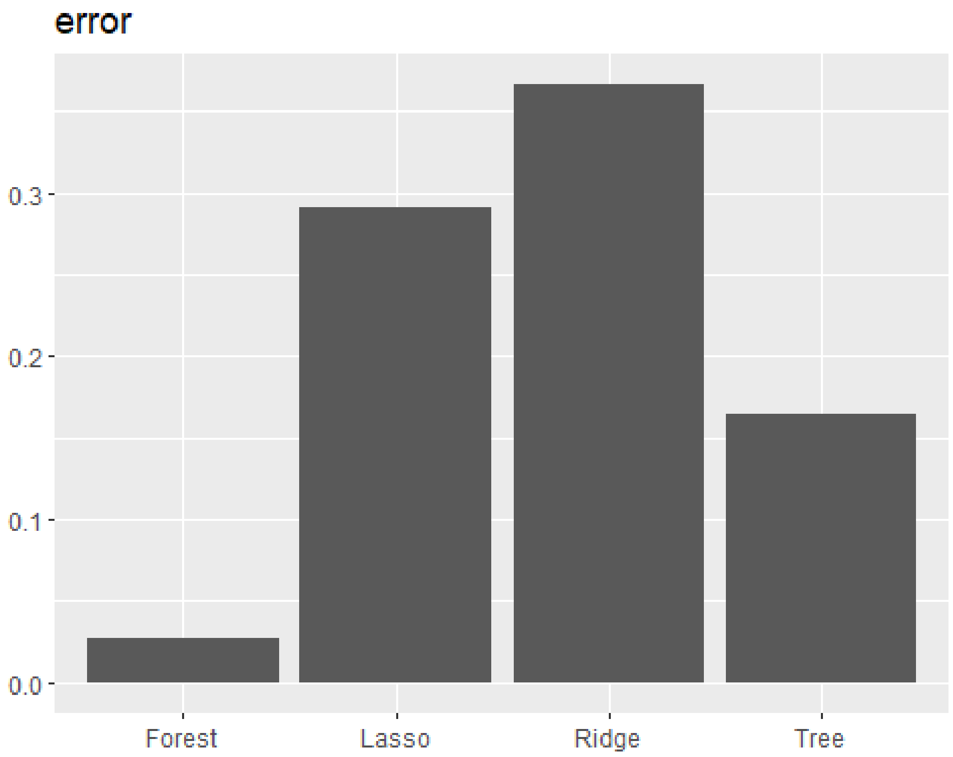
The confusion matrix of my random forest model is as following:

	Actual 0	Actual 1
Predicted 0	1457	42
Predicted 1	28	1047

```
## [1] 0.02719503
```

And the error rate is 2.7195027%. Not surprisingly, my random forest model outperforms my decision tree model. In fact, this is the most accurate model among all methods I used.

The histogram report the error rates of the 6 models that I have wxperimented with. then I get the result: Random Forests has the smallest error rate which is the best bankruptcy model in my project. In Random Forest model, top 3 important features are Attr27 (profit on operating activities / financial expenses), Attr21 (sales in this year / sales in last year) and attr24 (gross profit (in 3 years) / total assets). I can make the conclusion that activitis' profit, the quantity of sales and gross profit are the decisive factors for the company bankruptcy



From the results shown above, I can conclude that the best classification method is Random Forest.