Statistical Analysis of Bankruptcy based on Financial Indicators

Adrija Saha Sampurna Mondal
Shrayan Roy

Abstract

Bankruptcy prediction is the art of predicting bankruptcy and various measures of financial distress of public firms. It is a vast area of finance and accounting research. The importance of the area is due in part to the relevance for creditors and investors in evaluating the likelihood that a firm may go bankrupt. This report analyzes a dataset of four financial indicators to find a discriminant rule that predicts the likelihood of a firm going bankrupt in the next two years. The analysis also involves identifying variables that correspond to the same common factor and exploring any other relationships that may exist.

1 Introduction

Bankruptcy is a phenomenon that has severe financial repercussions for both companies and individuals. As such, predicting the likelihood of bankruptcy is a critical area of research for both finance and accounting professionals. By analyzing various financial indicators, researchers can develop models that can help predict whether a company is at risk of going bankrupt.

Statistical analysis of bankruptcy data involves the use of various financial ratios and indicators that are known to be highly correlated with financial distress. These ratios and indicators are typically categorized into liquidity, solvency, profitability, and efficiency ratios. By analyzing these ratios, researchers can identify patterns and trends that indicate financial distress and use this information to predict the likelihood of bankruptcy.

Furthermore, the statistical analysis of bankruptcy data can help researchers identify the key factors that contribute to financial distress. By identifying these factors, companies can take proactive measures to prevent financial distress, and investors can make informed decisions about which companies to invest in.

Overall, statistical analysis of bankruptcy data is a crucial tool in the fields of finance and accounting. It can help identify patterns and trends that indicate financial distress, predict the likelihood of bankruptcy, and help companies and investors make informed decisions to prevent financial distress. Keeping this in mind, here we will be

analysing a dataset where we have various financial indicators and will try to find a discriminant rule based on which we can predict how likely is the firm to go bankrupt in the next two years, also we will try to analyze which variables among those correspond to same common factor and we will try to explore out some other relationships, if exists.

NOTE: We have used *non-bankrupt* and *financially sound firm* interchangeably.

Some Short Notes on topics used in this project (Not covered in Class)

1. **Shapiro Wilk Test**:

   The test is based on checking whether a sample $X_1, X_2, \cdots, X_n$ is drawn from a normal population. Setup is like, suppose there are n sample points $X_1, X_2, \cdots, X_n$. To test,

   $$\mathcal{H}_0 : \text{Samples are from normal distribution}$$

   $$vs$$

   $$\mathcal{H}_1 : \text{Samples are not from normal distribution}$$

   The test statistic for this is given by,

   $$W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

   where, $\mathbf{a} = \frac{\mathbf{m}^T\mathbf{V}}{\mathbf{m}^T\mathbf{V}^{-1}\mathbf{V}^{-1}\mathbf{m}}$, with $\mathbf{m}$ and $\mathbf{V}$ are the mean vector and variance-covariance matrix of $(Z_(1), Z_(2), \cdots, Z_(n))^T$ when, $Z_1, Z_2, \cdots, Z_n \sim \mathcal{N}(0, 1)$. And we reject $\mathcal{H}_0$ at level α in favor of $\mathcal{H}_1$ for large values of W.

2. **Royston's H Test**:

   Royston's test uses the *Shapiro-Wilk/Shapiro-Francia* test statistic to test multivariate normality. If kurtosis of the data is greater than 3, then it uses the Shapiro-Francia test for leptokurtic distributions, otherwise it uses the Shapiro-Wilk test for platykurtic distributions.

   Let W_j be the *Shapiro-Wilk/Shapiro-Francia* test statistic for the j_{th} variable $(j = 1, 2, \cdots, p)$ and Z_j be the values obtained from the normality transformation proposed by *Royston*.

   $$\begin{aligned} \text{if } 4 \le n \le 11; \quad & x = n \quad && \text{and} \quad w_j = -\log\left[\gamma - \log(1 - W_j)\right] \\ \text{if } 12 \le n \le 2000; \quad & x = \log(n) && \text{and} \quad w_j = \log(1 - W_j) \end{aligned}$$

   Hence we can get

   $$Z_j = \frac{w_j - \mu}{\sigma}$$

where γ, μ and σ are derived from the polynomial approximations given below.

$$\begin{aligned}\gamma =& a_{0\gamma} + a_{1\gamma}x + a_{2\gamma}x^2 + \cdots + a_{d\gamma}x^d \\ \mu =& a_{0\mu} + a_{1\mu}x + a_{2\mu}x^2 + \cdots + a_{d\mu}x^d \\ \log(\sigma) =& a_{0\sigma} + a_{1\sigma}x + a_{2\sigma}x^2 + \cdots + a_{d\sigma}x^d\end{aligned}$$

The *Royston's H test statistic* for multivariate normality is as follows:

$$H = \frac{e\sum_{j=1}^{p}\psi_j}{p} \sim \chi_e^2$$

where e is the equivalent degrees of freedom and $\Phi(.)$ is the cumulative distribution function of the standard normal distribution such that,

$$\begin{aligned}e =& p/[1 + (p-1)\bar{c}] \\ \psi_j =& \left\{\Phi^{-1}[\Phi(-Z_j)/2]\right\}^2, \quad j = 1, 2, \cdots, p\end{aligned}$$

when

$$\bar{c} = \sum_i \sum_j \frac{c_{ij}}{p(1-p)}; \quad i \neq j$$

where

$$c_{ij} = \begin{cases} g(r_{ij}, n) & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$$

with r_{ij} be the correlation between i^{th} and j^{th} variables and the boundaries of $g(.)$ as $g(0, n) = 0$ and $g(1, n) = 1$. The function $g(.)$ is defined as:

$$g(r, n) = r^{\lambda}\left[1 - \frac{\mu}{v}(1-r)^{\mu}\right]$$

The unknown parameters μ, λ and v were estimated from a simulation study conducted by Ross. He found $\mu = 0.715$ and $\lambda = 5$ for sample size $10 \leq n \leq 2000$ and v is a cubic function which can be obtained as follows:

$$v(n) = 0.21364 + 0.015124x^2 - 0.0018034x^3$$

where, $x = \log(n)$. [Ref: Korkmaz, Goksuluk, and Zararsiz]

3. **Adjusted Chi-Square Q-Q Plot**:

   Adjusted Chi-Square Q-Q Plot is considered here to detect multivariate outliers. Multivariate outliers are the common reason for violating MVN assumption. In other words, MVN assumption requires the absence of multivariate outliers. Thus, it is crucial to check whether the data have multivariate outliers, before

starting multivariate analysis. The MVN includes two multivariate outlier detection methods which are based on robust Mahalanobis distances ($rMD(x)$). Mahalanobis distance is a metric which calculates how far each observation is to the center of joint distribution, which can be thought of as the centroid in multivariate space. Robust distances are estimated from minimum covariance determinant estimators rather than the sample covariance.

Adjusted Mahalanobis Distance: [Ref: Korkmaz, Goksuluk, and Zararsiz]

(a) Compute robust Mahalanobis distances ($rMD(x_i)$)

(b) Compute the 97.5 percent adjusted quantile (AQ) of the chi-Square distribution

(c) Declare $rMD(x_i) > AQ$ as possible outlier.

4. **Pillai's trace in MANOVA**: One of the test statistics that is produced by a MANOVA is Pillai's trace. It is a value that ranges from 0 to 1.The closer Pillai's trace is to 1, the stronger the evidence that the explanatory variable has a statistically significant effect on the values of the response variables. Pillai's trace, often denoted V, is calculated as:

$$V = trace(H(H + E)^{-}1)$$

where, H is the sum of squares and cross products matrix under H_0 and E The error sum of squares and cross products matrix.

2 Data Description

The dataset considered here, is an Annual financial data of financially sound firms and the firms which went bankrupt after two years. It contains 46 observations and 5 columns, where last column is the categorical response variable − 0, if the firm went *bankrupt* (21 data points) and 1, if the firm remains *financially sound* (25 data points). The rest of the four columns are predictor variables (continuous).

The predictor variables provided in this dataset are as follows (The terms used, in behalf of these variables, further in this project, are provided in bracket after each variable names):

- **Ratios of cash flow to total debt (CFTD)**
- **Ratios of net income to total assets (NITA)**
- **Ratios of current assets to total liabilities (CATL)**
- **Ratios of current assets to net sales (CANS)**

These are all financial terminologies, which are described briefly below, to increase the understanding of the data, and to assess on how they may affect the financial position of a firm.

1. Cash Flow: The term cash flow refers to the net amount of cash and cash equivalents being transferred in and out of a company. Cash received represents inflows, while money spent represents outflows.

2. Total Debt: Total debt includes long-term liabilities, such as mortgages and other loans that do not mature for several years, as well as short-term obligations, including loan payments, credit cards, and accounts payable balances.

3. Net Income: Net income, also called net earnings, is calculated as sales minus cost of goods sold, selling, general and administrative expenses, operating expenses, depreciation, interest, taxes, and other expenses.

4. Total Assets: Total assets are the representation of the worth of everything company owns, which can you calculate by adding its owner's equity to its liabilities. Equity is how much the company is worth, or its capital, and liabilities are what it owes.

5. Current Assets: The Current Assets account for all company-owned assets that can be converted to cash within one year. Current assets include cash, cash equivalents, accounts receivable, stock inventory, marketable securities, pre-paid liabilities, and other liquid assets.

6. Total Liabilities: Total liabilities are the combined debts and obligations that an individual or company owes to outside parties. Everything the company owns is classified as an asset and all amounts the company owes for future obligations are recorded as liabilities.

7. Net Sales: Net sales is the sum of a company's gross sales minus its returns, allowances, and discounts.

Understanding the meaning of predictor variables

Now as the terminologies, used in the predictor variables are briefed above, we are in a position of knowing what the actual predictors measure or how it may affect the financial condition of a firm.

1. Ratios of cash flow to total debt: The cash flow-to-total debt ratio is the ratio of a company's cash flow from operations to its total debt. This ratio is a type of coverage ratio and can be used to determine how long it would take a company to repay its debt if it devoted all of its cash flow to debt repayment.

   It is clear from the definition, **more the value of the ratio more financially stable it is**. A healthy ratio would generally fall between 1.0 and 2.0, with anything above 2.0 being considered very strong. Hence, the value of this ratio close to 1.0 is indicator of goof financial condition.

2. Ratios of net income to total assets: The net income-to-total asset ratio or **return on assets (ROA)** refers to a financial ratio that indicates how profitable a company is in relation to its total assets.

   A higher ROA means a company is more efficient. A ROA of over 5% is generally considered good and over 20% excellent.

3. Ratios of current assets to total liabilities: This ratio measures a company's ability to pay short-term obligations or those due within one year with its total assets.

   A small ratio indicates that the company's total liabilities are less probable to be payed by its assets, cash or other short-term assets expected to be converted to cash within a year or less. Hence, **higher value of this ratio indicates less probable of being bankrupt**.

4. Ratios of current assets to net sales: This ratio has a relation with **current assets turnover**.

$$\text{Current Assets Turnover} = \frac{\text{Net Sales}}{\text{Current Assets}} \times 100$$

   The current assets turnover ratio indicates how many times the current assets are turned over in the form of sales within a specific period of time. A higher asset turnover ratio means a better percentage of sales. That is why **the less the amount of current assets-net sales ratio, the better the ability of the company to generate sales**.

3 Exploratory Data Analysis

As we are well acquired with the dataset, it is time to explore the data visually. It has two groups, one is *bankrupt* (21 data points) and another group is *not bankrupt* (25 data points). Scatter plots, histograms, QQ plots and box plots are drawn to analyse the data visually and make some initial hypothesis which are shown statistically later.

3.1 Scatter Plots, Box Plots and Histograms(Density Curves)

- **Scatter plot of y vs *covariates***

  At first the scatter plots for each predictor against the response is plotted, to get an idea whether some predictor can classify the two groups well.
  Observation: It is seen from Fig:1 that CFTD, NITA and CATL may classify the two groups better than that of CANS (which was sort of intuitive from the data description as well). So, it is hard to distinguish. Moreover, it is seen that for high values of CFTD chance of not getting bankrupt is high. Same conclusions can be drawn for NITA and CATL as well.

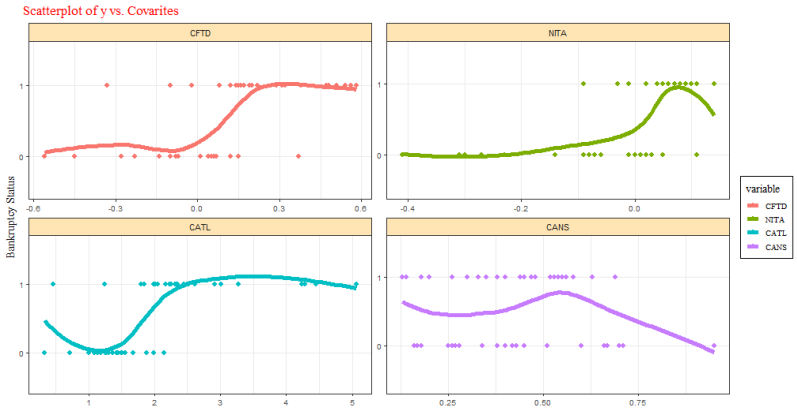


Figure 1: Scatter plot of y vs *covariates*

- **Box-plot of y vs *covariates***

  Now to re-emphasize our observation, box-plots are drawn and seen the difference between the median within the two groups for each of the predictor variables.

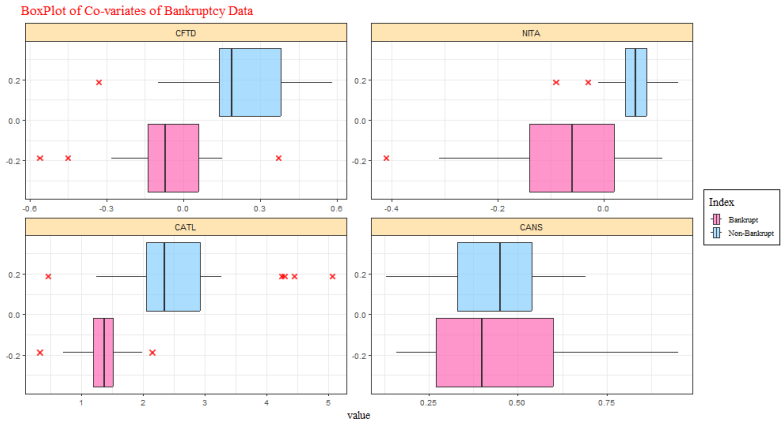


Figure 2: Box-plot of y vs *covariates*

Observation: Fig:2 again indicates that CFTD, NITA and CATL may have more loading to classify the two groups than that of CANS.

- **Pairwise comparison between the *covariates***

  Pairwise scatter plots are done to know if there exist any dependence between the predictor variables. Histograms and density plots of each predictor variables is also drawn to notice the separability between the groups.

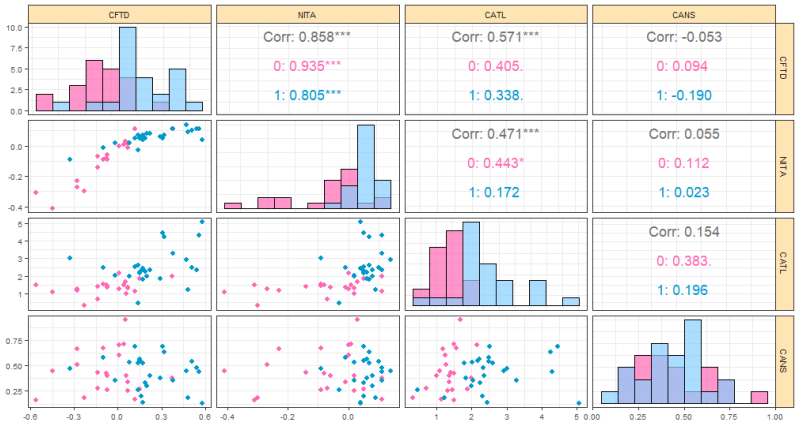


Figure 3: Pairwise Comparisons between *covariates* with Histograms

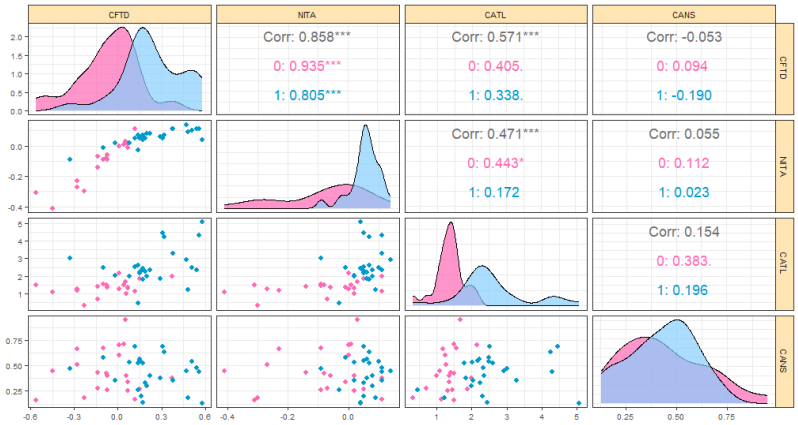


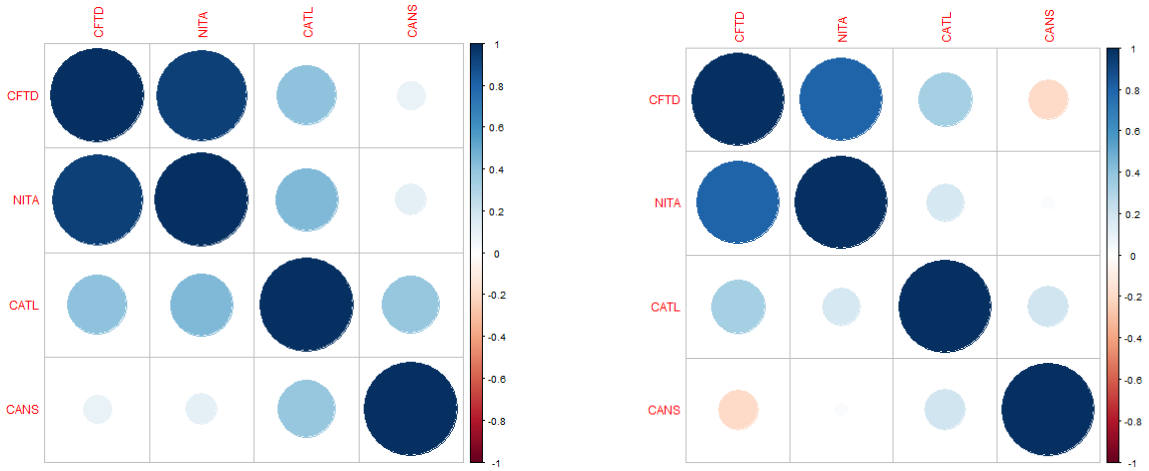
Figure 4: Pairwise Comparisons between *covariates* with Density

Observation: From Fig:3 and Fig:4 it is intuitively evident that separability of the two groups can be pointed out by CFTD, NITA and CATL rather than that of CANS. Because, in case of CANS the density plots (or the histograms) almost overlap each other for the two different groups. Here, another observation can be made that, for bankrupt firms the value of NITA is quite spread out throughout the x axis, but for non-bankrupt firms there are hardly any low value.

- **Correlation Plots**

  As we have seen form the previous pairwise scatter plots between the predictor variables, it is clear that CFTD and NITA has high linear dependence. To know

the perfect correlations between the predictor variables, the correlation plots for both the bankrupt and non bankrupt firms are plotted.



(a) Bankrupt Firm (b) Non-Bankrupt Firm

Figure 5: Correlation Plots for two groups

Observation: CFTD and NITA are highly correlated in both groups. CATL with CFTD and NITA are decently correlated. CANS is almost not correlated to any of the other predictors variables.

3.2 Checking Univariate Normality

First we try to see whether univariate normality exists for each variables in each of the two groups, crudely, through QQ plots. Then *Shapiro Wilk Test* has been done for checking univariate normality for each variables in both groups.

3.2.1 For bankrupt firms

- QQ Plots for each variable:

  The QQ plots for the four predictors are plotted as Fig:6. It is seen from the graphs that they may have univariate normality in each variables for this group.

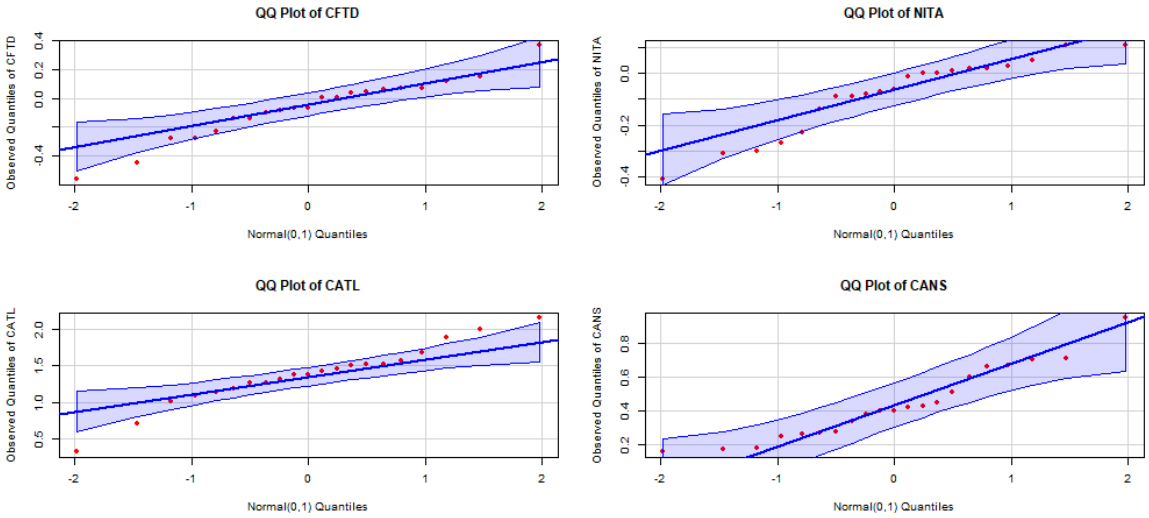


Figure 6: QQ Plots of predictors for Bankrupt Firms

- Shapiro Wilk Test for each variable:

  For ensuring the assumption from the QQ plot, Shapiro Wilk Test is performed and the result got from the tests are as followed.

	Test	Variable	Statistic	p value	Normality
1	Shapiro-Wilk	CFTD	0.9582	0.4800	YES
2	Shapiro-Wilk	NITA	0.9108	0.0571	YES
3	Shapiro-Wilk	CATL	0.9595	0.5057	YES
4	Shapiro-Wilk	CANS	0.9372	0.1921	YES

  Hence, univariate normality is there in each variable for bankrupt group with respect to Shapiro Wilk test.

3.2.2 For Non-bankrupt firms

- QQ Plots for each variable:

  The QQ plots for the four predictors are plotted as Fig:7. It is seen from the graphs that they may have univariate normality in each variables except for *CATL* for this group.

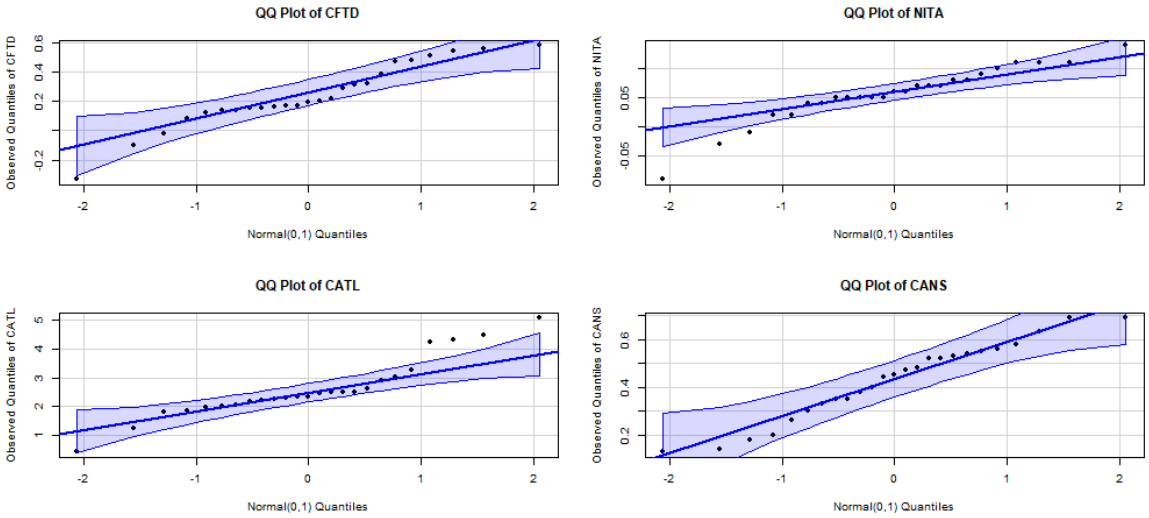


Figure 7: QQ Plots of predictors for Non-Bankrupt Firms

- Shapiro Wilk Test for each variable:

  For ensuring the assumption from the QQ plot, Shapiro Wilk Test is performed and the result got from the tests are as followed.

	Test	Variable	Statistic	p value	Normality
1	Shapiro-Wilk	CFTD	0.9417	0.1620	YES
2	Shapiro-Wilk	NITA	0.9238	0.0626	YES
3	Shapiro-Wilk	CATL	0.9074	0.0267	NO
4	Shapiro-Wilk	CANS	0.9614	0.4429	YES

  Hence, univariate normality is there in each variable for bankrupt group except for *CATL*, as expected from the QQ plot, with respect to Shapiro Wilk test.

3.3 Checking Multivariate Normality

For checking multivariate normality we have used *Royston test*. Chi square plot is not used for checking normality, as it is applicable only when $n - p \geq 30$, where n is the number of observations and p is the number of predictor variables. Hence, for multivariate normality we have leaned on to *Royston Test*, whose implementation is available in the `MVN package` in `R`. And for detecting outliers *Adjusted Chi square QQ plot* is considered.

3.3.1 For bankrupt firms

- Multivariate Normality Checking:

  For checking multivariate normality Royston test is performed and we got that it is multivariate normal for this group. The result is as follows:

	Test	H	p value	MVN
1	Royston	6.04872	0.129197	YES

- Outlier Detection:

  Now, we have used Adjusted Chi-Square Q-Q Plot to detect the outliers. The graph of the outliers is as follows:

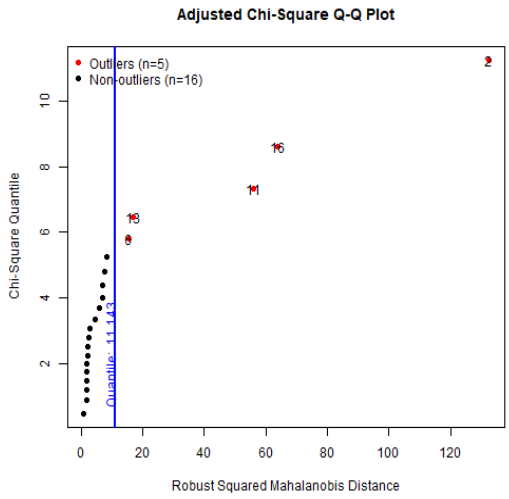


Figure 8: Adjusted Chi-Square Q-Q Plot for bankrupt firms

The outliers got from here are

	Observation	Mahalanobis Distance	Outlier
2	2	132.443	TRUE
16	16	63.885	TRUE
11	11	55.893	TRUE
13	13	16.726	TRUE
6	6	15.691	TRUE

3.3.2 For non-bankrupt firms

- Multivariate Normality Checking:

  For checking multivariate normality Royston test is performed and we got that it is not multivariate normal for this group. The result is as follows:

	Test	H	p value	MVN
1	Royston	12.45531	0.01239924	NO

- Outlier Detection:

  Now, we have used Adjusted Chi-Square Q-Q Plot to detect the outliers. The graph of the outliers is as follows:

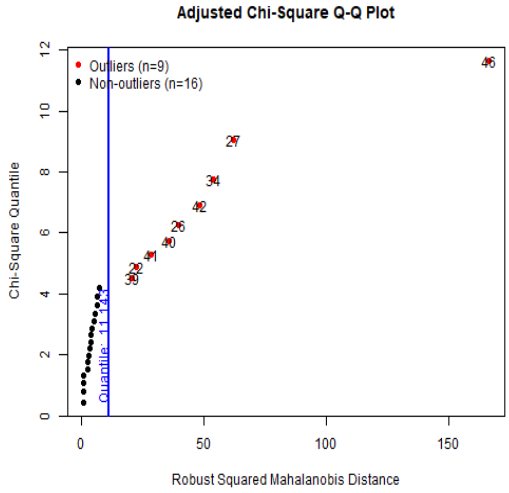


Figure 9: Adjusted Chi-Square Q-Q Plot for non-bankrupt firms

The outliers got from here are

	Observation	Mahalanobis Distance	Outlier
25	25	166.133	TRUE
6	6	62.055	TRUE
13	13	53.992	TRUE
21	21	48.329	TRUE
5	5	39.645	TRUE
19	19	35.802	TRUE
20	20	28.260	TRUE
1	1	22.397	TRUE
18	18	20.710	TRUE

NOTE: While trying to change it to normality, we have tried removing or estimating the outliers (some extreme one), but it is rather disturbing the univariate normality of the variables. Hence, these outliers are those few data from the outside region of the normal distribution, which is supposed to come normally. So, we proceed without disturbing the outliers.

4 Discriminant Analysis

As we are not getting multivariate normality for both the groups, first approach taken here is, *dropping variables*. Then univariate and multivariate transformations are done to the variables in each group. The results got from those approaches are elaborated below. And then according to what we have got, LDA or QDA is applied. Lastly, we have tried a visual approach to see whether the data can be said to be well separated, and whether LDA or QDA can be used without any transformation.

4.1 Dropping variables without transforming

While taking subsets of the predictor variables and checking multivariate normality, we have not considered the third variable $CATL$, as it has disturbed the univariate normality for the non-bankrupt group.

4.1.1 Two variables at a time

First we have chosen all the possible combinations of two variables and checked multivariate normality without trying any type of transformations to them to see if we can get normality. *Royston test* is used only for these checking.

Variables Included	**Bankrupt Firms**			**Non-Bankrupt Firms**		
	Test Statistic	**p-value**	**Decision**	**Test Statistic**	**p-value**	**Decision**
CFTD, NITA	3.151806	0.1168098	Accept	6.122089	0.0392588	Reject
CFTD, CANS	2.737765	0.2543941	Accept	2.735482	0.2545416	Accept
NITA, CANS	5.322533	0.0698239	Accept	5.146921	0.0762711	Accept

Table 1: Checking multivariate normality

Conclusion: Hence, we get that, we can take *CFTD, CANS* and *NITA, CANS* for further analysis such as LDA or QDA.

- CFTD & CANS:

  As multivariate normality is present in this case, **Box-M** test is performed (to check whether the covariance matrices are same). Because, if we get the covariance matrices are same, then we can perform *LDA*, otherwise we have to do *QDA*.

  Result of Box-M Test

```
        Box's M-test for Homogeneity of Covariance Matrices

        data:  Y
        Chi-Sq (approx.) = 2.3791, df = 3, p-value = 0.4975
```

  The *p-value* is found to be 0.4975. Hence, we fail to reject the null hypothesis that the population covariance matrix of the two groups are same. So, the next we check whether the mean of the two populations are different. If yes then the two populations are not different, and hence LDA cannot be applied. But if we get that the two populations have different means then we ca perform LDA.

  Result of MANOVA

```
           Df  Pillai approx F num Df den Df    Pr(>F)
y           1 0.34432    11.29      2     43 0.0001145 ***
Residuals 44
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

  The *p-value* is found to be 0.0001145. Hence, we reject the null hypothesis that the two populations have the same mean. So, we can proceed further to do *LDA*.

Result of LDA

```
Call:
lda(My.data[, -c(2, 3, 5)], My.data$y)

Prior probabilities of groups:
        0         1
0.4565217 0.5434783

Group means:
          CFTD     CANS
0 -0.06904762 0.437619
1  0.23520000 0.426800

Coefficients of linear discriminants:
            LD1
CFTD 4.67736451
CANS 0.01965838
```

As we know, we will get only one LDA scaling vector (number of groups = 2). We plot the LDA scores with respect to the scaling we get, viz., *LD1*. As we have only one scaling vector, we have used `jitter` to clearly see the graph.

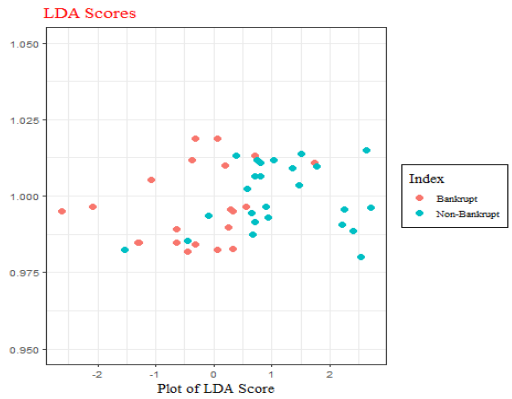


Figure 10: Plot of LDA Scores

From the plot we can say that, how LDA score separates Bankrupt firms from Financially Sound Firm (Non- Bankrupt). Here, we can say that the plot is not that good.

Now, we see the performance of this Discriminant rule in the *training set* and look at the *AER* with *Leave one out Cross Validation*.

Training Set Performance: We see the confusion matrix and we also see the partition plot to analyse the misclassification more elaborately.

		Predicted	
Actual		**0**	**1**
	0	15	6
	1	3	22

Table 2: Training Set Performance

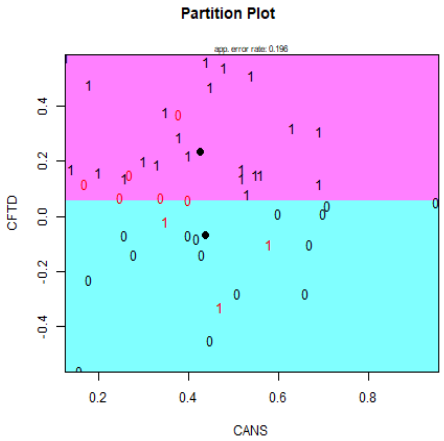


Figure 11: Partition Plot

This plot shows the boundary of LDA based on training set. We can see that the decision boundary is linear, which is expected also. Also, the red coloured observations denote the cases where LDA classifies the observations wrongly. It is reflecting the same as the LDA scores plot.

AER Estimate (Cross Validated): We get 11 misclassification out of 46 observations for leave one out cross validation. We get the *AER* as 0.2391304.

- NITA & CANS:

  As multivariate normality is present in this case, **Box-M** test is performed (to check whether the covariance matrices are same) as done in the previous case.
  Result of Box-M Test

```
Box's M-test for Homogeneity of Covariance Matrices

data:  Y
Chi-Sq (approx.) = 23.435, df = 3, p-value = 3.277e-05
```

The *p-value* is found to be 3.277e-05. Hence, we reject the null hypothesis that the population covariance matrix of the two groups are same. So, no need of doing MANOVA, as we cannot perform LDA here. Therefore, we perform *QDA* directly.

Result of QDA

```
Call:
qda(My.data[, -c(1, 3, 5)], My.data$y)

Prior probabilities of groups:
        0         1
0.4565217 0.5434783

Group means:
          NITA     CANS
0 -0.08142857 0.437619
1  0.05560000 0.426800
```

Now, we see the performance of this Discriminant rule in the *training set* and look at the *AER* with *Leave one out Cross Validation*.

Training Set Performance: We see the confusion matrix and we also see the partition plot to analyse the misclassification more elaborately.

		Predicted	
Actual		**0**	**1**
	0	12	9
	1	2	23

Table 3: Training Set Performance

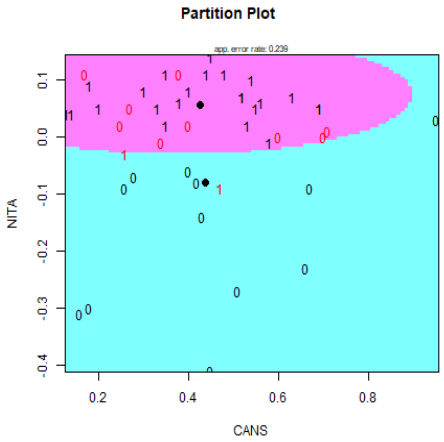


Figure 12: Partition Plot

This plot shows the boundary of QDA based on training set. We can see that the decision boundary is not linear, which is expected also. Also, the red coloured observations denote the cases where QDA classifies the observations wrongly. Compared to the previous LDA partition plot, the misclassification is more in training set.

AER Estimate (Cross Validated): We get 12 misclassification out of 46 observations for leave one out cross validation. We get the *AER* as 0.2608696.

4.1.2 Three variables at a time

Now we have chosen three variables and checked multivariate normality without trying any type of transformations to them to see if we can get normality. *Royston test* is used only for these checking. As we have dropped the third variable, which is ot univariate normal, we are left with only one choice, viz., *CFTD, NITA & CANS*.

Variables Included	Bankrupt Firms			Non-Bankrupt Firms		
	Test Statistic	p-value	Decision	Test Statistic	p-value	Decision
CFTD, NITA & CANS	4.823069	0.1128417	Accept	4.823069	0.1128417	Accept

Table 4: Checking multivariate normality

Hence we get **CFTD, NITA & CANS** are multivariate normal. So, we can further perform Box-M test.

Result of Box-M Test

```
Box's M-test for Homogeneity of Covariance Matrices

data:  Y
Chi-Sq (approx.) = 46.237, df = 6, p-value = 2.655e-08
```

The *p-value* is found to be 2.655e-08. Hence, we reject the null hypothesis that the population covariance matrix of the two groups are same. So, no need of doing MANOVA, as we cannot perform LDA here. Therefore, we perform *QDA* directly.

Result of QDA

```
Call:
qda(My.data[, -c(1, 3, 5)], My.data$y)

Prior probabilities of groups:
        0         1
0.4565217 0.5434783

Group means:
          CFTD        NITA     CANS
0 -0.06904762 -0.08142857 0.437619
1  0.23520000  0.05560000 0.426800
```

Now, we see the performance of this Discriminant rule in the *training set* and look at the *AER* with *Leave one out Cross Validation*.

Training Set Performance: We see the confusion matrix analyse the misclassification.

		Predicted	
Actual		**0**	**1**
	0	13	8
	1	3	22

Table 5: Training Set Performance

AER Estimate (Cross Validated): We get 10 misclassification out of 46 observations for leave one out cross validation. We get the *AER* as 0.2173913.

4.2 Transformation for Multivariate Normality

Now, as we are done with dropping variables without transforming, we try to transform the whole data to get multivariate normality. As we know **Box-cox Transformation** is used for transforming the data to get normality. But, in our case the variables can take negative values as well. So, we cannot apply our known Box-cox transformation. Therefore, we shift to a new transformation, **Yeo-Johnson Transformation**.

Yeo-Johnson suggested a generalized Box-cox transformation which can be used when the values are negative as well. We used this transformation in our case in different manner to try to get multivariate normality. As we recall the transformation looks like, [Ref: Yeo and Johnson]

$$\psi(y,\lambda) = \begin{cases} \frac{(y+1)^{\lambda}-1}{\lambda} & \text{if } y \geq 0, \lambda \neq 0 \\ \log(y+1) & \text{if } y \geq 0, \lambda = 0 \\ -\frac{(-y+1)^{2-\lambda}-1}{2-\lambda} & \text{if } y \geq 0, \lambda \neq 2 \\ -\log(-y+1) & \text{if } y \geq 0, \lambda = 2 \end{cases}$$

To obtain Optimal λ we will use likelihood based approach.

4.2.1 Transforming CATL for Non-bankrupt firms

As we dd not have univariate normality for only CATL for non-bankrupt firms, we first tried to use Yeo-Johnson Transformation just to this variable for non-bankrupt group. After finding the optimum λ we are to use that λ to transform the third variable in the bankrupt group as well. Then we will check whether we get multivariate normality. We plot the *log-likelihood vs* λ graph and find the optimal λ. The graph is as follows:

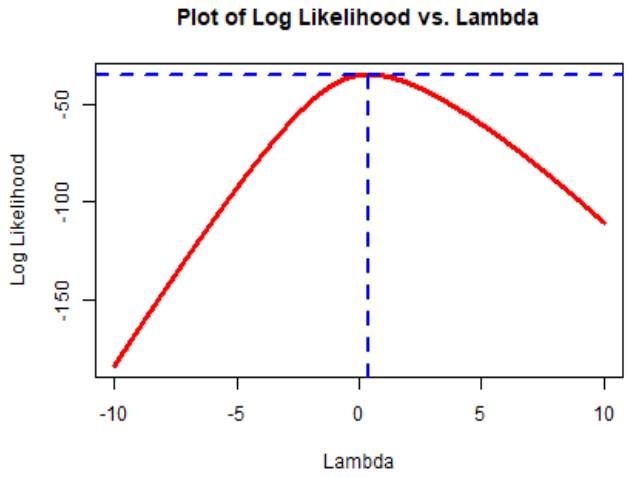


Figure 13: Transforming CATL for Non-bankrupt firms

And we get the optimal λ as 0.41 with the maximum log-likelihood to be -34.83718. We transform the data (first only the non-bankrupt group) and get that we are having univariate normality, but not having multivariate normality yet.

```
$multivariateNormality
     Test        H    p value MVN
1 Royston 11.19854 0.02137175  NO

$univariateNormality
          Test  Variable Statistic   p value Normality
1 Shapiro-Wilk   CFTD       0.9417    0.1620    YES
2 Shapiro-Wilk   NITA       0.9238    0.0626    YES
3 Shapiro-Wilk   CATL       0.9256    0.0688    YES
4 Shapiro-Wilk   CANS       0.9614    0.4429    YES
```

4.2.2 Transforming CATL maximising joint likelihood

Applying Yeo-Johnson family of power transformation is yielding univariate normality. But, we are not getting multivariate normality. Another approach could be to find the log likelihood (mentioned in Yeo-Johnson Paper) of two populations separately for same λ. And then maximize the sum of the log-likelihood as a function of λ. Then, we will get same lambda for both the population. So, we need to maximize -

$$l_{n_1,n_2}(\lambda|x_1,x_2) = l_{n_1}(\lambda|x_1) + l_{n_2}(\lambda|x_2)$$

where, for $i = 1, 2$

$$l_{n_i}(\lambda|x_i) = -\frac{n_i}{2}\log(2\pi) - \frac{n_i}{2}\log(\hat{\sigma}_i^2) - \frac{1}{2\hat{\sigma}_i^2}\sum_{j=1}^{n_i}\{\psi(\lambda, x_{ij}) - \hat{\mu}_i\}^2 + (\lambda - 1)\sum_{j=1}^{n_i} sgn(x_{ij}\log(|x_{ij}|+1))$$

when, $\hat{\mu}_i$ is the mean of $\psi(\lambda, x_{ij})$, $j = 1, 2, \cdots, n_i$ and $\hat{\sigma}_i^2 = \frac{1}{n}\sum_{j=1}^{n_i}\{\psi(\lambda, x_{ij}) - \hat{\mu}_i\}^2$ (symbols have usual meaning). In this case the log-likelihood is maximised at 0.72 with the corresponding log-likelihood to be -45.66093. The plot of log-likelihood vs λ is as follows:

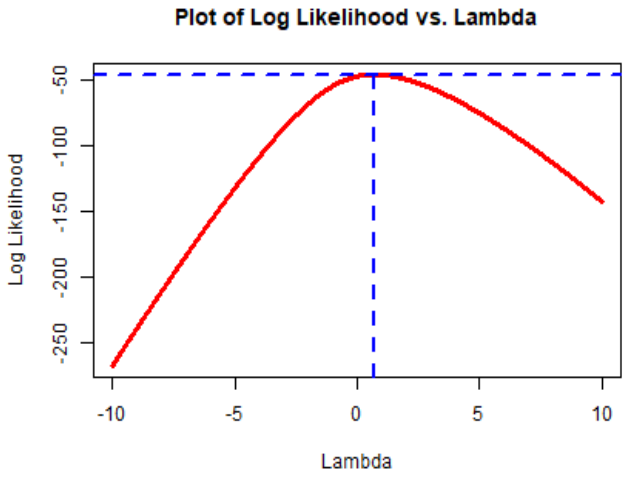


Figure 14: Transforming CATL maximising joint likelihood

But again for non-bankrupt firms, we could not achieve multivariate normality (though univariate normality was achieved).

```
$multivariateNormality
         Test        H    p value MVN
    1 Royston 11.46395 0.0190663  NO

    $univariateNormality
              Test  Variable Statistic   p value Normality
    1 Shapiro-Wilk   CFTD       0.9417    0.1620    YES
    2 Shapiro-Wilk   NITA       0.9238    0.0626    YES
    3 Shapiro-Wilk   CATL       0.9205    0.0527    YES
    4 Shapiro-Wilk   CANS       0.9614    0.4429    YES
```

4.2.3 Multivariate version of Yeo-Johnson family of Transformation

As univariate transformation is not helping much, we tried **Multivariate version of Yeo-Johnson Transformation**. Implementation is available in `R`, `powerTransform` function of `car` package. As we have multivariate normality to the bankrupt group, but not to the non-bankrupt group, we first apply the multivariate version of Yeo-Johnson Transformation to the non-bankrupt group. With that vector of optimum λ's we transform both the groups and checked multivariate normality.

- **Finding the optimum λ**:

  We get the optimum λ's as

```
Estimated transformation parameters
     CFTD      NITA      CATL      CANS
1.2504445 5.3233515 0.6919768 1.7079405
```

- **Checking Multivariate Normality to each group**:

  We check the multivariate normality for the transformed data and we get that multivariate normality is accepted along with univariate normality to both the groups.

  For Bankrupt Firms

```
$multivariateNormality
     Test       H  p value MVN
1 Royston 4.99441 0.200548 YES

$univariateNormality
          Test   Variable Statistic   p value Normality
1 Shapiro-Wilk CFTD_Trans    0.9637    0.5927    YES
2 Shapiro-Wilk NITA_Trans    0.9571    0.4596    YES
3 Shapiro-Wilk CATL_Trans    0.9474    0.3045    YES
4 Shapiro-Wilk CANS_Trans    0.9186    0.0813    YES
```

  For Non-Bankrupt Firms

```
$multivariateNormality
    Test        H   p value MVN
1 Royston 6.981046 0.1261358 YES

$univariateNormality
          Test   Variable Statistic   p value Normality
1 Shapiro-Wilk CFTD_Trans    0.9454    0.1970    YES
2 Shapiro-Wilk NITA_Trans    0.9714    0.6809    YES
3 Shapiro-Wilk CATL_Trans    0.9214    0.0552    YES
4 Shapiro-Wilk CANS_Trans    0.9663    0.5539    YES
```

  Hence, we get *multivariate normality with multivariate transformation.*

- **Box-M test**:

  Now, as we have multivariate normality in each group we can go further to check *Box-M test*.

```
Box's M-test for Homogeneity of Covariance Matrices

data:  Y
Chi-Sq (approx.) = 35.987, df = 10, p-value = 8.462e-05
```

  The $p-value$ is 8.462e-05. So, we reject the null hypothesis that the population covariance matrix of the two groups are same. So, no need of doing MANOVA, as we cannot perform LDA here. Therefore, we perform *QDA* directly.

- **QDA**:

```
Call:
qda(My.data[, -c(1, 3, 5)], My.data$y)

Prior probabilities of groups:
        0         1
0.4565217 0.5434783

Group means:
   CFTD_Trans  NITA_Trans CATL_Trans CANS_Trans
0 -0.06391799 -0.04291975   1.169570  0.5162760
1  0.24660291  0.06830228   2.028308  0.4970221
```

  Now, we see the performance of this Discriminant rule in the *training set* and look at the *AER* with *Leave one out Cross Validation*.

  Training Set Performance:

  We see the confusion matrix analyse the misclassification.

		Predicted	
Actual		**0**	**1**
	0	19	2
	1	1	24

Table 6: Training Set Performance

  AER Estimate (Cross Validated): We get 7 misclassification out of 46 observations for leave one out cross validation. We get the *AER* as 0.1521739. We get the lowest *AER* till now.

4.2.4 Dropping variables with transforming

Less number of variables in a model is always good unless and until we are sacrificing much on misclassification error rate. As we have already discussed some discriminant rules after dropping variables. Now, let us see after transformation how are the performances of some other rules. Here, we will judge based on Leave-one out cross-validation apparent error rate.

- **Transforming three variables**:

  We used multivariate Yeh-Johnson transformation to the subsets of the variables with cardinality three, i.e., three variables. We tried *univariate transformation, transformation through maximising likelihood* and *multivariate transformation.* Only, one set **NITA, CATL, CANS** is not giving us multivariate normality. So, only that is shown here as follows:

  - **Transforming CFTD,NITA & CATL**: First we tried to transform the *CATL* only in the non-bankrupt group and with that optimum λ transformed the whole data. But that could not give us multivariate normality. Hence, we tried the joint likelihood method. That also did not give us any fruitful result. So, lastly we tried to use the multivariate transformation. It yielded good result.

    The optimum λ:

```
Estimated transformation parameters
      CFTD      NITA      CATL
 1.1222010 5.4300750 0.6541604
```

    After Transformation:

    For bankrupt firms:

```
$multivariateNormality
     Test        H   p value MVN
1 Royston 2.705383 0.3148556 YES

$univariateNormality
    Test  Variable Statistic   p value Normality
1 Shapiro-Wilk CFTD^1.12    0.9611    0.5390    YES
2 Shapiro-Wilk NITA^5.43    0.9566    0.4510    YES
3 Shapiro-Wilk CATL^0.65    0.9457    0.2812    YES
```

For non-bankrupt firms:

```
$multivariateNormality
     Test        H   p value MVN
1 Royston 6.616813 0.07627823 YES

$univariateNormality
        Test  Variable Statistic   p value Normality
1 Shapiro-Wilk CFTD^1.12    0.9438    0.1809    YES
2 Shapiro-Wilk NITA^5.43    0.9720    0.6952    YES
3 Shapiro-Wilk CATL^0.65    0.9225    0.0585    YES
```

Hence, we get *multivariate normality with multivariate transformation.*

Box-M test:

Now, as we have multivariate normality in each group we can go further to check *Box-M test.*

```
Box's M-test for Homogeneity of Covariance Matrices

data:  Y
Chi-Sq (approx.) = 27.858, df = 6, p-value = 9.994e-05
```

The $p - value$ is 9.994e-05. So, we reject the null hypothesis that the population covariance matrix of the two groups are same. So, no need of doing MANOVA, as we cannot perform LDA here. Therefore, we perform *QDA* directly.

QDA:

```
Call:
qda(My.data[, -c(1, 3, 5)], My.data$y)

Prior probabilities of groups:
        0         1
0.4565217 0.5434783

Group means:
       CFTD        NITA     CATL
0 -0.06652109 -0.04223014 1.147983
1  0.24068187  0.06865672 1.970263
```

Now, we see the performance of this Discriminant rule in the *training set* and look at the *AER* with *Leave one out Cross Validation.*

Training Set Performance:
We see the confusion matrix analyse the misclassification.

		Predicted	
Actual		**0**	**1**
	0	17	4
	1	2	23

Table 7: Training Set Performance

AER Estimate (Cross Validated): We get 7 misclassification out of 46 observations for leave one out cross validation. We get the AER as 0.1521739. We get the same AER as that of using all the variables.

– **Transforming CFTD,CATL & CANS**: First we tried to transform the $CATL$ only in the non-bankrupt group and with that optimum λ transformed the whole data. But that could not give us multivariate normality. Hence, we tried the joint likelihood method. In this case joint likelihood gave us (with $\lambda = 0.72$)
For bankrupt firms:

```
$multivariateNormality
     Test        H   p value MVN
1 Royston 4.605265 0.2059711 YES

$univariateNormality
          Test  Variable Statistic   p value Normality
1 Shapiro-Wilk CFTD       0.9582    0.4800    YES
2 Shapiro-Wilk CATL       0.9487    0.3222    YES
3 Shapiro-Wilk CANS       0.9372    0.1921    YES
```

For non-bankrupt firms:

```
$multivariateNormality
     Test        H   p value MVN
1 Royston 7.449324 0.05910793 YES

$univariateNormality
         Test  Variable Statistic   p value Normality
1 Shapiro-Wilk CFTD       0.9417    0.1620    YES
2 Shapiro-Wilk CATL       0.9205    0.0527    YES
3 Shapiro-Wilk CANS       0.9614    0.4429    YES
```

Hence, we get *multivariate normality with multivariate transformation.*

Box-M test:

Now, as we have multivariate normality in each group we can go further to check *Box-M test.*

```
Box's M-test for Homogeneity of Covariance Matrices

data:  Y
Chi-Sq (approx.) = 16.082, df = 6, p-value = 0.01332
```

The $p-value$ is 9.994e-05. So, we reject the null hypothesis that the population covariance matrix of the two groups are same. So, no need of doing MANOVA, as we cannot perform LDA here. Therefore, we perform *QDA* directly.

QDA:

```
Call:
qda(My.data[, -c(1, 3, 5)], My.data$y)

Prior probabilities of groups:
        0         1
0.4565217 0.5434783

Group means:
       CFTD     CATL     CANS
0 -0.06904762 1.185911 0.437619
1  0.23520000 2.072756 0.426800
```

Now, we see the performance of this Discriminant rule in the *training set* and look at the *AER* with *Leave one out Cross Validation.*

Training Set Performance:

We see the confusion matrix analyse the misclassification.

		Predicted	
Actual		**0**	**1**
	0	19	2
	1	1	24

Table 8: Training Set Performance

AER Estimate (Cross Validated): We get 6 misclassification out of 46 observations for leave one out cross validation. We get the *AER* as 0.1304348. We get the best *AER* as of till now.

- **Transforming two variables:** Table of results of taking 2 variables at a time: (Which were not discussed previously)

Subsets	Transformation	Box-M p-Value	QDA CV AER estimate
CFTD & NITA	Multivariate	0.001785	0.2391304
CFTD & CATL	Not possible	NA	NA
NITA & CATL	Multivariate	0.015010	0.1304348
CATL & CANS	Univariate	0.002709	0.1521739

Table 9: For two variables

Only transformed NITA & transformed CATL is producing the lowest estimate of AER amongst all. So, for this rule the optimum λ for the transformation to get the multivariate normality is,

```
Estimated transformation parameters
      NITA      CATL
7.5873497 0.3379119
```

Analysis Using Transformed NITA and Transformed CATL

The QDA performed here

```
Call:
qda(My.data_trans51[, c(2, 3)], My.data_trans51[, 5])

Prior probabilities of groups:
        0         1
0.4565217 0.5434783

Group means:
         NITA      CATL
0 -0.02994461 0.9864808
1  0.07632512 1.5606233
```

The partition plot for this classification rule is as follows:

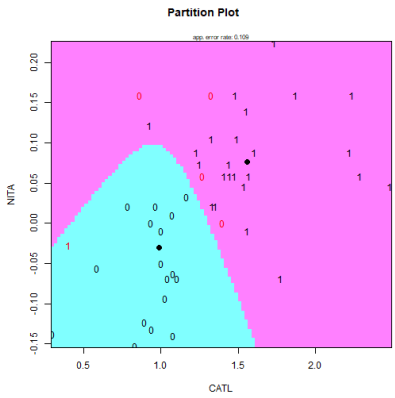


Figure 15: Partition plot

Observation: This plot shows the boundary of QDA based on training set. We can see that the decision boundary is not linear, which is expected also. Also, the red coloured observations denote the cases where QDA classifies the observations wrongly. We can see that the number of misclassified observation in training set is quite low.

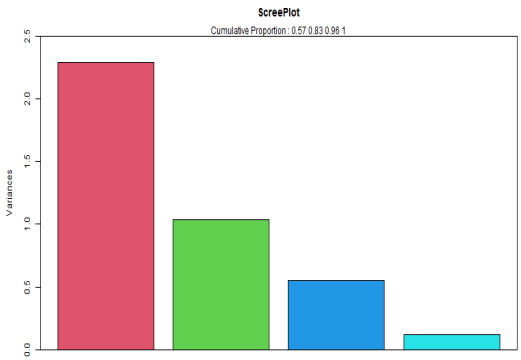
4.3 Analysis of the whole data without transforming

Now, we want to check if the data can be said to be well separated, so that we can apply *LDA or QDA* without even assuming multivariate normality (As discussed in class). We can maximum plot upto three dimension. But, here we have four explanatory variables. Plotting three variables at a time is not adequate. So, we are motivated to do **Principal Component Analysis**.

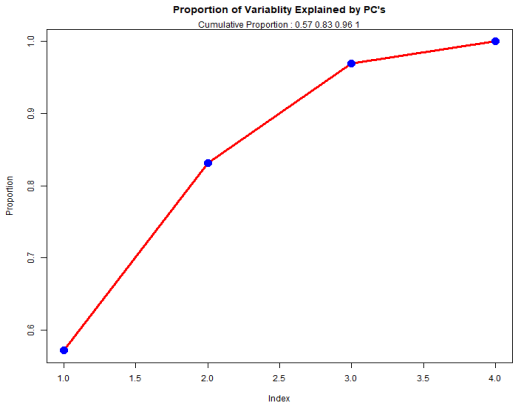
4.3.1 Principal Component Analysis

We apply PCA and get the following result

```
Standard deviations (1, .., p=4):
[1] 1.5121409 1.0187432 0.7437780 0.3498378

Rotation (n x k) = (4 x 4):
            PC1         PC2        PC3        PC4
CFTD 0.62014111 -0.17691691  0.1967033 -0.7385345
NITA 0.59989827 -0.08361003  0.4630444  0.6470868
CATL 0.50193544  0.20371413 -0.8266216  0.1525063
CANS 0.06006574  0.95927594  0.2521796 -0.1121929
```



(a) Prop of Variability explained by PC's



(b) Scree Plot

From the output and the plots above we can see that the cumulative proportion of variability explained by first three PCs is around 96%. Which is quite good.

Biplot: We drew the biplot as follows

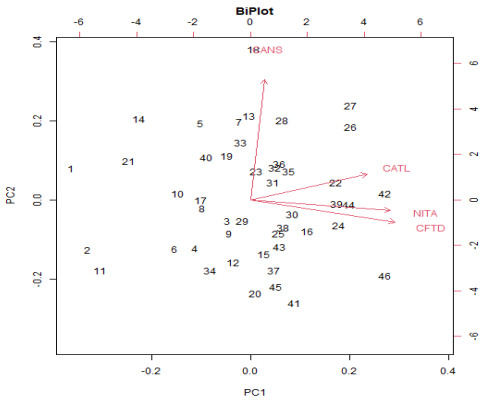


Figure 17: Biplot

Observation: In addition to the observations the plot shows the original variables as vectors (arrows). They begin at the origin [0,0] and extend to coordinates given by the loading vector. We can see that in the first principal component CFTD,NITA and CATL majorly contributes. While, in the second principal component CANS majorly contributes. Also, the small angle between NITA and CFTD indicates presence of high correlation between them.

3D Plot of first three Principal Components: We will plot the first three principal components, which explains about 96% of the variability of the data.

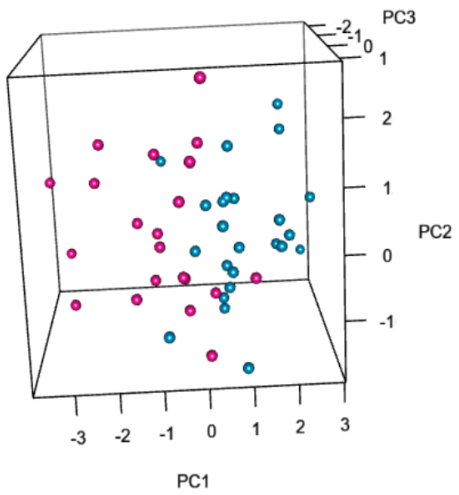


Figure 18: 3D Plot of first three Principal Components :

Here the *pink* dots indicates bankrupt firms and the *blue* dots indicates the non-bankrupt firms.

Observation: From the 3D plot of the first PCs we can clearly see that, with respect to four explanatory variables, the observations from Bankrupt firms and Fiancially sound firms are well separated.
Here, our original four explanatory variables are not multivariate normally distributed. So,LDA and QDA is not optimal here. But, since with respect to the four variables, the two populations are well separated. So, LDA and QDA should perform good.

4.3.2 LDA & QDA based on all un-transformed variables

- **LDA**

```
Call:
lda(My.data[, -5], My.data$y)

Prior probabilities of groups:
        0         1
0.4565217 0.5434783

Group means:
         CFTD        NITA     CATL     CANS
0 -0.06904762 -0.08142857 1.366667 0.437619
1  0.23520000  0.05560000 2.593600 0.426800
```

```
Coefficients of linear discriminants:
            LD1
CFTD  0.6612498
NITA  4.3935204
CATL  0.8872152
CANS -1.1785089
```

Training Set Performance:

We see the confusion matrix analyse the misclassification.

		Predicted	
Actual		**0**	**1**
	0	18	3
	1	1	24

Table 10: Training Set Performance

AER Estimate (Cross Validated):

We get 6 misclassification out of 46 observations for leave one out cross validation. We get the AER as 0.1304348.

- **QDA**

```
Call:
qda(My.data[, -5], My.data$y)

Prior probabilities of groups:
        0         1
0.4565217 0.5434783

Group means:
         CFTD        NITA     CATL     CANS
0 -0.06904762 -0.08142857 1.366667 0.437619
1  0.23520000  0.05560000 2.593600 0.426800
```

Training Set Performance:

We see the confusion matrix analyse the misclassification.

		Predicted	
Actual		**0**	**1**
	0	19	2
	1	1	24

Table 11: Training Set Performance

AER Estimate (Cross Validated):

We get 5 misclassification out of 46 observations for leave one out cross validation. We get the AER as 0.1086957. We get the best performance till now.

5 Factor Analysis

Factor Analysis is used to identify the underlying structure or patterns in a set of variables and to reduce their complexity into a smaller number of factors or components. But to proceed with factor analysis, we need to first test whether the variables are actually related. i.e, whether the Correlation matrix of the variables is an Identity matrix.

For that we will use **Bartlett Test of Sphericity**. The hypothesis is

$$\mathcal{H}_0 : R = I \;\; vs \;\; \mathcal{H}_1 : R \neq I$$

where R is the population correlation matrix. The test statistic is given by

$$-\log(det(R^*))\frac{(N-1-(2p+5))}{6}$$

where, R^* is the sample correlation matrix. N is the sample size, and p is the number of variables. It has asymptotic χ^2 distribution with d.f $\frac{p(p-1)}{2}$. The test is sensitive to deviation from normality.

5.1 Bartlett's Test

Since, Bartlett's test of sphericity is sensitive to deviations from normality. So, we cannot proceed with original data. Hence, we will use our transformed variables which are multivariate normally distributed.

```
$chisq
[1] 88.4141

$p.value
[1] 6.467218e-17

$df
[1] 6
```

As here Bartlett's test gets rejected, we can proceed with factor analysis using the four transformed variables. Now, an important question is how many factors we should choose. Originally in variance-covariance matrix we have $p(p+1)/2 = 10$ parameters and we are modelling the matrix using factor model with $p(m+1) = 4(m+1)$ parameters. (m is the number of factors chosen) So, m should not be more than 2. Since, Bartlett's test is rejected. So, m can be either 1 or 2. Now, choosing one factor is not that meaningful. So, we will proceed with $m = 2$. We have used two types of method to find the factors.

5.1.1 Using Principal Component Method & Varimax Rotation

The result after using factor analysis using Principal Component Method & Varimax Rotation is as follows:

```
Loadings:
              PA1    PA2
CFTD_Trans  1.035 -0.122
NITA_Trans  0.860
CATL_Trans  0.598  0.343
CANS_Trans         0.496

                 PA1   PA2
SS loadings    2.169 0.379
Proportion Var 0.542 0.095
Cumulative Var 0.542 0.637
```

The communalities are:

```
CFTD_Trans NITA_Trans CATL_Trans CANS_Trans
1.0866453  0.7393737  0.4755452  0.2465000
```

From the above output we can see that, for the first factor first three variables have high loadings and the sign is positive. While for the second factor last two variables contribute the most. The communality of first variable is high, so the variability of first variable is well explained by the first factor. Similarly, the variability of second variable is also well explained by factors.

5.1.2 Using Maximum Likelihood Method & Varimax Rotation

The result after using factor analysis using Maximum Likelihood Method & Varimax Rotation is as follows:

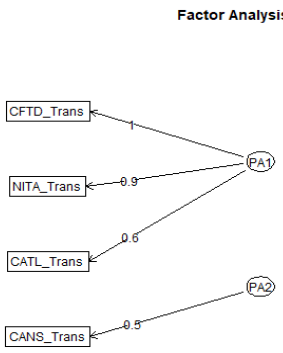
```
Loadings:
              ML1    ML2
CFTD_Trans  0.993
NITA_Trans  0.893
CATL_Trans  0.598  0.152
CANS_Trans         0.997

                 ML1   ML2
SS loadings    2.143 1.026
Proportion Var 0.536 0.257
Cumulative Var 0.536 0.792
```

The communalities are:

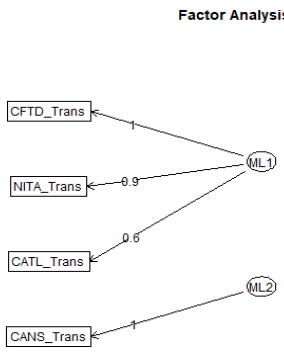
```
CFTD_Trans NITA_Trans CATL_Trans CANS_Trans
0.9950031  0.7985905  0.3802334  0.9950001
```

Here, also we are getting similar interpretation as in principal component method. But it seems that factors are well separating the variables in this case. The communality of first,second and fourth variable is high, so the variability of these variables are well explained by the factors.

5.1.3 Factor Diagrams



(a) Using Principal Component Method



(b) Using Principal Component Method

Figure 19: Factor Diagrams

The FA diagram indicates the above observations only. We can see that the first three variables are related to factor 1 and the fourth variables is related to factor 2. May be these factors have some nice financial indications.

5.2 Rotation does not Change Fitted-Matrix

5.2.1 Fitted Matrix

```
Fitted Matrix with no rotation :

            CFTD_Trans NITA_Trans CATL_Trans CANS_Trans
CFTD_Trans       1.000      0.889      0.579     -0.063
NITA_Trans       0.889      1.000      0.531      0.008
CATL_Trans       0.579      0.531      1.000      0.170
CANS_Trans      -0.063      0.008      0.170      1.000

Fitted Matrix with Varimax rotation :

           CFTD_Trans NITA_Trans CATL_Trans CANS_Trans
CFTD_Trans       1.000      0.889      0.579     -0.063
NITA_Trans       0.889      1.000      0.531      0.008
CATL_Trans       0.579      0.531      1.000      0.170
CANS_Trans      -0.063      0.008      0.170      1.000
```

We know from theory that, orthogonal rotation of the factors does not change the fitted matrix and also the residual matrix. Here, using data we have illustrated the same.

Using Maximum likelihood method with no rotation we have obtained the fitted matrix. Similar thing we have done using varimax rotation. We can see that the fitted matrices are same in both cases. Which justifies that by orthogonal factor rotation, fitted matrix does not change.

But, we should note that by orthogonal factor rotation the loadings change. And also, the factors change. Because, factor rotation is used to get cleaner interpretation of the data. We would really like to define new coordinate systems so that when we rotate everything, the points (loading vector) fall close to the vertices (end points) of the new axes. Such that same variables are not highly correlated with the same factor.

5.2.2 Graphical Illustration

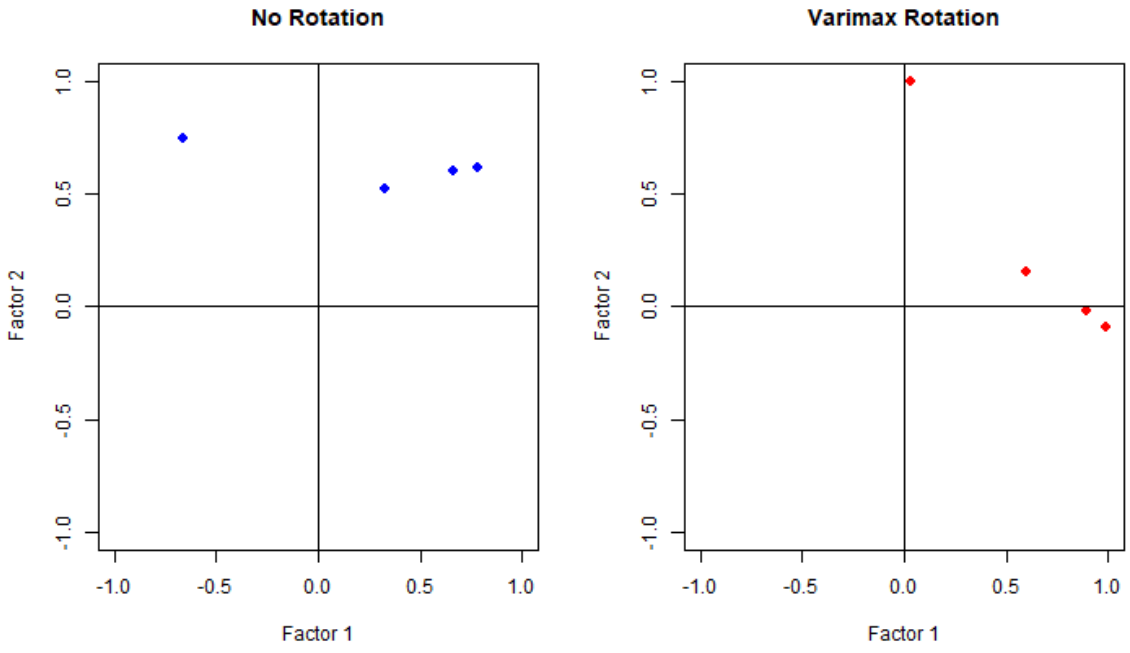


Figure 20: How factor changes by rotation

From the above graphs, we can see that using varimax rotation the points gets closer to the original co-ordinate axes which indicates the interpretability of factors increases.

6 Further Exploration

6.1 Logistic Regression

In LDA, QDA, we assume that X has mixture Gaussian distribution and group wise it has multivariate normal distribution. But in Logistic regression, we assume $X_{p\times 1}$ to be non-stochastic and we model

$$\mathbb{P}(Y = 1|x_1, x_2, ..., x_p) = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+...+\beta_p x_p}}$$

where, $\beta_0, \beta_1, ..., \beta_p$ are the parameters of the model.

6.1.1 Fitting Logistic Regression Model

```
Call:
glm(formula = y ~ ., family = binomial(link = "logit"),
data = My.data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.30416  -0.44545   0.00725   0.49102   2.62396

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -5.320      2.366  -2.248  0.02459 *
CFTD           7.138      6.002   1.189  0.23433
NITA          -3.703     13.670  -0.271  0.78647
CATL           3.415      1.204   2.837  0.00455 **
CANS          -2.968      3.065  -0.968  0.33286
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 63.421  on 45  degrees of freedom
Residual deviance: 27.443  on 41  degrees of freedom
AIC: 37.443

Number of Fisher Scoring iterations: 7
```

From the above output we can see that, if we choose 0.05 as level of significance, among four explanatory variables, only third variable is significant in the presence of the others. Also, comparing Null deviance and Residual Deviance we can say that the fitted model is quite good in terms of prediction.

6.1.2 Model Evaluation

We take 0.5 as our threshold value, i.e., if the predicted probabilities are greater than 0.5 we will predict the response variable as 1, otherwise as 0.

Training Set Performance:

We see the confusion matrix analyse the misclassification.

		Predicted	
Actual		**0**	**1**
	0	18	3
	1	1	24

Table 12: Training Set Performance

Error Estimate (Cross Validated):

We get 5 misclassification out of 46 observations for leave one out cross validation. We get the error estimate as 0.1086957. Again, we are getting 10.86% misclassification rate.

6.2 Profile Analysis

Profile Analysis is a multivariate data analysis technique that is applicable to situations in which p treatments are administrated to two or more groups of subjects. The question of equality of mean vectors is divided into several specific questions such as [Ref: Desjardins and Bulut]

1. Are the population profiles parallel?
2. Are they coincident? (Assuming they are parallel)
3. Are the profiles level? (Assuming they are coincident)

There are few assumptions in profile analysis which we will check one by one, so that we can understand whether profile analysis can be implemented here.

1. **Assumption 1**: The test scores should have a multivariate normal distribution.

   We can transform the data to retain multivariate normality. Hence, we can go ahead to check the next assumption.

2. **Assumption 2**: Homogeneity of the variance covariance matrix of test scores.

   Box-M Test rejected homogeneity assumption. Hence, we cannot go ahead further.

Therefore, **we cannot perform Profile Analysis here**.

7 Summary

- From EDA we have seen that, CFTD, NITA and CATL can separate bankrupt firms from financially sound firms well. From Factor analysis, we have got that these three are contributing to the first factor and CANS is contributing to the second factor.
- Also from EDA, we have seen that CFTD and NITA are very highly correlated.
- Plotting first three principal components, we visualized that the data is well separated, so we applied LDA or QDA even without multivariate normality.
- Finally, we have seen QDA to the original data and Logistic regression are yielding lowest Misclassification Rate(Leave one out cross validated), i.e., 11% approx.
- Further, if we only take transformed NITA and CATL, then also we are not sacrificing much on Misclassification Rate(Leave one out cross validated), i.e., 13% approx.

References

C. D. Desjardins and O. Bulut. Profile analysis of multivariate data in r: An introduction to the profiler package.

Korkmaz, Goksuluk, and Zararsiz. Mvn:anrpackageforassessing multivariate normality.

I.-K. Yeo and R. A. Johnson. A new family of power transformations to improve normality or symmetry.