MAY 2023

FA-582 : Foundations of Financial Data Science
Instructor: Dragos Bozdog

Project Report

BANKRUPTCY PREDICTION

Group 1 : Tashveen Kaur, Abhinav Ganguly, Clinton Nwokike

ABSTACT

In this project, we explore the effectiveness of several machine learning algorithms in predicting bankruptcy of companies, using a dataset of financial ratios for bankrupt and non-bankrupt firms. The algorithms we tested for our project include Logistic Regression, Linear Discriminant Analysis, Quadratic Discriminant Analysis, and K-Nearest Neighbors (KNN) with values of k being 2 and 8. The data was obtained from the UCI Machine Learning Repository and was provided by Deron Liang and Chih-Fong Tsai from National Central University in Taiwan.

Our results indicate that the KNN model with k=2 outperforms other models, achieving an accuracy of 95% in predicting bankruptcy. The KNN model is a non-parametric algorithm that works by finding the k closest data points in the training set to a new input point, and assigning a prediction based on average value of its neighbors. This model's high accuracy suggests that it could be an effective tool for investors, creditors, and other stakeholders in identifying companies at risk of financial distress and bankruptcy.

We also found that several financial ratios were highly correlated and could contribute to the prediction of bankruptcy. These attributes include Per Share Net Profit Before Tax, EPS in the Last 4 seasons, Net Value per Share A, Net Value per Share B, and Net Value per Share C. Identifying these highly correlated attributes could potentially help in developing more accurate predictive models in addition to providing valuable insights to investors and creditors for assessing a company's financial health.

In conclusion, our study demonstrates the potential of machine learning algorithms in predicting bankruptcy and identifying companies at risk of financial distress.

INTRODUCTION

Background of the project

Bankruptcy or corporate failure can hurt both the individual company and the global economy. Business practitioners, investors, governments, and academic academics have long sought techniques to identify the risk of business failure in order to reduce the economic losses associated with bankruptcies.

Research Question: Can we accuarately predict bankruptcy using Machine Learning?

Overview of the project

The project aims to analyze financial market data by applying data collection, data preparation, feature extraction, data cleaning, analytical processing and algorithms. Data clustering, data classification, outlier analysis, and data mining techniques may be implemented. The data can be obtained from the financial markets (stocks, financial statements, etc.) or from text data sources (twitter, news, etc.). You may implement new methods and argue the advantage of them over traditional methods

Significance of the project

Forecasting insolvency is an important task for many financial institutions. The purpose is to forecast the likelihood of a corporation going bankrupt. Effective prediction models are required by financial institutions in order to make suitable lending decisions.

Several models for predicting bankruptcy were able to be developed thanks to recent developments in machine learning (ML).

A recent glance at these bankruptcy activities is the bankruptcy of Silicon Valley Bank recently and Signature Banks. These recent bank failures led us to solve the issue that was in the mind of various investors about the prediction of these banks and companies and to create a model that could accurately predict the failure.

WHY BANKRUPTCY?

- One of the primary objectives of credit risk assessment is predicting business insolvency.

- Bankruptcy or corporate failure can hurt both the individual company and the global economy.

- Particularly since the financial crisis of 2007–2008, it has elevated in importance for most financial institutions, professionals, and scholars.

- Rising interest rates and inflation have created more probability of bankruptcy in banks and institutions.

RECENT NEWS ARTICLES

Business

Cash-strapped biotech firm Codiak files for bankruptcy protection

March 27, 2023

Business

Virgin Orbit bankruptcy casts shadow over Japan's space dreams

April 13, 2023

The Last Haul From Bed Bath & Beyond

Shoppers flooded into the stores after the retailer filed for Chapter 11 bankruptcy.

By JEENAH MOON

British Battery Start-Up Files for Bankruptcy

The failure of Britishvolt is a blow to Britain's plans to promote the manufacture of electric cars, and threatens the future of its automotive industry.

By STANLEY REED

• The DATA

Description of the data

The data set we have chosen was collected from the Taiwan Economic Journal from 1999 to 2009. The data contains Bankrupt companies that were identified based on the business regulations of the Taiwan Stock Exchange.

Features of the data

- The data runs across different industries (electronic manufacturing, retail, shipping, tourism...) Each industry has a sufficient amount of companies in similar size in order to do the comparison
- There are 95 features (X1-X95, business regulations of Taiwan Stock Exchange) and 1 label (bankrupt or not)

Source of the data

Deron Liang and Chih-Fong Tsai, deronliang '@' gmail.com; cftsai '@' mgt.ncu.edu.tw, National Central University, Taiwan

The data was obtained from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Taiwanese+Bankruptcy+Prediction

In our data as the table below would show the data set contains various ratios like (ROA, Gross Margin, Operating Profit, etc.) of said banks and other quantitative metrics like (Net Income or Equity to Liability) that are used as variables.

• The DATA

Features of the data

- Target variable: Bankrupt (0 or 1)

- How many 0 and how many 1?

```
> table(df$Bankrupt.)

   0    1
6599  220
```

- How many 0 and how many 1 after SMOTE and oversampling?

```
> table(new_df$Bankrupt.)

   0    1
3442 3377
```

- Data contains different industries: Manufacturing, Shipping, retail, Tourism, Pharmaceuticals

- Number of years of data: 10 years (From 1999 to 2009)

- Some of the features from our Dataset are discussed in detail in the next page

Variable Name	Description	Data Type
Bankrupt.	Whether the company had gone bankrupt, in 0 and 1	Factor
ROA.C..before.interest.and.depreciation.before.interest	Return on assets, calculated as earnings before interest and taxes divided by total assets.	Numeric
ROA.A..before.interest.and...after.tax	Return on assets, calculated as earnings before interest divided by total assets, then multiplied by (1 - tax rate)	Numeric
ROA.B..before.interest.and.depreciation.after.tax	Return on assets, calculated as earnings before interest and taxes divided by total assets	Numeric
Operating.Gross.Margin	Gross margin, calculated as gross profit divided by revenue, showing the percentage of revenue left after deducting cost of goods sold	Numeric
Realized.Sales.Gross.Margin	Gross margin on actual sales, calculated as gross profit on actual sales revenue minus cost of goods sold, divided by actual sales revenue.	Numeric
Operating.Profit.Rate	Operating profit divided by revenue.	Numeric
Pre.tax.net.Interest.Rate	Pre-tax net profit divided by revenue.	Numeric

EXPLORATORY DATA ANALYSIS:

Exploratory Data Analysis

Data scientists utilize exploratory data analysis (EDA) to study and investigate data sets and describe their essential properties, frequently using data visualization approaches. It aids in determining how to best modify data sources to obtain the answers required, making it easier for data scientists to uncover patterns, detect anomalies, test hypotheses, and validate assumptions.

EDA is largely used to discover what data can reveal beyond the formal modeling or hypothesis testing tasks, and it provides a deeper understanding of data set variables and their interactions.

EDA was used to perform the following functions on our dataset

- Produce graphical displays of high-dimensional data comprising multiple variables.
- Each field in the raw dataset is visualized in univariate form, along with summary statistics.
- Dividing the data into training and testing and looking for biasness in the data.
- To perform normalisation and standardisation in the dataset
- Predictive models, such as linear regression, rely on statistics and data to make predictions.
- Remove and detect outliers
- Analyse the data type and observations of all the variables.
- Determine a ranked list of the variables
- Cleaning the data
- Removing missing values and for mapping and understanding interactions between different fields in the data.

EXPLORATORY DATA ANALYSIS: SUMMARY STATISTICS

Outcomes of our EDA -

- Zero Missing Values
- Data type for all features are numeric
- The Bankrupt Variable was changed into Factor using as.factor()
- 96 Features
- 6819 Observations

Skim Function -

- It provides summary statistics for each variable in the dataset, including data type, missing values, minimum and maximum values, quartiles, and other information. In a tabular style, I
- t provides a thorough summary of the dataset's numerical and category variables.
- It assists in swiftly identifying any errors or anomalies in the data, such as missing numbers, outliers, or unexpected data kinds. It is frequently used as a first step in exploratory data analysis to get preliminary insights into the dataset and inform further data cleaning and analysis operations.

EXPLORATORY DATA ANALYSIS: SUMMARY STATISTICS

Outcome of the Skim Function(10 observations) -

```
> skim(new_df)
── Data Summary ────────────────────────
                           Values
Name                       new_df
Number of rows             6819
Number of columns          95
_________________________
Column type frequency:
  numeric                  95
_________________________
Group variables            None

── Variable type: numeric ──────────────────────────────────────────────────────────────────────────────────────────────
   skim_variable                                                n_missing complete_rate    mean      sd
 1 Bankrupt.                                                            0             1 4.95e-1 5.00e-1
 2 ROA.C..before.interest.and.depreciation.before.interest              0             1 4.64e-1 8.43e-2
 3 ROA.A..before.interest.and...after.tax                               0             1 5.10e-1 1.03e-1
 4 ROA.B..before.interest.and.depreciation.after.tax                    0             1 5.09e-1 9.11e-2
 5 Operating.Gross.Margin                                               0             1 6.04e-1 1.79e-2
 6 Realized.Sales.Gross.Margin                                          0             1 6.04e-1 1.79e-2
 7 Operating.Profit.Rate                                                0             1 9.99e-1 1.21e-2
 8 Pre.tax.net.Interest.Rate                                            0             1 7.97e-1 1.02e-2
 9 After.tax.net.Interest.Rate                                          0             1 8.09e-1 9.56e-3
10 Non.industry.income.and.expenditure.revenue                          0             1 3.03e-1 9.66e-3
```

ANALYSIS OF THE SKIM FUNCTION -

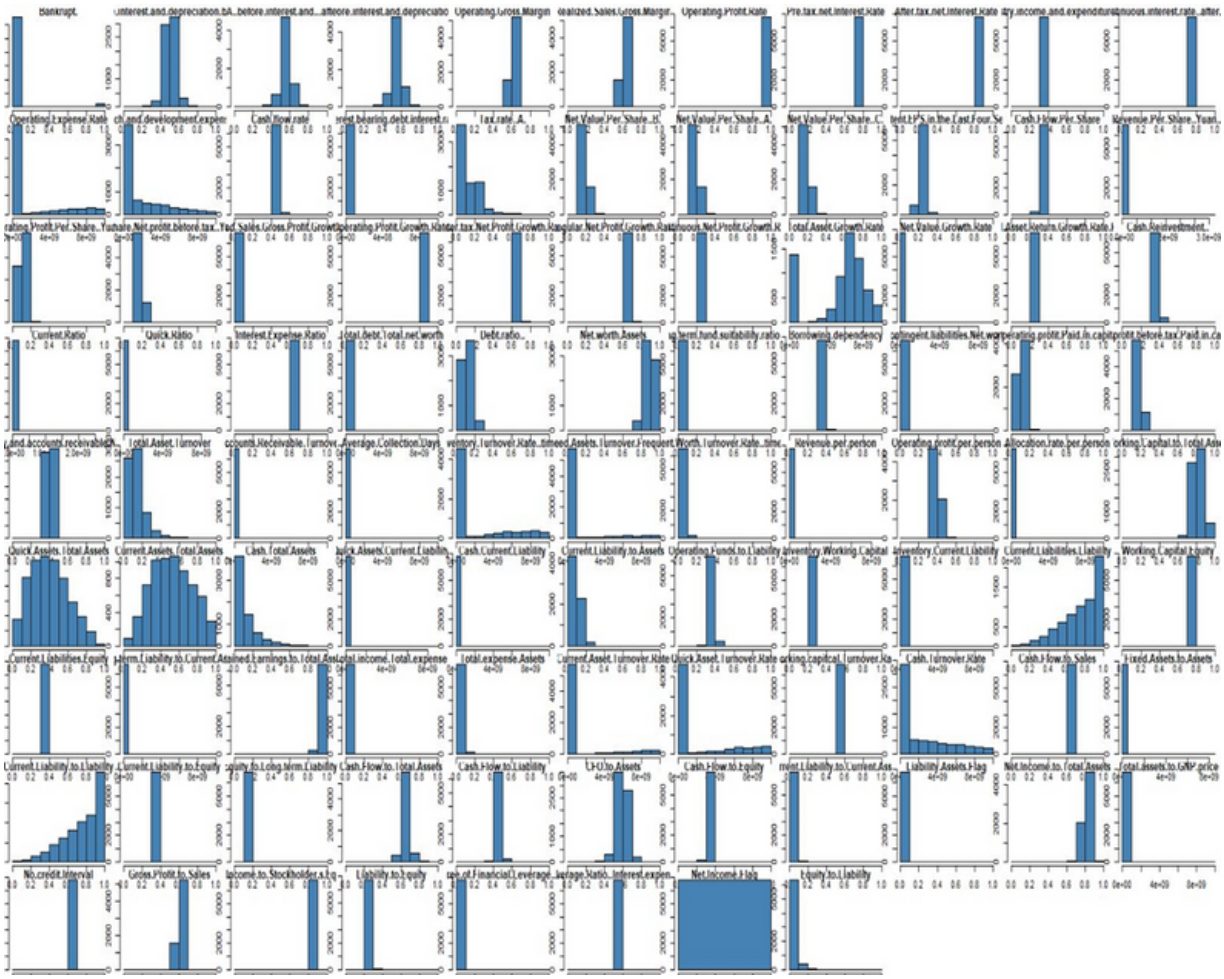
- According to the summary, there are no missing values in any of the columns.
- There are several variables with a wide range of values, such as columns 12, 13, 30, 31, 57, 64, and 68.
- Some of these variables contain outliers.
- It also displays a histogram for each variable, which gives an overview of the distribution of values.

EXPLORATORY DATA ANALYSIS: DATA VISUALISATION

Data Visualisation -

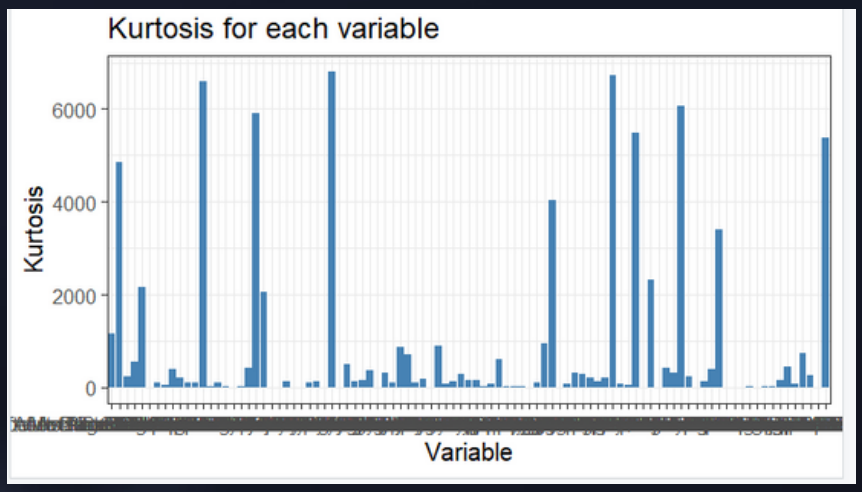
- Visualizing data allows you to gain a better understanding of the underlying patterns, trends, and relationships within the data
- Visualizations make it easier to communicate complex information and tell a compelling story.
- This interactive and exploratory approach helps you discover hidden patterns, investigate relationships, and gain a deeper understanding of the data.
- Identify errors or anomalies in the data.
- It helps transform raw data into actionable knowledge and empowers individuals and organizations to make more informed decisions.

OUTPUT OF DATA VISULISATION -

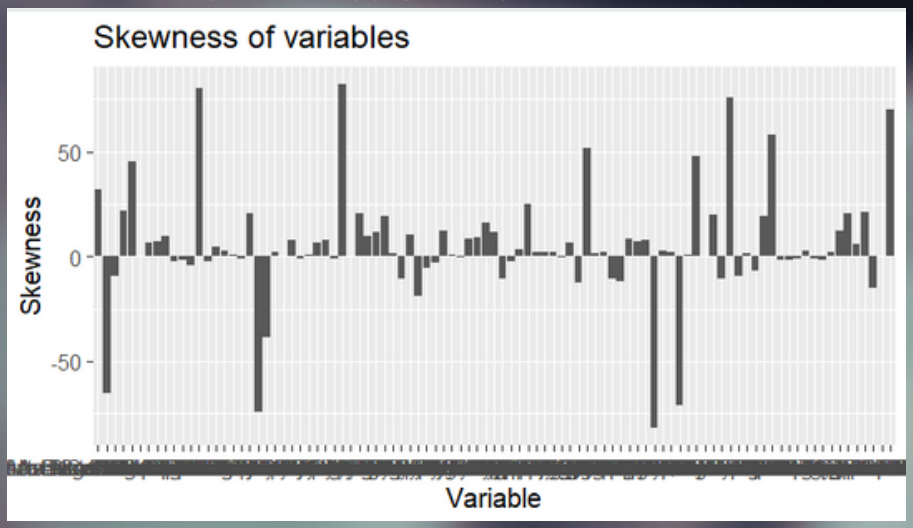


BAR GRAPHS/SCATTER PLOTS /HISTOGRAMS ETC.

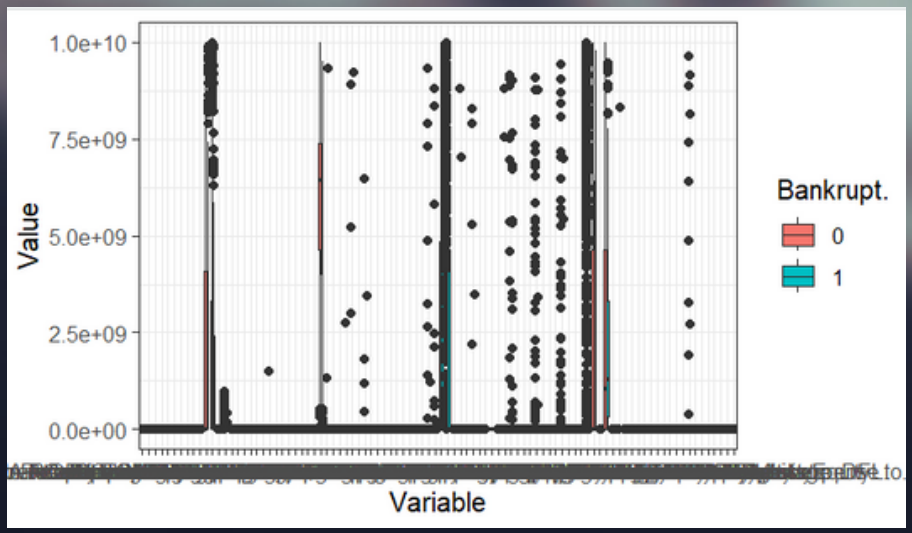
KURTOSIS OF EACH VARIABLE



SKEWNESS OF EACH VARIABLE

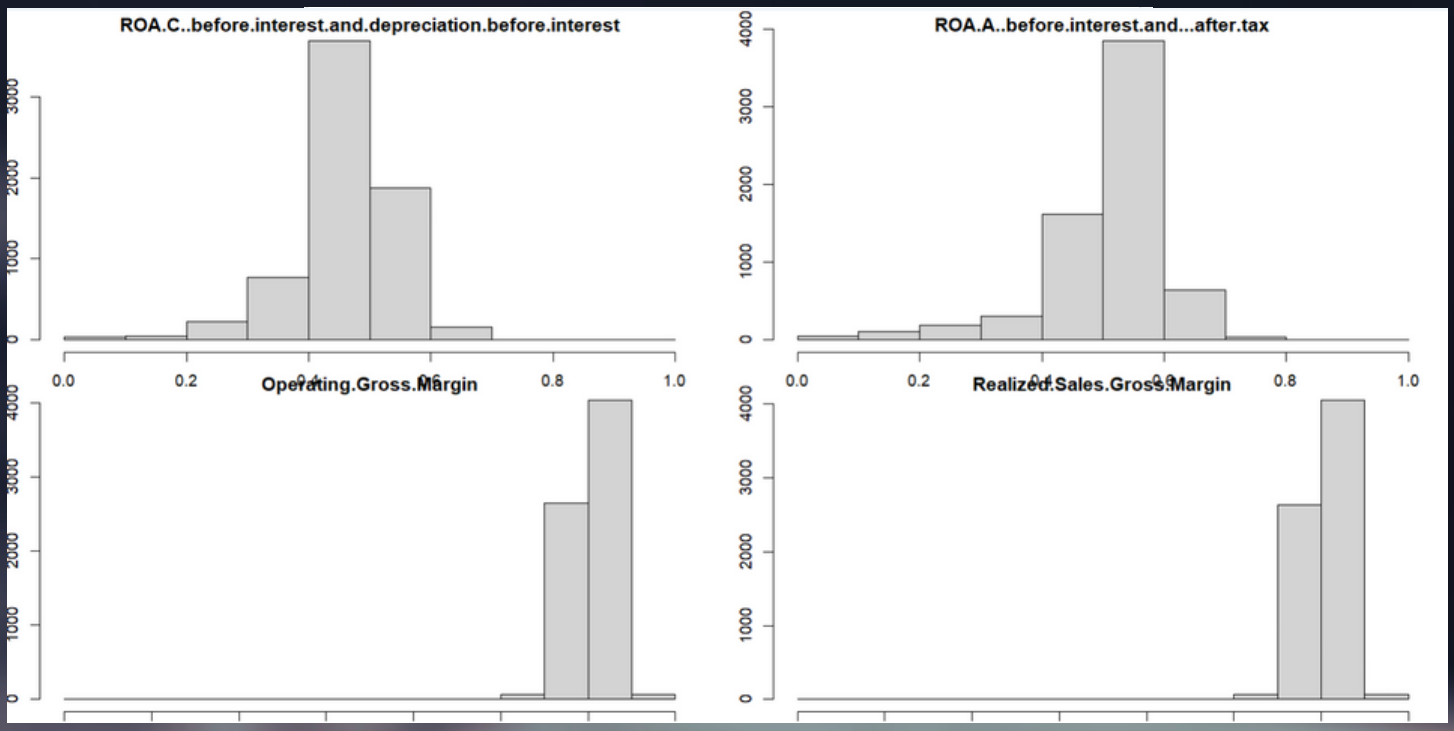


SCATTERPLOT OF EACH VARIABLE

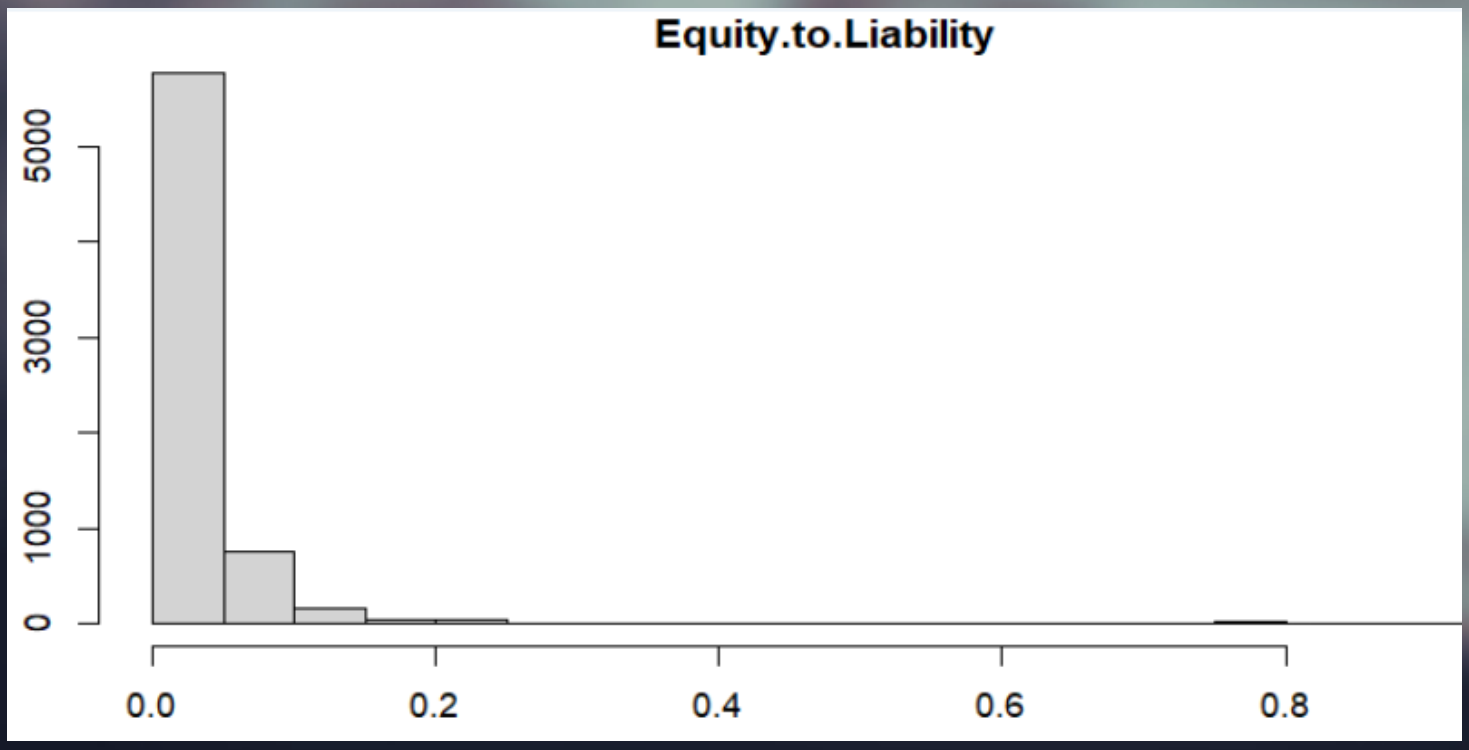


BAR GRAPHS/SCATTER PLOTS /HISTOGRAMS ETC.

HISTOGRAMS FOR THE SOME SELECTED VARIABLES



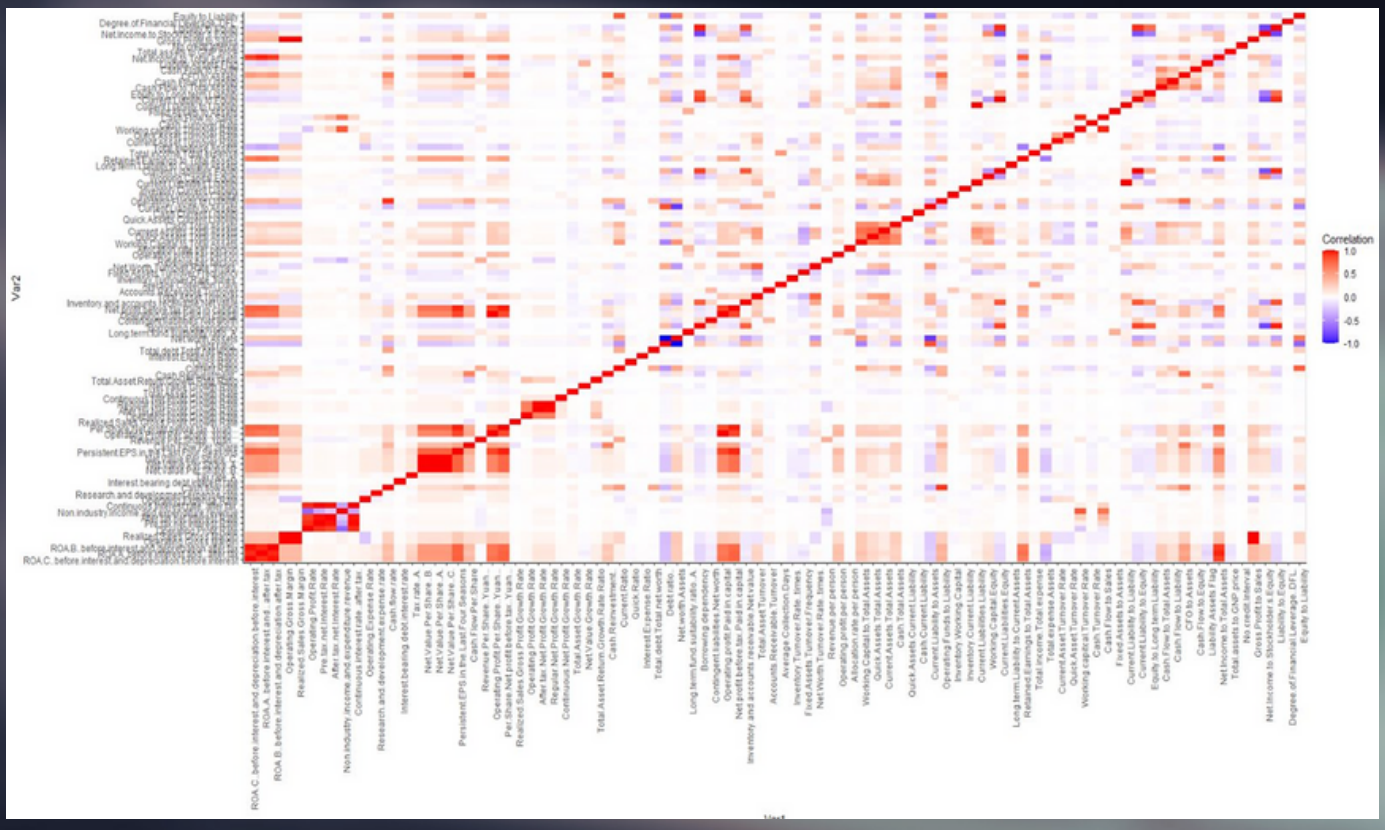
HISTOGRAM FOR EQUITY TO LIABILITY RATIO



HEATMAP

A HEATMAP IS A GRAPHICAL REPRESENTATION OF DATA WHERE THE VALUES IN A MATRIX ARE REPRESENTED AS COLORS. IT IS A WAY TO VISUALIZE AND ANALYZE PATTERNS AND RELATIONSHIPS IN THE DATA.

A HEATMAP IS A GRAPHICAL REPRESENTATION OF DATA WHERE THE VALUES IN A MATRIX ARE REPRESENTED AS COLORS. IT IS A WAY TO VISUALIZE AND ANALYZE PATTERNS AND RELATIONSHIPS IN THE DATA.
HEATMAPS ARE COMMONLY USED IN VARIOUS FIELDS, INCLUDING DATA ANALYSIS, GENOMICS, FINANCE, AND IMAGE PROCESSING.
IN A HEATMAP, EACH CELL OF THE MATRIX IS ASSIGNED A COLOR BASED ON ITS VALUE. THE COLORS TYPICALLY REPRESENT A GRADIENT FROM LOW TO HIGH VALUES, ALLOWING YOU TO EASILY IDENTIFY PATTERNS AND VARIATIONS IN THE DATA.



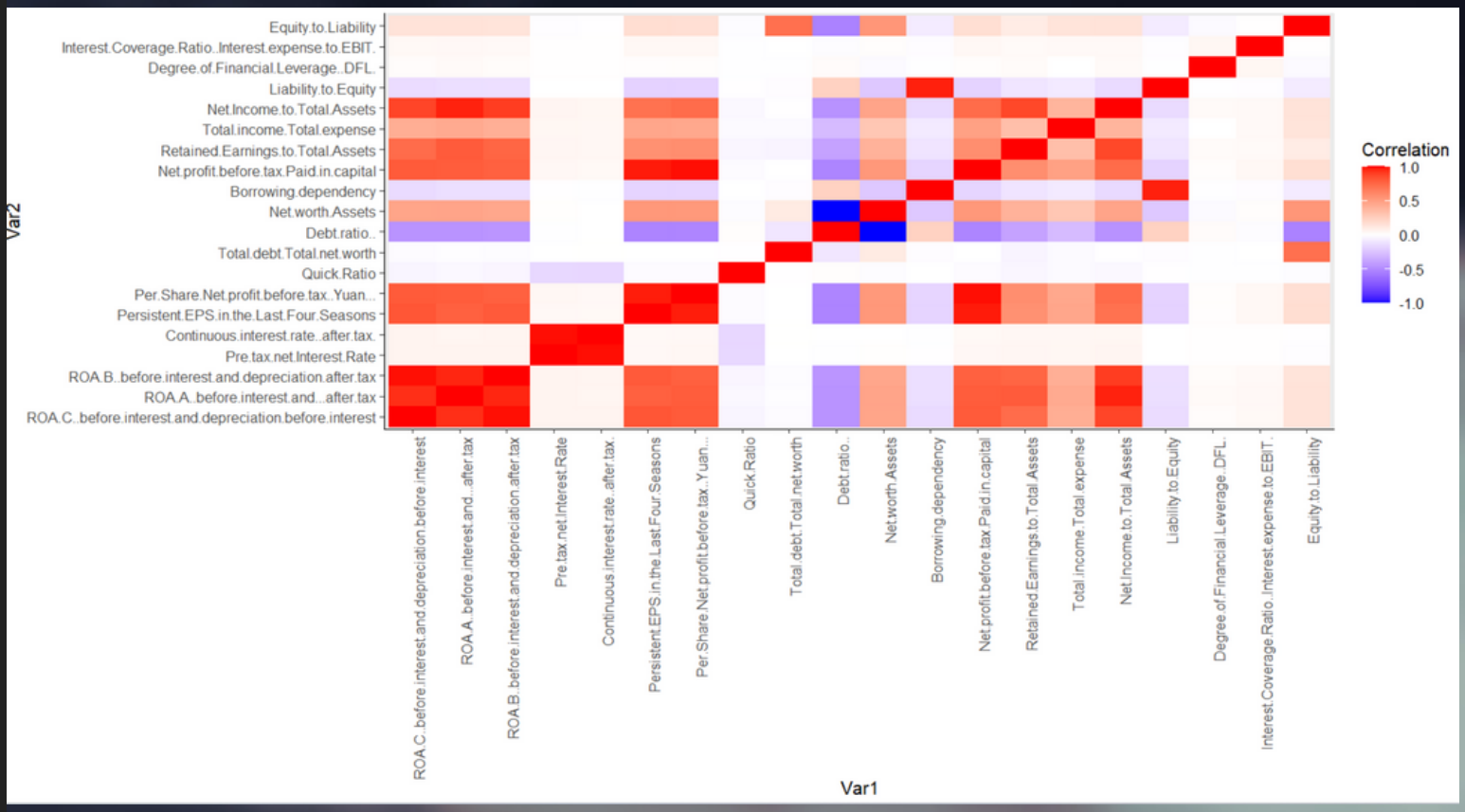
The data seems to be extremely cluttered.

SO WHAT DO WE DO NOW?

HEATMAP

From the given correlation heatmap, we can infer that the following attributes are highly correlated:

- Per Share Net Profit Before Tax
- EPS in the Last 4 seasons Net Value per Share A
- Net Value per Share B
- Net Value per Share C



We will use these most correlated variables in the LDA /QDA / KNN models

SMOTE FUNCTION

- SMOTE is an abbreviation for Synthetic Minority Oversampling Technique.
- SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line.
- The data we're working with is skewed, with higher numbers for non-bankrupt corporations than for bankrupt ones.
- As a result, we needed to oversample the infrequent observations in order to accurately predict bankrupcy using machine learning models.

```r
#Oversampling the bankrupt (response)variable
install.packages("ROSE")
library(ROSE)

# Split the data into predictors and target variable
X <- df[, !(names(df) %in% c("Bankrupt."))]
y <- df$Bankrupt.

# Apply SMOTE using ROSE
oversampled_data <- ovun.sample(Bankrupt. ~ ., data = df, method = "both", p = 0.5, seed = 123)

# Access the oversampled data
new_df <- oversampled_data$data

table(new_df$Bankrupt.)
```

RESULT OF SMOTE (OVERSAMPLING) -

```
> table(df$Bankrupt.) > table(new_df$Bankrupt.)

   0    1                    0    1
6599  220 ------------------> 3442 3377
```

METHODOLOGY

Machine Learning Algorithms Used:

- **Random Forest**
- **Logistic Regression**
- **Linear Discriminant Analysis (LDA)**
- **Quadratic Discriminant Analysis (QDA)**
- **K-Nearest Neighbors**

RANDOM FOREST

What is Random Forest?

The random forest, as the name implies, is made up of a huge number of individual decision trees that work together as an ensemble. Each individual tree in the random forest produces a class prediction, and the class with the most votes becomes the prediction of our model.

why we used random forest ?

- It performed both regression and classification tasks.
- A random forest in our case produced good predictions that could be understood easily.
- This approach handled large datasets efficiently.
- The random forest algorithm provided a higher level of accuracy in predicting outcome over the decision tree algorithm
- It helped us identify important variables for the prediction of bankrupcy

METHODOLOGY:

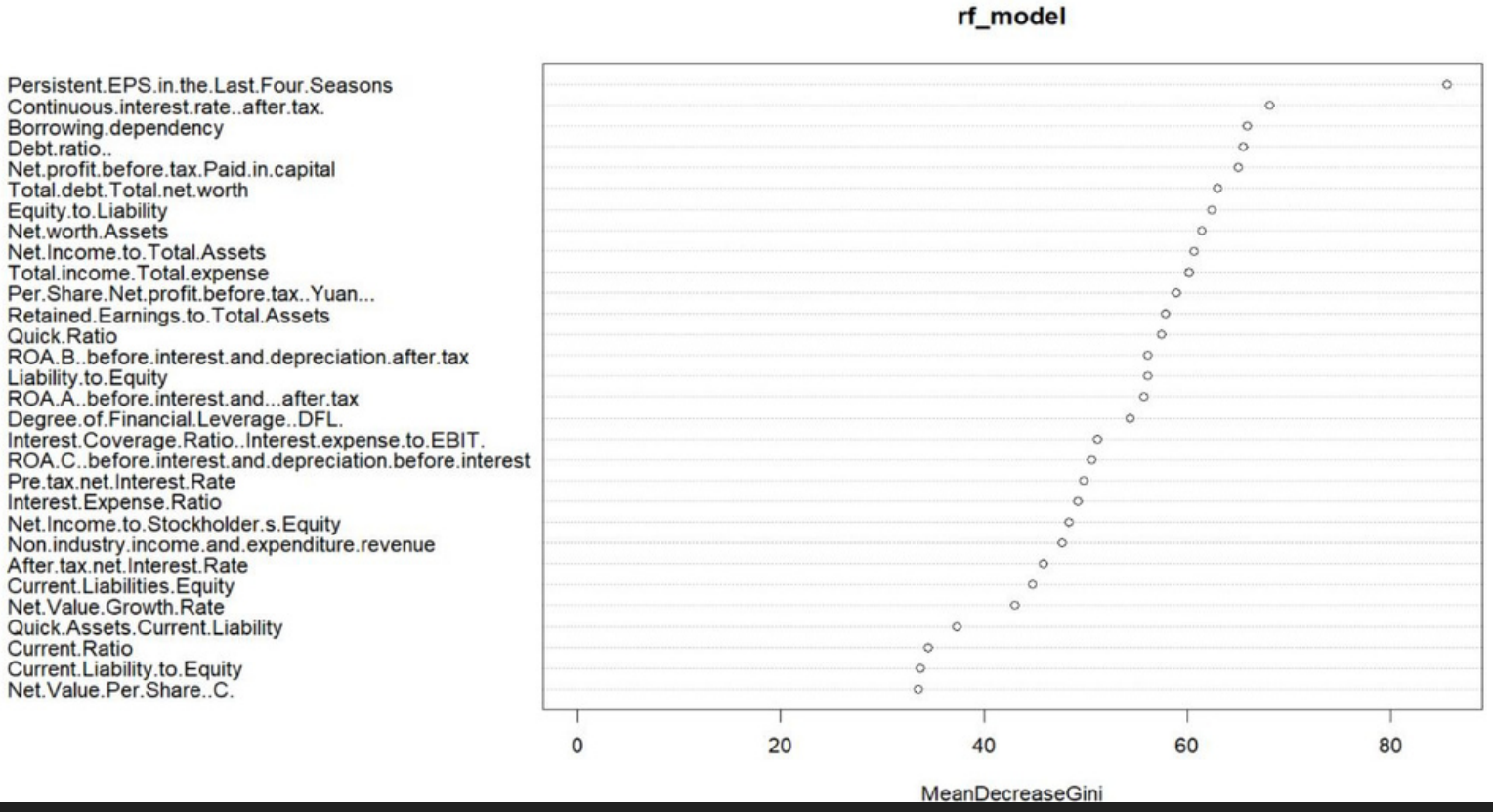
Random Forest:

- Rf model divided the data set into training and testing sets
- We then set the seed
- Post setting seeds, we created the random forest model
- Predicted the target variables
- Extracted most important features using importance function
- Sorted the important features in decreasing order
- Created a list of top 15 important functions that we will use in our Machine Learning Models

```
#Creating the model
rf_model <- randomForest(formula, data = train_set, ntree = 800, mtry = 3)
```

OUTPUT:

Implementing a Random Forest model to list out the 20 most important features from our dataset



ANALYSIS:

- Implementing a Random Forest model to list out the 20 most important features from our dataset
- The top 20 features, ranked by their importance measures, significantly impact the prediction in descending order of importance.
- The model suggests that among the 15 variables, persistent EPS in the last 4 seasons is most important while net value per share is least.

LOGISTIC REGRESSION

What is Logistic Regression?

Logistic regression is a statistical model used to predict the probability of a binary outcome based on one or more predictor variables.t is commonly used for classification tasks where the dependent variable is categorical and has two levels (e.g., yes/no, true/false, success/failure).

why we used Logistic Regression ?

- Logistic regression is particularly useful for bankruptcy prediction because it can handle non-linear relationships between the predictors and the dependent variable. It is also robust to outliers and can handle missing data, which are common issues in financial data.
  Hence, we used Logistic Regression.

METHODOLOGY:

Logistic Regression:

- Used the top 15 important variables in LR model
- From random forest, we then created a subset of the bankrupt variable.
- With the above 15 features that we extracted, we then created a logistic regression model

OUTPUT:

```
Call:
glm(formula = Bankrupt. ~ ., family = "binomial", data = train_subset,
    weights = weights)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.2640  -0.4218  -0.0005   0.5003   2.4460

Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                             -9.636e+00  3.612e+00  -2.668  0.00764 **
Net.Income.to.Stockholder.s.Equity       3.678e+01  2.895e+00  12.703  < 2e-16 ***
Persistent.EPS.in.the.Last.Four.Seasons -8.008e+01  9.409e+00  -8.511  < 2e-16 ***
Net.Value.Growth.Rate                   -2.608e-10  4.217e-10  -0.618  0.53626
Net.Income.to.Total.Assets              -2.104e+00  1.301e+00  -1.617  0.10585
Net.profit.before.tax.Paid.in.capital   -2.833e+01  1.802e+01  -1.572  0.11594
Per.Share.Net.profit.before.tax..Yuan...  3.081e+01  1.403e+01   2.196  0.02810 *
Equity.to.Liability                      5.656e+00  7.891e-01   7.167 7.64e-13 ***
Net.Value.Per.Share..B.                 -2.859e+02  5.031e+01  -5.682 1.33e-08 ***
Degree.of.Financial.Leverage..DFL.       3.004e+00  2.083e+00   1.442  0.14921
Interest.Expense.Ratio                   4.622e+00  2.601e+00   1.777  0.07552 .
Borrowing.dependency                     3.948e+01  2.725e+00  14.486  < 2e-16 ***
Net.Value.Per.Share..C.                  2.799e+02  5.002e+01   5.595 2.21e-08 ***
Net.worth.Assets                        -2.372e+01  1.098e+00 -21.605  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7562.1  on 5454  degrees of freedom
Residual deviance: 3525.6  on 5441  degrees of freedom
AIC: 3553.6

Number of Fisher Scoring iterations: 7
```

```
> calculate_metrics(model, test_subset)
[1] "F1Score:  0.869627507163324"
          Reference
Prediction   0   1
         0 607  88
         1  94 575
[1] "Accuracy: 0.87"
```

Analysis:

- The accuracy we received for the LR model was 0.87
- The f1 score is 0.869
- The non-bankrupt companies have been predicted correctly 575 times and the bankrupt companies have been predicted correctly 607 times
- These are the non-significant variables obtained from regression Net.Income.to.Total.Assets, Net.profit.before.tax.Paid.in.capital (p-value = 0.11594) Degree.of.Financial.Leverage..DFL. (p-value = 0.14921) and Interest.Expense.Ratio

REGRESSION AFTER REMOVING NON-SIGNIFICANT VARIABLES:

INPUT

```
model2 <- glm(Bankrupt. ~
                Net.Income.to.Total.Assets +
                Net.profit.before.tax.Paid.in.capital +
              Per.Share.Net.profit.before.tax..Yuan... +
                Degree.of.Financial.Leverage..DFL. +
                Interest.Expense.Ratio,
              family = "binomial",
              data = train_subset2)
```

OUTPUT

```
Call:
glm(formula = Bankrupt. ~ Net.Income.to.Total.Assets + Net.profit.before.tax.Paid.in.capital +
    Per.Share.Net.profit.before.tax..Yuan... + Degree.of.Financial.Leverage..DFL. +
    Interest.Expense.Ratio, family = "binomial", data = train_subset2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.8089  -0.7168  -0.0005   0.7430   2.5451

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                               12.172      1.677   7.257 3.95e-13 ***
Net.Income.to.Total.Assets                 2.788      1.149   2.426  0.01525 *
Net.profit.before.tax.Paid.in.capital   -136.558     12.516 -10.910  < 2e-16 ***
Per.Share.Net.profit.before.tax..Yuan...   31.014     10.981   2.824  0.00474 **
Degree.of.Financial.Leverage..DFL.          9.001      3.144   2.863  0.00420 **
Interest.Expense.Ratio                      4.652      2.479   1.876  0.06064 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7562.1  on 5454  degrees of freedom
Residual deviance: 4698.3  on 5449  degrees of freedom
AIC: 4710.3

Number of Fisher Scoring iterations: 6
```

```
> calculate_metrics(model2, test_subset2)
[1] "F1Score:  0.812411847672779"
          Reference
Prediction   0   1
         0 576 141
         1 125 522
[1] "Accuracy: 0.8"
```

LINEAR DISCRIMINANT ANALYSIS (LDA)

What is LDA?

LDA is a classification algorithm that projects the input features onto a lower-dimensional space, while maximizing the separation between the classes. It assumes normally distributed data with equal covariance matrices and is often used for feature extraction and dimensionality reduction.

why we used LDA ?

LDA is effective for bankruptcy prediction as it helps to identify key variables that differentiate between bankrupt and non-bankrupt companies, enabling a better understanding of factors contributing to financial distress and the development of more accurate prediction models.

METHODOLOGY:

LDA:

- Utilized the most correlated variables to create the LDA model.
- To calculate the heatmap, we removed variables with near-zero variance.
- Calculated the correlation matrix.
- Installed the reshaped library.
- Melted the correlation matrix.
- Plotted the heatmap
- Found the highly correlated variables
- Created a subset of highly correlated variables
- Created the LDA model with these variables Computed the confusion matrix, f1 score and accuracy

OUTPUT:

```
lda_class   0   1
        0 680 366
        1  21 297
> print(paste("F1 score:", round(f1_score, 2)))
[1] "F1 score: 0.61"
> accuracy
[1] 0.7162757
```

Analysis:

- The accuracy we received for the LDA model was 0.7162
- The f1 score is 0.604
- Recall: 0.934
- Precision: 0.448
- The non-bankrupt companies has been predicted correctly 297 times and the bankrupt companies has been predicted correctly 680 times

QUADRATIC DISCRIMINANT ANALYSIS (QDA)

What is QDA?

QDA is a classification algorithm that assumes normally distributed data with different covariance matrices per class. It finds a quadratic decision boundary that best separates the classes, allowing for more flexible modeling but requiring more parameters to be estimated.

why we used QDA ?

QDA is preferred in bankruptcy prediction due to its ability to handle non-linear relationships between variables and to estimate class probabilities, providing a useful tool for risk assessment in financial data analysis.

METHODOLOGY:

QDA:

- Extract the top 6 features list from the sorted features importance from the random forest model
- We formulated a heatmap from these 20 features and then extracted 6 features that were highly correlated, and omitted them due to muti-collinearity
- Created a subset of these 6 important variables and formulated a QDA model
- Extracted predictions and confusion matrix
- Calculated accuracy and f1 score

OUTPUT:

```
> accuracy
[1] 0.5945748
> # Print F1 score
> print(paste("F1 score:", f1_score))
[1] "F1 score: 0.295541401273885"
> conf_mat

qda_class   0   1
        0 695 547
        1   6 116
>
```

Analysis:

- The QDA model achieves an F1 score of 0.295541401273885, indicating its performance in balancing precision and recall.
- The accuracy of the model is reported as 0.5945748.
- The precision and recall values can be calculated based on the confusion matrix, as described above.
- The non-bankrupt companies has been predicted correctly 116 times and the bankrupt companies has been predicted correctly 695 times

K - NEAREST NEIGHBORS

What is KNN?

KNN is a non-parametric machine learning algorithm for classification and regression tasks. It works by finding the k nearest neighbors in the training set to a new input point, and making a prediction based on the majority or average value of its neighbors.

why we used KNN?

KNN can be useful in bankruptcy prediction because it does not assume any underlying distribution of the data, which is especially important when dealing with financial data that may not follow a normal distribution

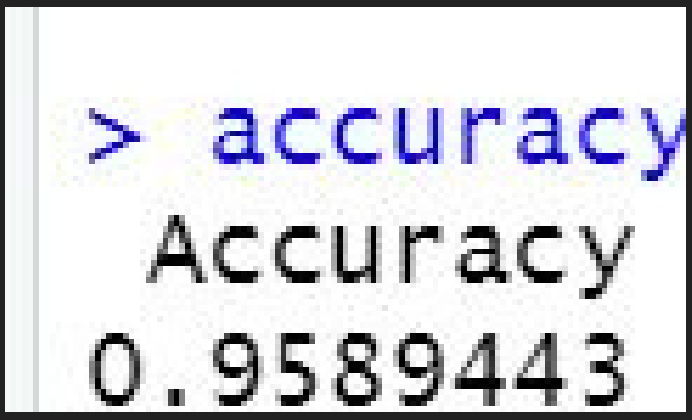
METHODOLOGY:

KNN, K = 2:

- Utilized the same 6 features for the KNN model
- Extracted predictions and confusion matrix
- Calculated accuracy and f1 score
- Tried using different values of K and checked the accuracy which was almost the same for every k value

OUTPUT:

```
> confusion_matrix
Confusion Matrix and Statistics

                Reference
Prediction      Not Bankrupt Bankrupt
  Not Bankrupt           645        0
  Bankrupt                56      663

               Accuracy : 0.9589
                 95% CI : (0.947, 0.9688)
    No Information Rate : 0.5139
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.918

 Mcnemar's Test P-Value : 1.987e-13

            Sensitivity : 0.9201
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.9221
             Prevalence : 0.5139
         Detection Rate : 0.4729
   Detection Prevalence : 0.4729
      Balanced Accuracy : 0.9601

       'Positive' Class : Not Bankrupt
```



Analysis:

- The accuracy of the model is 0.9589, which indicates that it correctly predicts the class of 95.89% of the instances in the test set.
- The 95% confidence interval for the accuracy is (0.947, 0.9688), which means that we can be 95% confident that the true accuracy of the model falls within this range.
- The NIR represents the accuracy achieved by always predicting the majority class. In this case, the NIR is 0.5139, which indicates that the majority class ("Not Bankrupt") occurs in 51.39% of the instances. The KNN model significantly outperforms the NIR.

KNN, K = 8:

OUTPUT:

```
> f1_score
Pos Pred Value
     0.8787879
```

```
> accuracy
Accuracy
0.888563
```

```
> confusion_matrix
Confusion Matrix and Statistics

                Reference
Prediction      Not Bankrupt Bankrupt
  Not Bankrupt           551        2
  Bankrupt               150      661

               Accuracy : 0.8886
                 95% CI : (0.8707, 0.9048)
    No Information Rate : 0.5139
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7783

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7860
            Specificity : 0.9970
         Pos Pred Value : 0.9964
         Neg Pred Value : 0.8150
             Prevalence : 0.5139
         Detection Rate : 0.4040
   Detection Prevalence : 0.4054
      Balanced Accuracy : 0.8915

       'Positive' Class : Not Bankrupt
```

Analysis:

- The accuracy of the KNN model is 0.8886, which means it correctly predicted the class of approximately 88.86% of the samples.

- The specificity is approximately 0.9970, indicating that the model can identify non-bankrupt companies with a high accuracy.

- The precision for the "Bankrupt" class is approximately 0.8788, indicating that when the model predicts a company as "Bankrupt," it is correct around 87.88% of the time.

- The KNN accuracy decreased when we increased the k value from 2 to 8, because of the bias in our dataset

RESULT TABLE

Algorithm Implemented	Accuracy	F1 Score
Logistic Regression	0.87	0.87
Logistic Regression (After removing non-significant Variables)	0.8	0.81
Linear Discriminant Analysis	0.71	0.61
Quadratic Discriminant Analysis	0.59	0.29
K- Nearest Neighbors (K = 2)	0.95	0.95
K- Nearest Neighbors (K = 8)	0.88	0.87

CONCLUSION

As we can see, The KNN model comes out as the best model to predict bankruptcy in companies. Hence, we recommend to use the KNN model as an answer to our research question

- Some important challenges to our project were feature selection, dimension reduction, data preprocessing, and model selection.
- Oversampling the bankrupt variable
- Possibility of correlation between various dimensions
- In-depth homogenization would occur for some features with higher correlation, making it really difficult to conduct a thorough analysis of the data and creating a challenge to defining features
- As the dimension increases, the possible combinations of clusters grow, and the clustering will be hard to define

Clinton Nwokike - Challenges, Table of Data Description in final and progress report

Tashveen Kaur - Coding, Progress Reports, Presentation slides, Final Report

Abhinav Ganguly - Coding,
Progress Reports, Presentation slides, Final Report

REFERENCES

Liang, D., Lu, C.-C., Tsai, C.-F., and Shih, G.-A. (2016) Financial Ratios and Corporate Governance Indicators in Bankruptcy Prediction: A Comprehensive Study. European Journal of Operational Research, vol. 252, no. 2, pp. 561-572. https://www.sciencedirect.com/science/article/pii/S0377221716000412