Guoyu Dong, Zack He, Haoyu Wang

**Bankruptcy Alert**

**Business Problems and Hypothesis**

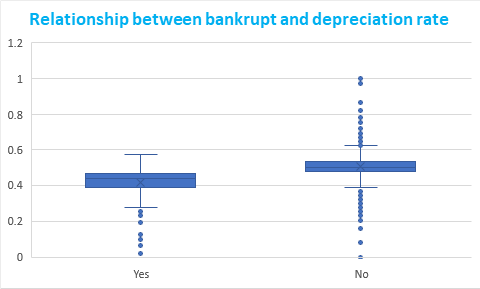
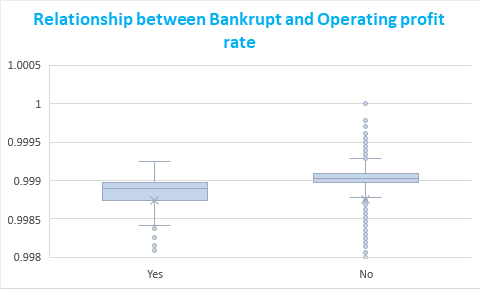
For businesses, the most crucial thing is making profits, while making sure performing with a bottom line in mind. But when they stop making profits and enter the realm of losses to a certain degree when all hopes are lost, they are facing the ultimate nightmare -- bankruptcy. Granted, there are numerous factors that will contribute to a firm’s success, there are also numerous factors that will lead to a firm’s demise. Financial terms such as Return on Assets (ROA), Current Ratio, Debt Ratio, and Total Assets Turnover can all be indicators for businesses and investors to predict the future performance of that company. However, with so many useful indicators at hand, is there a list of factors that play a more important and decisive role in a firm’s performance? Or, outside of the numbers, are there things that cannot be reflected on paper, such as human factors, that will result in the eventual bankruptcy of a firm? The goal for this report then is to build data analytics models to predict, to our best capabilities, that if a company will go bankrupt given certain criteria. Our initial hypothesis is that there will be some factors that heavily impact financial health of companies and it might be a combination of factors that work together to reflect a firm’s performance, in other words, these factors will decide if a company will go bankrupt. Additionally, we wanted to see what types of companies will have a higher chance to be bankrupt, for example, it might be due to things like, greater competition, lower entry barriers, and higher cost. Therefore, we expect that a knn classification can help us predict it.

**Dataset**

To start our analysis, we found a dataset collected from the Taiwan Economic Journal for the years 1999 to 2009, which was released publicly on Kaggle, an online community of data scientists and machine learning practitioners (the dataset can be found here: <https://www.kaggle.com/fedesoriano/company-bankruptcy-prediction>). Company bankruptcy in the dataset was defined based on the business regulations of the Taiwan Stock Exchange. There are more than 6,000 observations and a total of 96 columns, including the response variable “Bankrupt”. Before we started building our models, we examined the dataset and found that three of the variables did not fit with our analysis, which were: Research and development expense rate, tax rate, and total asset growth rate, because the value in these column are either extremely large or 0, which might cause negative effects to our model. After eliminating the unwanted variables, we changed the data type of our response variable, Bankrupt, from numeric into binary form, so that we can correctly run the intended models which we will discuss in detail in the next section. For better accuracy and re-examination on our model, we also choose to split the data into training, which consists of 6,500 observations, and validation, which has the remaining 319 observations, parts. Doing these steps allowed us to perform further analysis and improve the quality of our model.

**Descriptive Analysis**

We did some descriptive analysis for our datasets. At the beginning, we had great doubts about our factors, because we were worried that some variables could not affect whether bankrupt or not. So we created boxplots to check the relationship between the response variable, Bankrupt, and variables that we have doubts about. Based on our results, we know that the depreciation rate is relatively highly correlated with bankruptcy because the average depreciation rate for Bankrupt is much lower than the depreciation rate for not bankrupt. Additionally, we know that the operating profit rate has a positive relationship with the bankrupt variable. After descriptive analysis, we decide to keep all variables to begin our data analysis works.

** **

**Methodologies**

**Model 1: Subset Selection**

To start our analysis, we implemented the subset selection method, trying to find the set of features or variables that would contribute to the model we were seeking, which is to prove that there is a certain set of criteria for us to predict if a company would go bankrupt. However, as we were doing it, we started to encounter problems. For example, since there are 93 variables we were using, best subset method was not returning the most accurate results, because there were too many variables for R to run. So we had to discard it and continue with forward and backward stepwise selection methods. We set the maximum number of variables to be included at 50, because we thought it would be a good place to begin with. Then, we tried to find the adjusted r^2 for both forward and backward methods. However, for both methods, the highest adjusted r^2 was around 20%, which is not really significant. After getting this result, we suddenly realized that since our response variable is binary, adjusted r^2 does not mean anything, and so is the linear regression model. Instead, we should be running logistic regression models.

**Model 2: Logistic Regression**

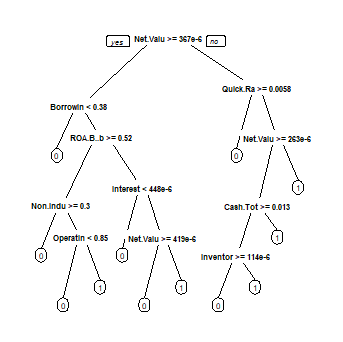
The next model we decided to run was the logistic regression, which is a very important model in the data mining work. We used glm function to create two regression models (full regression and null regression) in order to do stepwise analysis. Then , we used the step function to generate a logistic regression for us. The final model includes 23 variables. 19 out of 23 variables are individual variables and 2 out of 23 are interaction variables and the rest of two are quadratic variables. Then we tested our model by creating the final model in our validation dataset and then, we found that 18 out of 23 variables are statistically significant, which means that most variables contribute to explain our target variable, bankruptcy. Our final model r-squared is relatively low (0.38), and we believe that the main reason for that is our dataset included over 90 variables and some variables are highly correlated with other variables. Additionally, our validation dataset only includes 300 records, so there is a certain possibility that some important variables become insignificant variables in our final test model.

**Model 3: K-NN Classification**

The next model we wanted to run was the knn classification, which is a part of unsupervised learning that will help us predict if a company will go bankrupt or not. In order to get the same results, we used the set.seed(1) function at first. Then, we split the data in half randomly, with half of them being the test data and the other being the training data. We picked all the rows and columns except for Bankrupt and stored them as training and test datasets separately, and also stored the vector of Bankrupt (which is the response variable) into training and test datasets. After all the preparations were done, we were able to use the knn functions to calculate the results. Since there are only two possible outcomes, i.e., bankrupt or not, we believe that we should use k=2 for this model, which gave us a prediction accuracy of about 94.5%.

**Model 4: Decision Tree**

Our last model is the decision tree model. We implemented a decision tree to test which variables contribute the most to explain our target variable. We set up the maximum depth equal to 5 because we believe if we did not set up maximum depth, our model would become overwhelming and would not provide too much valuable information for us. Based on our result, we can know that when the borrowing dependency rate is smaller than 0.38 and operating profit rate is smaller than 0.85, it is more easy to get bankrupt. The decision tree provided some key criteria for us to judge bankrupt or not quickly.



**Results and Conclusion**

Based on our four model results, we decided to use the logistic regression model as our selected model to explain bankruptcy. Our variables include Persistent EPS in the Last Four Seasons: EPS-Net Income, Debt Ratio, Total Assets, Net Value Growth Rate, Cash Turnover Rate, Cash Flow to Liability, Accounts Receivable Turnover, Operating profit per person, Accounts Receivable Turnover, Operating profit per person, Total debt, Fixed Assets to Assets, Per Share Net profit before tax, Inventory and Accounts Receivable Net Value, Current Assets Over Total Assets, Borrowing dependency: Cost of Interest-bearing Debt, before interest and depreciation before interest: Return On Total Assets, Cash Reinvestment Ratio, Liability to Equity Ratio, Cash Flow Per Share, Liability to Asset Flag, Working Capital over Equity, Operating Profit Rate, Continuous Interest Rate After Tax, Cash Flow Rate, and Quick Ratio.

Combining with our decision tree model, we can conclude more details that when the borrowing dependency rate is smaller than 0.38, operating profit rate is smaller than 0.85 and cash reinvestment ratio is larger than 0.013, it is easier to get bankrupt. It seems that the models work well together and they can help us accomplish our goal.

However, as we were delving into the analysis, we realized that, like all other analysis with mathematical models, there are missing factors that cannot be explained through data, for example, human actions cannot be quantified into data numbers and thus cannot be expressed in our analysis. Even though we obtained a logistic regression model that seems to answer our hypothesis, we did not account for the human factors, which can shift the results dramatically. After all, if an omnipotent model exists for solving bankruptcy, the financial world will be so much different now.

Appendix - Coding

library(dplyr)

library(rpart)

library(pROC)

library(ggplot2)

library(rattle)

library(caret)

library(rpart.plot)

library(nnet)

data1 <- read.csv("data.csv",na.strings = "?")

summary(data1)

library(tidyverse)

## Changing data type for dependent variable

data1$Bankrupt. <- as.factor(data1$Bankrupt.)

summary(data1$Bankrupt.)

## Splitting into training and validation dataset

data\_training <- data1[c(1:6500),]

data\_validation <- data1[c(6501:6819),]

## Model 1: Subset selection

library(ISLR)

library(leaps)

# regfit.full=regsubsets(Bankrupt.~., data = data\_training,nvmax=20,really.big=T)

# summary(regfit.full)$adjr2

regfit.forward=regsubsets(Bankrupt.~., data = data\_training,nvmax=50,really.big=T,method="forward")

summary(regfit.forward)$adjr2

regfit.backward=regsubsets(Bankrupt.~., data = data\_training,nvmax=20,really.big=T,method="backward")

summary(regfit.backward)$adjr2

## Model 2 Logistic Regression

fullreg <- glm(Bankrupt.~.,data=data\_training,family = "binomial")

nullreg <- glm(Bankrupt.~1,data = data\_training,family = "binomial")

model1 <- step(nullreg, list(lower = formula(nullreg), upper = formula(fullreg)),

data = data\_training, direction = "both", trace = 0)

f1 <- formula(model1)

model2 <- glm(Bankrupt.~(Persistent.EPS.in.the.Last.Four.Seasons + Debt.ratio.. +

Net.Worth.Turnover.Rate..times. + Cash.Total.Assets + Net.Value.Per.Share..B. +

Net.Value.Per.Share..C. + Cash.Turnover.Rate + Cash.Flow.to.Liability +

Accounts.Receivable.Turnover + Operating.Profit.Per.Share..Yuan.?.. +

Fixed.Assets.Turnover.Frequency + Total.debt.Total.net.worth +

Per.Share.Net.profit.before.tax..Yuan.?.. + Fixed.Assets.to.Assets +

Inventory.and.accounts.receivable.Net.value + Current.Assets.Total.Assets +

Borrowing.dependency + ROA.C..before.interest.and.depreciation.before.interest +

Cash.Reinvestment.. + Liability.to.Equity + Cash.Flow.Per.Share +

Liability.Assets.Flag + Working.Capital.Equity + Operating.Profit.Rate +

Continuous.interest.rate..after.tax. + Cash.flow.rate + Quick.Ratio),data = data\_training,family = "binomial")

summary(model2)

## Model 3: knn Classification

library(class)

attach(data1)

set.seed(1)

train = sample(1:nrow(data1), nrow(data1)/2)

company.train = data1[train,-1]

company.test = data1[-train,-1]

Bankrupt.train = Bankrupt.[train]

Bankrupt.test = Bankrupt.[-train]

knn.pred.1 = knn(company.train,company.test,Bankrupt.train,k=1)

table(knn.pred.1,Bankrupt.test)

mean(knn.pred.1==Bankrupt.test)

knn.pred.2 = knn(company.train,company.test,Bankrupt.train,k=2)

table(knn.pred.2,Bankrupt.test)

mean(knn.pred.2==Bankrupt.test)

knn.pred.3 = knn(company.train,company.test,Bankrupt.train,k=3)

table(knn.pred.3,Bankrupt.test)

mean(knn.pred.3==Bankrupt.test)

## Model 4: Decision Tree

tree1 <- rpart(Bankrupt.~.,data= data1,control = rpart.control(maxdepth=5,cp=0.0001))

rpart.plot(tree1)

prp(tree1)