# Assignment 4: Big Data Analytics
# Survival Analysis
# By: Valay Shah

**Purpose**

The purpose of this report is to present two models to the firm, one with logistic regression to see how some financial variables impact a company's probability of default, and another model to predict the expected time to bankruptcy for any given company.

**Logistic Regression – Probability of Default**

The first thing to do when determining a logistic regression was to find the relevant variables which might have an impact on the probability of default for any given company. For this reason, the following variables were chosen: Net_Debt_EBITDA, Profit_Margin, Current_Ratio, Debt_to_Cash_Flow_From_Ops, and EBITDA. The reason these five variables were chosen was because they relate to bankruptcy by being variables that are profitability related (Profit Margin, EBITDA), or liquidity/solvency related (Net_Debt_EBITDA, Current_Ratio, Debt_to_Cash_Flow_From_Ops) which have a direct bearing on the going concern of a company.

Once these variables were finalized, summary statistics for these variables were pulled:

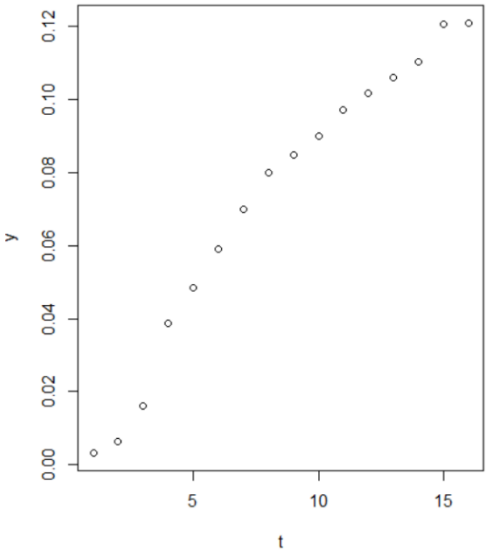
```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -2.003e+00  7.157e-02 -27.987  < 2e-16 ***
Net_Debt_EBITDA          3.652e-02  6.060e-03   6.026 1.68e-09 ***
Profit_Margin           -1.177e+00  2.506e-01  -4.698 2.63e-06 ***
Current_Ratio           -3.389e-02  1.883e-02  -1.800   0.0718 .
Debt_to_Cash_Flow_From_Ops -1.293e-03  1.258e-03  -1.028   0.3042
EBITDA                  -1.184e-04  6.173e-05  -1.917   0.0552 .
```

From the p-values we can observe that only Debt_to_Cash_Flow_From_Ops has a large p-value but besides that all other variables are strong indicators of determining whether a company will go bankrupt or not. From the coefficients we can observe that the higher the ratio for debt to EBITDA, the higher the chance of a company going bankrupt which makes sense as more debt to earnings will lead a company to spiral into bankruptcy faster. Along with this, we also see that higher profit margins and a higher EBITDA will lead to a lower chance of a company going bankrupt which also makes sense as the more profitable a company is, the less the chance of the company going bankrupt. Lastly, we can observe that a higher current ratio will cause a company to reduce its chances of going bankrupt which once again passes the eye test because a higher current ratio means a company has sufficient means to meet their current obligations. These 4 variables, Net Debt to EBITDA, Profit Margin, Current Ratio, and EBITDA can therefore be used to predict whether a company will go bankrupt with good accuracy. From this a confusion matrix can be utilized to produce a model which predicts which companies which are bound to go bankrupt (3rd page appendix).

**Survival Analysis – Expected Time to Bankruptcy**

For survival analysis, there are three methodologies which can be implemented. These are the exponential, weibull, and cox methods. For the purposes of this report, the cox methodology was utilized as the cox model is the least restrictive of the three models and it can be applicable to almost any dataset as it is applicable for usage for the widest class of distributions.

Using the cox model, we can plot the companies that go into bankruptcy (y-axis) with the years it takes for the companies to go bankrupt (x-axis). The log of y is taken to put together a survival analysis plot. This can be summarized in the plot below (left):




| | coef | exp(coef) | se(coef) | z | Pr(>\|z\|) | |
|---|---|---|---|---|---|---|
| Net_Debt_EBITDA | 2.326e-02 | 1.024e+00 | 2.812e-03 | 8.270 | < 2e-16 | *** |
| Profit_Margin | -1.047e+00 | 3.508e-01 | 1.781e-01 | -5.881 | 4.08e-09 | *** |
| Current_Ratio | -4.176e-02 | 9.591e-01 | 1.893e-02 | -2.206 | 0.0274 | * |
| Debt_to_Cash_Flow_From_Ops | -7.229e-04 | 9.993e-01 | 1.021e-04 | -7.082 | 1.42e-12 | *** |
| EBITDA | -1.237e-04 | 9.999e-01 | 6.181e-05 | -2.001 | 0.0454 | * |

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

| | exp(coef) | exp(-coef) | lower .95 | upper .95 |
|---|---|---|---|---|
| Net_Debt_EBITDA | 1.0235 | 0.977 | 1.0179 | 1.0292 |
| Profit_Margin | 0.3508 | 2.850 | 0.2474 | 0.4974 |
| Current_Ratio | 0.9591 | 1.043 | 0.9242 | 0.9953 |
| Debt_to_Cash_Flow_From_Ops | 0.9993 | 1.001 | 0.9991 | 0.9995 |
| EBITDA | 0.9999 | 1.000 | 0.9998 | 1.0000 |

From the graph at the top left, we can see that as the time increases from 0 to 15 years, the number of companies going bankrupt also increases. This makes sense as companies are more likely to go bankrupt over the course of several years as their debt and liabilities accumulate over time. After this, we can apply the cox survival method to give us the summary outputs, which are displayed as tables in the top right.

From the summary output in the first table, we can determine that all five variables chosen for survival analysis are good predictors for determining how long a company has until it will go bankrupt due to their p-values all being quite low. After this, we can look at the second table which has the exp(coef) column which conveys the effect that each variable has on the bankruptcy rate. For the first variable, Net_Debt_EBITDA, this ratio indicates to us that having a higher Net debt to EBITDA ratio by 1 increases the chances of bankruptcy by 2%. For the profit margin, having this ratio increase by 1 will reduce the chances of bankruptcy by a whopping 65%. Current Ratio has a les profound effect as increasing this ratio by 1 will decrease the chances of bankruptcy by around 4%. The impact that Debt_to_Cash_Flow_From_Ops and EBITDA have on bankruptcy risk is negligible. From the survival analysis, we can therefore conclude that Profit Margin is the paramount ratio to consider when determining whether a company will remain a going concern in the foreseeable future. Additionally, Net Debt to EBITDA and Current Ratio can also be used as indicators to do a sense-check and validate the findings of what the Profit Margin reveals.

**Recommendation**

From the two analyses conducted, Jennifer can present to her firm two models, one which uses logistic regression to predict if a company will go bankrupt in the future based on four variables, Net Debt to EBITDA, Profit Margin, Current Ratio, and EBITDA, and another model which uses survival analysis to predict how many years it will take until a company goes bankrupt which will utilize 3 variables, Profit Margin, Net Debt to EBITDA, and Current Ratio.

**Code**:

Logistic Regression:

```
bankruptcy<-read.csv("C:/Users/valay/OneDrive/Desktop/Ivey/Winter Term 2021/Big Data Analytics/Class 9/Assignment/Bankruptcy Data - Ha

head(bankruptcy)

#educ1 changes categorical variable to binary; if 12th or above then 1 or else 0

bankruptcy_glm <- glm (Bankruptcy~ +Net_Debt_EBITDA +Profit_Margin +Current_Ratio +Debt_to_Cash_Flow_From_Ops +EBITDA, data=bankruptcy
summary(bankruptcy_glm)
#0.05 p-value
```

Confusion Matrix:

```
pred <- predict(bankruptcy_glm, bankruptcy, type = "response")
Output <- data.frame(actual = bankruptcy$Bankruptcy, predicted = pred)

classification <- ifelse(pred > 0.2, 1, 0)
#0.2 threshold based on confusion matrix specificity number

actual = bankruptcy$Bankruptcy
Output <- data.frame(actual, predprob=pred, classification)

install.packages("caret")
install.packages('e1071', dependencies=TRUE)
library(caret)

confusionMatrix(as.factor(classification), as.factor(actual))
#Confusion matrix gives you your predictions vs. actual (reference)
```

Survival Analysis (Cox):

```
library(survival)
#Survival Curves

#Exponential code
banksurv <- read.csv("C:/Users/valay/OneDrive/Desktop/Ivey/Winter Term 2021/Big Data Analytics/Class 9/Assignment/Bankruptcy Da

out<-survfit(Surv(Years_to_Bankrupt, Bankruptcy) ~ 1, data=banksurv)
y <- out$surv
y

y<- -log(y)
t<-out$time
plot(t,y)

#This gives expected number of failures over the time t
#this plot should be a straight line or else it indicates that the relationship is not exponential

#after exponential, use Weibull model
#in Weibull you plot the minimum or maximum of independent random variables
#Difference between Weibull and Exponential is that gamma = 1 for exponential model

#This is the weibull model code
cox.out<-coxph(Surv(Years_to_Bankrupt, Bankruptcy) ~ Net_Debt_EBITDA +Profit_Margin +Current_Ratio +Debt_to_Cash_Flow_From_Ops
summary(cox.out)
```

Exhibit (Confusion Matrix for Logistic Regression):

At pred>0.2:

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4034  487
         1   95   44

               Accuracy : 0.8751
                 95% CI : (0.8653, 0.8845)
    No Information Rate : 0.8861
    P-Value [Acc > NIR] : 0.9906
```