Data Analysis: Logistic Regression

Overview Bankruptcy Data:
The Bankruptcy dataset contains the financial information and bankruptcy status from a variety of companies where 85% of the records are from the year 1999. There is 5436 observations and 13 variables. The variable DLRSN is the symbol for Bankruptcy and Non-Bankruptcy. 1 stands for Bankruptcy and 0 for Non-Bankruptcy companies. The following is descriptions for variables R1through R10.

The variable description is as following:

- *R1: Working Capital/Total Asset;*
- *R2: Retained Earning/Total Asset;*
- *R3: Earning Before Interest & Tax/Total Asset;*
- *R4: Market Capital/Total Liability;*
- *R5: SALE/Total Asset;*
- *R6: Total Liability/Total Asset*
- *R7: Current Asset/Current Liability;*
- *R8: Net Income/Total Asset;*
- *R9: log(SALE);*
- *R10: log(Market Cap)*

```
# Read in Bankruptcy Data:
bank.data <- read.csv("bankruptcy.csv")
```

Exploratory Data Analysis:

Summary (Bankruptcy Data)

```
     DLRSN             R1                R2                R3                 R4
 Min.   :0.0000   Min.   :-4.3828   Min.   :-2.2418   Min.   :-2.06423   Min.   :-0.42712
 1st Qu.:0.0000   1st Qu.:-0.7501   1st Qu.:-1.0805   1st Qu.:-1.07887   1st Qu.:-0.38752
 Median :0.0000   Median :-0.2220   Median : 0.1337   Median : 0.06858   Median :-0.30754
 Mean   :0.1428   Mean   :-0.2352   Mean   :-0.2915   Mean   :-0.24411   Mean   : 0.23903
 3rd Qu.:0.0000   3rd Qu.: 0.4821   3rd Qu.: 0.5103   3rd Qu.: 0.50070   3rd Qu.: 0.02746
 Max.   :1.0000   Max.   : 2.0234   Max.   : 1.4854   Max.   : 2.14246   Max.   : 6.69536
       R5                R6                  R7                 R8                 R9
 Min.   :-1.3639   Min.   :-1.505889   Min.   :-1.23340   Min.   :-2.2082   Min.   :-2.76356
 1st Qu.:-0.8748   1st Qu.:-0.656827   1st Qu.:-0.77996   1st Qu.:-1.0054   1st Qu.:-0.66606
 Median :-0.3508   Median : 0.003493   Median :-0.43205   Median : 0.2053   Median : 0.06219
 Mean   :-0.1338   Mean   : 0.195707   Mean   :-0.09612   Mean   :-0.2272   Mean   : 0.02538
 3rd Qu.: 0.2878   3rd Qu.: 0.608376   3rd Qu.: 0.17578   3rd Qu.: 0.5289   3rd Qu.: 0.81215
 Max.   : 4.0362   Max.   : 5.110424   Max.   : 2.87648   Max.   : 2.0006   Max.   : 2.17918
      R10
 Min.   :-2.2140
 1st Qu.:-0.6413
 Median : 0.1230
 Mean   : 0.1806
 3rd Qu.: 0.9878
 Max.   : 2.4846
```

- No NA values from the data set

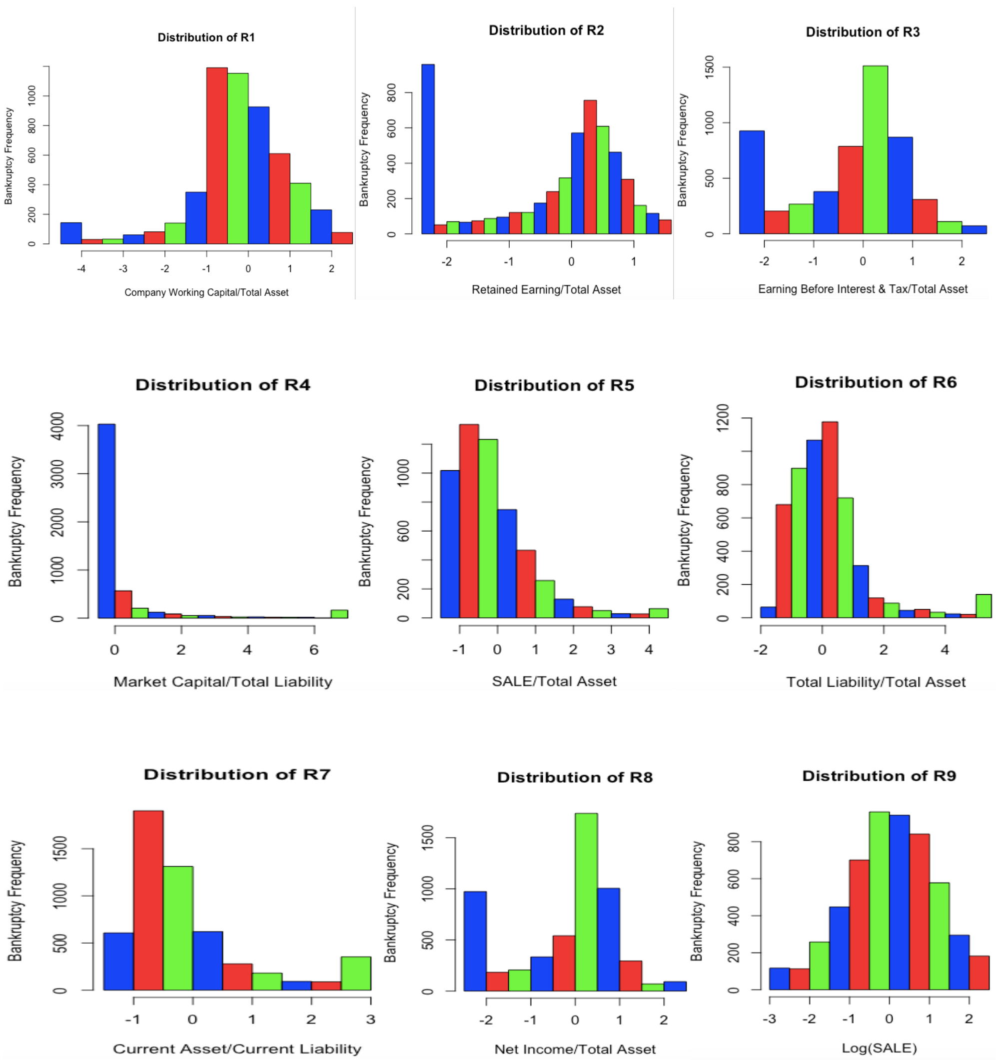
```
> sum(is.na(bank.data))
[1] 0
```

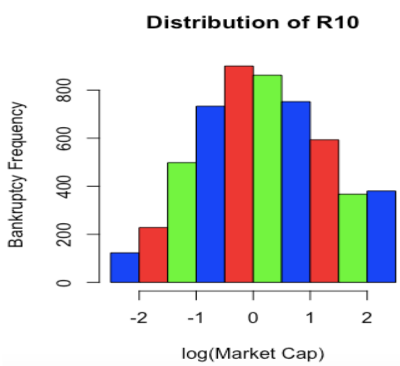
Structure of Data Set:

```
> str(bank.data)
'data.frame':	5436 obs. of  11 variables:
 $ DLRSN: int  0 0 0 1 0 0 0 0 1 0 ...
 $ R1   : num  0.307 0.761 -0.514 -0.466 2.023 ...
 $ R2   : num  0.887 0.592 0.338 0.371 0.215 ...
 $ R3   : num  1.648 0.453 0.299 0.496 0.183 ...
 $ R4   : num  -0.1992 -0.3699 -0.0291 -0.3734 6.6954 ...
 $ R5   : num  1.093 0.186 -0.433 -0.267 -1.148 ...
 $ R6   : num  -0.3133 0.0396 0.83 0.9778 -1.5059 ...
 $ R7   : num  -0.197 0.327 -0.708 -0.611 2.876 ...
 $ R8   : num  1.207 0.428 0.476 0.457 0.287 ...
 $ R9   : num  0.282 1.107 2.179 0.152 -0.986 ...
 $ R10  : num  0.1589 0.7934 2.4846 0.0478 0.7911 ...
```

- Type class of variables R1-10 are numeric.
- Variables CUSIP and FYear have been removed

Distribution of Variables:





- R10 and R9 where the only variables that returned as a normal distribution curve.

<u>What is the overall bankruptcy probability?</u>

```
# Overall bankruptcy probability
mean(bank.data$DLRSN) * 100
```

```
> mean(bank.data$DLRSN) * 100
[1] 14.2752
```

- Overall bankruptcy probability is about 14%.

<u>Splitting Dataset:</u>

- Split Data to 80% training and 20% test.
- We will build a logistic model using the training data set with all predictors.
- Calculate mean and standard deviation for Bankrupted and Non-Bankrupted.
- Calculate AIC, BIC, and Mean Residual Deviance.

```
# Split data to 80% training and 20% test
index <- sample(nrow(bank.data), 0.8 * nrow(bank.data))
bank.train <- bank.data[index,]
bank.test <- bank.data[-index,]
```

```
# Building Logistic Regression Model for training sample using all predictors
# Logistic Model will estimate the "probability" of the outcome
bank.glm <- glm(DLRSN~., data = bank.train, family = "binomial")
```

```
# Bankruptcy and Non-Bankruptcy companies
bank1 <- bank.data[bank.data$DLRSN == 1,] # Companies that went bankrupted
bank2 <- bank.data[bank.data$DLRSN == 0,] # Companies that didn't go bankrupted
```

```
# Calculate mean and standard deviation
b1 <- c(bank1$DLRSN, bank1$R1, bank1$R2, bank1$R3,
        bank1$R4, bank1$R5, bank1$R6, bank1$R7, bank1$R8,
        bank1$R9, bank1$R10)
mean(b1)
sd(b1)
b2 <- c(bank2$DLRSN, bank2$R1, bank2$R2, bank2$R3,
       bank2$R4, bank2$R5, bank2$R6, bank2$R7, bank2$R8, bank2$R9, bank2$R10)
mean(b2)
sd(b2)
```

	Bankrupted	Non-Bankrupted
Mean	-0.33	0.008
Standard Deviation	1.30	1.10

```
# Calculate AIC, BIC, and Mean Residual Deviance
bank.glm$deviance
AIC(bank.glm)
BIC(bank.glm)
```

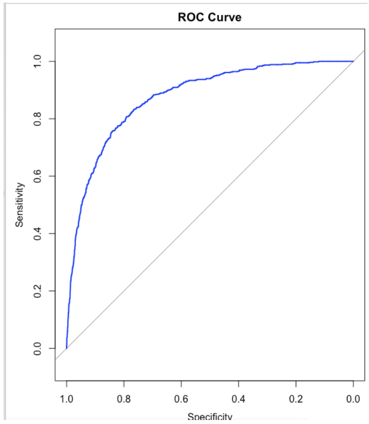
	Logistic Model
Mean Residual Deviance	2421.026
AIC	2443.026
BIC	2513.178

<u>Exploring Dependent Variable DLRSN:</u>

```
# Explore Dependent Variable:
# 1 Companies that went Bankrupted
# 0 Companies that didn't go Bankrupted
table(bank.train$DLRSN)
```

```
   0    1
3730  618
```

<u>Create ROC Curve and find Area Under the Curve:</u>




```
> auc(ROC) * 100
[1] 87.52766
```

Model Selection with BIC:

- Create full model, null model and conduct stepwise model using BIC to find best model selection.

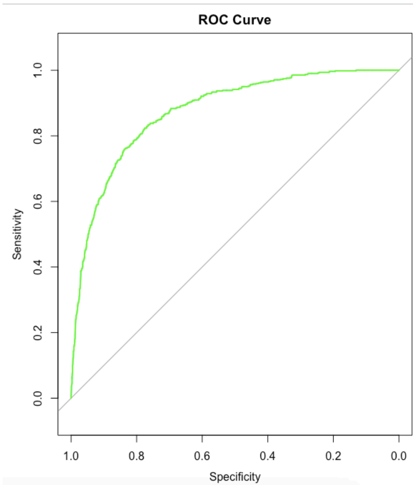
```
# Specify a null model with no predictors
null_model <- glm(DLRSN ~ 1, data = bank.train, family = "binomial")

# Specify the full model
full_model <- glm(DLRSN ~ ., data = bank.train, family = "binomial")

# Use stepwise algorithm to build model with BIC criterion
step_model <- step(null_model, scope = list(lower = null_model, upper = full_model), direction = "both", k = log(nrow(bank.train)))
summary(step_model)
```

```
> summary(step_model)

Call:
glm(formula = DLRSN ~ R10 + R7 + R8 + R2 + R9 + R3 + R6 + R1,
    family = "binomial", data = bank.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1417  -0.4551  -0.2422  -0.1120   3.1723

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.54299    0.07719 -32.946  < 2e-16 ***
R10         -1.55434    0.08193 -18.973  < 2e-16 ***
R7          -0.54319    0.10218  -5.316 1.06e-07 ***
R8          -0.32563    0.08487  -3.837 0.000125 ***
R2           0.59533    0.07790   7.642 2.14e-14 ***
R9           0.35832    0.08555   4.188 2.81e-05 ***
R3          -0.46325    0.09950  -4.656 3.23e-06 ***
R6           0.28083    0.05459   5.144 2.68e-07 ***
R1           0.25088    0.07394   3.393 0.000691 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3555.1  on 4347  degrees of freedom
Residual deviance: 2422.7  on 4339  degrees of freedom
AIC: 2440.7

Number of Fisher Scoring iterations: 6
```

Summary of (step_model) tells us that using predictors R1, R2, R3, R6, R7, R8, R9 and R10 is best predictors with BIC criterion for our model.

ROC and AUC of Final_Model:

```r
# Create and Plot final model. Find ROC curve of final model and find AUC
final_model <- glm(DLRSN ~ + R1 + R2 + R3 + R6 + R7 + R8 + R9 + R10, family = "binomial", data = bank.train)
summary(final_model)
final_prob <- predict(final_model, type = "response")
ROC_Final <- roc(bank.train$DLRSN, final_prob)
plot(ROC_Final, col = "green", main = "ROC Curve")
auc(ROC_Final)
```




```
Area under the curve: 0.8749
> auc(ROC_Final) * 100
[1] 87.49
```

We want our ROC curve as close to the left as possible and AUC as close to 1 as possible. Our first model with all predictors is slightly better than our Final Model, However, both models are sufficient with their selections of predictors as both models AUC is at least above 70%.