Kristine Plunkett
Costa Mesa, CA | 949-610-3981 | kplunkett13@gmail.com | www.linkedin.com/in/kristineplunkett

Determining Credit Worthiness - R Programming

This project aimed to analyze and gauge the credit worthiness of consumers included within the German.csv dataset. Several methods of descriptive and predictive statistics tests were performed within R programming in order to determine those most qualified to obtain credit.

Initial Preparation and Data Exploration

Initial Preparation Includes: Setting working directory, creating data frame, installing libraries

* The following packages and libraries have been installed and loaded: arules, arulesViz, rpart, rpart.plot, caret, gains, glm2

```
setwd("/Users/User1/Documents/")
credit.df <- read.csv("GermanCredit.csv", header = TRUE)
dim(credit.df) # shows dimension information of the dataset

## [1] 1000   32

head(credit.df) # shows the first six rows

##   OBS. CHK_ACCT DURATION HISTORY NEW_CAR USED_CAR FURNITURE RADIO.TV
## 1    1        0        6       4       0        0         0        1
## 2    2        1       48       2       0        0         0        1
## 3    3        3       12       4       0        0         0        0
## 4    4        0       42       2       0        0         1        0
## 5    5        0       24       3       1        0         0        0
## 6    6        3       36       2       0        0         0        0
##   EDUCATION RETRAINING AMOUNT SAV_ACCT EMPLOYMENT INSTALL_RATE
MALE_DIV
```

```
## 1        0         0   1169       4         4          4        0
## 2        0         0   5951       0         2          2        0
## 3        1         0   2096       0         3          2        0
## 4        0         0   7882       0         3          2        0
## 5        0         0   4870       0         2          3        0
## 6        1         0   9055       4         2          2        0
##   MALE_SINGLE MALE_MAR_or_WID CO.APPLICANT GUARANTOR PRESENT_RESIDENT
## 1          1             0           0        0              4
## 2          0             0           0        0              2
## 3          1             0           0        0              3
## 4          1             0           0        1              4
## 5          1             0           0        0              4
## 6          1             0           0        0              4
##   REAL_ESTATE PROP_UNKN_NONE AGE OTHER_INSTALL RENT OWN_RES
NUM_CREDITS
## 1          1            0  67           0    0      1          2
## 2          1            0  22           0    0      1          1
## 3          1            0  49           0    0      1          1
## 4          0            0  45           0    0      0          1
## 5          0            1  53           0    0      0          2
## 6          0            1  35           0    0      0          1
##   JOB NUM_DEPENDENTS TELEPHONE FOREIGN RESPONSE
## 1   2            1        1       0       1
## 2   2            1        0       0       0
## 3   1            2        0       0       1
## 4   2            2        0       0       1
## 5   2            2        0       0       0
## 6   1            2        1       0       1
```

```
credit.df[5, 1:10] # shows  fifth row of the 1st 10 columns

##   OBS. CHK_ACCT DURATION HISTORY NEW_CAR USED_CAR FURNITURE RADIO.TV
## 5    5        0       24       3       1        0         0        0
##   EDUCATION RETRAINING
## 5         0          0

credit.df$AMOUNT[1:10] # shows the first 10 rows of the column named "AMOUNT"

##  [1] 1169 5951 2096 7882 4870 9055 2835 6948 3059 5234

summary(credit.df) # finds the summary statistics for each column

##       OBS.           CHK_ACCT        DURATION       HISTORY
##  Min.   :   1.0   Min.   :0.000   Min.   : 4.0   Min.   :0.000
##  1st Qu.: 250.8   1st Qu.:0.000   1st Qu.:12.0   1st Qu.:2.000
##  Median : 500.5   Median :1.000   Median :18.0   Median :2.000
##  Mean   : 500.5   Mean   :1.577   Mean   :20.9   Mean   :2.545
##  3rd Qu.: 750.2   3rd Qu.:3.000   3rd Qu.:24.0   3rd Qu.:4.000
##  Max.   :1000.0   Max.   :3.000   Max.   :72.0   Max.   :4.000
##     NEW_CAR         USED_CAR       FURNITURE        RADIO.TV
##  Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.00
##  1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.00
##  Median :0.000   Median :0.000   Median :0.000   Median :0.00
##  Mean   :0.234   Mean   :0.103   Mean   :0.181   Mean   :0.28
##  3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:1.00
##  Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :1.00
##    EDUCATION      RETRAINING        AMOUNT         SAV_ACCT
##  Min.   :0.00   Min.   :0.000   Min.   :  250   Min.   :0.000
##  1st Qu.:0.00   1st Qu.:0.000   1st Qu.: 1366   1st Qu.:0.000
##  Median :0.00   Median :0.000   Median : 2320   Median :0.000
```

```
##  Mean   :0.05   Mean   :0.097   Mean   : 3271   Mean   :1.105
##  3rd Qu.:0.00   3rd Qu.:0.000   3rd Qu.: 3972   3rd Qu.:2.000
##  Max.   :1.00   Max.   :1.000   Max.   :18424   Max.   :4.000
##    EMPLOYMENT     INSTALL_RATE      MALE_DIV     MALE_SINGLE
##  Min.   :0.000   Min.   :1.000   Min.   :0.00   Min.   :0.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:0.00   1st Qu.:0.000
##  Median :2.000   Median :3.000   Median :0.00   Median :1.000
##  Mean   :2.384   Mean   :2.973   Mean   :0.05   Mean   :0.548
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:0.00   3rd Qu.:1.000
##  Max.   :4.000   Max.   :4.000   Max.   :1.00   Max.   :1.000
##  MALE_MAR_or_WID  CO.APPLICANT     GUARANTOR     PRESENT_RESIDENT
##  Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :1.000
##  1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:2.000
##  Median :0.000   Median :0.000   Median :0.000   Median :3.000
##  Mean   :0.092   Mean   :0.041   Mean   :0.052   Mean   :2.845
##  3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:4.000
##  Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :4.000
##   REAL_ESTATE    PROP_UNKN_NONE       AGE        OTHER_INSTALL
##  Min.   :0.000   Min.   :0.000   Min.   :19.00   Min.   :0.000
##  1st Qu.:0.000   1st Qu.:0.000   1st Qu.:27.00   1st Qu.:0.000
##  Median :0.000   Median :0.000   Median :33.00   Median :0.000
##  Mean   :0.282   Mean   :0.154   Mean   :35.55   Mean   :0.186
##  3rd Qu.:1.000   3rd Qu.:0.000   3rd Qu.:42.00   3rd Qu.:0.000
##  Max.   :1.000   Max.   :1.000   Max.   :75.00   Max.   :1.000
##       RENT          OWN_RES       NUM_CREDITS        JOB
##  Min.   :0.000   Min.   :0.000   Min.   :1.000   Min.   :0.000
##  1st Qu.:0.000   1st Qu.:0.000   1st Qu.:1.000   1st Qu.:2.000
##  Median :0.000   Median :1.000   Median :1.000   Median :2.000
```

```
##  Mean   :0.179   Mean   :0.713   Mean   :1.407   Mean   :1.904
##  3rd Qu.:0.000   3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :1.000   Max.   :1.000   Max.   :4.000   Max.   :3.000
##  NUM_DEPENDENTS    TELEPHONE       FOREIGN        RESPONSE
##  Min.   :1.000   Min.   :0.000   Min.   :0.000   Min.   :0.0
##  1st Qu.:1.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.0
##  Median :1.000   Median :0.000   Median :0.000   Median :1.0
##  Mean   :1.155   Mean   :0.404   Mean   :0.037   Mean   :0.7
##  3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:0.000   3rd Qu.:1.0
##  Max.   :2.000   Max.   :1.000   Max.   :1.000   Max.   :1.0
```

```
credit.df[1:10, ] # first 10 rows of each of the columns
```

```
##    OBS. CHK_ACCT DURATION HISTORY NEW_CAR USED_CAR FURNITURE RADIO.TV
## 1     1       0       6      4      0       0       0       1
## 2     2       1      48      2      0       0       0       1
## 3     3       3      12      4      0       0       0       0
## 4     4       0      42      2      0       0       1       0
## 5     5       0      24      3      1       0       0       0
## 6     6       3      36      2      0       0       0       0
## 7     7       3      24      2      0       0       1       0
## 8     8       1      36      2      0       1       0       0
## 9     9       3      12      2      0       0       0       1
## 10   10       1      30      4      1       0       0       0
##    EDUCATION RETRAINING AMOUNT SAV_ACCT EMPLOYMENT INSTALL_RATE
MALE_DIV
## 1         0         0   1169       4         4          4       0
## 2         0         0   5951       0         2          2       0
## 3         1         0   2096       0         3          2       0
## 4         0         0   7882       0         3          2       0
```

```
## 5         0         0   4870       0         2          3       0
## 6         1         0   9055       4         2          2       0
## 7         0         0   2835       2         4          3       0
## 8         0         0   6948       0         2          2       0
## 9         0         0   3059       3         3          2       1
## 10        0         0   5234       0         0          4       0
##    MALE_SINGLE MALE_MAR_or_WID CO.APPLICANT GUARANTOR PRESENT_RESIDENT
## 1            1               0            0         0                4
## 2            0               0            0         0                2
## 3            1               0            0         0                3
## 4            1               0            0         1                4
## 5            1               0            0         0                4
## 6            1               0            0         0                4
## 7            1               0            0         0                4
## 8            1               0            0         0                2
## 9            0               0            0         0                4
## 10           0               1            0         0                2
##    REAL_ESTATE PROP_UNKN_NONE AGE OTHER_INSTALL RENT OWN_RES
NUM_CREDITS
## 1            1              0  67             0    0       1           2
## 2            1              0  22             0    0       1           1
## 3            1              0  49             0    0       1           1
## 4            0              0  45             0    0       0           1
## 5            0              1  53             0    0       0           2
## 6            0              1  35             0    0       0           1
## 7            0              0  53             0    0       1           1
## 8            0              0  35             0    1       0           1
## 9            1              0  61             0    0       1           1
```

```
## 10          0             0  28            0    0       1          2
##    JOB NUM_DEPENDENTS TELEPHONE FOREIGN RESPONSE
## 1    2             1         1       0        1
## 2    2             1         0       0        0
## 3    1             2         0       0        1
## 4    2             2         0       0        1
## 5    2             2         0       0        0
## 6    1             2         1       0        1
## 7    2             1         0       0        1
## 8    3             1         1       0        1
## 9    1             1         0       0        1
## 10   3             1         0       0        0
```

Logistic Regression Model

Fitting a Logistic Regression Model (First Partitioning the data into training and validation sets)

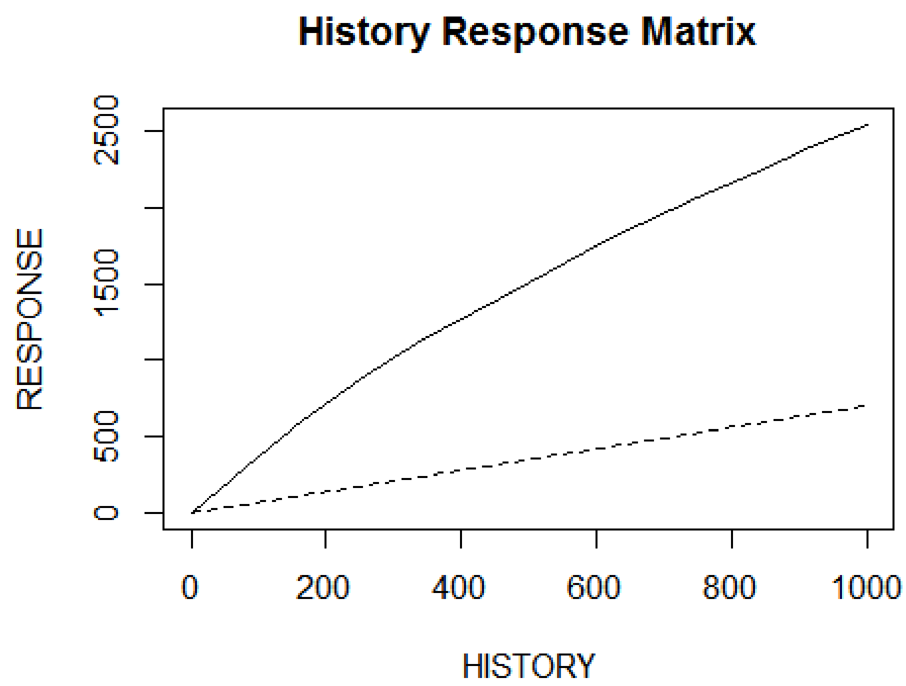
```
#partition the data
set.seed(2) # set seed for reproducing the partition
train.index <- sample(c(1:dim(credit.df)[1]), dim(credit.df)[1]*0.6)
train.df <- credit.df[train.index, ]
valid.df <- credit.df[-train.index, ]


# running the logistic regression using glm() to fit a logistic regression.
logit.reg <- glm(RESPONSE ~ ., data = train.df, family = "binomial")
options(scipen=999)
summary(logit.reg)
```

```
##
## Call:
```

```
## glm(formula = RESPONSE ~ ., family = "binomial", data = train.df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9003  -0.7064   0.3944   0.6947   2.4171
##
## Coefficients:
##                     Estimate  Std. Error z value    Pr(>|z|)
## (Intercept)       1.63504202  1.18094272   1.385    0.166198
## OBS.              0.00025867  0.00038552   0.671    0.502247
## CHK_ACCT          0.47326685  0.09272650   5.104 0.000000333 ***
## DURATION         -0.03576241  0.01123555  -3.183    0.001458 **
## HISTORY           0.41477296  0.11654809   3.559    0.000373 ***
## NEW_CAR          -1.14662854  0.57952285  -1.979    0.047864 *
## USED_CAR          0.73108814  0.69707071   1.049    0.294270
## FURNITURE        -0.19198673  0.59155660  -0.325    0.745525
## RADIO.TV         -0.36150337  0.57629037  -0.627    0.530467
## EDUCATION        -1.07504582  0.71455348  -1.505    0.132453
## RETRAINING       -0.21707088  0.64323367  -0.337    0.735764
## AMOUNT           -0.00012208  0.00005438  -2.245    0.024766 *
## SAV_ACCT          0.23793412  0.08064257   2.950    0.003173 **
## EMPLOYMENT        0.11607146  0.10053610   1.155    0.248285
## INSTALL_RATE     -0.39801149  0.11739652  -3.390    0.000698 ***
## MALE_DIV         -0.61625959  0.50612073  -1.218    0.223371
## MALE_SINGLE       0.51341977  0.27318513   1.879    0.060192 .
## MALE_MAR_or_WID   0.32346077  0.42830888   0.755    0.450126
## CO.APPLICANT     -0.91204553  0.52749935  -1.729    0.083809 .
## GUARANTOR         1.29558319  0.70533697   1.837    0.066235 .
```

```
## PRESENT_RESIDENT  0.05041849  0.11067016   0.456    0.648696
## REAL_ESTATE       0.17899001  0.27464827   0.652    0.514591
## PROP_UNKN_NONE   -0.64619979  0.48341107  -1.337    0.181304
## AGE               0.02073335  0.01166135   1.778    0.075411 .
## OTHER_INSTALL    -0.59045558  0.25890715  -2.281    0.022574 *
## RENT             -0.55774400  0.61178680  -0.912    0.361946
## OWN_RES          -0.37196368  0.58319741  -0.638    0.523603
## NUM_CREDITS      -0.33436027  0.21202554  -1.577    0.114800
## JOB               0.05845154  0.19153277   0.305    0.760231
## NUM_DEPENDENTS   -0.50652914  0.32481301  -1.559    0.118890
## TELEPHONE         0.18292369  0.25653001   0.713    0.475803
## FOREIGN           2.35270482  1.09074288   2.157    0.031008 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 738.05  on 599  degrees of freedom
## Residual deviance: 536.78  on 568  degrees of freedom
## AIC: 600.78
##
## Number of Fisher Scoring iterations: 6
```

Confusion Matrix Visualization



Classification Tree Model

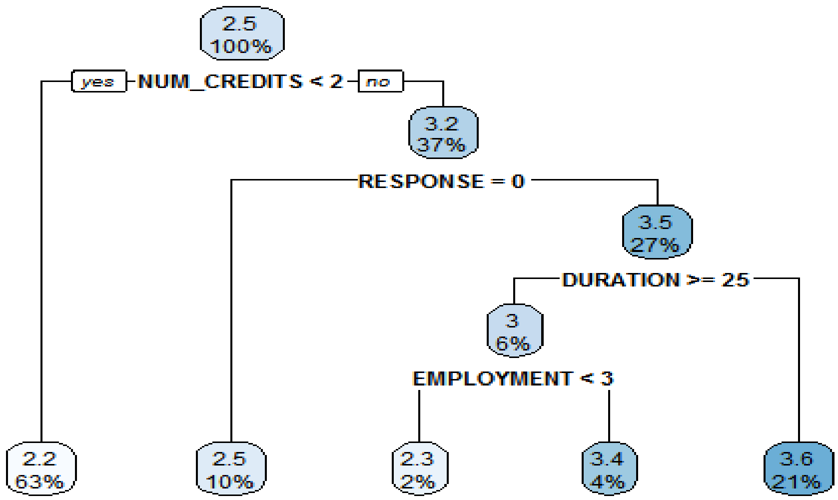
Fitting a Classification Tree Model * First there is a creation of new data frame * Next partition the data into training and validation sets

```
CreditTree.df <- read.csv("GermanCredit.csv")

CreditTree.df <- CreditTree.df


#partition the data

set.seed(1)

train.index <- sample(c(1:dim(CreditTree.df)[1]), dim(CreditTree.df)[1]*0.6)

train.df <- CreditTree.df[train.index, ]

valid.df <- CreditTree.df[-train.index, ]
```

Classification Tree Visualization

Classification Tree

```
## Warning: package 'rpart.plot' was built under R version 3.5.3
```



In efforts to better understand the data and perform the necessary exploration of the dataset, an initial run through of the standard data exploration functions took place. Consequently, it was the use of boxplot that started to bring correlations and significance within the data to light. It was uncovered that a direct correlation exists between credit history and credit rating response. Moving onto the logistic regression model, a great deal of significance with the credit history factor came to light and this would be useful in gauging credit worthiness within the dataset. The following boxplot illustrated that those who had no credit taken (0) and those who had paid all creditors back successfully (1) were those with the highest credit worthy significance.

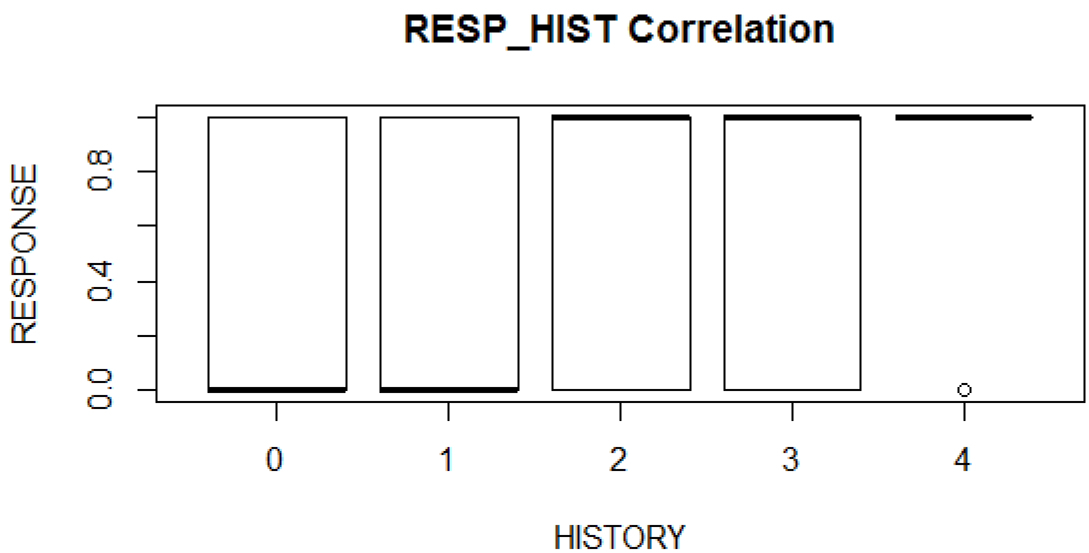


Figure 1. RESP_HIST Correlation. This visualization was run using boxplot and illustrates the correlation between credit rating REPSPONSE and Credit HISTORY fields within the german credit data set.

Final Thoughts and Conclusions

Running through the logisitc regression classification model highlighted where the p-value, Pr(>|t|), also showed strong significance with the credit history factor. This was illustrated by the triple asterisks next to the p-value for each factor. This led to focusing on that factor for analysis. Finally a classification tree model was utilized and gave further telling results on the dataset. It shows that in addition to credit history, there is a strong correlation for the number of existing credits, NUM_CREDITS. In conclusion, it appears that those with less than 2 existing credits at the bank encompass 63% of the credit worthy customers. The best approach would be for the bank to offer credit to those customers who have a combination of less than 2 existing credits along with credit history in good standing.