Portfolio Project Option 1

Jon Brookins

9/12/2021

MIS 510 Portfolio Project requires me to examine data related to the credit worthiness of German citizens. I will use my current knowledge in R programming to examine factors to determine creditworthiness.

In this part, GermanCredit.csv is loaded and formatted the data to be examined. #### Load German.csv

credit <- read.csv("GermanCredit.csv")

#### Create data frame

credit.df <- data.frame(credit, check.rows = TRUE, check.names = TRUE)

#### Remove OBS column

credit.df <- credit.df[-1]  
credit <- credit[-1]

#### Summary of data

summary(credit.df)

## CHK\_ACCT DURATION HISTORY NEW\_CAR USED\_CAR   
## Min. :0.000 Min. : 4.0 Min. :0.000 Min. :0.000 Min. :0.000   
## 1st Qu.:0.000 1st Qu.:12.0 1st Qu.:2.000 1st Qu.:0.000 1st Qu.:0.000   
## Median :1.000 Median :18.0 Median :2.000 Median :0.000 Median :0.000   
## Mean :1.577 Mean :20.9 Mean :2.545 Mean :0.234 Mean :0.103   
## 3rd Qu.:3.000 3rd Qu.:24.0 3rd Qu.:4.000 3rd Qu.:0.000 3rd Qu.:0.000   
## Max. :3.000 Max. :72.0 Max. :4.000 Max. :1.000 Max. :1.000   
## FURNITURE RADIO.TV EDUCATION RETRAINING AMOUNT   
## Min. :0.000 Min. :0.00 Min. :0.00 Min. :0.000 Min. : 250   
## 1st Qu.:0.000 1st Qu.:0.00 1st Qu.:0.00 1st Qu.:0.000 1st Qu.: 1366   
## Median :0.000 Median :0.00 Median :0.00 Median :0.000 Median : 2320   
## Mean :0.181 Mean :0.28 Mean :0.05 Mean :0.097 Mean : 3271   
## 3rd Qu.:0.000 3rd Qu.:1.00 3rd Qu.:0.00 3rd Qu.:0.000 3rd Qu.: 3972   
## Max. :1.000 Max. :1.00 Max. :1.00 Max. :1.000 Max. :18424   
## SAV\_ACCT EMPLOYMENT INSTALL\_RATE MALE\_DIV MALE\_SINGLE   
## Min. :0.000 Min. :0.000 Min. :1.000 Min. :0.00 Min. :0.000   
## 1st Qu.:0.000 1st Qu.:2.000 1st Qu.:2.000 1st Qu.:0.00 1st Qu.:0.000   
## Median :0.000 Median :2.000 Median :3.000 Median :0.00 Median :1.000   
## Mean :1.105 Mean :2.384 Mean :2.973 Mean :0.05 Mean :0.548   
## 3rd Qu.:2.000 3rd Qu.:4.000 3rd Qu.:4.000 3rd Qu.:0.00 3rd Qu.:1.000   
## Max. :4.000 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

Examing the data, there are 1000 observations with 32 variables. No missing data is observed at this time.

The data contains numerical, categorical, and binary variable types. I change each variable to a factor using the as.factor function. #### Manually change variable type to factors

credit.df$CHK\_ACCT <- as.factor(credit.df$CHK\_ACCT)   
credit.df$HISTORY <- as.factor(credit.df$HISTORY)   
credit.df$SAV\_ACCT <- as.factor(credit.df$SAV\_ACCT)   
credit.df$EMPLOYMENT <- as.factor(credit.df$EMPLOYMENT)   
credit.df$PRESENT\_RESIDENT <- as.factor(credit.df$PRESENT\_RESIDENT)  
credit.df$JOB <- as.factor(credit.df$JOB)  
credit.df$NEW\_CAR <- as.factor(credit.df$NEW\_CAR)   
credit.df$USED\_CAR <- as.factor(credit.df$USED\_CAR)   
credit.df$FURNITURE <- as.factor(credit.df$FURNITURE)   
credit.df$RADIO.TV <- as.factor(credit.df$RADIO.TV)   
credit.df$EDUCATION <- as.factor(credit.df$EDUCATION)   
credit.df$RETRAINING <- as.factor(credit.df$RETRAINING)   
credit.df$MALE\_DIV <- as.factor(credit.df$MALE\_DIV)   
credit.df$MALE\_SINGLE <- as.factor(credit.df$MALE\_SINGLE)   
credit.df$MALE\_MAR\_or\_WID <- as.factor(credit.df$MALE\_MAR\_or\_WID)   
credit.df$CO.APPLICANT <- as.factor(credit.df$CO.APPLICANT)   
credit.df$GUARANTOR <- as.factor(credit.df$GUARANTOR)   
credit.df$REAL\_ESTATE <- as.factor(credit.df$REAL\_ESTATE)  
credit.df$PROP\_UNKN\_NONE <- as.factor(credit.df$PROP\_UNKN\_NONE)   
credit.df$OTHER\_INSTALL <- as.factor(credit.df$OTHER\_INSTALL)   
credit.df$RENT <- as.factor(credit.df$RENT)  
credit.df$OWN\_RES <- as.factor(credit.df$OWN\_RES)   
credit.df$TELEPHONE <- as.factor(credit.df$TELEPHONE)   
credit.df$FOREIGN <- as.factor(credit.df$FOREIGN)   
credit.df$RESPONSE <- as.factor(credit.df$RESPONSE)

## Explore data

In this section I use R functions to explore the data. To simply the presentation, I created a table for statistical functions that includes mean, median, standard deviation, min and max for variables duration, amount, and age.

#### Create a table of all the important statistical measures

data.frame(mean = sapply(credit.df[,c(2,10,22)], mean, na.rm=TRUE), median = sapply(credit.df[,c(2,10,22)], median, na.rm=TRUE), sd = sapply(credit.df[,c(2,10,22)], sd, na.rm=TRUE),  
 min = sapply(credit.df[,c(2,10,22)], min, na.rm=TRUE),  
 max = sapply(credit.df[,c(2,10,22)], max, na.rm=TRUE))

## mean median sd min max  
## DURATION 20.903 18.0 12.05881 4 72  
## AMOUNT 3271.258 2319.5 2822.73688 250 18424  
## AGE 35.546 33.0 11.37547 19 75

The chart shows that the age range range for applicants is between 19 years old and 75 years old. The average age is ~35 years old. The amount variable has the largest standard deviation - ~2822. The minimum amount borrowed is $250 and the maximum is $18,424. It would be interesting to compare this to American’s amount borrowed. All of the information appears to useful to further examine the data.

## Data Visualization

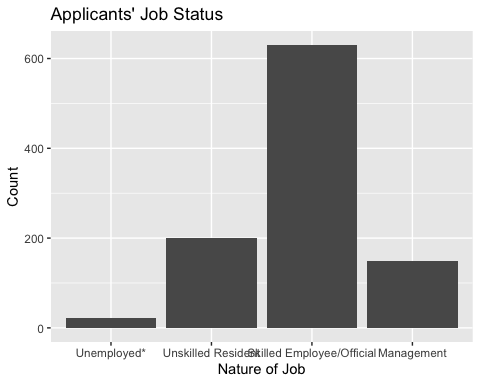
library(GGally)

## Loading required package: ggplot2

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

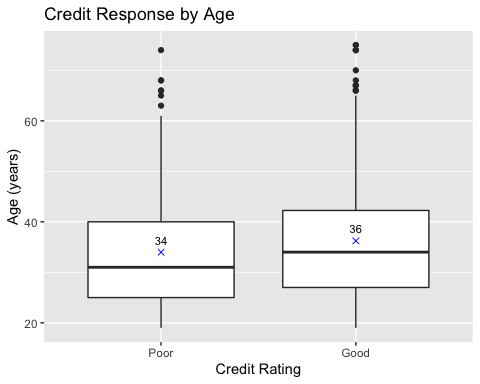
In this section I will build some visual models to represent the data. I examine the variables job status that is counted by level of skill - unemployed, unskilled resident, skilled employee, and management. I use an asterisk for unemployed because the code description label is Unemployed, unskilled - non-resident.

ggplot(data = credit.df, aes(x = factor(credit.df$JOB))) +  
 geom\_bar() +  
 labs(x = "Nature of Job", y = "Count") +  
 scale\_x\_discrete(labels = c("Unemployed\*", "Unskilled Resident", "Skilled Employee/Official", "Management")) +  
 ggtitle("Applicants' Job Status") +  
 annotate("text", x=1.4, y=100, label = "Professional Status", size = 0)

 The bar graph depicts that most borrowers are in the Skilled Employee/Official category. As expected the Unemployed category has the fewest applicants.

In this box plot, I examine the relationship of rating to age. The Response variable is related to the customers credit rating with 0 = No and 1 = Yes. No credit score numbers were given to classify the customer good or bad.

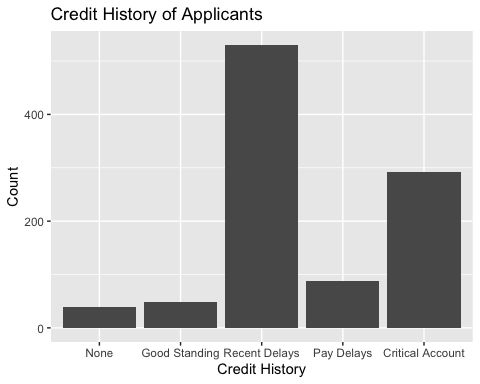
ggplot(data = credit.df, aes(x=factor(credit.df$RESPONSE), y=credit.df$AGE)) +   
 geom\_boxplot() +   
 stat\_summary(fun=mean, geom="point", color="blue", shape = 4, fill = "blue", size = 2) +  
 stat\_summary(aes(label=round(..y..,0)), fun=mean, geom="text", size=3, vjust=-1) + # add value of mean  
 labs(x = "Credit Rating", y = "Age (years)", title = "Credit Response by Age") + # label plot and axes  
 scale\_x\_discrete(labels=c("Poor", "Good"))



The credit by response box plot depicts that the mean age between customers with good or bad credit only varied by 2 years. The outliers for poor and good credit were another sign that age and credit ratings are similar between the two groups. This would tell us that customer age is not a good factor when determining credit rating.

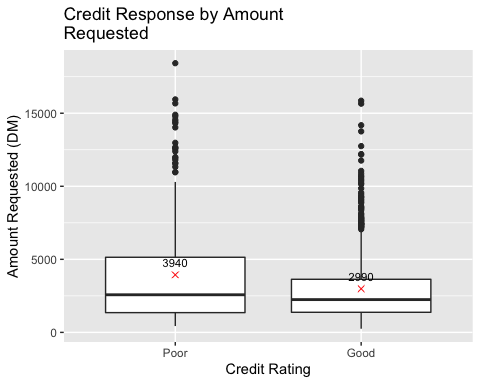
ggplot(data = credit.df, aes(x = factor(credit.df$HISTORY))) +  
 geom\_bar() +  
 labs(x = "Credit History", y = "Count") +  
 scale\_x\_discrete(labels = c("None", "Good Standing", "Recent Delays", "Pay Delays",  
 "Critical Account")) +  
 ggtitle("Credit History of Applicants")

## Warning: Use of `credit.df$HISTORY` is discouraged. Use `HISTORY` instead.



This boxplot examines the relationsip beween credit rating and amount requested.

ggplot(data = credit.df, aes(x=factor(credit.df$RESPONSE), y=credit.df$AMOUNT)) + geom\_boxplot() +  
 stat\_summary(fun.y=mean, geom="point", color="red", shape = 4, fill = "red", size = 2) + stat\_summary(aes(label=round(..y..,-1)), fun.y=mean, geom="text", size=3, vjust=-1) + labs(x = "Credit Rating", y = "Amount Requested (DM)", title = "Credit Response by Amount  
Requested") + scale\_x\_discrete(labels=c("Poor", "Good"))

 The mean amount requested by customers with a poor credit rating was $950 more than customers with a good credit rating. This would portend that customers with poor credit would go further into debt when borrowing or lack the funds for an adequate downpayment when purchasing a car.

To model the data I must partition the data into training and validation data.I do this by pationing the data into %60 train and 40% validation.

set.seed(5)  
credit.train.index <- sample(c(1:dim(credit.df)[1]), dim(credit.df)[1]\*0.6)   
credit.train.df <- credit.df[credit.train.index, ]  
credit.valid.df <- credit.df[-credit.train.index, ]

library(rpart)

#### Regression Model

library(caret)

## Loading required package: lattice

I created a regression model for the variable Response.

tree <- rpart(RESPONSE ~ ., data=credit.train.df, method="class")   
reg.pred <- predict(tree, credit.valid.df, type = "class")  
mat<- confusionMatrix(reg.pred, credit.valid.df$RESPONSE)   
mat

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 51 56  
## 1 63 230  
##   
## Accuracy : 0.7025   
## 95% CI : (0.6551, 0.7469)  
## No Information Rate : 0.715   
## P-Value [Acc > NIR] : 0.7304   
##   
## Kappa : 0.2563   
##   
## Mcnemar's Test P-Value : 0.5823   
##   
## Sensitivity : 0.4474   
## Specificity : 0.8042   
## Pos Pred Value : 0.4766   
## Neg Pred Value : 0.7850   
## Prevalence : 0.2850   
## Detection Rate : 0.1275   
## Detection Prevalence : 0.2675   
## Balanced Accuracy : 0.6258   
##   
## 'Positive' Class : 0   
##

This regression model has a accuracy of 70% with a sensitivity of .4474 and a specificity of .8042.

library(caret)

In this section I will use the rpart().

tree <- rpart(RESPONSE ~ ., data=credit.train.df, method="class", parms=list(split="information", loss=matrix(c(0,5,1,0), byrow=TRUE, nrow=2)))  
reg.pred <- predict(tree, credit.valid.df, type = "class")  
mat <- confusionMatrix(reg.pred, credit.valid.df$RESPONSE)  
mat

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 82 139  
## 1 32 147  
##   
## Accuracy : 0.5725   
## 95% CI : (0.5224, 0.6215)  
## No Information Rate : 0.715   
## P-Value [Acc > NIR] : 1   
##   
## Kappa : 0.1819   
##   
## Mcnemar's Test P-Value : 5.23e-16   
##   
## Sensitivity : 0.7193   
## Specificity : 0.5140   
## Pos Pred Value : 0.3710   
## Neg Pred Value : 0.8212   
## Prevalence : 0.2850   
## Detection Rate : 0.2050   
## Detection Prevalence : 0.5525   
## Balanced Accuracy : 0.6166   
##   
## 'Positive' Class : 0   
##

The accuracy is reduced to 57% with a sensitivity of .7193 and specificity of .5140.

In this section I use the data to form and validate a Random Forest regression tree.

library(randomForest)

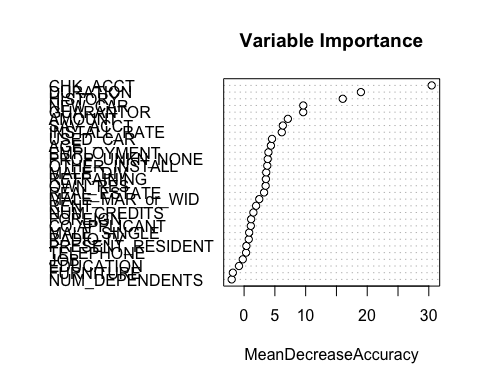
## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':  
##   
## margin

rf <- randomForest(as.factor(RESPONSE) ~ ., data = credit.train.df, ntree = 1000, importance = TRUE)  
varImpPlot(rf, type = 1, main = "Variable Importance")

 This jumbled mess shows Checking Account is the most important variable. I tried a million ways to fix the plot but came up empty or with this mess of a plot.

Next I created a Random Forest using the variables checking account status, history of payments, and duration of credit. # build random forest 2

rf2 <- randomForest(as.factor(RESPONSE) ~ CHK\_ACCT + HISTORY + DURATION + SAV\_ACCT, data = credit.train.df, ntree = 1000, importance = TRUE)

# show confusion matrix

rf.pred <- predict(rf, credit.valid.df)  
rf.pred2 <- predict(rf2, credit.valid.df)   
confusionMatrix(rf.pred, as.factor(credit.valid.df$RESPONSE))

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 44 20  
## 1 70 266  
##   
## Accuracy : 0.775   
## 95% CI : (0.7309, 0.815)  
## No Information Rate : 0.715   
## P-Value [Acc > NIR] : 0.003963   
##   
## Kappa : 0.364   
##   
## Mcnemar's Test P-Value : 2.404e-07   
##   
## Sensitivity : 0.3860   
## Specificity : 0.9301   
## Pos Pred Value : 0.6875   
## Neg Pred Value : 0.7917   
## Prevalence : 0.2850   
## Detection Rate : 0.1100   
## Detection Prevalence : 0.1600   
## Balanced Accuracy : 0.6580   
##   
## 'Positive' Class : 0   
##

Using all of the variables provided an accuracy measure of 70% while using the three variables resulted in an accuracy of 77.5%.

#### Conclusion

The regression model using all the variables has an accuracy of 70% with a sensitivity of .4474 and specificity of .8042. I was able to use a RandomForest regression model to narrow the variables down to CHK\_ACCT, HISTORY, DURATION, and SAV\_ACCT. Using the three variables improved accuracy to 77.5% with sensitivity reduced to .38 and specificity increased to .93. Sensitivity is the percentage of Good Risk customers properly identified while specificity is the opposite – the percentage of Poor Risk customers properly identified. Knowing the costs of the these changes can help determine the appropriate varaibles to included in the decision making.