MIS 510 Portfolio Project Option 1

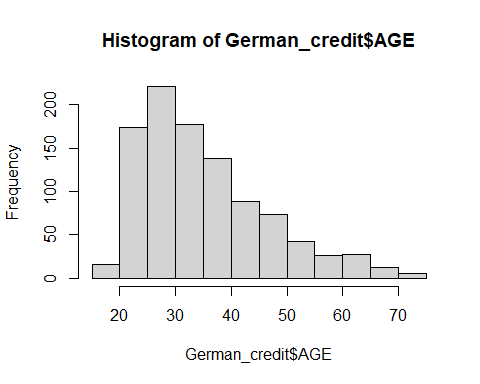
Madelaine Bass

November 22, 2020

## Data Exploration

In this section, import the data and run at least 5 appropriate data exploration functions.

#Import the data  
German\_credit <- read.csv("C:\\Users\\bassg\\Documents\\Grad School\\MIS510\\GermanCredit.csv", header = TRUE)  
  
#Data Exploration Functions  
hist(German\_credit$AGE)



range(German\_credit$AGE)

## [1] 19 75

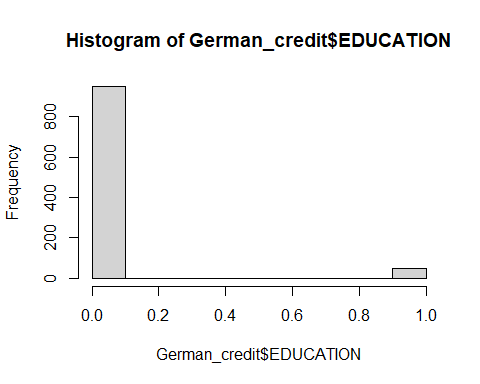
mean(German\_credit$AMOUNT)

## [1] 3271.258

range(German\_credit$DURATION)

## [1] 4 72

hist(German\_credit$EDUCATION)



range(German\_credit$INSTALL\_RATE)

## [1] 1 4

## Training and Validation Partitions

In this section, divide the data into training and validation partitions.

#Partition Data  
set.seed(2)  
train.index <- sample(c(1:dim(German\_credit)[1]), dim(German\_credit)[1]\*0.6)  
train.df <- German\_credit[train.index, ]  
valid.df <- German\_credit[-train.index, ]

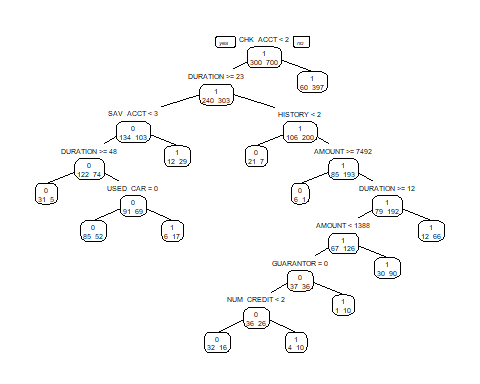
## Classification Models

In this section, explore classification using one of three data mining techniques: logistic regression, classification trees, or neural networks. Analyze the results and include any appropriate visualizations.

# Logistic Regression  
logit.reg <- glm(RESPONSE ~ CHK\_ACCT + DURATION + HISTORY + NEW\_CAR + USED\_CAR + FURNITURE + RADIO.TV + EDUCATION + RETRAINING + AMOUNT + SAV\_ACCT + EMPLOYMENT + INSTALL\_RATE + MALE\_DIV + MALE\_SINGLE + MALE\_MAR\_or\_WID + CO.APPLICANT + GUARANTOR + PRESENT\_RESIDENT + REAL\_ESTATE + PROP\_UNKN\_NONE + AGE + OTHER\_INSTALL + RENT + OWN\_RES + NUM\_CREDITS + JOB + NUM\_DEPENDENTS + TELEPHONE + FOREIGN, data = train.df, family = "binomial")  
summary(logit.reg)

##   
## Call:  
## glm(formula = RESPONSE ~ CHK\_ACCT + DURATION + HISTORY + NEW\_CAR +   
## USED\_CAR + FURNITURE + RADIO.TV + EDUCATION + RETRAINING +   
## AMOUNT + SAV\_ACCT + EMPLOYMENT + INSTALL\_RATE + MALE\_DIV +   
## MALE\_SINGLE + MALE\_MAR\_or\_WID + CO.APPLICANT + GUARANTOR +   
## PRESENT\_RESIDENT + REAL\_ESTATE + PROP\_UNKN\_NONE + AGE + OTHER\_INSTALL +   
## RENT + OWN\_RES + NUM\_CREDITS + JOB + NUM\_DEPENDENTS + TELEPHONE +   
## FOREIGN, family = "binomial", data = train.df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.6963 -0.5903 0.3649 0.7156 2.2628   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.204e+00 1.144e+00 1.052 0.292836   
## CHK\_ACCT 5.516e-01 9.656e-02 5.712 1.12e-08 \*\*\*  
## DURATION -4.645e-02 1.261e-02 -3.685 0.000229 \*\*\*  
## HISTORY 4.896e-01 1.315e-01 3.724 0.000196 \*\*\*  
## NEW\_CAR -1.123e+00 4.955e-01 -2.266 0.023476 \*   
## USED\_CAR 8.601e-01 6.559e-01 1.311 0.189734   
## FURNITURE -5.780e-01 5.207e-01 -1.110 0.266959   
## RADIO.TV -4.536e-01 4.994e-01 -0.908 0.363650   
## EDUCATION -1.258e+00 6.855e-01 -1.835 0.066470 .   
## RETRAINING -5.899e-01 5.602e-01 -1.053 0.292359   
## AMOUNT -5.295e-05 5.942e-05 -0.891 0.372879   
## SAV\_ACCT 2.471e-01 7.968e-02 3.101 0.001929 \*\*   
## EMPLOYMENT 2.238e-01 1.014e-01 2.207 0.027335 \*   
## INSTALL\_RATE -2.330e-01 1.133e-01 -2.056 0.039821 \*   
## MALE\_DIV -4.767e-01 5.073e-01 -0.940 0.347429   
## MALE\_SINGLE 3.285e-01 2.711e-01 1.212 0.225620   
## MALE\_MAR\_or\_WID -4.849e-02 3.974e-01 -0.122 0.902887   
## CO.APPLICANT 6.646e-02 5.686e-01 0.117 0.906956   
## GUARANTOR 1.731e+00 6.129e-01 2.824 0.004736 \*\*   
## PRESENT\_RESIDENT -1.548e-02 1.148e-01 -0.135 0.892742   
## REAL\_ESTATE 2.617e-01 2.819e-01 0.928 0.353217   
## PROP\_UNKN\_NONE -5.628e-01 4.552e-01 -1.236 0.216307   
## AGE 1.358e-02 1.148e-02 1.182 0.237052   
## OTHER\_INSTALL -4.493e-01 2.754e-01 -1.632 0.102770   
## RENT -1.535e-01 5.871e-01 -0.262 0.793681   
## OWN\_RES 8.991e-02 5.538e-01 0.162 0.871034   
## NUM\_CREDITS -5.488e-01 2.293e-01 -2.393 0.016702 \*   
## JOB -2.046e-01 1.896e-01 -1.080 0.280342   
## NUM\_DEPENDENTS -1.979e-01 3.523e-01 -0.562 0.574327   
## TELEPHONE 4.130e-01 2.580e-01 1.600 0.109488   
## FOREIGN 1.060e+00 6.384e-01 1.661 0.096750 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 726.13 on 599 degrees of freedom  
## Residual deviance: 522.84 on 569 degrees of freedom  
## AIC: 584.84  
##   
## Number of Fisher Scoring iterations: 5

#Classification tree  
library(rpart)  
library(rpart.plot)  
class.tree <- rpart(RESPONSE ~ CHK\_ACCT + DURATION + HISTORY + NEW\_CAR + USED\_CAR + FURNITURE + RADIO.TV + EDUCATION + RETRAINING + AMOUNT + SAV\_ACCT + EMPLOYMENT + INSTALL\_RATE + MALE\_DIV + MALE\_SINGLE + MALE\_MAR\_or\_WID + CO.APPLICANT + GUARANTOR + PRESENT\_RESIDENT + REAL\_ESTATE + PROP\_UNKN\_NONE + AGE + OTHER\_INSTALL + RENT + OWN\_RES + NUM\_CREDITS + JOB + NUM\_DEPENDENTS + TELEPHONE + FOREIGN, data = German\_credit, control = rpart.control(maxdepth = 30), method = "class")  
prp(class.tree, type = 1, extra = 1, split.font = 1, varlen = -10)



## Logistic Equation

## Analysis

Based on the summary output of the logistic regression function, I would include the variables CHK\_ACCOUNT, DURATION, HISTORY, NEW\_CAR, SAV\_ACCT, EMPLOYMENT, INSTALL\_RATE, GUARANTOR, and NUM\_CREDITS based on an alpha of 0.05 compared to each variable's p-value. Of course this could change if using some sort of step-wise regression. The equation when using every single variable can be found above. When credit is good, i.e. RESPONSE = 1, then the impact of the intercept would be 1.204, CHK\_ACCOUNT 0.5516, DURATION -0.04645, HISTORY 0.4896, NEW\_CAR -1.123, USED\_CAR 0.8601, FURNITURE -0.578, RADIO.TV -0.4536, EDUCATION -1.258, RETRAINING -0.5899, AMOUNT -5.295e-5, SAV\_ACCT 0.2471, EMPLOYMENT 0.2238, INSTALL\_RATE -0.233, MALE\_DIV -0.4767, MALE\_SINGLE 0.3285, MALE\_MAR\_or\_WID -0.04849, CO.APPLICANT 0.06646, GUARANTOR 1.1731, PRESENT\_RESIDENT -0.01548, REAL\_ESTATE 0.2617, PROP\_UNKN\_NONE -0.5628, AGE 0.01358, OTHER\_INSTALL -0.4493, RENT -0.1535, OWN\_RES 0.08991, NUM\_CREDITS -0.5488, TELEPHONE 0.413, and FOREIGN 1.06.

As for the classification tree, we can see that it determined which two variables were necessary: CHK\_ACCT and DURATION. Just to be sure that this wasn't an error, when taking out the varaible CHK\_ACCT the new decision tree will use HISTORY and SAV\_ACCT. But when changing the maximum depth to 30-- which is the true maximum depth that can be chosen-- only 11 variables are used. These variables are CHK\_ACCT, DURATION, HISTORY, SAV\_ACCT, DURATION (again but at greater than or equal to 48 and not just 23), USED\_CAR, AMOUNT, DURATION (at greater than or equal to 12 instead of 23), AMOUNT (at less than 1388 and not at greater than or equal to 7492), GUARANTOR, and NUM\_CREDIT.

## Reflection

The first section of this assignment was to import the data into RStudio and run a minimum of five data exploration funcitons. Data exploration functions give the opportunity to explre the data and get a feel for waht it is telling you. The results of these data exploration functions can help guide you towards what type of analysis you should conduct. In this section, I chose to run a histogram of the variables AGE and EDUCATION. I chose these so that I could see how the data was spread. I wanted to see what age most of these people were and if most of them were educated or not. I can tell that based on the information, most of these people are between the age of 20 and 40 and that most of these people are not educated -- under the assumption that 1 is a "yes" for education and a 0 is a "no" for education. I also chose to find the range of the variables AGE, DURATION, and INSTALL\_RATE. I wanted to see what range of ages this data set looked at, the duration of credit per person, and the installment rate. I finally found the mean of the AMOUNT variable. From this I can tell that the average credit amount is about 3271. This section posed no issues to complete. These data exploration functions have been conducted in numerous assignments by now and therefore pose little to no difficulty.

The second part in this assignment was to partition the data into training and validation parts. This is typically used to divide the data set into two parts, one to train a model and one to validate it's accuracy. If the model that was created using the training data can accurately predict the validation set, then you have a good model. I used the code in table 10.2 to do this. I was nervous that this section might cause issues due to my unfamiliarity utilizing this code. However, despite my apprehension, I had no issues running this section. In fact, once I found the section to model my code off of, I had no issues. It ran the first time with no errors.

The final section of this assignment was to explore classification models using two of three techniques. I chose to use logistic regression and a classifiation tree. Since we have previously run logistic regression functions, I was confident in my ability to run this function. I ran this function the first time with no issues. As I stated above, I would suggest using only significant predictors in the model. There are many reasons for this; the main two being that the model is incredibly long and that you wouldn't be completely accurate in your predictions by utilizing insignificant predictors. While, in this case, the significant predictors often had larger coefficients, that is not always the case. While sometimes a long list of variables are significant predictors, in this case that is not true. It kind of feels like over kill with how many variables are used that aren't significant. The classification tree is a good way to visually see a model and how to get to a final result using various levels within variables. I liked this classification model because I could visually see which requirements got you to each outcome; good or bad credit. I was not aware that a classification tree could use the same variable more than once. This could be seen through the model above how DURATION and AMOUNT were used more than once. I also learned what the "Depth" meant in the code. I originally thought that there were only two variables needed to predict the RESPONSE. This was not true when I changed the depth from 2 to 30. Figuring this out helped me create the model above. When creating the logistic regression and decision tree models, the only issues were that initially I included "OBS" as a variable and that I had to figure out which packages to install to get the classification tree to run properly. I could not quite figure out how to only look at select variables so instead I had to create my models not using OBS as a variable.

Overall this project was not super difficult to complete. Other than figuring out which packages to install, minimal errors presented themselves. This was also a good way to show the different classification models that can be used on the same data set. Some are more visual than others. A classification tree is a visual way to display how to come to the conclusion of which scenarios mean a person has good or bad credit. A logistic regression function can show how to calculate if a person has good or bad credit.