Jonothan Meyer
MTH 3270, Data Science
Midterm/Project 2
05/01/21

1. Carry out a multiple regression analysis. You may choose any response variable (Y) for your model, but it must be a numerical variable (not categorical). Likewise, you may use any explanatory (X) variables, but they too must be numerical (not categorical). Note that a categorical variable that's been coded using integer values is still considered to be a categorical variable.

Data Wrangling:

Data wrangling for this assignment was fairly straightforward and simple. It mostly just involved utilization of the 'dplyr' library with the 'select()', 'mutate()', and 'filter()' functions. 'PRMINC1' values were converted from their original form to 0's and 1's to ensure appropriate interpretation of the logistic regression model in the second portion of this assignment. In a few instances NA's were filtered out of the data as well. For a couple of the models a training sample of the data was used because otherwise the computational load the entire data set put on the computer caused the program to crash.

- Summarize your fitted model by reporting the estimated model coefficients.

For this model I wanted to look at the relationship between predictors that could represent the supposed size of a business and construct and multiple linear model to determine their effect on the percentage of the business owned by the first owner. I speculated prior to building the model that the size of a business has a negative relationship on this response. The thinking behind this was that as a business grows in size the more likely it is to be owned by more owners because of the need for capital investment as a business grows, and also the necessity to partition the head responsibilities to more people as those responsibilities grow with the size of the business.

The multiple regression linear model initially used 'PCT1' as the response, and 'EMPLOYMENT_NOISY', 'PAYROLL_NOISY', and 'RECIPTS_NOISY' as the predictors. However, after building the model the first predictor, 'EMPLOYMENT_NOISY' was deemed to have an insignificant effect on the response (>.20) and was therefore removed. Although somewhat surprising intuitively this does make some sense: as the amount of employee's at a business grows the need for managers and supervisors will proportionally grow with it, but it does not necessarily scale with the actual owners of the company in any significant way.

The final model was then built using the following command in R: lm(PCT1~EMPLOYMENT_NOISY+PAYROLL_NOISY, data = sb) where 'sb' is the 'small business' data set. Both of the predictors were determined to have a very significant effect on the response, both independently with a p-value ~2x10^-16, and a model combined p-value of 2x10^-16. The model results suggest that my initial inclination about the relationship of the size of business on the response does seem to have a significant effect.

- Interpret the estimated model (coefficients).

Coefficients	**Estimate**
Intercept	82.72

PAYROLL_NOISY	-0.0001864
RECEIPTS_NOISY	-0.00001157

When holding the other predictor constant, we can expect that for every 1 unit increase in 'PAYROLL_NOISY' we can expect that the percentage the first owner of the business holds goes down by -0.0001864. Also, when holding the other predictor constant, we can expect that for every 1 unit increase in 'RECEIPTS_NOISY' that the percentage the first owner of the business holds goes down by 0.00001157.

Both of these coefficient estimates make sense within the context of the data set, and we would expect them both to be very small. This is because a variable dealing with percentage will always exist within the range of 0 to 100, which is relatively very small when compared to the amount of money a business is paying out via payroll, or pulling in with receipts. Therefore the actual value of the coefficient is not expected to be especially accurate or relevant, but the sign of it is more illuminating. Since both of these are negative it does support my initial inclination that the size of the business tends to negatively affect the percentage the first owner holds.

- Report the value of at least one measure of how well the model fits the data (e.g. the R2).

For this model the coefficient of determination was found to be $R^2 = .0031$. That a statistical measure that represents the proportion of the variance for the response variable that's explained by the predictor variables. R^2 was found to be extremely small in this instance, and leads me to believe this model in its current form might not accurately predict percent ownership by the first owner. However, the data set has a large amount of observations, and has a wide distribution. So it is somewhat expected that this value will be rather low, and may be the case that the model is still fairly accurate. More investigative work as to the accuracy and utility of the model would need to be examined.

2. Carry out a logistic regression analyses for predicting whether a business is the primary source of income for the first owner based on other explanatory variables from the data set. For the response (Y) variable, you'll use the dichotomous PRMINC1 variable taking the value 1 if yes and 2 if no. You may use any explanatory (X) variable(s), but they must be numerical (not categorical).

A logistic regression was carried out using the 'glm()' function in R's library with the 'binomial' family para meter. The values in the 'PRMINC1' were converted from 1 and 2 to 0 and 1 using the 'mutate()' function in the dyplr library within R, with 1 meaning the business was considered the owners primary form of income. The logistic regression model was built using 'PRMINC1' as the response and 'HOURS1', 'EDUC1', 'EMPLOYMENT_NOISY', 'PAYROLL_NOISY', 'RECEIPTS_NOISY', and 'AGE1'. This was able to provide some insight into what variables influence whether the first owner of a business holds the business as their primary or secondary form of income. It is speculated that among the predictors used that the amount of hours worked at the business by the owner would be one of the best predictors for determine the response, and was measured by the 'HOURS1' column. Additionally it is speculated that the size of the business, which was determined by number of employees ('EMPLOYMENT_NOISY'), the amount of employee expenses ('PAYROLL_NOISY'), and the general amount of revenue ('PAYROLL_NOISY') would also have a significant effect on the response. The age of the owner ('AGE1'), as well as their education level (EDUC1) is speculated to have a relatively smaller effect on the response, but were included for exploratory purposes. The final logistic regression model was built, analyzed, and interpreted using the

following R commands:

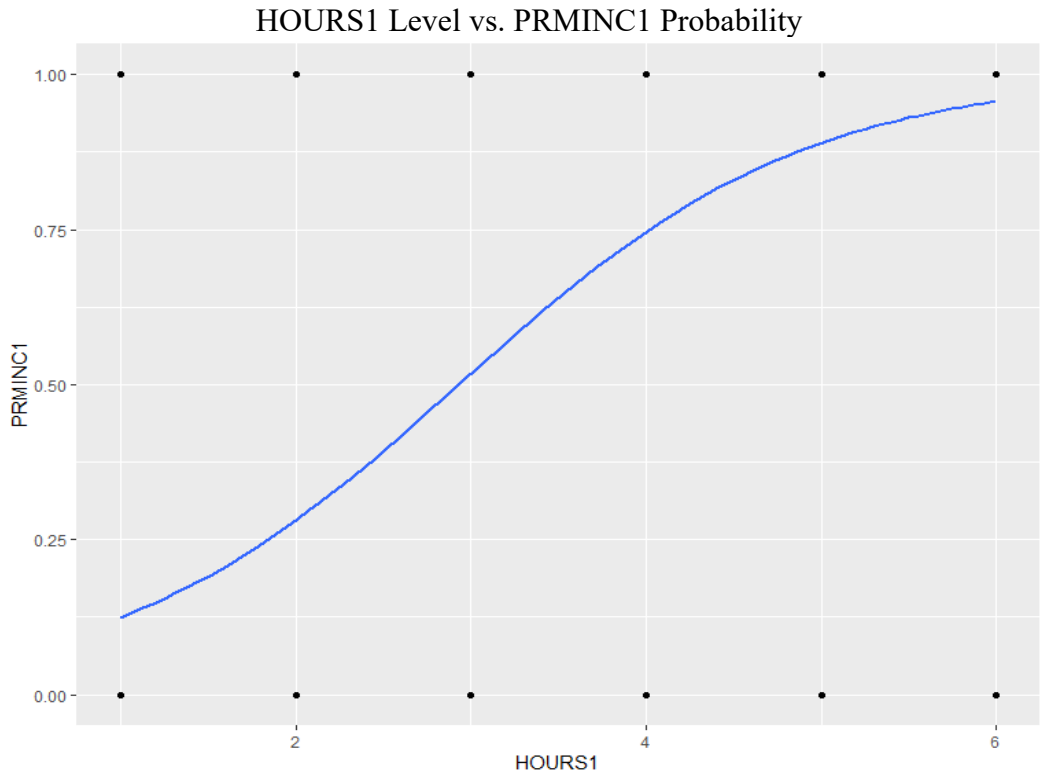
```
employ.glm <-glm(PRMINC1~HOURS1+EDUC1+EMPLOYMENT_NOISY+PAYROLL_NOISY+
        RECEIPTS_NOISY+AGE1, data = sb, family = "binomial")
summary(employ.glm)
confint(employ.glm)
```

-Summarize your fitted model by reporting the estimated model coefficients.

All of the predictors other than 'RECEIPTS_NOISY' were found to have a significant p-value with $\alpha = .95$. This is somewhat to be expected with a data set as large as the one used, so the significance of some variables should be taken with a grain of salt, and need further examination to deem them significant on the response. However, that is outside the scope and requirements of this project, and therefore was not explored further. The following coefficients were found for the predictors in the following table:

Coefficient	Estimate
Intercept	-2.5100000
HOURS1	1.0000000
EDUC1	0.6480000
EMPLOYMENT_NOISY	0.0004230
PAYROLL_NOISY	0.0000726
RECIPTS_NOISY	0.0000001
AGE1	-0.0387000

The amount of hours a business owner works per week holds the most weight within our model for determining if that business is the owners primary form of income. Within the context of this model and the value of the intercept this means that when holding all other predictors constant that it is estimated with a probability >50% that an owner has their business as a primary form of income if they work level 3 or more hours per week. This can also be observed from plotting the relationship between 'HOURS1' vs. 'PRMINC1' using 5000 random samples:



The results of the regression model and this observation of the data are relatively consistent. They both indicate that when an owner works level 3 or more hours at their business that it becomes more probable than not that the business is their primary source of income.

To test the accuracy of the logistic regression model 5000 observations from the data set were sampled and used to predicted the response. The predicted value was then compared against the actual value to determine its accuracy across the sample size. This was performed using the following R code:

```
pred_index <- sample(1:nrow(sb), 5000, replace = FALSE) # Pulling x amount of random samples
y_pred <- predict(employ.glm, sb[pred_index,], type = "response") # Making predictions
y_pred <- round(y_pred) # Rounding those predictions to 0 or 1
act_pred <- data.frame(Actual = sb[pred_index,]$PRMINC1, Predicted = y_pred) # Finding actual
confusion <- table(act_pred) # Building a confusion table with the predictions and answers
length_values <- act_pred %>% filter(!is.na(Actual), !is.na(Predicted)) # Removing NA rows
sum(diag(confusion)) / nrow(length_values) # Finding accuracy of model
```

Generally this model is ~80% accurate with the false predictions fairly evenly spread over false positives, and false negatives as can be observed in the following confusion matrix:

	Predicted	
Actual	0	1
0	671	282
1	229	1601

3. Carry a machine learning classification procedure (decision tree, random forest, k nearest

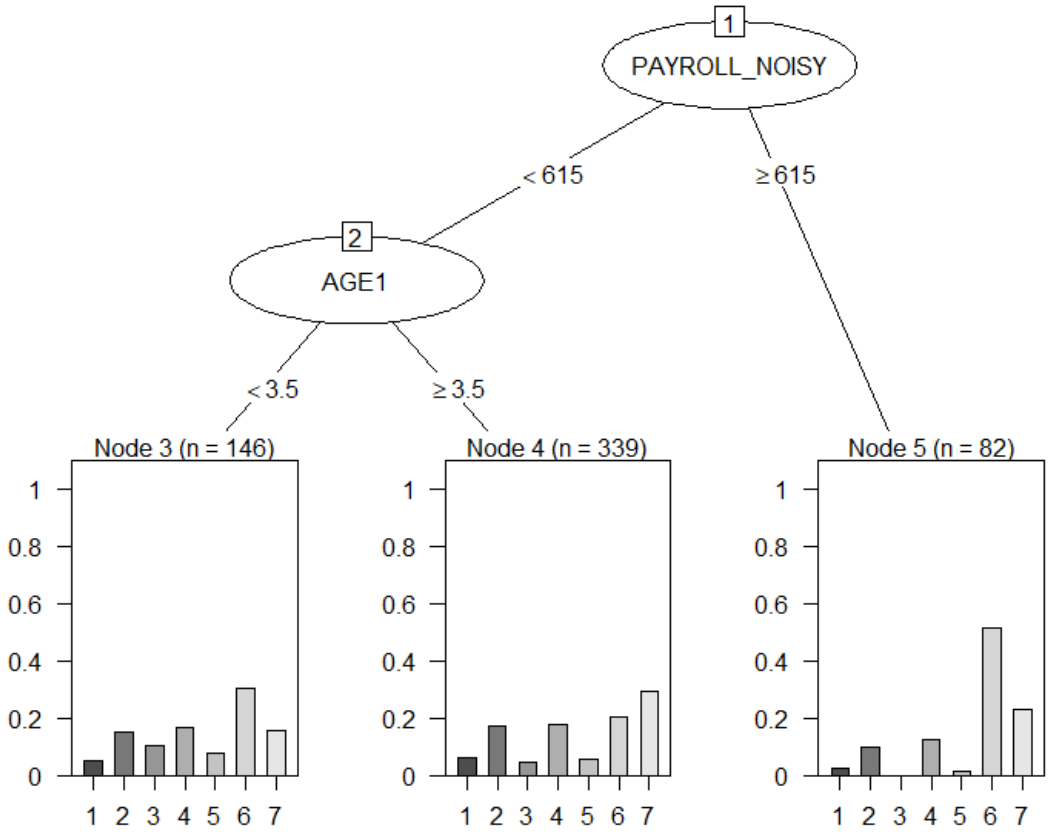
neighbor, or artificial neural network { your choice) for predicting one of the following categorical variables (your choice). You may use any explanatory (X) variables, but they must be numerical (not categorical).

Education level of the first business owner EDUC1. You'll need to convert the 1, 2, ..., 7 values to "character" (so they won't be treated as numerical responses by the model-fitting function in R) using as.character() with mutate().

Two separate machine learning methods were performed to explore the affect of various numerical predictors in determining the education level of the first owner of small business's. The objectives outlined in the assignment are satisfied using the second method. Random Partition was the first machine learning method used, and was primarily performed for the purposes of practicing, building my data acumen, as well as understanding the relationship between the predictors on the response. The predictors used were 'EMPLOYMENT_NOISY', 'PAYROLL_NOISY', 'RECEIPTS_NOISY', 'AGE1', and 'PCT1'. The conversion method for converting the response to a characters was not working properly, and for this reason an alternative method ('factor(EDUC1)') was used to ensure that R did not interpret them as numerical. The computer used to build the decision tree was greatly struggling when attempting to build it using the whole data set. Since this was a consistent issue that caused many crashes the decision was made to sample from the data for the purpose of lowering the computational load on the computer. The sampling procedure, construction of the decision tree, and visualization was performed using the following R code:

```
sb <- small_business %>% select(AGE1, PAYROLL_NOISY, EMPLOYMENT_NOISY,
                  RECEIPTS_NOISY, PRMINC1, EDUC1,HOURS1,PCT1, FAMILYBUS)
index <- sample(1:nrow(sb), 1000, replace = FALSE)
my.tree <- rpart(factor(EDUC1) ~ EMPLOYMENT_NOISY+PAYROLL_NOISY+
                                  RECEIPTS_NOISY+AGE1+PCT1, data = sb[index,])
my.party.tree <- as.party(my.tree)
plot(my.party.tree)
```

This procedure was used many times to get a general idea of the structure of decision tree consistently produced. In general it seems as though the larger a business and the greater the age of the owner typically means that they have a higher level of education. The plot below serves as a general representation of a trend that was observed over many different samples:



The x-axis on each of the lower plots shows the proportion within the constant given above of the breakdown of amount of owners education level (1-7 as factors). In general it seems as though business owners with a payroll above 615 tend to have an education level above an associate degree.

The other machine learning procedure constructed was a random forest. This was done with the same response and predictors used previously other than 'AGE1', but differs in that it was made for predicting a business owner's level of education. Similarly to the previous method used, the computer had trouble building the forest using the entire data set, and for this reason a random sample of 5000 training datum were used to build the forest. 800 tree's were chosen for the forest with all of the variables available for splitting at each tree node. The reason these values were chosen was because over many iterations they seemed to provide the highest and most consistent accuracy for the testing data. The forest was built and predictions made using the following R commands:

```
sb <- small_business %>% select(AGE1, PAYROLL_NOISY, EMPLOYMENT_NOISY,
        RECEIPTS_NOISY, PRMINC1, EDUC1,HOURS1, PCT1) # Grabbing relevant data
sb1 <- sb[complete.cases(sb),] # Removing NA's
index <- sample(1:nrow(sb1), 5000, replace = FALSE) #Pulling training sample
my.forest <- randomForest(factor(EDUC1) ~ EMPLOYMENT_NOISY+PAYROLL_NOISY+
                          RECEIPTS_NOISY+AGE1+PCT1,
                          data = sb1[index,],
                          ntree = 800,
                          mtry = 5) #Building forest
index2 <- sample(1:nrow(sb1), 1500, replace = FALSE) #Pulling testing data
```

```
pred <- predict(my.forest, newdata = sb1[index2,]) # Making predictions
act_pred <- data.frame(Actual = sb1[index2,]$EDUC1, Predicted = pred) #Organizing data
confusion <- table(act_pred) # Building confusion matrix
sum(diag(confusion)) / nrow(sb1[index2,]) # Finding accuracy
```

Generally this procedure had an accuracy of ~25%. At first glance this might seem low, but considering the response has 7 different levels this is actually fairly decent. If a program or person were to randomly guess the education level of any given owner with no information they would be correct ~14.29% of the time. So this is the benchmark by which the program must beat in order to be considered more accurate in its predictions than totally random. Considering the forest training data was built with only 5000 observations its accuracy is actually fairly impressive.

Using this forest we can predict that an owner with 1 employee, pays 0 in payroll, has 50 in receipts, and owns 10% of the business will have an education level of 2, or a high school diploma. We would also predict that an owner with 100 employees, pays 5,000 in payroll, has 10,000 in receipts, and owns 99% of the business will have an education level of 6, or a bachelors degree. We input these values that resulted in the subsequent predictions using the following R commands:

```
newdata <- data.frame(EMPLOYMENT_NOISY = c(2, 100), PAYROLL_NOISY = c(50, 5000),
                      RECEIPTS_NOISY = c(50, 10000), PCT1 = c(10, 99))
predict(my.forest, newdata = newdata, type = "class")
```