Jonothan Meyer
Data Science - MTH 3270
05/13/21
Final Project

Your first task is to separate the small business data set randomly into 75% training and 25% testing sets, then fit one of the above machine learning models (your choice) to the training set using using at least three different values of the tuning parameter, and compare the effectiveness of each model for classifying individuals in the test set. Your model should predict one of the following categorical variables (your choice). You may use any explanatory (X) variables, but they must be numerical (not categorical).

The data was wrangled and separated into training/testing sets using the following R commands:

```
# Separating data into training/testing
sb_knn <- small_business %>% select(AGE1, PAYROLL_NOISY, EMPLOYMENT_NOISY,
                                    RECEIPTS_NOISY, PCT1, EDUC1)
rm(small_business) # This variable uses up a ton of memory. So after I have what I need it is removed
                   # from the global environment
sb_knn <- sb_knn[complete.cases(sb_knn),] #Removing all NA's

set.seed(1) # Making sure sample draws are consistent
data_set_size <- round(floor(nrow(sb_knn))*.75) # Getting size for training data (75%)
indexes <- sample(1:nrow(sb_knn), size = data_set_size) # Getting Indices for training data
training <- sb_knn[indexes,] # Pulling training data
testing <- sb_knn[-indexes,] # Pulling testing data
```

Summarize your procedure: Indicate which classification procedure you used, which (categorical) response variable you were predicting, and which explanatory variables you used.

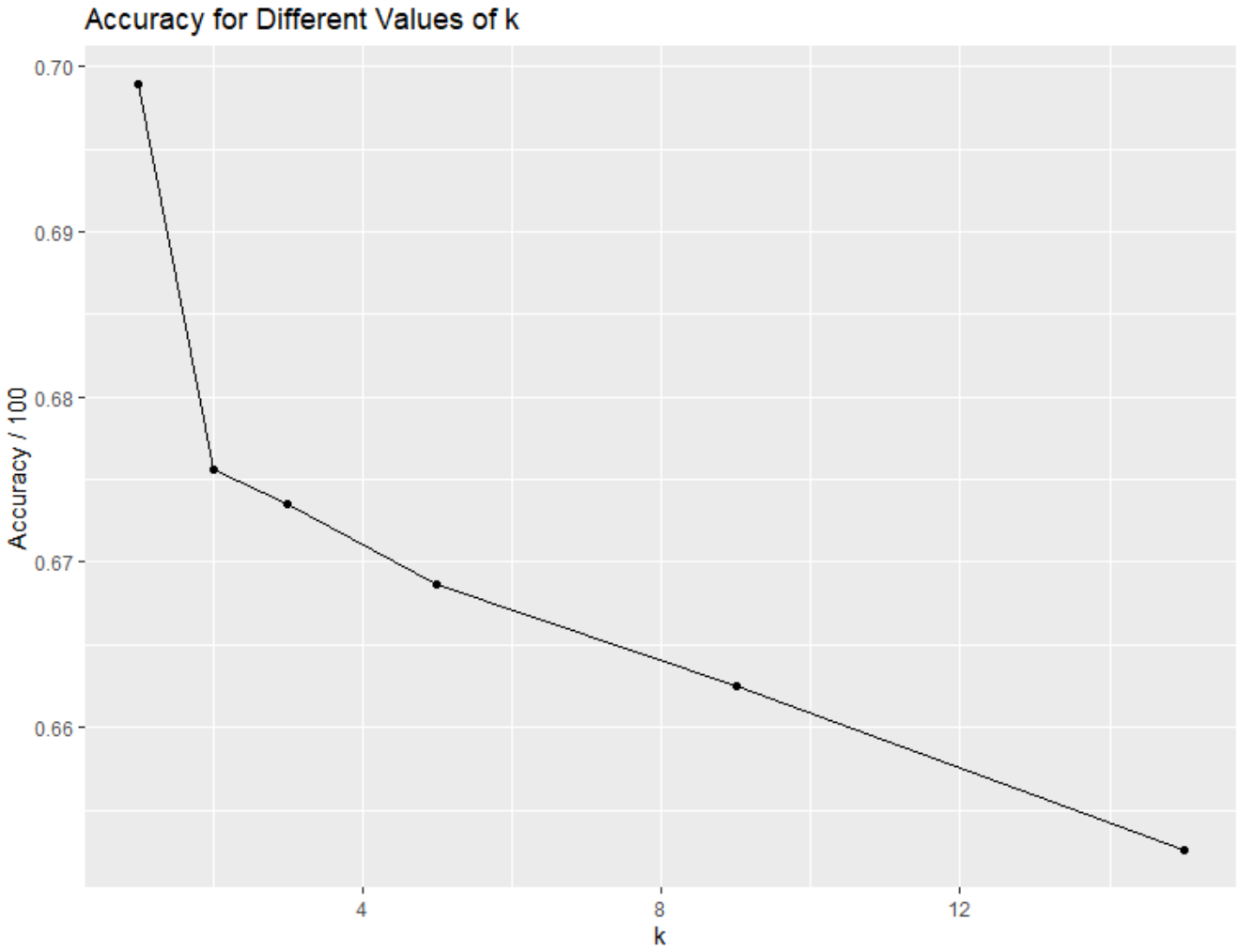
For this project I wanted to explore the education level (EDUC1) of the first owner of a business more in depth, and chose that as my response. KNN was used as the classification procedure with most of the numerical variables as the predictors or explanatory variables. These variables included age of the first owner (AGE1), how many people were employed at the business (EMPLOYMENT_NOISY), percentage of the business owned by the first owner (PCT1), age of the first owner (AGE1), amount paid to employee's (PAYROLL_NOISY), and income of the business (RECIEPTS_NOISY). All these variables were used to build the KNN classifier with different values for k for the purpose of attempting to predict the education level of the first owner of the business. This procedure was carried out using the following R commands:

```
nrow(training)/(nrow(training)+nrow(testing)) # Making sure training data proportions are correct
nrow(testing)/(nrow(training)+nrow(testing)) # Making sure testing data proportions are correct

# BUILDING AND ASSESSING KNN W/VARIOUS VALUES OF K
timer1 <- createTimer() #Want to time how long all the predictions take
timer1$start("Making y_pred, diff k's") # Starting timer
```

```
# Knn w/ k = 1 has a 69.88% accuracy rate
y_pred1 <- knn(train = training, test = testing, cl = factor(training$EDUC1), k = 1)
conf1 <- confusionMatrix(y_pred1, factor(testing$EDUC1))
c1 <- conf1$overall[1]
c1 # .6988

# This process was repeated several times using different values for k
```

Report the values of the tuning parameter you evaluated, and which of these values resulted in the best classifier of individuals in the testing set, e.g. which one had the highest correct classification rate for this set.

The tuning parameter used to refine the classifier and find the best accuracy was k using values 1, 2, 3, 5, 9, and 15. Gold was struck with the very first classifier, k = 1, which was able to predict with the highest accuracy of 69.88% correct classification! The classification rate and accuracy were obtained by adding up the 'correct' classifications from the resulting confusion matrix and dividing by the total amount of attempted classifications. Since this process was performed on the entire data set (without NA's), as well as used different values of k, the process was very computationally extensive for the computer. When timed it was determined that the whole process took ~9 minutes to carry out. The performance of all classifiers can be observed in the plot below.



k = 1 drastically outperformed all other values of k, with a sharp drop off in accuracy with k = 2 and beyond. It appears as though the most accurate way with with to predict the level of education of a business owner is to look at the education level of the most similar business in terms of these explanatory variables.

k Value	1	2	3	5	9	15
Accuracy	0.6988	0.6756	0.6735	0.6686	0.6625	0.6526

2. Your second task is to carry out a cluster analysis (hierarchical or k means, your choice) to group the businesses into k clusters, where k is in the range 2-5 (your choice). You must use four or more explanatory (X) variables in the cluster analysis, and they must be numerical (not categorical). It's your choice which ones to use.

The clustering procedure was conducted using the following R commands:

```
sb1 <- small_business %>% filter(SECTOR == 54 & FRANCHISE == 1) %>%
  select(EMPLOYMENT_NOISY, PAYROLL_NOISY, RECEIPTS_NOISY, PCT1, PRMINC1)
sb1 <- sb1[complete.cases(sb1),] # Removing NA's
nrow(sb1) # Checking how many observations are left
center <- 3 # Choosing the 'amount' of clusters
sb1_clust <- sb1 %>% kmeans(centers = center) %>% fitted("classes") %>% as.character() #clustering
clust <- kmeans(sb1, centers = center)
sb1 <- sb1 %>% mutate(cluster = sb1_clust) # Adding cluster classification column to data
sb1 %>% group_by(cluster) %>% summarize (n = n()) # Summarizing clustering
sb2 <- sb1 %>% select(-cluster)
pairs(sb2,
      col = clust$cluster,
      main = "Scatterplot Matrix of Small Business",
      pch = 19) # Looking for clustering classification in terms of PRMINC1
```

Summarize your procedure: Indicate which cluster analysis procedure you used, how many groups k you used, and which explanatory variables you used, and how many observations ended up being in each of the k clusters (groups).

k means was chosen as the type of cluster analysis that would be performed on the data. For this analysis a value of k = 3 was chosen for the amount of clusters within the data set. The explanatory variables chosen for this analysis were 'EMPLOYMENT_NOISY', 'PAYROLL_NOISY', 'RECEIPTS_NOISY', and PCT1. After filtering out NA's and selecting the data based upon SECTOR == 54, and FRANCHISE == 1 the cluster analysis was performed on 255 observations. Of these observations 24 were grouped in cluster 1, 228 were grouped into cluster 2, and 3 were grouped into cluster 3.

Cluster	1	2	3
n	24	228	3

Report the results of your assessment of whether the clusters seem to correspond to businesses that largely are or aren't the primary source of income for the first owner.

For cluster groups 1 and 2 it seems as though the clustering was independent of whether the business was the owners primary form of income. However, the third cluster group exclusively contained owners who used their business as their primary form of income. Looking at the pairs() output below it is observed that clustering largely seems to correspond with the size of a business where the black dots are cluster 1, the green dots are cluster 2, and the pink dots are cluster 3. The most defined separation between clusters can be observed on the RECEIPTS_NOISY variable. Looking at row 5, column 3 shows that cluster 1 consists of business's that have RECEIPTS_NOISY <= ~1000, while cluster 2 contains business's between ~4500 >= RECEIPTS_NOISY > 1000, and cluster 3 is the business's > ~4500, and is all owners who use their business as a primary form of income.