CYO-reddit

Scott Boersma

12/30/2020

#Introduction

The data set was downloaded from https://www.kaggle.com/michaelkitchener/mbti-type-and-digital-footprints-for-reddit-users. It is called the MBTI type and digital footprint for reddit users. Reddit is a website that is "the front page of the internet" it has many subsections as well. Each row contains anonymized reddit user's MBTI personality type and each column represents how much a user posts or comments in a particular subreddit. Specifically, the 'posts_ examplesubreddit'is how many users from the top 100 posts of all time are in 'posts_ examplesubreddit', and 'comments_examplesubreddit' is how many users from the top 100 posts of all time are in 'posts_ examplesubreddit'.

The data was obtained using the PRAW (Reddit's API wrapper for python) to scrape a list of users who comment on the r/mbti subreddit along with their self identified MBTI type (as is shown with their flair). Then, each user whose MBTI type we are are aware of, we go through their top 100 posts and newest 100 comments to record teh frequency of interactions with various subreddits. This then creates the user-footprint matrix.

The purpose of this data set is to see how well MBTI personality types (or even just specific traits i.e. extroversion vs. introversion) can be predicted on the basis of a user's subreddit interactions.

The scientific community regards the MBTI as an illegitimate personality test. Thus said, both extroversion/introversion and sensing/intuition are correlated strongly with extroversion and openess as measured by the widely accepted big 5 model of personality. With this in mind, it was determined that instead of attempting to predict all 16 types, only predicting introversion or extroversion.

The goal of this project was to see how easily it is to predict if a person is an extrovert or an introvert based upon their reddit interactions.

The document consists of an introduction, an overview, a summary, method and analysis.

Overview

The prompt of this project is to create your own. The reasoning behind this particular data set is because I am a reddit user and the ability to get information from a website and turn that into actionable tasks is a must have skill. It is possible for one to come to the conclusion that nearly any information can be found on the internet with the right tools.

This data set has 3,586 different users scrapped data from 27,091 different posts and comments from reddit. The first column is a four letter string consisting of one of the 16 MBTI types. All the rest of the data is a number of interactions from 0 to 100.

The goal of this project is to see how well a simple agorithm can be build to predict introversion or extroversion.

1. Data preparation: download, parse, import, and prepare the data to be processed and analyzed.

2. Data exploration: explore the data to understand the variables, relationships between them, and where possible predictors lie.
3. Data analysis and modeling: creating the model based on insights from the exploration of the data set.
4. Results
5. Conclusion

Executive Summary

The data is first downloaded and is already in tidy format and due to the simplicity of the data requires no cleaning. The initial exploration was done using the original data set. This was done to ensure that no data was missing from the exploration as some of the types only have a small number of observations. Additionally with the over 27,000 variables, a principle component analysis was done to reduce the dimensions of the data. Several algorithms were trained and tested to come up with the best combination or singular algorithm.

Method

Since MBTI is widely rejected by the scientific community, only extroversion and introversion were predicted. This also increases the amount of data for each type.

The main work of this project was in the dimension reduction. Here a principle component analysis was done. The types were split off and measurements were taken of column in order to see the spread in the data. The standard deviations were plotted against each PC to decide where the cut off was as far as how much of the variance was accounted. The amount of variance to take was by getting the most variance with the least amount of components. Based on the amounts it was decided to cover for 80 percent of the variance. Before this cut off point there is still a steep curve and after this point it gets much closer to a straight line.

The data was split into a train set of 80 percent and a test set of 20 percent because this is a widely accepted split and to ensure the algorithms have ample data to train. Then finally several algorithms were trained and put to the test individually and as an ensemble.

```
if(!require(tidyverse)) install.packages("tidyverse", repos = "http://cran.us.r-project.org")
if(!require(caret)) install.packages("caret", repos = "http://cran.us.r-project.org")

library(tidyverse)
library(caret)

reddit <- read.csv(unzip("./data/reddit_psychometric_data.zip"))
```

Set up

Inital Data Exploration Exploration starts with looking at the dimensions of the data set. The head of the data set was also looked at and due to the large number of variables only the first six columns were viewed.

```
dim(reddit)
```

```
## [1]  3586 27091
```

```
head(reddit[1:6])
```

```
##   mbti_type post_AskRedditAfterDark post_unpopularopinion post_introvert
## 1      INTP                       3                     31              3
## 2      ENTP                       0                      0              0
## 3      INFJ                       0                      0              0
## 4      INFP                       0                      0              0
## 5      ENTP                       0                      0              0
## 6      ENFP                       0                      0              0
##   post_UnpopularFacts post_changemyview
## 1                   1                 4
## 2                   0                 0
## 3                   0                 0
## 4                   0                 0
## 5                   0                 0
## 6                   0                 0
```

```
types_table <- reddit %>%
  group_by(mbti_type) %>%
  summarise(total = n()) %>%
  arrange(total)
types_table
```

```
## # A tibble: 16 x 2
##    mbti_type total
##    <chr>     <int>
##  1 ESFJ         13
##  2 ESTJ         19
##  3 ESFP         29
##  4 ISFJ         32
##  5 ESTP         44
##  6 ISTJ         46
##  7 ENFJ         53
##  8 ISFP         77
##  9 ISTP        147
## 10 ENTJ        148
## 11 ENFP        320
## 12 INFJ        367
## 13 ENTP        409
## 14 INTJ        493
## 15 INFP        596
## 16 INTP        793
```

Through this initial look at the data, it is seen the first column is users self identified types. To better understand the overall set of the data, the first column is removed from the table and more exploration is done. It is also seen the distribution of each of the 16 types are widely distributed.

```
# storing the types
mb <- reddit[1]

# storing the interactions
r_dat <- reddit[-1]
```

```
namez <- colnames(r_dat)


# getting an idea of number of interactions
# per person and per sub_reddit
s_col <- colSums(r_dat)
s_row <- rowSums(r_dat)
```

```
# understanding the total number of interactions per redit
max(s_col)
```

```
## [1] 39682
```

```
min(s_col)
```

```
## [1] 1
```

```
mean(s_col)
```

```
## [1] 16.44758
```

```
sd(s_col)
```

```
## [1] 277.9007
```

```
# working with per user
max(s_row)
```

```
## [1] 200
```

```
min(s_row)
```

```
## [1] 1
```

```
mean(s_row)
```

```
## [1] 124.2513
```

```
sd(s_row)
```

```
## [1] 48.15418
```

This shows the extreme variability in the features and in the observations. There are some features that have only one observation while the max is nearly 40k. The features are less variable however they do cover a wide range of 1 to 100.

```
col_df <- as.data.frame(cbind(columns = namez, totals = as.numeric(s_col)))
head(col_df)
```

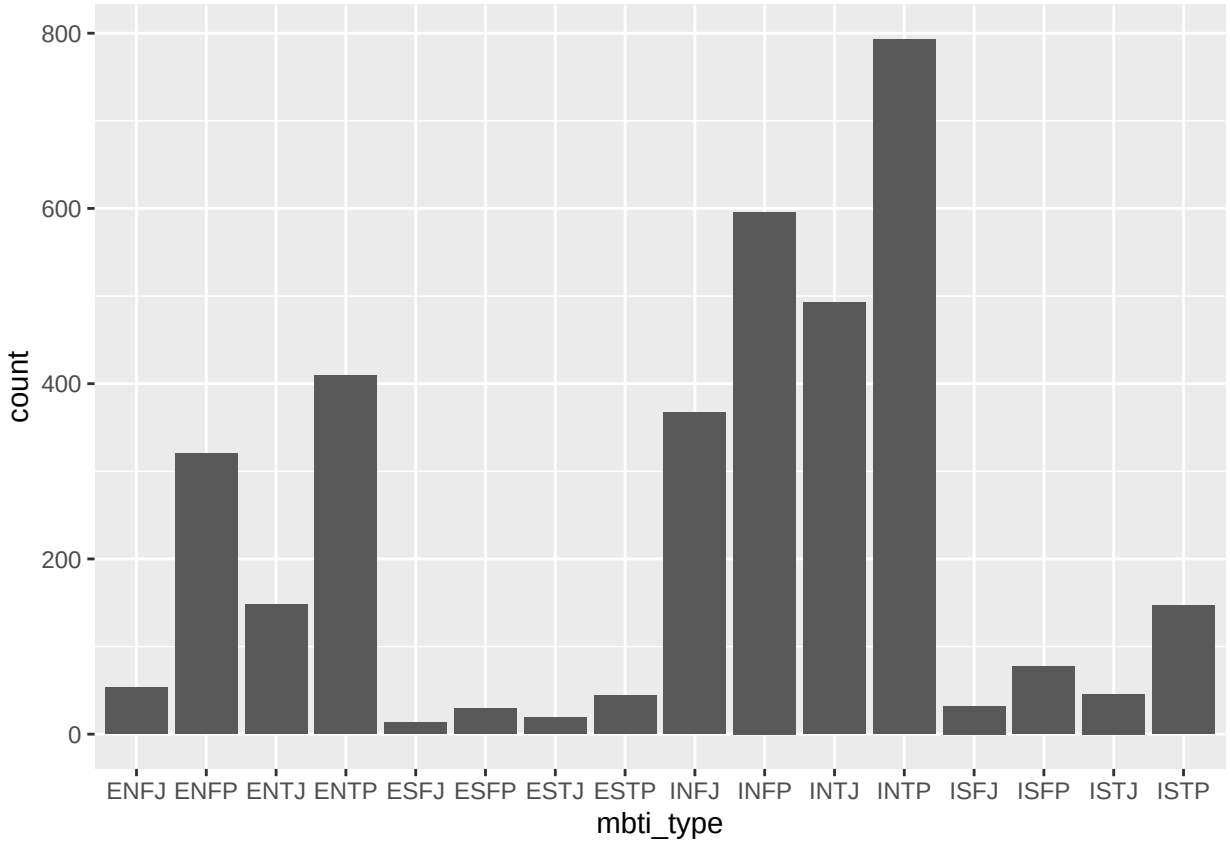
Visualizing

```
##                   columns totals
## 1 post_AskRedditAfterDark      8
## 2   post_unpopularopinion    613
## 3          post_introvert     85
## 4     post_UnpopularFacts      2
## 5       post_changemyview     54
## 6           post_nihilism     15
```

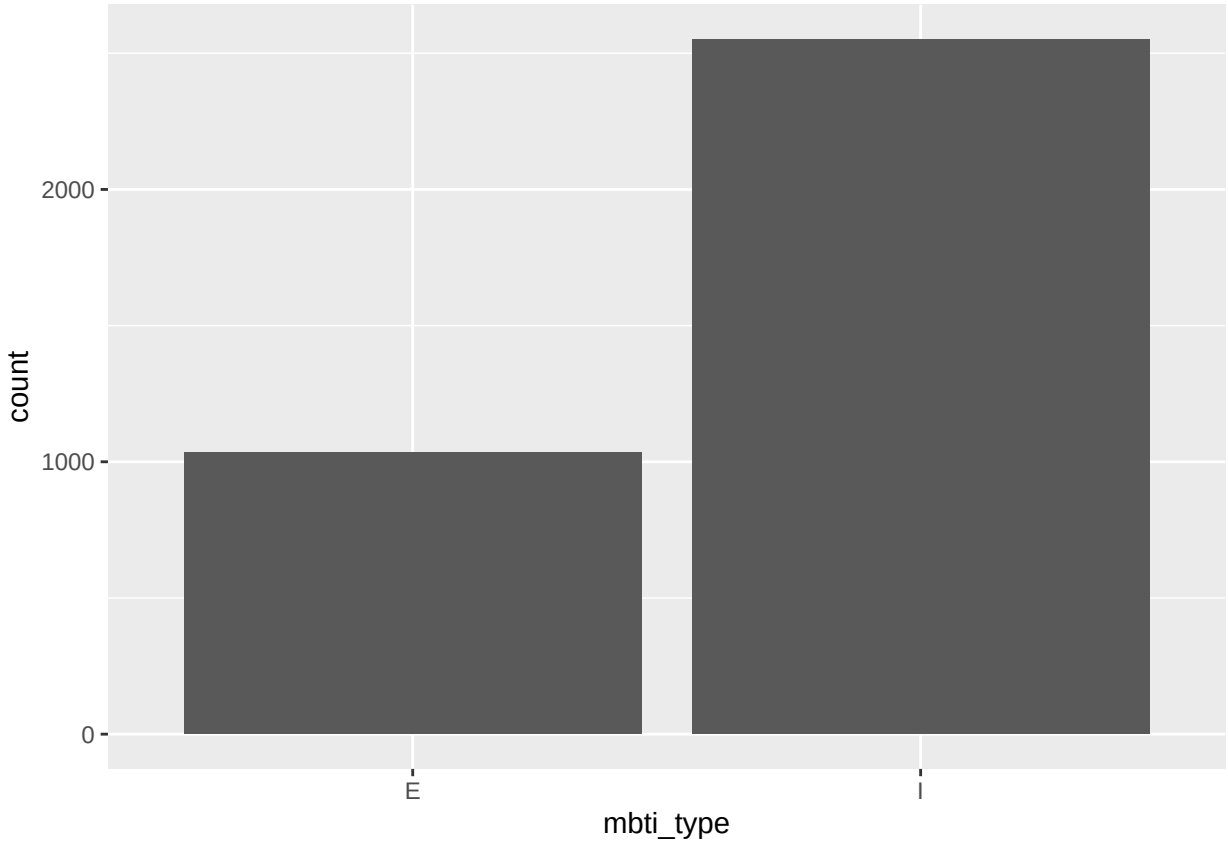
```
col_df %>%
  group_by(totals) %>%
  summarise(total = n()) %>%
  arrange(totals)
```

```
## # A tibble: 356 x 2
##    totals total
##    <chr>  <int>
##  1 1       9987
##  2 10       408
##  3 100        7
##  4 101        4
##  5 102        4
##  6 103       12
##  7 10329      1
##  8 104        5
##  9 105        6
## 10 106        6
## # ... with 346 more rows
```

```
# create data frames with totals
type_totals <- data.frame(mbti = mb,total = s_row)

type_totals %>%
  ggplot(aes(mbti_type)) +
  geom_bar()
```




```
# strip the mbti type to E for Extroversion and I for Introversion
mb_extro <- mb %>% mutate(mbti_type = str_extract(mbti_type, "[A-Z]")) %>% select(mbti_type)

ie_totals <- data.frame(mbti = mb_extro,total = s_row)

ie_totals %>%
  ggplot(aes(mbti_type)) +
  geom_bar()
```



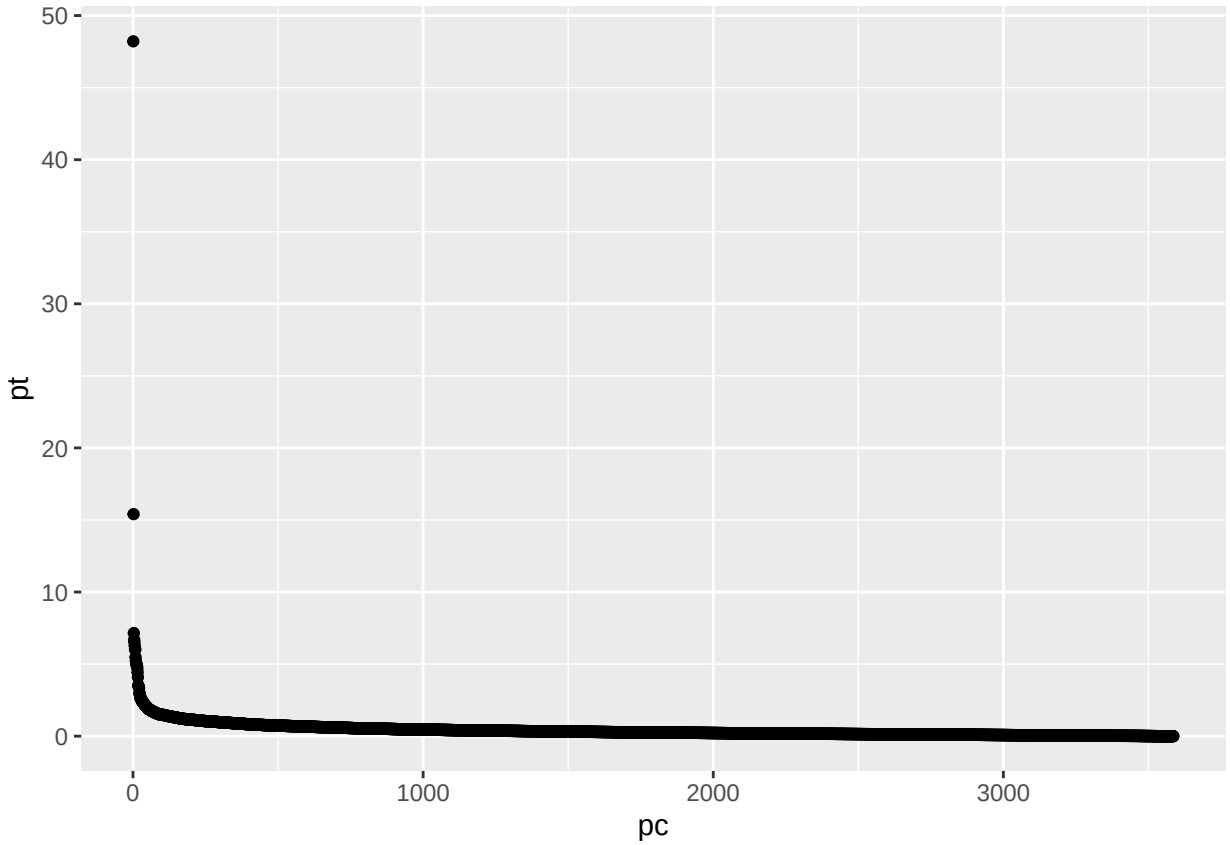
looking at the total extroverts and introverts. Roughly speaking introverts interact twice as much as extroverts. This can be rationalized by the depending on the definition of each type, extroverts tend to want to be with other people and less comfortable sitting alone at a computer alone. On the other hand introverts are much more likely to enjoy time alone and spend time with their thoughts. Additionally introverts tend to recharge alone and be more introspective leading them to be more curious about their MBTI.

With this line of thinking, it was decided to add the totall number of interactions per user to the original data set before performing a principle componet analaysis.

```
# adding a column with the total interations per user.
r_dat_totals <- cbind(r_dat, totals = s_row)

# perfomring PCA
pca_red_t <- prcomp(r_dat_totals)
```

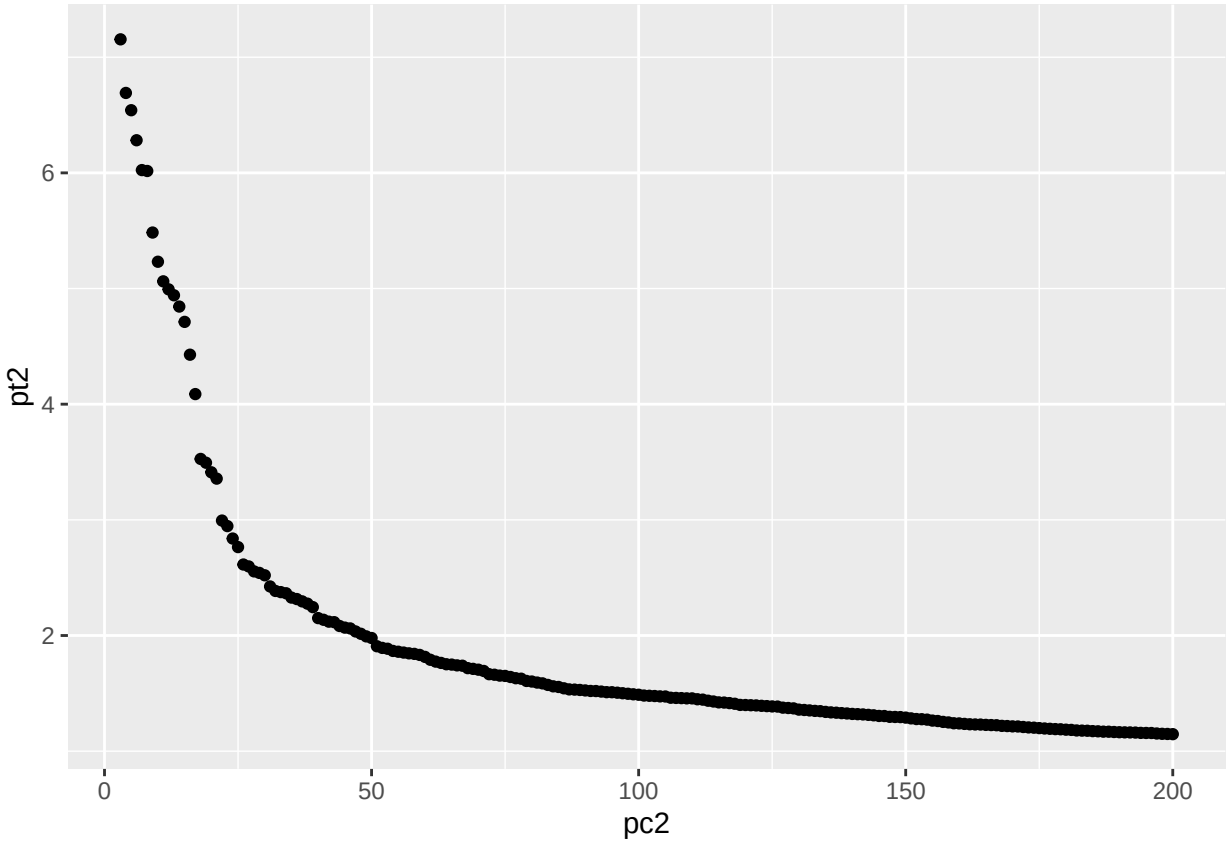
All the PC's were plotted to see how much of the variance each component accounted for

```
# See the breakdown of sd's
pc <- 1:length(pca_red_t$sdev)
pt <- pca_red_t$sdev
qplot(pc, pt)
```



This first graph shows the first few account for the majority amount of data and the PCs seem to account for less and less towards the end, leveling off. These outliers were eliminated in the second iteration of the graph to get a better understanding of what would be a cut off for the number of PCs and variance would be used in the training of models.

```
# zooming in to where it starts to flatten
pc2 <- 3:200
pt2 <- pca_red_t$sdev[3:200]
qplot(pc2, pt2)
```




```
# Looked for percise cut off

# summary(pca_red_t)$importance[,1:100]
# summary(pca_red_t)$importance[,100:400]
# 70% of data = 10, 75% =  21, 80% = 62, 85% = 157, 90% = 338
```

After taking a closer look at the actual cumulative Proportion. It is decided to account for 80% of the variance. This look also reveals an insignificant change in standard deviation and porportion of variance. This means 62 PCs will be used in the machine learning phase of this project.

```
fin_redd <- cbind(mb_extro, pca_red_t$x[,1:62])
# if using R 3.6 or later
set.seed(2020, sample.kind = "Rounding")
test_index <- createDataPartition(y = fin_redd$mbti_type, times = 1, p = 0.2, list = FALSE)
test <- fin_redd[test_index,]
train <- fin_redd[-test_index,]
```

After deciding to split the train and test sets 80/20, a simultanious train of 8 algorithms is used to decide which is the best one or if a combination of them will be better.

```
i <- seq(1, 8)

models <- c("glm", "lda", "naive_bayes", "svmLinear", "knn", "multinom", "qda", "rf")

fits <- lapply(models, function(model){
```

```
  train(mbti_type ~ ., method = model, data = train)
})

names(fits) <- models
```

After each has been trained, a closer look at the individual accuracies is noted. The highest one is generalized linear model and there are a couple above the 80 percent mark to attempt to beat glm in an ensemble.

```
# See the all accuricies of each individual model.
acc_train <- sapply(fits, function(fit){
  fit$results$Accuracy
})
acc_train
```

```
## $glm
## [1] 0.8472028
##
## $lda
## [1] 0.8010155
##
## $naive_bayes
## [1] 0.5052215 0.7132496
##
## $svmLinear
## [1] 0.8421257
##
## $knn
## [1] 0.7299729 0.7378095 0.7444364
##
## $multinom
## [1] 0.8444319 0.8445843 0.8444343
##
## $qda
## [1] 0.5939677
##
## $rf
## [1] 0.7946384 0.8389723 0.8348998
```

Finding the minimum accuracies for each model, checking which one is the highest and what the average is among them.

```
# picking the lowest accuracies from the models
acc_hat <- sapply(fits, function(fit){
  min(fit$results$Accuracy)
})
which.max(acc_hat)
```

```
## glm
##   1
```

```r
mean(acc_hat)
```

```
## [1] 0.744822
```

Results

```r
# Final Testing -----

# all models
pred_ensemble <- sapply(fits, function(object){
  predict(object, newdata = test)})


# First using the best performing single model.
glm_pred <- predict(fits$glm, test)
glm_acc <- mean(glm_pred == test$mbti_type)


# Using the models that performed better than 80%
ind <- acc_hat >= 0.8
ind
```

```
##         glm         lda naive_bayes   svmLinear         knn    multinom
##        TRUE        TRUE       FALSE        TRUE       FALSE        TRUE
##         qda          rf
##       FALSE       FALSE
```

```r
votes <- rowMeans(pred_ensemble[,ind] == "E")
y_hat <- ifelse(votes > 0.5, "E", "I")
ensem_acc <- mean(y_hat == test$mbti_type)

ensem_acc < glm_acc
```

```
## [1] TRUE
```

From the training of each of the algorithms, it is determined that the glm is the best performing. This is then put up against the ensemble created from using algorithms with better than an 80% prediction rate. Comparing the two determines the ensemble algorithm created from glm, lda, svmLinear, and multinom does not beat the glm model. It wins out by less than one percentage point.

Conclusion

It is possible to predict introversion or extroversion based on reddit users postings with an accuracy over 80%.This is also only about 10% better than assuming each person is an introvert.

The limitations of this analysis are time and the available resources to process data on a personal computer. The big five personality theory, works based off of a spectrum. This then creates a gray area where people are really toward the middle of the introversion/extroversion scale. The middle of this scale is ambiversion.

Knowing what is known about statistics, it is likely that the majority of the population lands somewhat close to this middle ground. When it comes to the data itself, there are always issues with self reporting in how these people came to the conclusion they were an INFJ or an ISTP. Additionally there is no scientific evidence that MBTI even exists in humans. Reddit is also a huge length of subreddits creating a very narrow scope of who even has a possibility of being able to make such a prediciton. As seen with this data, there were some subreddits with only one comment or post on them.

Finally this data set does not seem to have much actionable information. Predicting introvert or extrovert is helpful and really needs to be done as an additional factor instead of the only predictor.