# BDS516 Homework #9

Sophie Bass, Julianna Catania, Uri Federman, Farzana Khan, Madeline Mauboussin, and Quinn Rhodes

4/21/2021

#Introduction ##Although they achieved fame in different ways, Kanye West and Barack Obama are two of the most influential people of our time. Kanye West has amassed over 30 million followers, mostly from his incredibly successful entertainment career and relationship with mega-socialite family, the Kardashians, but has also received recent attention for his most recent foray into the world of politics as a 2020 candidate running on a conservative platform. On the other hand, Barack Obama established his prominence as an American politician and lawyer. He served as the 44th president of the United States, serving from 2008-2016, and was the first African-American president of the U.S. He is a member of the Democratic party and has a relatively progressive political platform. Since his presidency, Obama continues to remain an influential figure, both in politics and pop culture. One factor that makes this comparison interesting is that both Kanye and Obama are politicians (to differing degrees). However, they have very different political ideologies, are in different phases of their careers, differ in personality, and differ in demographic characteristics (such as age). These characteristics influence the characteristics of their followers which may, in turn, influence the content of their tweets and how their followers engage with their tweets. Therefore, this may affect factors such as the timing, source, content, sentiment, and engagement level of their tweets. For this reason, our analysis aims to address the following question: can an algorithm based on tweet source, use of quotes, use of photos, and sentiment scores successfully predict whether a tweet was made by Barack Obama or Kanye West?

#Methods ##Methods Twitter data from both Kanye West and Barack Obama was collected through a Twitter Developer account. After receiving a twitter API, tweets from both @kanyewest and @BarackObama were extracted from twitter in R Studio using the rtweet package. Other packages used to run models, manage and clean data, create visuals, text mine, run text analyses, and calculate statistics include: httpuv, tidytext, tidyverse, dplyr, lubridate, ggplot, scales, readr, syuzhet, mlbench, caret, vtreat, and InformationValue. Of all tweets posted, 3,200 tweets from both @BarackObama and @kanyewest were sampled randomly on April 21st, 2021. Dataframes containing the two sets of tweets were cleaned to only contain the source (which device the tweet was made on), status_id (ID of the tweet), text (text content of the tweet), created_at (time when tweet was created), retweet_count (number of retweets), favorite_count (number of favorites), is_retweet (whether or not the tweet is a retweet), and screen_name (screen name of the individual who tweeted it).Prior to modeling, a number of summary statistics were calculated. To analyze differences in source of tweets, the count function was used to examine from which device both Kanye and Obama tweeted from most often. For time of day, the percent of total tweets sampled tweeted at each hour of the day was calculated for both Kanye and Obama and graphed as a line graph. Time tweeted was standardized using Eastern Standard Time (EST). Furthermore, the percentage of total tweets sampled starting with a quotation mark was calculated. Tendencies to include pictures or links in tweets was captured by calculating the total percentage of tweets containing either a picture, a link, or both among the tweets sampled. Furthermore, retweets and favorites were analyzed by calculating the percentage of tweets that are retweets out of the total sample, the average number of retweets on each of their tweets among the tweets sampled, and the average number of favorites on each of their tweets among the tweets sampled. Lastly, average sentiment for tweets were calculated for both Kanye and Obama. This was performed by removing all non-alphabetic characters using the str_replace_all command. Sentiment scores were calculated using the NRC Sentiment Dictionary. By running the command get_nrc_sentiment, sentiment scores were calculated for tweet texts. The average for both Kanye and Obama was then calculated for each of the following sentiments and valence: anticipation, fear, disgust, joy, sadness, surprise, trust, negative valence, and positive valence. Notably, percentages were used for comparisons because Kanye had fewer than 3,200 tweets while Obama had enough tweets to reach

3,200 sample tweets. Therefore, this was done to account for the difference in total number of tweets in the dataset for each person. Multiple logistic regression was used to develop the predictive algorithm, using the glm function in R. The following predictors were included: whether the tweet uses a quotation mark, whether the tweet features a picture or link, and sentiment scores for the sentiments anticipation, fear, joy, trust, and positive valence. These factors were chosen due to a notable difference in behaviors observed between Kanye and Obama based on the summary statistics. The outcome variable was captured as being tweeted by Obama or not tweeted by Obama (i.e. tweeted by Kanye). The algorithms predictive power was then tested on a new set of tweets made by both Kanye and Obama. A train-test split was created with 70% of tweets from both Kanye and Obama included in the train set and 30% from each included in the test set. Model diagnostics were calculated, including the optimal prediction probability cutoff, the misclassification error, the sensitivity level, and the specificity level. An AUC-ROC curve was created to demonstrate the ability to distinguish between the positive (i.e. Obama's tweets) and negative class (i.e. Kanye's tweet). Lastly, the model was used on a set of tweets from an unrelated user. A random sample of 3,200 of Drake's tweets (@Drake) run through the model. Examples of tweets by Drake that were classified as Obama and Kanye were then analyzed by examining how the characteristics of the tweets related to each component included in the model.

```
library(rtweet)
library(httpuv)
library(tidytext)
library(tidyverse)
library(dplyr)
library(lubridate)
library(ggplot2)
library(scales)
library(readr)
library(syuzhet)
library(mlbench)
library(caret)
library(vtreat)
library(InformationValue)
```

### Packages

```
## Loading in csv files
obama <- read_csv("obama_tweets.csv")
kanye <- read_csv("kanye_tweets.csv")
```

### Set up

### Feature Extraction

```
obama %>% count(source) %>% arrange(-n)
kanye %>% count(source) %>% arrange(-n)
```

#### *(1.) Source*

```
# A tibble: 5 x 2
  source                   n
  <chr>                <int>
```

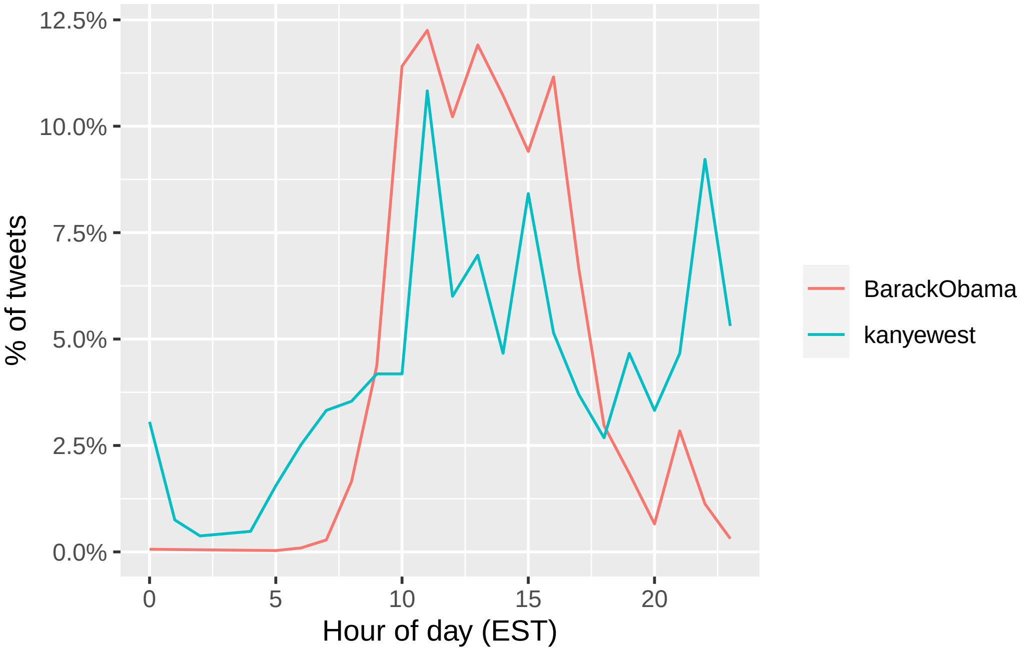
```
1 Twitter Web Client    2444
2 Twitter for iPhone     473
3 Twitter Web App        200
4 Twitter Media Studio    77
5 Thunderclap              5
# A tibble: 2 x 2
  source                 n
  <chr>              <int>
1 Twitter for iPhone  1827
2 Twitter Web App       38
```

#Primary Tweet Source ##When analyzing the source of Kanye and Obama's tweets, we see that Kanye primarily tweets from an iPhone (n=1830 or ~98%) while Obama primarily tweets from a desktop/laptop computer (n=2446 or ~76.5%).

```
merged_df <- rbind(obama, kanye)

merged_df %>% group_by(screen_name) %>%
  count(hour = hour(with_tz(created_at, "EST"))) %>%
  mutate(percent = n/sum(n)) %>%
  ggplot(aes(x = hour, y = percent, color = screen_name)) +
  labs(x = "Hour of day (EST)", y = "% of tweets", color = "") +
  scale_y_continuous(labels = percent_format()) +
  geom_line()
```

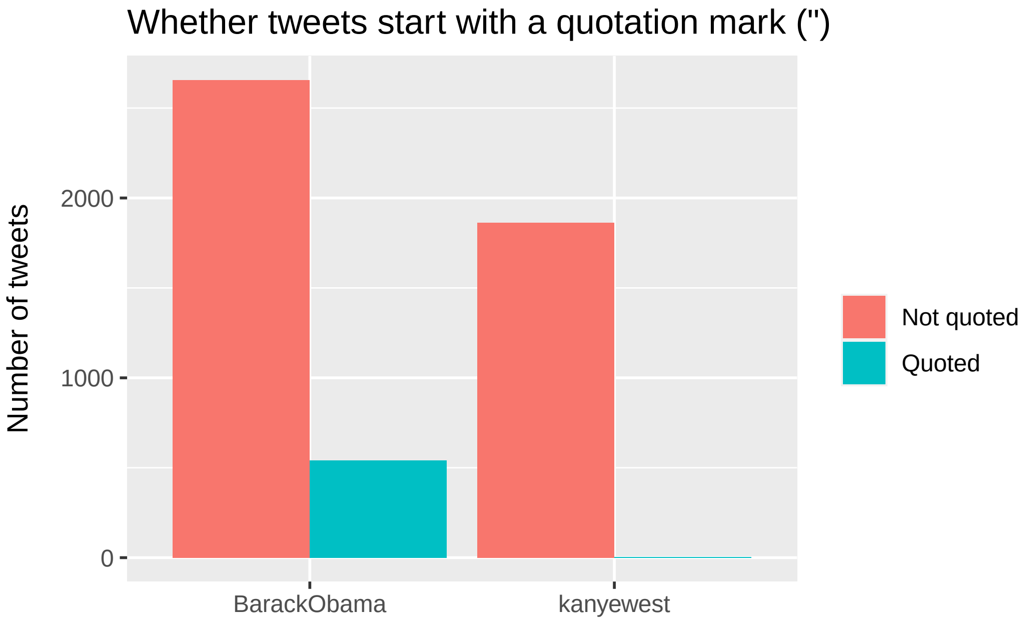
***(2.) Time of Day***



#Time of Day ##The majority of Obama's tweets were made between 10am and 4pm with very few made before 7 am or after 9 pm. This is potentially due to staffers making many of his tweets on his behalf, especially for tweets made while he was in office. On the contrary, Kanye's time of tweets is much more variable, with many tweets being made in the late morning/early afternoon (~11 am to 3 pm) but a considerable number being made late at night (~9 pm to 1 am). This may indicate fewer individuals making tweets on Kanye's behalf/a more personal use of twitter since these tweets are occurring outside the typical workday.

```
## Plot of tweets with quotes vs. no quotes
merged_df %>% group_by(screen_name) %>%
  count(quoted = ifelse(str_detect(text, '^"'), "Quoted", "Not quoted")) %>%
  ggplot(aes(x = screen_name, y = n, fill = quoted)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "", y = "Number of tweets", fill = "") +
  theme(axis.title.y = element_text(margin = margin(t = 0, r = 10, b = 0, l = 0))) +
  ggtitle('Whether tweets start with a quotation mark (")')
```

*(3.) Quotes*



```
# Table of tweets with quotes vs. no quotes
merged_df %>% group_by(screen_name) %>%
  count(quoted = ifelse(str_detect(text, '^"'), "Quoted", "Not quoted")) %>%
  mutate(percent_quote = n/sum(n)*100)
```

```
# A tibble: 4 x 4
# Groups:   screen_name [2]
  screen_name quoted         n percent_quote
  <chr>       <chr>      <int>         <dbl>
1 BarackObama Not quoted  2657        83.1
2 BarackObama Quoted       542        16.9
3 kanyewest   Not quoted  1862        99.8
4 kanyewest   Quoted         3         0.161
```

#Percentage of Tweets Starting with Quotation Marks ##While neither Kanye nor Obama begin tweets with quotation marks particularly often, Obama does appear to use them significantly more since Kanye uses them so infrequently. While 17% of Obama's tweets use quotes at the beginning, fewer than 1% of Kanye's tweets use quotation marks.

```
merged_df %>%
  group_by(screen_name) %>%
  filter(!str_detect(text, '^"')) %>%
  count(picture = ifelse(str_detect(text, "t.co"),
                       "Picture/link", "No picture/link")) %>%
  mutate(percent_picture = n/sum(n)*100)
```

### *(4.) Pictures*

```
# A tibble: 4 x 4
# Groups:   screen_name [2]
  screen_name picture             n percent_picture
  <chr>       <chr>           <int>           <dbl>
1 BarackObama No picture/link   209            7.87
2 BarackObama Picture/link     2448           92.1
3 kanyewest   No picture/link   841           45.2
4 kanyewest   Picture/link     1021           54.8
```

#Percentage of Tweets Containing Pictures or Links ##Whether or not a tweet contains pictures may also be useful information for predicting whether a tweet was made by Obama or Kanye. Out of this sample, 92% of Obama's tweets contain a picture or link whereas only 55% of Kanye's tweets do.

```
merged_df %>% group_by(screen_name) %>%
  count(is_retweet) %>%
  mutate(perc_retweet = n/sum(n)*100)
```

### *(5.) Re-tweets*

```
# A tibble: 4 x 4
# Groups:   screen_name [2]
  screen_name is_retweet     n perc_retweet
  <chr>       <lgl>      <int>        <dbl>
1 BarackObama FALSE       2847         89.0
2 BarackObama TRUE         352         11.0
3 kanyewest   FALSE       1670         89.5
4 kanyewest   TRUE         195         10.5
```

#Percentage of Tweets that are Re-tweets ##Obama and Kanye have nearly the same percentage of tweets that are retweets. 11% of Obama's tweets are retweets and 10.6% of Kanye's tweets are retweets.

```
merged_df %>% group_by(screen_name) %>%
  summarize(avg_retweet = mean(retweet_count),
            avg_fav = mean(favorite_count))
```

### *(6.) Re-tweet Counts & Favorite Counts*

```
# A tibble: 2 x 3
  screen_name avg_retweet avg_fav
* <chr>             <dbl>   <dbl>
1 BarackObama      11364.  58088.
2 kanyewest         9319.  48951.
```

#Retweet/Favorite Counts on Original Tweets ##Overall, Obama gets 11,400 retweets on average and 58,068 favorites. Meanwhile, Kanye gets 9,314 retweets on average and 48,920 favorites. However, it is important to note that Obama has 130 million followers and Kanye only has 30 million followers. This may indicate a higher engagement rate with Kanye's tweets compared to Obama's. Considering we are looking at average retweets/favorites regardless of follower count, the similarities in retweet and favorite counts for Kanye and Obama may indicate that these characteristics may not be particularly predictive.

```
merged_sentiment <- merged_df %>%
  mutate(text2 = str_replace_all(text, "[^[:alpha:]]", " "), # removes all non-alphabetic characters
         get_nrc_sentiment(text2)) # getting nrc scores for tweet texts
```

```
merged_sentiment %>%
  group_by(screen_name) %>%
  summarize(anger = mean(anger),
            anticipation = mean(anticipation),
            fear = mean(fear),
            disgust = mean(disgust),
            joy = mean(joy),
            sadness = mean(sadness),
            surprise = mean(surprise),
            trust = mean(trust),
            negative = mean(negative),
            positive = mean(positive))
```

***(7.) Sentiment***

```
  screen_name anger anticipation  fear disgust   joy sadness surprise trust
1 BarackObama 0.316        0.738 0.453   0.100 0.566   0.233    0.284 1.247
2   kanyewest 0.149        0.305 0.213   0.071 0.376   0.153    0.120 0.414
  negative positive
1    0.517    1.779
2    0.269    0.743
```

#Sentiment ##On average, it seems that Obama expressed more emotion through his tweets than Kanye. Obama's tweets had a higher sentiment score across all sentiments included in this analysis (i.e. anticipation, fear, disgust, joy, sadness, surprise, trust, negative valence, positive valence) compared to Kanye. Obama's tweets scored particularly higher for "anticipation", "trust", and "positive valence".

# Part A

***Develop an algorithm that allows to predict who of the politicians tweeted using just the information in the text of the tweet and the time of the tweets. You are not allowed to use the information about the user. You can use sentiments, individual words, punctuation and anything else as a source of features.***

```
## Setting up data frame for logistic regression
obama_kanye <- merged_sentiment

# Changing names of sources (before filtering)
obama_kanye$source[obama_kanye$source=="Twitter Web Client"] <- "web"
obama_kanye$source[obama_kanye$source=="Twitter for iPhone"] <- "iphone"

obama_kanye2 <- obama_kanye %>%
  select(screen_name, source, created_at, text, status_id,
         anger, anticipation, fear, disgust, joy, sadness, surprise, trust, negative, positive) %>%
  # filtering twitter sources for only web/iPhone
  filter(source %in% c("web", "iphone")) %>%
  # creating variables for time of day, whether the tweet uses a quote, and whether
  # there is a picture or link in the tweet
  mutate(hour = hour(with_tz(created_at, "EST")),
         quoted = ifelse(str_detect(text, '^"'), "quote", "NO_quote"),
         picture = ifelse(str_detect(text, "t.co"), "picture_link", "NO_picture_link"),
         is_obama = case_when(screen_name == "BarackObama" ~ 1,
                              screen_name == "kanyewest" ~ 0))

# Selecting variables for regression
obama_kanye3 <- obama_kanye2 %>%
  select(is_obama, screen_name, source, hour, quoted, picture,
         anger, anticipation, fear, disgust, joy, sadness, surprise, trust, negative, positive)
```

**Based off the feature extraction above, we believe that the features which most contribute to the prediction of whether a tweet was authored by Obama vs. Kanye are: source, quotes, pictures, and sentiment scores. We now will develop a classification algorithm using logistic regression model to predict the probability of a tweet being authored by Obama. As such, the *outcome variable* will be a tweet by Obama (yes or no) and the *predictor variables* will be some combination of the features mentioned above. To that end, we will run several logistic regression models, but only include the model with the greatest predictive power.**

```
model <- glm(is_obama ~  factor(quoted) + factor(picture) +
                anticipation + fear + joy + trust + positive,
             family = "binomial",
             data = obama_kanye3)

summary(model)
```

### Logistic Regression Model

```
Call:
glm(formula = is_obama ~ factor(quoted) + factor(picture) + anticipation + 
    fear + joy + trust + positive, family = "binomial", data = obama_kanye3)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-4.2690  -0.5607   0.1226   0.6564   2.7439  

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)    
(Intercept)                 -3.74117    0.14644 -25.547  < 2e-16 ***
factor(quoted)quote          6.68990    0.72081   9.281  < 2e-16 ***
factor(picture)picture_link  3.40999    0.13460  25.334  < 2e-16 ***
anticipation                 0.55093    0.08229   6.695 2.15e-11 ***
fear                         0.58056    0.07469   7.773 7.68e-15 ***
joy                         -1.14234    0.09478 -12.052  < 2e-16 ***
trust                        0.73085    0.07298  10.014  < 2e-16 ***
positive                     0.81770    0.06215  13.157  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 6323.9  on 4743  degrees of freedom
Residual deviance: 3745.6  on 4736  degrees of freedom
AIC: 3761.6

Number of Fisher Scoring iterations: 8
```

```
exp(model$coefficients)
```

```
                (Intercept)         factor(quoted)quote 
                 0.02372623                804.24385232 
factor(picture)picture_link                anticipation 
                30.26492569                  1.73487349 
                       fear                         joy 
                 1.78703885                  0.31907270 
                      trust                    positive 
                 2.07683779                  2.26527822 
```

***Interpretation of Model***

- “**quoted**”
  - All else equal, tweets with quotes have a ~80,000% greater odds of being Obama’s tweets.
  - *Calculation:* odds = (797.47143283 - 1)*100 = 79647.14
- “**picture_link**”
  - All else equal, tweets with pictures or links have a ~3,000% greater odds of being Obama’s tweets.
  - *Calculation:* odds = (30.04893338 - 1)*100 = 2904.893
- “**anticipation**”
  - All else equal, a one-unit increase in the sentiment score for anticipation increases the odds of the tweet being authored by Obama by 74%.
  - *Calculation:* odds = (1.74400434 - 1)*100 = 74.40043
- “**fear**”
  - All else equal, a one-unit increase in the sentiment score for fear increases the odds of the tweet being authored by Obama by 78%.
  - *Calculation:* odds = 1.78329704 - 1)*100 = 78.3297
- “**joy**”
  - All else equal, a one-unit increase in the sentiment score for joy decreases the odds of the tweet being authored by Obama by 68%.
  - *Calculation:* odds = (0.31620393 - 1)*100 = -68.37961
- “**trust**”
  - All else equal, a one-unit increase in the sentiment score for trust increases the odds of the tweet being authored by Obama by 108%.
  - *Calculation:* odds = (2.07902310 - 1)*100 = 107.9023
- “**positive**”
  - All else equal, a one-unit increase in the sentiment score for positive increases the odds of the tweet being authored by Obama by 127%.
  - *Calculation:* odds = (2.27180811 - 1)*100 = 127.1808

# Part B

***Apply the algorithm to new tweets from both users to estimate how well the predictions work.***

**Given that our logistic regression model was developed using *all* of the tweets ever posted by Kanye West (n = 1,868), rather than applying the algorithm to a new tweets, we will evaluate the algorithm using a train-test split.**

### Train-Test Split Evaluation

```
# Checking for class bias
table(obama_kanye3$is_obama)
```

```
##
##    0    1
## 1827 2917
```

```
## Creating train and test data

# Ensuring Train Data draws equal proportions of Obama (1) and Kanye (0))
set.seed(04917)
input_ones <- obama_kanye3[which(obama_kanye3$is_obama == 1), ]  # all 1's
input_zeros <- obama_kanye3[which(obama_kanye3$is_obama == 0), ]  # all 0's

# 1's for training
input_ones_training_rows <- sample(1:nrow(input_ones), 0.7*nrow(input_ones))
training_ones <- input_ones[input_ones_training_rows, ]

# 0's for training. Pick as many 0's as 1's
input_zeros_training_rows <- sample(1:nrow(input_zeros), 0.7*nrow(input_zeros))
training_zeros <- input_zeros[input_zeros_training_rows, ]

#Row bind the 1's and 0's
train.data <- rbind(training_ones, training_zeros)

# Creating Test Data
test_ones <- input_ones[-input_ones_training_rows, ]
test_zeros <- input_zeros[-input_zeros_training_rows, ]

# Row bind the 1's and 0's
test.data <- rbind(test_ones, test_zeros)

## Building Logistical Model and Predicting on Test Data
model_train <- glm(is_obama ~  factor(quoted) + factor(picture) +
                anticipation + fear + joy + trust + positive,
                data=train.data,
                family=binomial(link="logit"))

predicted <- predict(model_train, test.data, type="response")
```

### Model Diagnostics

```
# Optimal prediction probability cutoff
optCutOff <- optimalCutoff(test.data$is_obama, predicted)
optCutOff #  = 0.52
```

```

misClassError(test.data$is_obama, predicted, threshold = optCutOff)
```

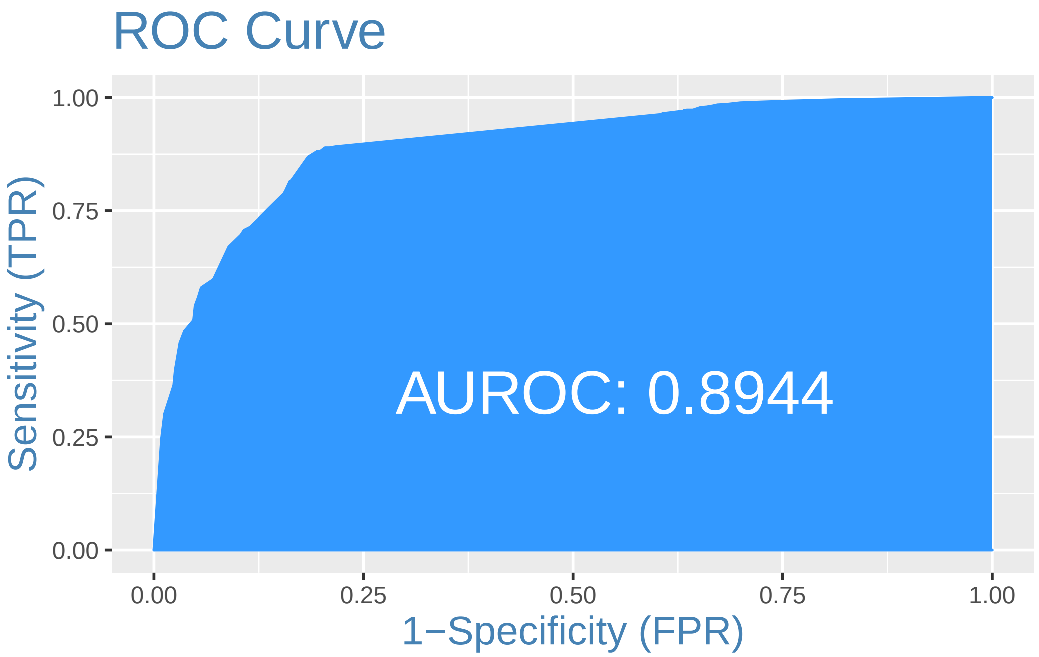
***Misclassification Error***

```
[1] 0.1467
```

The model's misclassification error (i.e. the percentage of incorrectly classified instances) is 14%.

```
plotROC(test.data$is_obama, predicted)
```

***AUC-ROC Curve***



Our model has an AUC of .9, meaning there is a ~90% chance that the model will be able distinguish between positive class (i.e. Obama's tweets) and negative class (i.e. Kanye's tweets)

```
sensitivity(test.data$is_obama, predicted, threshold = optCutOff)
specificity(test.data$is_obama, predicted, threshold = optCutOff)
```

***Sensitivity and Specificity***

```
[1] 0.8892694
[1] 0.7959927
```

The model's true positive rate (i.e. sensitivity) and true negative rate (i.e specificity) are both about 85%.

**All in all, our model has fairly strong predictive ability**

#Predictive Algorithm ##All predictors in our logistic regression model were significant at the p<.001 level when holding the other predictors constant. Whether or not the tweet included a quotation mark had an odds ratio of 797. This means that tweets containing a quotation mark have a 797x greater odds of being Obama's tweets than Kanye's tweets when all other predictors in the model are held constant. Furthermore, tweets with a picture or link included have a 30x greater odds of being Obama's tweet than being Kanye's tweet when all other predictors are held constant. Sentiments do not demonstrate such a strong relationship with the author of the tweet, but do demonstrate a meaningful relationship. The odds ratios for anticipation, fear, trust, joy, and positive valence are 1.74, 1.78, 2.08, 0.316, and 2.27, respectively. This means that tweets expressing more anticipation, fear, trust, and positive valence have a higher odds of being tweeted by Obama. Nonetheless, tweets with higher joy scores have a lower odds of being authored by Obama. Unfortunately, Kanye's total number of tweets did not sum up to 3,200. Therefore, there were no new tweets to run the algorithm on. Therefore, the model was evaluated on a 70-30 train-test split. Using the test data, the optimal prediction probability cutoff was found to be 0.52. Furthermore, the misclassification error was 0.15, indicating that the percentage of incorrectly classified instances is 15% and the AUROC was 0.8916, indicating that there is a 90% chance that the model will be able to distinguish between the positive class (i.e. Obama's tweets) and the negative class (i.e. Kanye's tweets). Furthermore, the sensitivity and specificity were 0.88 and 0.79, meaning that the model correctly identified Obama's tweets as being tweeted by Obama 88% of the time and the model correctly identified tweets by Kanye as being tweeted by Kanye 79% of the time. Overall, these diagnostics indicate that our model has strong predictive power in distinguishing between tweets by Kanye and tweets by Obama.

# Part C

***Try the prediction algorithm with a different set of tweets from unrelated users. Discuss how the algorithm works / breaks in this case.***

**In the following section, we will apply our prediction algorithm (i.e. the trained logistic model created above) to a set of tweets posted by the rapper, Drake.**

```r
## Creating a data set for Drake Tweets

# Extracting data from Twitter
setwd("~/Documents/Penn/Spring 2021/Data Analysis")
drake_raw <- get_timeline("@Drake", n = 3200)


# Cleaning the data sets
drake_clean <- drake_raw %>% select("source", "status_id", "text","created_at",
                                "retweet_count", "favorite_count", "is_retweet",
                                "screen_name")

drake <- as.data.frame(drake_clean)
```

```r
## Getting sentiment scores
drake_sentiment <- drake %>%
  mutate(text2 = str_replace_all(text, "[^[:alpha:]]", " "), # removes all non-alphabetic characters
         get_nrc_sentiment(text2)) # getting nrc scores for tweet texts
```

```r
## Preparing data set for testing
drake_test <- drake_sentiment

drake_test2 <- drake_test %>%
  # creating variables for whether the tweet uses a quote & whether there is a picture/link
  mutate(quoted = ifelse(str_detect(text, '^"'), "quote", "NO_quote"),
         picture = ifelse(str_detect(text, "t.co"), "picture_link", "NO_picture_link")) %>%
  select(screen_name, text2, quoted, picture,
         anticipation, fear, joy, trust, positive)
```

```
## Sanity check
head(drake_test2)
```

### ***Preparing a data set of Drake Tweets***

```
# A tibble: 6 x 9
  screen_name text2       quoted picture anticipation  fear   joy trust positive
  <chr>       <chr>       <chr>  <chr>          <dbl> <dbl> <dbl> <dbl>    <dbl>
1 Drake       It s the b~ NO_qu~ pictur~            2     1     1     0        2
2 Drake       Tune in  h~ NO_qu~ pictur~            0     0     0     0        0
3 Drake       Fry Yiy ht~ NO_qu~ pictur~            0     0     0     0        0
4 Drake       What s Nex~ NO_qu~ pictur~            0     0     0     1        0
5 Drake       SCARY HOUR~ NO_qu~ pictur~            0     0     0     0        0
6 Drake       Going live~ NO_qu~ pictur~            1     0     0     0        2
```

```
## Predicting the train logistical regression model on the drake data
predicted_drake <- predict(model_train, drake_test2, type="response")
predicted.classes <- ifelse(predicted_drake > 0.5, "Obama", "Kanye")
table(predicted.classes)
```

### ***Applying Prediction Algorithm on Drake Tweets***

```
predicted.classes
Kanye Obama 
 1584   164 
```

When the original prediction algorithm was used on a data set of tweets posted by Drake, the algorithm classified 91% of the tweets as being Kanye West's tweets and 9% being Obama's. Given this result, it would be interesting to look at examples of Drake's tweets that were classified as Kanye's vs. Obama's.

```
drake_test3 <- drake_test2
drake_test3$predictions <- predicted.classes
drake_test3 <- drake_test3 %>%
  mutate(n = row_number())
```

```
drake_test3 %>% filter(n == 36) %>%
  pull(text2)
```

### *Example:* **Predicted Classification of Tweet = Obama**

```
[1] "Drake When to Say When  amp  Chicago Freestyle  Video  https   t co ZIAX R UCY"
```

```
drake_test3 %>% filter(n == 121) %>%
  pull(text2)
```

### *Example:* **Predicted Classification of Tweet = Kanye**

```
[1] "Seventh Annual OVOFEST https   t co Y KeKSHt R"
```

#Running the Algorithm on @Drake's Tweets ##When the original prediction algorithm was used on a data set composed of tweets posted by just Drake, the algorithm classified 91% of those tweets as being Kanye West's tweets and 9% being Obama's. When pulling a tweet from Drake that was classified as an Obama tweet, we found that the tweet did not begin with a quote or link but included a picture. It had an anticipation score of 4; fear score of 1; joy score of 2; trust score of 3; and positive score of 5. Given that our algorithm found that pictures and sentiments of anticipation, fear, trust, and positive *all* increase the odds of a tweet belonging to Obama instead of Kanye, it is unsurprising that a tweet by Drake consisting of an image and with such sentiment scores was coded as an Obama tweet. We then pulled a Drake tweet that was classified as a Kanye tweet. This tweet also did not begin with a quote, and included a picture. However the sentiment scores were different. It had an anticipation score of 0; fear score of 0; joy score of 1; trust score of 0; and positive score of 0. While our algorithm found that a one unit increase in the sentiment score for joy *decreases* the odds of the tweet being authored by Obama by 68%, it also found that tweets with pictures and links have a 30 times odd of being an Obama tweet. As such, this classification seems to be rather perplexing given the differing magnitude of these variables.

#Conclusions ##We found that the tweets from Obama's account were more likely to be created from a desktop or laptop computer compared to the tweets from Kanye's account which were from an iPhone. This may, in part, explain why we also found that Obama was more likely to tweet between 11am and 3pm - within standard work hours - compared to the greater variability in the timing of tweets from Kanye. It is also important to note that, as a performer, Kanye may be working late at night (i.e. when concerts tend to occur) which may also explain why his tweet behavior looks differently than someone like Obama who may, along with his staffers, work closer to the conventional work day. Finally, it is important to note that both Kanye and Obama likely travel often, leading to variability in the time zones in which they make tweets. This is a weakness of looking at when tweets are made in EST since the increase in Kanye's tweet time variability may reflect more frequent travel rather than a meaningful difference in which tweets are made. Nonetheless, although this may not directly measure the hour at which each of them tweet, this may still be a useful input in our model to predict which tweets belong to whom.We also found that Obama's tweets conveyed more emotion compared to Kanye's. This may not be particularly surprising as Obama and Kanye may be tweeting to different audiences and have different pressures when crafting their tweets. It is also possible that Obama's tweets are registered as being more emotive as it may be considered inappropriate or risky for him to use more complex language, like sarcasm. As someone in office, he likely had to be careful that his tweets were literal so that they were less likely to be misconstrued by the public which would then make it easier to associate his tweets with specific emotions. Kanye may not face quite as much public scrutiny as an entertainer, so he may use language that is considered to be more neutral by the NRC algorithm or use more sarcastic language. There may also be less pressure on Kanye to relate or empathize with his twitter audience whereas there may be more of an expectation for a president to do so. This raises an interesting question if Kanye's sentiment levels changed following his announcement that he would run for president (although, not every politician conforms to society's expectations of presidential conduct). As noted above, our model has an AUC of .9, indicating that there is a 90% chance that any given tweet will be correctly diagnosed as either Obama's or Kanye's. Further, the model's true positive rate and true negative rate are both about .85, indicating that our model has strong predictive power. When we tried our prediction algorithm on a third user's tweets (Drake's), it overwhelmingly classified his tweets as having originated from Kanye's account. This is not very surprising, given that Kanye and Drake are entertainers of a similar (43 vs. 34) age. Overall, we feel that our predictive model does an impressive job of classifying tweet origin. Barack Obama and Kanye West are both African American men with a significant cultural impact. Yet, it is unsurprising that they tweet in different ways, considering their difference in personal characteristics, backgrounds, and primary audiences. Nonetheless, it is noteworthy that those differences can be ascertained by our model and put to use on test data.