Russell Valenzuela

STAT 576: Analyzing Lonzo Ball's On-Court Performance and Tweets

Introduction

Since being launched in 2006, Twitter has grown into one of the most used social media websites around the world. The main usage for Twitter began as a microblogging with users being limited to share their thoughts in 140 or less characters. As the years went on, companies began to see Twitter's potential and began using it for their own advertisements or sharing their own news. Twitter's popularity saw the company register 330 million active users as of October 2017. Many of those users often share thoughts during live events during real time. Just as the Super Bowl is one of the most watched sporting events on television, it is also one of the most tweeted about events. It would then make sense to see how the events of a live game impact the tweets being published in Twitter.

In this study, the performance of Los Angeles Lakers rookie Lonzo Ball is considered. The Lakers originally drafted Ball with their second overall pick in the 2017 NBA Draft and quickly became one of the most scrutinized rookie in NBA history. Most of the attention came as a byproduct of his father's pre-draft hype and not because of on-court performance. For his actual, basketball play, Ball is highly regarded as a gifted passer with excellent court vision but carries an unorthodox shooting form. Ball also grew up in nearby Chino Hills, CA and spent one year at UCLA making his selection to the Lakers something highly anticipated. The goal of this study is to examine how Ball's play during games influences the types of tweets users say about him during those games.

Data

Three regular season games were tracked. The settings for each game were as follows. The first game was a nationally televised home game against the Golden State Warriors on a Wednesday night. The second game had the Lakers travel for a road game against the Denver Nuggets while the final game saw the team matchup back in Staples Center.

During every two minutes of regulation game time as well as every minute elapsed during overtime periods, 1000 tweets were collected. A sample of the tweets as well as other information obtained is shown below in Figure 1. For this study, only the tweet content is used. Each tweet collected revolved around the search term "Lonzo Ball." Since one game out of the three went into overtime, 77,000 total tweets were gathered.

In addition to the tweets, Ball's statistics were logged. Player performance statistics taken into consideration were field goal percentage, three-point percentage, points, assists, rebounds, blocks, steals, turnovers, and whether or not Ball was on the court during the time in question. The team performance was also highlights with the team point differential also being tracked whenever tweets were collected. A sample of in game statistics is also shown below in Figure 2. Similarly to the number of tweets collected, there was a total of 77 observations or time periods logged over the court of the three games in question.

	text	favorited	favoriteCount	replyToSN	created	truncated	replyToSID	id	replyToUID
1	Warriors bout to put the almighty ass whoopin on the...	FALSE	0	NA	2017-11-30 03:42:37	FALSE	NA	9.360780e+17	NA
2	Lonzo Ball is on TV	FALSE	0	TobiBryant	2017-11-30 03:42:10	FALSE	9.335406e+17	9.360779e+17	2.711442e+08
3	Lavar Ball said Lonzo was better than Curry. Can't wai...	FALSE	1	NA	2017-11-30 03:42:08	FALSE	NA	9.360779e+17	NA
4	I liked a @YouTube video https://t.co/TQLU68uqMt L...	FALSE	0	NA	2017-11-30 03:42:02	FALSE	NA	9.360778e+17	NA
5	Stephen Curry, Kevin Durant have Lonzo Ball's back, s...	FALSE	0	NA	2017-11-30 03:42:00	FALSE	NA	9.360778e+17	NA
6	Lonzo Ball Vs Steph Curry	FALSE	0	NA	2017-11-30 03:41:57	FALSE	NA	9.360778e+17	NA
7	NBA app headline: "warriors take on Lonzo ball" what:...	FALSE	0	NA	2017-11-30 03:41:51	FALSE	NA	9.360778e+17	NA
8	Lonzo \"Bust\" Ball will be dropping -81 tonight. #Let...	FALSE	0	NA	2017-11-30 03:41:50	FALSE	NA	9.360778e+17	NA
9	Twitter gonna explode tonite Lavar Ball said Lonzo Ba...	FALSE	0	NA	2017-11-30 03:41:42	FALSE	NA	9.360778e+17	NA
10	Lonzo Ball look like the monkey that turned on Cesar.	FALSE	0	NA	2017-11-30 03:41:05	FALSE	NA	9.360776e+17	NA
11	Regardless of whether or not he's comparable to Step...	FALSE	0	NA	2017-11-30 03:40:53	FALSE	NA	9.360775e+17	NA
12	Lonzo ball you ain't got no beard duke shave your la...	FALSE	0	NA	2017-11-30 03:40:42	FALSE	NA	9.360775e+17	NA
13	Lonzo Ball is one ugly MoFo ������������...	FALSE	0	NA	2017-11-30 03:40:41	FALSE	NA	9.360775e+17	NA
14	Lonzo and Steph bout to square off. Lavar Ball said Lo...	FALSE	0	NA	2017-11-30 03:40:36	TRUE	NA	9.360775e+17	NA
15	Lonzo ball is getting a triple double tonight. Mark my...	FALSE	1	NA	2017-11-30 03:40:26	FALSE	NA	9.360774e+17	NA
16	Match up of the year lonzo ball vs steph curry ���...	FALSE	0	NA	2017-11-30 03:39:37	FALSE	NA	9.360772e+17	NA
17	Tonight, @NickDePaula & I got the first look at L...	FALSE	9	NA	2017-11-30 03:39:31	TRUE	NA	9.360772e+17	NA
18	I liked a @YouTube video https://t.co/kFYaJ1ObyX LO...	FALSE	0	NA	2017-11-30 03:39:31	FALSE	NA	9.360772e+17	NA
19	Whenever I need a good laugh, I just go back and loo...	FALSE	0	NA	2017-11-30 03:39:21	FALSE	NA	9.360772e+17	NA
20	I liked a @YouTube video https://t.co/4lSWMYHsgh St...	FALSE	0	NA	2017-11-30 03:39:19	FALSE	NA	9.360772e+17	NA

Figure 1: Sample of Tweets Collected

	opp	QtimeLeft	timeLeft	timeElapsed	date	pacTime	onCourt	onCourtCode	pointDef	FGM	FGA	FGPCT	TPM	TPA	TPPCT	PTS	AST	REB	BLK	STL	TO
1	GSW	10	46	2	2017-11-29	2017-11-29 19:43:00	Y	1	-3	0	0	0.0000000	0	0	0.0000000	0	0	0	0	0	0
2	GSW	8	44	4	2017-11-29	2017-11-29 19:45:00	Y	1	-2	0	0	0.0000000	0	0	0.0000000	0	0	0	0	0	0
3	GSW	6	42	6	2017-11-29	2017-11-29 19:50:00	Y	1	-6	0	0	0.0000000	0	0	0.0000000	0	0	0	0	0	0
4	GSW	4	40	8	2017-11-29	2017-11-29 19:55:00	Y	1	-7	0	0	0.0000000	0	0	0.0000000	0	0	0	0	0	0
5	GSW	2	38	10	2017-11-29	2017-11-29 20:00:00	N	0	-7	0	1	0.0000000	0	0	0.0000000	0	1	0	0	1	0
6	GSW	EQ1	36	12	2017-11-29	2017-11-29 20:04:00	N	0	-10	0	2	0.0000000	0	1	0.0000000	0	2	0	0	1	0
7	GSW	10	34	14	2017-11-29	2017-11-29 20:09:00	N	0	-8	0	2	0.0000000	0	1	0.0000000	0	2	0	0	1	0
8	GSW	8	32	16	2017-11-29	2017-11-29 20:16:00	N	0	-7	0	2	0.0000000	0	1	0.0000000	0	2	0	0	1	0
9	GSW	6	30	18	2017-11-29	2017-11-29 20:19:00	N	0	-8	0	2	0.0000000	0	1	0.0000000	0	2	0	0	1	0
10	GSW	4	28	20	2017-11-29	2017-11-29 20:27:00	Y	1	-7	1	3	0.3333333	0	1	0.0000000	2	3	0	0	1	0
11	GSW	2	26	22	2017-11-29	2017-11-29 20:33:00	Y	1	-1	1	3	0.3333333	0	1	0.0000000	2	4	0	0	1	0
12	GSW	EQ2	24	24	2017-11-29	2017-11-29 20:36:00	Y	1	4	1	3	0.3333333	0	1	0.0000000	4	5	1	0	1	1
13	GSW	10	22	26	2017-11-29	2017-11-29 20:54:00	Y	1	1	1	3	0.3333333	0	1	0.0000000	4	6	1	0	1	1
14	GSW	8	20	28	2017-11-29	2017-11-29 20:57:00	Y	1	2	3	6	0.5000000	2	4	0.5000000	10	6	1	0	1	1
15	GSW	6	18	30	2017-11-29	2017-11-29 21:05:00	Y	1	3	4	7	0.5714286	2	4	0.5000000	12	6	1	0	1	1
16	GSW	4	16	32	2017-11-29	2017-11-29 21:09:00	Y	1	4	5	8	0.6250000	3	5	0.6000000	15	7	1	0	1	2
17	GSW	2	14	34	2017-11-29	2017-11-29 21:16:00	Y	1	2	5	9	0.5555556	3	6	0.5000000	15	7	1	0	1	2
18	GSW	EQ3	12	36	2017-11-29	2017-11-29 21:19:00	Y	1	0	5	9	0.5555556	3	6	0.5000000	15	7	1	0	1	2
19	GSW	10	10	38	2017-11-29	2017-11-29 21:26:00	N	0	0	5	9	0.5555556	3	6	0.5000000	15	7	1	0	1	2
20	GSW	8	8	40	2017-11-29	2017-11-29 21:33:00	Y	1	2	5	9	0.5555556	3	6	0.5000000	15	7	1	0	1	2
21	GSW	6	6	42	2017-11-29	2017-11-29 21:36:00	Y	1	0	5	10	0.5000000	3	6	0.5000000	15	8	1	0	1	2
22	GSW	4	4	44	2017-11-29	2017-11-29 21:43:00	Y	1	-2	5	11	0.4545455	3	6	0.5000000	15	8	1	0	1	2
23	GSW	2	2	46	2017-11-29	2017-11-29 21:49:00	Y	1	0	5	11	0.4545455	3	6	0.5000000	15	8	1	0	1	2

Figure 2: Sample of On-Court Statistics

Methods

Sentiment analysis of the 77,000 tweets was done using the RSentiment package in R. The goal of the sentiment analysis was to count the number of tweets that fall into certain categories. By using the RSentiment package, tweets were separated into six categories: Positive, Negative, Very Positive, Very Negative, Neutral, and Sarcastic. The manner in which RSentiment performs this task is done by calculating a sentiment score for each sentence within a vector. Here, a vector containing the 1,000 tweets at each time period was passed through a function which

gave the number of tweets in each of the six categories. However, this study only looked at four of the six with the four of interest being Positive, Negative, Very Positive, and Very Negative. Neutral and sarcastic tweets were of little interest in the study which looked to see if there were any big differences in the number of Positive and Negative tweets in relation to Ball's performance. The decision to keep the separate Positive and Very Positive tweets was made because the difference between the two was also of interest, likewise for Negative and Very Negative.

With the number of Positive, Very Positive, Negative, and Very Negative tweets being found. A Random Forest was utilized to predict the number of tweets in each of the four categories. Thus, four total random forests were created with one being for Positive, another for very Positive, and the two others for the negative counterparts. In a random forest, a number of decision trees are created. In contrast with bagging, a random forest uses a random subset of predictors at each decision node. Bagging uses all the predictors. In this study, the predictors were Ball's game stats, team point differential, and the number of minutes elapsed during the game. By default, each node random selects the total number of predictors divided by three rounded down. Since there were a total of 11 predictors, a random subset of three predictors was used at each node. A random subset of 70% of the observations was used to create each of the random forests.

Results

The sentiment analysis resulted in the number of Positive, Very Positive, Negative, and Very Negative tweets at each of the 77 time periods. Histograms displayed in Figure 3 shows the distributions for the four categories. As seen from the histograms, there is no clear distribution for the types of tweets. Interestingly each distribution is multimodal and look to be clustered around these modes. Separating the types of tweets per game reveals the series shown in Figure 4. The horizontal axis of each series represents the time elapsed per game while the vertical axis represents the number of tweets in that particular tweet category.

The number of positive tweets for each game converged between 190 and 200 per 1000 tweets as the game went on for all three games despite Ball having different types of performances in each game. His performance against the Golden State Warriors was his best. The number of positive tweets declined in the first half before increasing again during the third quarter when Ball began to make impactful plays. For this game, the number of negative tweets appropriately decreased as Ball's performance improved. However, the number of very positive tweets decreased while the number of very negative tweets increased during the same time span.

As for the other two games, Ball's on-court performance declined with each game. The game against the Denver Nuggets saw Ball nearly post a double-double as he finished with nine points and nine assists which can be somewhat be seen with the increase of positive tweets and decrease of negative tweets. The number of very positive and very negative tweets generally stay around their initial levels. Against the Houston Rockets, the Lakers and Ball struggled as a whole. The game ended in a blowout with Ball not playing the fourth quarter. The number of negative and very negative tweets took a high jump around halftime. Interestingly, despite having his best performance against the Warriors, the number of very positive tweets was the lowest for this game. The highest number of very positive tweets was against the Nuggets.

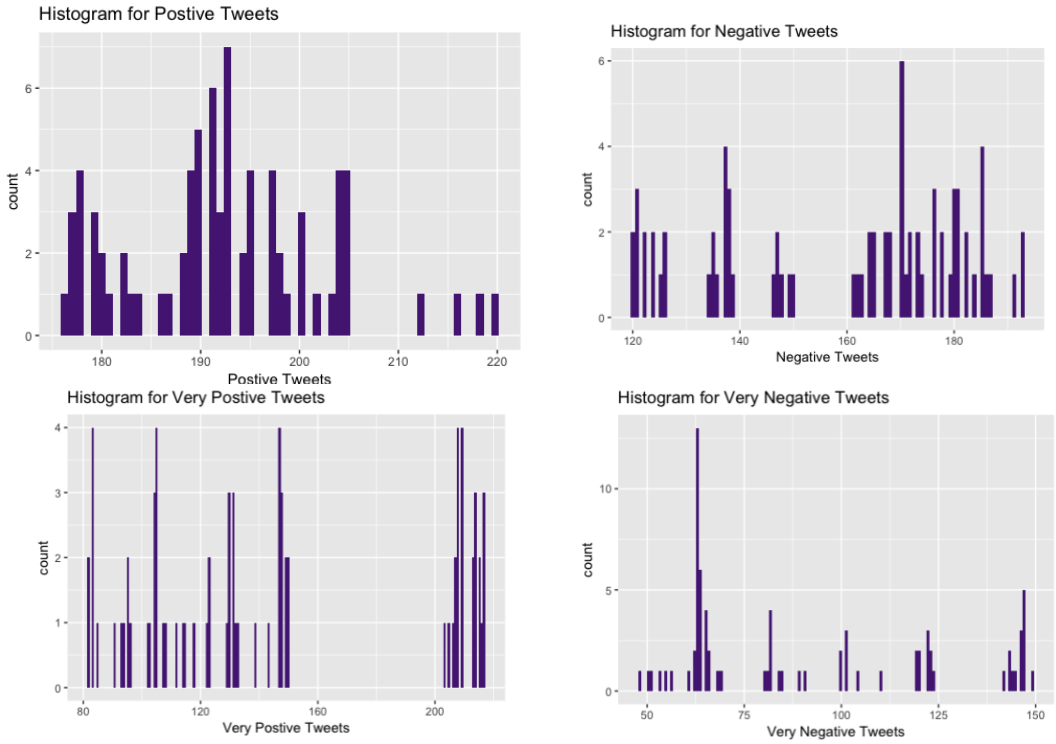

Figure 3: Histograms for Positive, Negative, Very Positive, and Very Negative Tweets

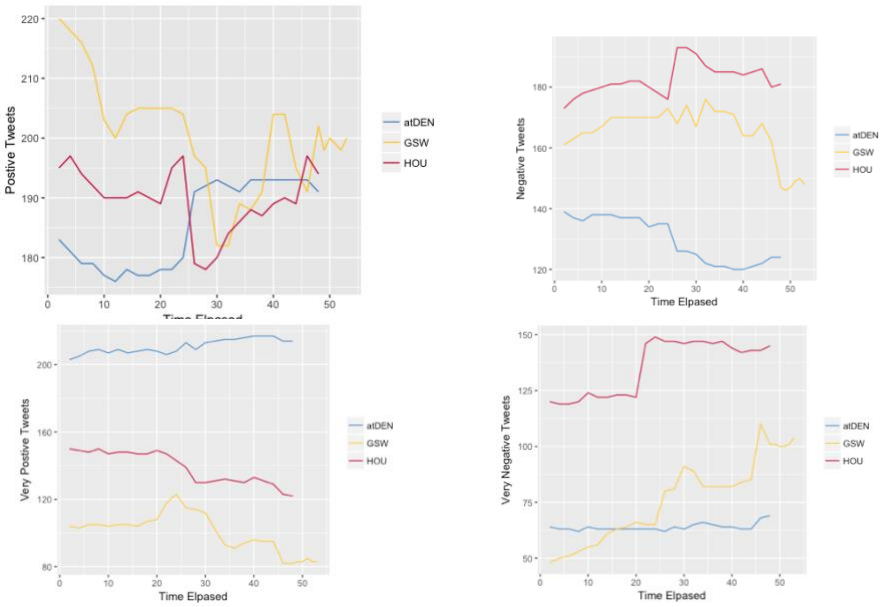

Figure 4: Number of Tweets Per Game

Tweet Type	% Variation Explained	MSE	Top 3 Important Variables
Postive	62.26	31.23432	TO, FGPCT, timeElapased
Negative	88.64	59.30674	STL, FGPCT, REB
Very Postive	89.40	237.1062	STL, REB, PTS
Very Negative	75.55	258.5846	pointDef, STL, BLK

Table 1: Random Forest Results for Each Tweet Type

Using Random Forests results in the percent variation explained, mean squared error, and top three important inputs listen in Table 1. Postive tweets had the least amount of variation explained while very positive tweets had the highest amount. Ball's field goal percentage was an important factor in determining positive and negative tweets. This makes sense since Ball has received plenty of criticism for his shooting form and efficiency. The number of steals was important in three of the four types with steals not being in the top three for positive tweets. Ball's defense has been an underrated portion of his game with the rookie displaying good off-ball intelligence.

Conclusion

Through the three games analyzed, it appears Ball's performance does have an impact on the tweets. As he had more positive impacts on the game, the number of positive tweets increased while the number of negative tweets decreased during the game against the Warriors. This game is noteworthy since it was a nationally broadcast game on ESPN which allowed for a wider audience. The other two games were weekend games only shown on regional networks which did not have the same number of viewers.

The overall variation in positive tweets was additionally not as well explained with a random forest. Despite not a perfect fit, the most important results from the random forest lies in what the model deemed as important factors in determining tweets. Since there was much talk concerning Ball's shooting form and inefficient shooting, it would make sense for his field goal percentage to have impact on tweets which it did for positive and negative tweets. Another interesting outcome involves the team's point differential to have a strong impact on the number of very negative tweets.

Limitations and Future Research

The most glaring limitation of this study comes with only three games being used. Only those three games were considered because of time limitations. Despite having 77 total observations, one of the games produced flawed tweet results. The game against the Nuggets had the same type of tweet constantly repeated due to the way the Twitter search was conducted, although there was check for retweets. A high number of tweets were of the form "I liked a YouTube

video" followed by a URL. The presence of the word "liked" most likely led to the inflated amount of very positive tweets that occurred during the game.

Another limitation comes in just on technique being used. Future research would like to involve utilizing additional techniques as well as more games. For example, a support vector machine can also be used to predict the number of tweets in each category with a comparison with random forests being explored.

R CODE

```
options(java.parameters = "-Xmx8g")

library(twitteR)
library(tm)
library(ggplot2)
library(wordcloud)
library(RColorBrewer)
library(RSentiment)
library(randomForest)
library(readxl)


consumer_key <-"XXX"
consumer_secret <- "XXX"
access_token <- "XXX "
access_secret <- "XXX"

setup_twitter_oauth(consumer_key,consumer_secret,access_token,access_secret)

opponent <- "HOU"

#Lonzo tweets
LonzoTweets <-searchTwitter("Lonzo Ball -filter:retweets", n=1000, lang="en")
LonzoTweets_df <- twListToDF(strip_retweets(LonzoTweets,
strip_manual=TRUE,strip_mt=TRUE))
#LonzoTweets_df <-twListToDF(LonzoTweets)

currentTime <-Sys.time()
currentTime <- format(currentTime,tz="America/Los_Angeles")
write.table(LonzoTweets_df,
file=paste("LonzoTweets.csv",opponent,currentTime),sep=",",row.names=F)

 #Remove Emojis
 LonzoTweets$text <- iconv(LonzoTweets$text, "latin1", "ASCII", sub="")


 #Create Corpus for LA Tweets
 LonzoTweets_corpus <- Corpus(VectorSource(LonzoTweets$text))

 #convert to lower case
 LonzoTweets_corpus <- tm_map(LonzoTweets_corpus, content_transformer(tolower))
```

```
  #remove URLs
  removeURL <-function(x) gsub("http(s?)[^[:space:]]*","",x)
  LonzoTweets_corpus <- tm_map(LonzoTweets_corpus, content_transformer(removeURL))

  #remove Mentions
  removeMention <-function(x) gsub("@\\w+","",x)
  LonzoTweets_corpus <- tm_map(LonzoTweets_corpus,
content_transformer(removeMention))

  #remove anything other than English letters or space
  removeNumPunct <-function(x) gsub("[^[:alpha:][:space:]]*","",x)
  LonzoTweets_corpus <- tm_map(LonzoTweets_corpus,
content_transformer(removeNumPunct))

  #Create document term matrix
  additional_stopwords <- c("lonzo","ball","lakers","los","angeles","shot","like","just",
                    "balls", "nba", "video")

  tdm_Lonzo <- TermDocumentMatrix(LonzoTweets_corpus,
control=list(removePunctuation=TRUE,
                                          stopwords=c(additional_stopwords,stopwords("english")),
                                          removeNumbers=TRUE, tolower=TRUE))

  #Obtain words and frequencies
  m_Lonzo <- as.matrix(tdm_Lonzo)
  word_freqs_Lonzo <- sort(rowSums(m_Lonzo), decreasing = TRUE)
  dm_Lonzo <- data.frame(word=names(word_freqs_Lonzo), freq=word_freqs_Lonzo)

  #Plot WordCloud
  wordcloud(dm_Lonzo$word,dm_Lonzo$freq, scale=c(4,.2), max.words=50, random.order =
FALSE, colors=brewer.pal(5,"Purples"))

  #Sentiment Analysis
  sentiments_Lonzo <- calculate_total_presence_sentiment(LonzoTweets$text)

  #Seperating Categories from Counts
  sentiment_cats_Lonzo <-sentiments_Lonzo[c(TRUE,FALSE)]
  sentiment_counts_Lonzo <-as.numeric(sentiments_Lonzo[c(FALSE,TRUE)])

  #Converting to DataFrame
  sentiments_Lonzo_df <- data.frame(sentiment_cats_Lonzo,sentiment_counts_Lonzo)
  sentiments_Lonzo_df$sentiment_cats_Lonzo <- as.character(sentiment_cats_Lonzo)
  sentiments_Lonzo_df$sentiment_cats_Lonzo <-
factor(sentiments_Lonzo_df$sentiment_cats_Lonzo,
```

```
                                levels = unique(sentiments_Lonzo_df$sentiment_cats_Lonzo))

 #Graphing Sentiment Counts
 ggplot(sentiments_Lonzo_df, aes(sentiment_cats_Lonzo,sentiment_counts_Lonzo)) +
  geom_bar(stat = "identity", fill="#FDB927") +
  labs(x="Emotion", y="Number of Tweets") +
  ggtitle("Lonzo Ball Tweet Sentiment")

LonzoTweetStats <- read_excel("~/Desktop/LonzoTweetStats.xlsx")

set.seed(2)

train=sample(1:nrow(LonzoTweetStats),54)

LonzoRF_pos <- randomForest(posTweets ~ onCourtCode + timeElapsed
                  + PTS + AST + REB + STL + TO + BLK + pointDef + FGPCT + TPPCT,
                  data=LonzoTweetStats, subset = train, ntree=500)
LonzoRF_neg <- randomForest(negTweets ~ onCourtCode + timeElapsed
                  + PTS + AST + REB + STL + TO + BLK + pointDef + FGPCT + TPPCT,
                  data=LonzoTweetStats, subset = train, ntree=500)
LonzoRF_vpos <- randomForest(vposTweets ~ onCourtCode + timeElapsed
                  + PTS + AST + REB + STL + TO + BLK + pointDef + FGPCT + TPPCT,
                  data=LonzoTweetStats,subset = train, ntree=500)
LonzoRF_vneg <- randomForest(vnegTweets ~ onCourtCode + timeElapsed
                  + PTS + AST + REB + STL + TO + BLK + pointDef + FGPCT + TPPCT,
                  data=LonzoTweetStats,subset = train, ntree=500)


print(LonzoRF_pos)
print(LonzoRF_neg)
print(LonzoRF_vpos)
print(LonzoRF_vneg)

varImpPlot(LonzoRF_pos, sort=TRUE)
plot(LonzoRF_pos)
varImpPlot(LonzoRF_neg, sort=TRUE)
plot(LonzoRF_neg)
varImpPlot(LonzoRF_vpos, sort=TRUE)
plot(LonzoRF_vpos)
varImpPlot(LonzoRF_vneg, sort=TRUE)
plot(LonzoRF_vneg)
```