NFL Play by Play Analysis, Linear, Logarithmic, and Random Forest Models

Conor Fogarty and Brennan Frost

2024-05-03

1.0 Introduction

1.1 Project Description For our project, we will be using play-by-play data across one if not more NFL seasons and trying to see what "out-of-the-box" insights we can gain. Naturally, the domain of our project is sports, specifically the NFL (National Football League). The NFL is a professional American football league located in the United States that consists of 32 franchises that compete annually for the Super Bowl Championship. We have 4 guiding questions that serve as our foundational thinking of this project.

1. How does the specific formation of quarterback impact certain results?
2. How does the direction in which the running back rushed through impact certain results?
3. How does the type of pass impact certain results?
4. Is there a trend with the measured impact through different years?

We will use linear regression to predict the yards gained or lost from both the rush direction and pass type.

We will use logistic regression to predict three different variables – whether the play resulted in a touchdown, whether the play carried out was a passing play, and whether the play carried out was a running play from the formation of the quarterback at the snap. We will also predict whether the play resulted in a touchdown from the rush direction, and whether the play resulted in a touchdown from the type of pass play.

We will use Random Forest modeling as another form of a predictive model to help evaluate whether a play results in a touchdown or not, whether the play carried out results in a running play, and whether the play carried out results in a passing play

We will then see if the results from those analyses can be plotted as trends using other years as reference, observing whether the accuracy of the models increased, decreased, or remained stagnant.

1.2 Background/2.0 Data Description We combined these two sections becasue it was easier to write through and understand when combined into one section Our data comes from NFLsavant.com, a website that houses play-by-play data from every game of every year for the past 10 NFL seasons. These are split up into 10 different .csv files (we used the 2023 .csv for our Data Exploration section). Each .csv file contains 45 different variables. Here are descriptions of the important variables that we know could used in some shape or form of the project;

Yards - int: The amount of yards gained or lost on the play Formation - chr (will be factored): Where the quarterback was in relation to the football before the ball was snapped Play Type - chr (will be factored): Which type of play was carried out by the offense? IsRush - int (will be factored): Whether the play resulted in a run IsPass - int (will be factored): Whether the play resulted in a pass IsTouchdown - int (will be factored): Whether the play resulted in a touchdown PassType - chr (will be factored): Where the quarterback throws the ball from the line of scrimmage relative to the field. RushDirection - chr (will be factored): The player on the offensive line that the running back ran behind if the play was a run.

```
data.df <- read.csv("pbp-2023.csv")
head(data.df[,1:5], 5)
```

```
##       GameId   GameDate Quarter Minute Second
## 1 2023121101 2023-12-11       3      1     28
## 2 2023121101 2023-12-11       3      1     35
## 3 2023121101 2023-12-11       3      2     19
## 4 2023121101 2023-12-11       3      2     56
## 5 2023121101 2023-12-11       3      3     43
```

```
data.df$Formation <- factor(data.df$Formation)
formation.tbl<-table(data.df$Formation)
kable(formation.tbl, col.names = c("Formation", "Frequency of Plays"))
```

Formation	Frequency of Plays
	636
FIELD GOAL	824
NO HUDDLE	381
NO HUDDLE SHOTGUN	2302
PUNT	1746
SHOTGUN	18038
UNDER CENTER	15545

```
data.df <- read.csv("pbp-2023.csv")
head(data.df, 10)
```

```
##        GameId   GameDate Quarter Minute Second OffenseTeam DefenseTeam Down
## 1  2023121101 2023-12-11       3      1     28         NYG          GB    0
## 2  2023121101 2023-12-11       3      1     35         NYG          GB    3
## 3  2023121101 2023-12-11       3      2     19         NYG          GB    2
## 4  2023121101 2023-12-11       3      2     56         NYG          GB    1
## 5  2023121101 2023-12-11       3      3     43         NYG          GB    1
## 6  2023121101 2023-12-11       3      4     29         NYG          GB    2
## 7  2023121101 2023-12-11       3      6     30         NYG          GB    0
## 8  2023121101 2023-12-11       3      7     31         NYG          GB    0
## 9  2023121101 2023-12-11       3      7     31         NYG          GB    0
## 10 2023121100 2023-12-11       2      0     23         TEN         MIA    0
##    ToGo YardLine  X SeriesFirstDown X.1 NextScore
## 1     0       85 NA               1  NA         0
## 2     7       92 NA               1  NA         0
## 3    11       88 NA               0  NA         0
## 4    10       89 NA               0  NA         0
## 5    10       64 NA               1  NA         0
## 6     3       55 NA               1  NA         0
## 7     0      100 NA               1  NA         0
## 8     0       65 NA               1  NA         0
## 9     0      100 NA               1  NA         0
## 10    0      100 NA               1  NA         0
##
## 1                                                                  46-R.BULLOCK EXTRA POINT IS GOOD, CENT
## 2                                                (1:35) (SHOTGUN) 15-T.DEVITO PASS SHORT RIGHT TO 1
## 3                                                          (2:19) (SHOTGUN) 26-S.BARKLEY RIGHT GUAR
## 4                                                       (2:56) (SHOTGUN) 15-T.DEVITO UP THE MIDDLE
## 5  (3:43) 15-T.DEVITO PASS DEEP LEFT TO 17-W.ROBINSON TO GB 11 FOR 25 YARDS (59-D.CAMPBELL). FLEA FL
## 6                                                                  (4:29) (SHOTGUN) 31-M.BREIDA RIGHT GUAR
## 7
```

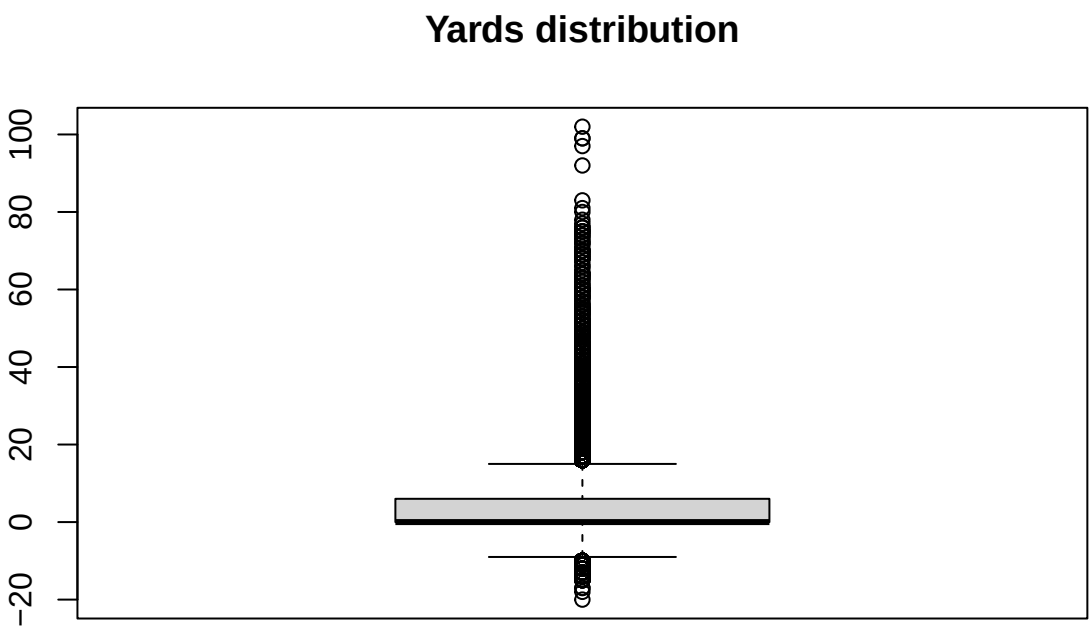
```
## 8                                                                      17-A.CARLSON KICKS 65 YARDS
## 9
## 10
##    TeamWin X.2 X.3 SeasonYear Yards    Formation    PlayType IsRush IsPass
## 1        0  NA  NA       2023     0 UNDER CENTER EXTRA POINT      0      0
## 2        0  NA  NA       2023     8      SHOTGUN        PASS      0      1
## 3        0  NA  NA       2023     4      SHOTGUN        RUSH      1      0
## 4        0  NA  NA       2023    -1      SHOTGUN        RUSH      1      0
## 5        0  NA  NA       2023    25 UNDER CENTER        PASS      0      1
## 6        0  NA  NA       2023     9      SHOTGUN        RUSH      1      0
## 7        0  NA  NA       2023     0 UNDER CENTER                  0      0
## 8        0  NA  NA       2023     0 UNDER CENTER    KICK OFF      0      0
## 9        0  NA  NA       2023     0 UNDER CENTER                  0      0
## 10       0  NA  NA       2023     0 UNDER CENTER     TIMEOUT      0      0
##    IsIncomplete IsTouchdown    PassType IsSack IsChallenge IsChallengeReversed
## 1             0           0                  0           0                   0
## 2             0           1 SHORT RIGHT      0           0                   0
## 3             0           0                  0           0                   0
## 4             0           0                  0           0                   0
## 5             0           0   DEEP LEFT      0           0                   0
## 6             0           0                  0           0                   0
## 7             0           0                  0           0                   0
## 8             0           0                  0           0                   0
## 9             0           0                  0           0                   0
## 10            0           0                  0           0                   0
##    Challenger IsMeasurement IsInterception IsFumble IsPenalty
## 1          NA             0              0        0         0
## 2          NA             0              0        0         0
## 3          NA             0              0        0         0
## 4          NA             0              0        0         0
## 5          NA             0              0        0         0
## 6          NA             0              0        0         0
## 7          NA             0              0        0         0
## 8          NA             0              0        0         0
## 9          NA             0              0        0         0
## 10         NA             0              0        0         0
##    IsTwoPointConversion IsTwoPointConversionSuccessful RushDirection
## 1                     0                              0
## 2                     0                              0
## 3                     0                              0   RIGHT GUARD
## 4                     0                              0        CENTER
## 5                     0                              0
## 6                     0                              0   RIGHT GUARD
## 7                     0                              0
## 8                     0                              0
## 9                     0                              0
## 10                    0                              0
##    YardLineFixed YardLineDirection IsPenaltyAccepted PenaltyTeam IsNoPlay
## 1             15               OPP                 0                    0
## 2              8               OPP                 0                    0
## 3             12               OPP                 0                    0
## 4             11               OPP                 0                    0
## 5             36               OPP                 0                    0
## 6             45               OPP                 0                    0
```

```
## 7              0              OPP                0                   0
## 8             35              OPP                0                   0
## 9              0              OPP                0                   0
## 10             0              OPP                0                   0
##    PenaltyType PenaltyYards
## 1                         0
## 2                         0
## 3                         0
## 4                         0
## 5                         0
## 6                         0
## 7                         0
## 8                         0
## 9                         0
## 10                        0
```

```
summary(data.df$Yards)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -20.00    0.00    0.00    3.74    6.00  102.00
```

```
boxplot(data.df$Yards, main='Yards distribution')
```



3.0 Analysis

3.1 Methods (The Plan) We have 4 guiding questions that serve as our foundational thinking of this project.

1. How does the specific formation of quarterback impact certain results?
2. How does the player on the offensive line that the running back ran behind impact certain results?
3. How does the type of pass impact certain results?
4. Is there a trend with the measured impact through different years?

With these questions in mind, we will use three different modeling techniques to help us answer them; linear regression, logistic regression, and random forest modeling.

LINEAR REGRESSION We will use linear regression to predict the yards gained or lost from both the rush direction and pass type.

We will have to create a subset of just the pass plays from the existing dataframe, and another subset of just the rush plays from the existing dataframe. We will then factor the independent variables – PassType and RushDirection – into levels.

PassType is where the quarterback throws the ball from the line of scrimmage relative to the field. The dataset has six types of pass plays for the variable PassType; SHORT RIGHT, SHORT MIDDLE, SHORT LEFT, DEEP RIGHT, DEEP MIDDLE, and DEEP LEFT. For example, the NFL defines a deep pass 15 yards or more downfield from the line of scrimmage, so DEEP LEFT means that the quarterback threw the ball 15 yards or more down his left side of the field from the line of scrimmage.

RushDirection is the player on the offensive line that the running back ran behind if the play was a run. The goal of a running back is to find holes in the offensive line to run through as they are smashing into the defensive line. These holes are created by players at the specific positions.The dataset has seven different positions for the variable Rush Direction; RIGHT TACKLE, RIGHT GUARD, RIGHT END, LEFT TACKLE, LEFT GUARD, LEFT END, and CENTER.

We will also divide the data into training and testing sets. The training set is what each linear regression model will learn from, and the testing set is what each linear regression model will be evaluated on to see how well it performs. It does this division randomly, with 75% of the data used for training and 25% for testing.

We then plan to run two linear regression models.

1. One uses Rush Direction to predict the total amount of yards gained (either positive or negative) on the play. We will use the training and test data accordingly to judge overall accuracy.
2. Another uses PassType to predict the total amount of yards gained (either positive or negative) on the play. We will use the training and test data accordingly to judge overall accuracy.

The analysis of the linear regression model that used Rush Direction to predict the total amount of yards gained (either positive or negative) on the play can be seen in the Results section.

LOGISTIC REGRESSION We plan to use logistic regression for three processes. One is to predict three different variables – whether the play resulted in a touchdown (isTouchdown), whether the play carried out was a passing play (isPass), and whether the play carried out was a running play (isRush) from the formation of the quarterback at the snap (Formation). Another is to predict whether the play resulted in a touchdown (isTouchdown) from the rush direction (RushDirection). The last process is to predict whether the play resulted in a touchdown (isTouchdown) from the type of pass play (PassType).

We will create an additional subset of just plays that have a specific quarterback formation from the existing dataframe.

We will then factor the independent variable – Formation – into levels. Formation explains the specific formation of the quarterback relative to where the ball is on the line of the scrimmage. The dataset has six types of pass plays for the variable PassType; UNDER CENTER, SHOTGUN, NO HUDDLE SHOTGUN, NO HUDDLE, and PUNT. UNDER CENTER means that the quarterback squats behind the center (who is waiting to snap the football) ready to take the ball from the center on his command. SHOTGUN means that the quarterback is three to five yards behind the center waiting to catch the snap from the center. NO HUDDLE means that the offense is not huddling after a play has ended, wasting little time to get off the next play. NO HUDDLE SHOTGUN means the offense is running no huddle and the quarterback is in the SHOTGUN formation. This implies that NO HUDDLE also means that the quarterback is UNDER CENTER. PUNT is just the standard punt formation when the offensive team is getting ready to punt away the ball to the defensive team. The RushDirection and PassType subsets are described in the LINEAR REGRESSION section.

Like the linear regression models, we will divide the data into training and testing sets. The training set is what each logistic regression model will learn from, and the testing set is what each logistic regression model

will be evaluated on to see how well it performs. It does this division randomly, with 75% of the data used for training and 25% for testing.

We then plan to run five logistic regression models.

1. One uses Formation to predict whether the play carried out ended up as a touchdown or not.
2. Another uses Formation to predict whether the play carried out ended up as a running play or not.
3. Another uses Formation to predict whether the play carried out ended up as a passing play or not.
4. Another uses RushDirection to predict whether the play carried out ended up as a touchdown or not.
5. The last one uses PassType to predict whether the play carried out ended up as a touchdown or not.

The analysis of the logistic regression model that used Formation to predict whether the play carried out ended up as a touchdown or not can be seen in the Results section.

RANDOM FOREST MODELING For our third modeling tool, we plan to build and evaluate three Random Forest models.

1. One Random Forest model builds two Random Forest sub-models to predict whether a play results in a touchdown or not; one sub-model is trained on the original, imbalanced data. The other is trained on the data after applying the oversampling technique.
2. Another Random Forest model builds two Random Forest sub-models to predict whether a play results in a running play; one sub-model is trained on the original, imbalanced data. The other is trained on the data after applying the oversampling technique.
3. Another Random Forest model builds two Random Forest sub-models to predict whether a play results in a passing play; one sub-model is trained on the original, imbalanced data. The other is trained on the data after applying the oversampling technique.

Once the models are trained, it uses them to make predictions on the test set, which contains data the models haven't seen before. Then, it evaluates the performance of the models using confusion matrices.

The analysis of one Random Forest model that built two Random Forest sub-models to predict whether a play results in a touchdown or not can be seen in the Results section.

YEAR-BY-YEAR TREND We plan to run these same models over years of the same NFL play-by-data. Meaning, we will carry out the same process, producing the same models, just with different seasons. For example, the models carried out in the Results section all take from data from the year 2023 (pbp-2023.csv). We will run these same models with the 2022 dataset, 2021, 2020, etc.

We are trying to see if there is a trend in which the accuracy of the model increases, decreases, or stays the same over an interval of time. Different plots and graphs will be generated to show this insight.

```
df <- read.csv('pbp-2023.csv')
columns_to_keep <- c(1, 2, 8:9, 20:28, 33:38, 41, 45)
df_subset <- df[, columns_to_keep]
```

```
#Factor Section
df_subset$Down <- factor(df_subset$Down)
df_subset$Formation <- factor(df_subset$Formation)
df_subset$PlayType <- factor(df_subset$PlayType)
df_subset$IsRush <- factor(df_subset$IsRush)
df_subset$IsPass <- factor(df_subset$IsPass)
df_subset$IsIncomplete <- factor(df_subset$IsIncomplete)
df_subset$IsTouchdown <- factor(df_subset$IsTouchdown)
df_subset$PassType <- factor(df_subset$PassType)
df_subset$IsSack <- factor(df_subset$IsSack)
```

```r
df_subset$IsInterception <- factor(df_subset$IsInterception)
df_subset$IsFumble <- factor(df_subset$IsFumble)
df_subset$IsPenalty <- factor(df_subset$IsPenalty)
df_subset$IsTwoPointConversion <- factor(df_subset$IsTwoPointConversion)
df_subset$IsTwoPointConversionSuccessful <- factor(df_subset$IsTwoPointConversionSuccessful)
df_subset$RushDirection <- factor(df_subset$RushDirection)
df_subset$IsPenaltyAccepted <- factor(df_subset$IsPenaltyAccepted)
```

```r
df_form <- df_subset %>% arrange(Formation)
df_form <- df_form[-c(1:636), ]
```

```r
#train/test for linear and log regression
RNGversion("4.1.2")
set.seed(123456)
#train data for Lin Reg
index.yards.train <- createDataPartition(y = df_form$Yards, p = 0.75, list = FALSE)
train.yards.data <- df_form[index.yards.train,]
test.yards.data <- df_form[-index.yards.train,]
```

```r
#Linear Regression Model, response = yards, input = RushDirection
yardsLM <- lm(Yards ~ RushDirection, data = train.yards.data)
summary(yardsLM)
```

3.1.1 Results (Tables/Plots and their explainations)

```
##
## Call:
## lm(formula = Yards ~ RushDirection, data = train.yards.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.924  -3.658  -3.658   1.659  95.342
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                3.65809    0.05283  69.248  < 2e-16 ***
## RushDirectionCENTER        0.13769    0.18181   0.757   0.4489
## RushDirectionLEFT END      1.26579    0.25599   4.945 7.67e-07 ***
## RushDirectionLEFT GUARD    0.27858    0.25347   1.099   0.2717
## RushDirectionLEFT TACKLE   0.61435    0.25898   2.372   0.0177 *
## RushDirectionRIGHT END     1.68324    0.26735   6.296 3.10e-10 ***
## RushDirectionRIGHT GUARD   0.45357    0.24844   1.826   0.0679 .
## RushDirectionRIGHT TACKLE  0.22434    0.27795   0.807   0.4196
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.757 on 29120 degrees of freedom
## Multiple R-squared:  0.002309,   Adjusted R-squared:  0.002069
## F-statistic: 9.628 on 7 and 29120 DF,  p-value: 5.126e-12
```

In the linear regression model, the significant predictors of variable Yards with in the variable RushDirection, are RightEnd(P-value:3.10e-10), LeftEnd(P-value: 7.67e-07), and LeftTackle(0.0177).These are presented in their level of importance based on P-value. The other

RushDirections did not reach any statistically significant marks.

```
#Prediction of LM - Yards
#pred_yardsLM <-predict(yardsLM, test.yards.data, interval="prediction", level=0.95)
#plot(pred_yardsLM)
```

```
#Log reg, isTouchdoum, response = isTouchdoum, input = Formation
log_reg_model <- glm(IsTouchdown ~ Formation, data = train.yards.data, family = binomial(link = "logit"
summary(log_reg_model)
```

```
##
## Call:
## glm(formula = IsTouchdown ~ Formation, family = binomial(link = "logit"),
##     data = train.yards.data)
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -6.4184     0.9994  -6.422 1.34e-10 ***
## FormationNO HUDDLE            3.5060     1.0339   3.391 0.000697 ***
## FormationNO HUDDLE SHOTGUN    2.7861     1.0107   2.757 0.005842 **
## FormationPUNT                 0.8334     1.0952   0.761 0.446705
## FormationSHOTGUN              3.1799     1.0004   3.179 0.001480 **
## FormationUNDER CENTER         2.4536     1.0017   2.449 0.014310 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7279.6  on 29127  degrees of freedom
## Residual deviance: 7117.3  on 29122  degrees of freedom
## AIC: 7129.3
##
## Number of Fisher Scoring iterations: 8
```

In the logarithmic regression model, the significant predictors of variable isTouchdown with in the variable Formation, are No Huddle (P-value: 0.000697), Shotgun (P-value: 0.001480), No Huddle Shotgun (P-value:0.005842), and finally Under Center (P-value: 0.014310). These are presented in their level of importance based on P-value.

```
#train/test for Random Forest
RNGversion("4.1.2")
set.seed(123456)
#train data for Lin Reg
index.linearyards.train <- createDataPartition(y = df_form$IsTouchdown, p = 0.75, list = FALSE)
train.linearyards.data <- df_form[index.linearyards.train,]
test.linearyards.data <- df_form[-index.linearyards.train,]
```

```
#ROSE
train.linearyards.data_rose <- ovun.sample(IsTouchdown ~ ., data = train.linearyards.data, method = "ov

# Extract the oversampled data
train.linearyards.data_rose <- train.linearyards.data_rose$data
```

```
#Random Forest Model building
rf_model_imb <- randomForest(IsTouchdown ~ ., data = train.linearyards.data, ntree = 500)
rf_model_ROSE <- randomForest(IsTouchdown ~ ., data = train.linearyards.data_rose, ntree = 500)
# Predict on the test set
```

```r
predictions_imb <- predict(rf_model_imb, test.linearyards.data)
predictions_ROSE <- predict(rf_model_ROSE, test.linearyards.data)

confusionMatrix(predictions_imb, test.linearyards.data$IsTouchdown)$table
```

```
##            Reference
## Prediction    0    1
##          0 9441  242
##          1    3   22
```

```r
confusionMatrix(predictions_ROSE, test.linearyards.data$IsTouchdown)$table
```

```
##            Reference
## Prediction    0    1
##          0 8375   70
##          1 1069  194
```

In the RandomForest model, were were usign all variables to be able to predict if a play was going to be a touchdown via the isTouchdown variable. We created to train/test data pools, one of the real data (considered imbalanced data) and one using the ROSE methadology (considered balanced data). The imbalanced data model corrected predicted touchdowns with 97.5%, and correctly predicted no touchdown with 88.8%, with a total accurately total accuracy of 97.49%. While the "balanced" ROSE data model corrected predicted touchdowns with 99.15%, and correctly predicted no touchdown with 84.8%, with a total accurately total accuracy of 88.17%. As expected with balanced data sets, the 'positive' (touchdown scored) prediction accuracy increased, however the 'negative' (no touchdown) prediction accuracy significantly decreased.

4.0 Conclusions

From the linear regression model that we ran that used Rush Direction to predict the total amount of yards gained (either positive or negative), we found that LEFT END, LEFT TACKLE, and RIGHT END were the most significant factors that the model produced. The adjusted R-squared value is very low, but we think that running the model with just the significant variables will produce a better model to predict yards.

From the logistic regression model that used Formation to predict whether the play carried out ended up as a touchdown or not, we found that both NO HUDDLE's, SHOTGUN, and UNDER CENTER were the most significant variables. Their low p-values suggest that the model is good to predict isTouchdown, and further analysis can be conducted predicting other variables. The results show that Formation could be a good predictor of touchdowns, showing promise that it can also predict running plays and passing plays.

From the random forest model that used all varaibles to predicted whether a play was going to result in a touchdown or not, we found that we could get 97.5% accuracy in correctly predicting this with a imbalanced model, and using a ROSE model we were able to get that accuracy to 99.15%, however this came at the cost of decreasing no touchdown prediction accuracy from 88.8% to 84.8%, this is to be expected when using "balanced" datasets.

5.0 References

https://nflsavant.com/about.php