**MSDS 6372 Project 2**

*Zackary Gill, Johnny Gipson, Satish Mylapore*

1. **Introduction**

Every NFL season there are numerous games where field goals are a significant decider in who wins a game. The ability to predict whether a field goal will be successfully scored would help teams decide if a field goal should be attempted. Our team will first create an easily interpretable logistic model. Next we will use all tools at our disposal to create predictive models using PCA, LDA, Random Forests, and a complex logistic model. To accomplish this goal we have a dataset that we will use to determine what predictors are good and use a combination of the accuracy, sensitivity, and specificity to show the correctness of our models.

1. **Data Description**

The data set includes the results of all NFL regular season field goal attempts for the 2008 season. The original data set has 23 variables with 1039 rows. The data set has three binary outcomes for field goals (of which we will obviously use only one as a response variable). That variable ‘GOOD’ is represented by a 1 if the field goal was successful and a 0 if it missed. The source of the data set is from advancedfootballanalytics.com. The data set was obtained from the university of Florida data science website. To see a list of data variables and data description please refer to the appendix. For more information go to: http://users.stat.ufl.edu/~winner/datasets.html

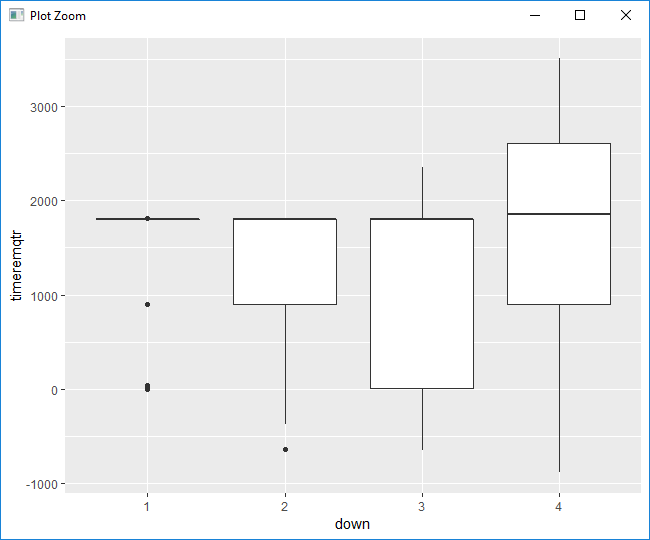
1. **Exploratory Data Analysis**

Using our football knowledge, usually a missed field goal is usually caused by the long distance and usually the time left remaining in the game. Investigating a plot of the two variables, we definitely see that the missed occur at all distance but have a heavy cluster between 40 plus yards but we also a lot field goals made in the 40 to 50 yd range (please refer to appendix).

In our data set in addition to our response variable ‘GOOD’, there are two variables ‘Missed’ and ‘Blocked’ that overlap in its indication. We have removed these. The dataset contained 2 rows that each contained 2 columns of missing data. Upon examining those rows we used information from pro-football-reference.com and our knowledge of football to determine the values of that missing data. For our analysis we decided that we did not want to limit the predictive capabilities to just the 2008 season and to those kickers, we removed these columns: ‘GameDate’, ‘name’, ‘kicker’, ‘season’, ‘HomeTeam’, ‘AwayTeam’, ‘kickteam’, and ‘def’.



**[CHART 3.1]**

**[CHART 3.2]**

Looking at the correlation of continuous variables **[CHART 3.1]** we noticed that ‘ydline’ and ‘distance’ are nearly identical (0.99 correlation) and so we removed ‘ydline’. ‘timerem’, ‘min’, and ‘sec’ were correlated so we created a new variable ‘timeremqtr’ that indicates how much time is remaining in the quarter and removed ‘min’ and ‘sec’. ‘kickdiff’ is correlated with ‘offscore’ and ‘defscore’ so we kept ‘kickdiff’ and removed the other two. At this point there was no more multi-collinearity for the continuous variables.

Next we made tables of the categorical variables (qtr, down, homekick, and GOOD) in groupings of two. We discovered that ‘qtr’ had several groupings of data that had zero information and also had very few values during overtime. Because this would impede our ability to train/test we removed ‘qtr’.

Looking plots of continuous vs categorical (example in **[CHART 3.2]**), we came across three instances of multi-collinearity. They were between [timeremqtr, down], [timerem, down], and [togo, down]. Running a Kruskal-Wallis ANOVA test resulted in there being definitive evidence of this (all p-values < 3.637e-07). Because ‘down’ was one of the culprits every time, that variable was removed. Finally, a vif (variance inflation factor) was performed on the rest of the variables. The vif quantifies the severity of multicollinearity in regression analysis. As a rule of thumb is that if a VIF value exceeds the number 10 then we have multicollinearity. The rest of potential variables have values less than two (please refer to Appendix Table#) so there is no collinearity between the variables. At this point the cleaning portion of the data analysis was completed.

1. **Formation of Test and Training Set**

**Short and brief paragraph on the technique chosen and why.**

1. **Objective 1**
2. **Problem Restatement**

Every NFL season there are numerous games where field goals are a significant decider in who wins a game. The ability to predict whether a field goal will be successfully scored would help teams decide if a field goal should be attempted. Our team will create a logistic model to predict field goal success in the NFL. First ,we build a simplistic logistic model. Whatdo we mean by simplistic logistic model? The below equation will be used as a basis for our simplistic logistic model:

β0+β1X

We will be using model selection techniques to do our variable selection and these computer model selection do not take into complex terms like interaction or transformations. We will be using the model selection techniques of forward selection, stepwise, and LASSO for variable selection. We will be using estimator of relative quality such as AIC (Akaike Information Criteria), accuracy, sensitivity and specificity to determine what is the best model. Once a model is selected then an interpretation of the of the model regression coefficients including hypothesis testing and confidence intervals. We will also comment on the practical and statistical significance of the important factors. We will not be including interaction terms at this time.

1. **Model Selection**

**b.i Forward, Stepwise, LASSO**

Forward Selection is a model selection technique that begins with just the intercept and then sequentially adds the effect that most improves the fit. The process terminates when no significant improvement can be obtained by adding variables. The forward selection process produce the following model.

6.221+9.8573e-3(togo)-1.092e-1(distance)-3.076e-1(homekick)+ 3.003e-3(kickdiff)-3.207e-5(timerem)+ 5.629e-4(timeremqtr)

In the stepwise model selection, a variable is considered for addition to or subtraction from the set explanatory variables based on some prescribed criteria. In our stepwise model, five fold cross validation was used for our training set and the prescribed criteria was AIC. The stepwise model selection produce a model where the two predictors which were ‘distance’ and ‘homekick1’.

6.311-0.11(distance)

LASSO (least absolute shrinkage and selection operator) is a penalized regression model technique that is used to select a subset of variables. The computed LASSO model is

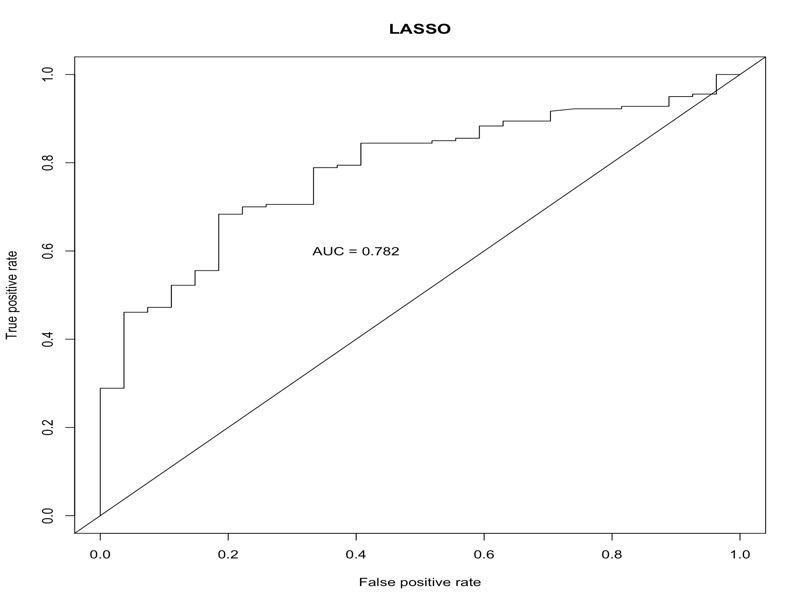
5.85+.013(togo)-.099(distance)-.161(homekick1)+ .0002(timeremqtr)

Table 1: Model Results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **AIC** | **AUC** | **Accuracy** | **Sensitivity** | **Specificity** |
| Forward | 560.75 | .77 | .86 | .03 | .99 |
| Stepwise | 554.48 | .78 | .86 | .03 | .99 |
| LASSO | 511 | .782 | .83 | .11 | .94 |

Based on quality estimators, the team selected the XXXX model.

Talk about AUC



**b.ii Assumptions and Influential points** (actually just EDA stuff + Cook’s D)

**b.iii Interpretation include confidence interval.**

Var Coef Z-score P-value LowConfPt UpConfPt LowTailArea UpTailArea

# 1 0.010 0.386 0.699 -0.202 0.045 0.000 0.049

# 2 -0.109 -8.148 0.000 0.083 0.131 0.049 0.049

# 3 -0.308 -1.379 0.168 -0.244 0.676 0.049 0.048

# 4 0.003 0.266 0.791 -0.125 0.016 0.050 0.049

# 5 0.000 -0.284 0.776 -0.001 0.000 0.050 0.050

# 6 0.001 1.358 0.174 0.000 0.001 0.050 0.048

#

1. **Conclusion**
2. **Objective 2**
3. **Introduction**

We have a created a simple logistic model to predict the success of NFL field goal kicks based on a set of predictors. Using this simple model as a baseline, we will complex logistic model using interaction terms and will also investigate competing models such as LDA and Random Forest.

1. **Main Analysis**

**b.i Complex Logistic**

**b.ii LDA**

**b.iii PCA**

**b.iv Random Forest**

1. **Conclusion**

The conclusion should reprise the questions and conclusions of objective 2 with recommendations of the final model, what could be done to help analysis and model building in the future, and any insight as to why one method outshined all the rest if that is indeed the case. If they all are similar why did you go with your final model?

1. **Appendix**

**Data Set** **Variable Names/Description:**

GameDate

AwayTeam

HomeTeam

qtr (quarter, 5=overtime)

min (minutes remaining)

sec (seconds remaining, added to minutes)

kickteam (Team kicking field goal)

def (Defending Team)

down

togo (Yards to go for 1st down)

kicker (ID #)

ydline (yardline of kicking team)

name (kicker's name)

distance (yards)

homekick (1 if kicker at Home, 0 if Away)

kickdiff (Kicking team lead +, or deficit -, prior to kick)

timerem (Time remaining in seconds, negative ==> overtime)

offscore (kicking team's score prior to kick)

defscore ( defense " " " " ")

season (2008)

GOOD (1 if Success, 0 if Miss)

Missed (Missed, not blocked =1, 0 ow)

Blocked (1 if Blocked, 0 ow)

**Data Cleansing:** In our data set we found NA’s in our data set for two games. With us having the games date and the 20078 schedule, we were able to obtain the box score information and stats for each game.

Link for the missing information:

Detroit vs Green Bay: <https://www.pro-football-reference.com/boxscores/200812280gnb.htm>

New York Giants vs Arizona: <https://www.pro-football-reference.com/boxscores/200811230crd.htm>

**VIF Table**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **To go** | **Distance** | **Homekick** | **Kickdiff** | **Timerem** | **timeremqtr** |
| **VIF** | **1.08** | **1.05** | **1.05** | **1.07** | **1.03** | **1.05** |

