**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**

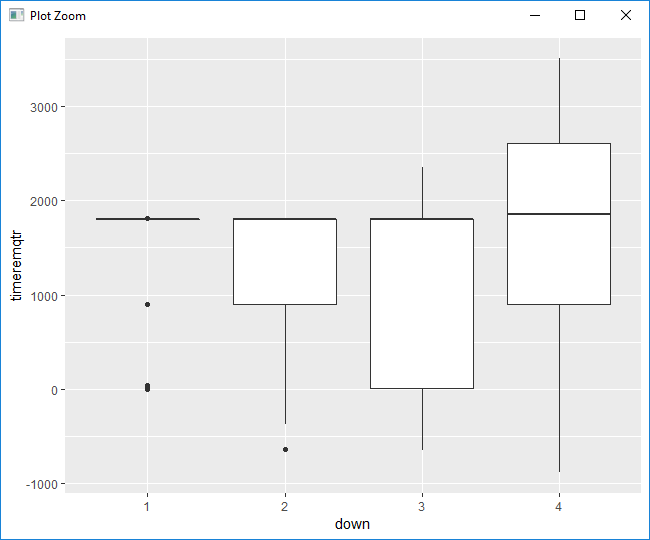
The data set includes the results of all NFL regular season field goal attempts for the 2008 season. The original data set has 23 variables with 1039 rows. The data set has three binary outcomes for field goals (of which we will obviously use only one as a response variable). That variable ‘GOOD’ is 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. For more information go to: http://users.stat.ufl.edu/~winner/datasets.html

1. **Exploratory Data Analysis**

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. At this point the cleaning portion of the data analysis was completed.

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

…and the overall approach to solve it

1. **Model Selection**

**b.i Lasso and Stepwise**

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

**b.iii Interpretation**

1. **Conclusion**
2. **Objective 2**
3. **Introduction**
4. **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**