**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

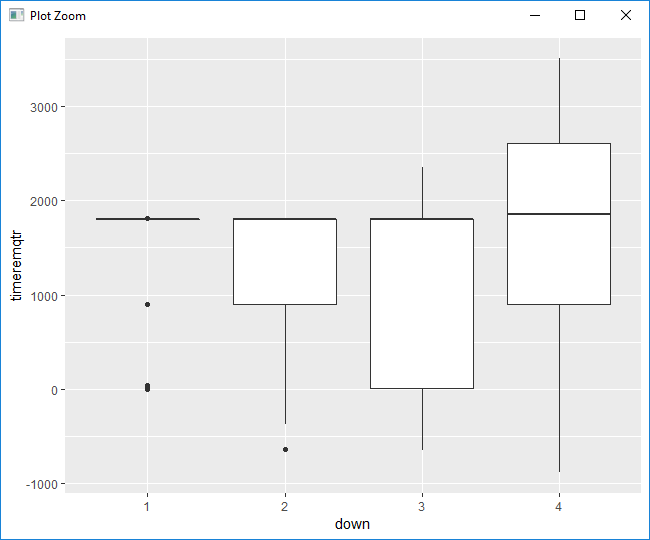
1. **Exploratory Data Analysis**

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 which we confirmed by looking at the variance inflation factor (VIF) and determining that the output values were less than 5.

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 (nonparametric test, normality is not a requirement) 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. **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. What do 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. **Formation of Training and Test Set**

For this objective, the typical way of doing random sampling of the data set to obtain train and test was not the best option. In below summary statistics table **[TABLE 4.b.1]**, you can see our categorical response ‘GOOD’ has a very small number of missed field goals (0’s) with respect to the number of successful field goals (1’s).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Summary Stats | **Togo** | **Distance** | **Homekick** | **Kickdiff** | **Timerem** | **Good** | **Timeremqtr** |
| Min | 1.0 | 18.0 |  | -41.0 | -887.0 |  | 1.0 |
| 1st Quartile | 4.0 | 28.0 |  | -6 | 895.5 |  | 113.0 |
| Medial | 6.0 | 37.0 |  | 0 | 1808.0 |  | 326.0 |
| Mean | 6.7 | 36.7 |  | 0.4 | 1704.6 |  | 352.5 |
| 3rd Quartile | 9 | 44.0 |  | 6.0 | 2556.5 |  | 579.5 |
| Max | 25 | 76.0 |  | 44.0 | 3507.0 |  | 900.0 |
| 0 (No) |  |  | 525 |  |  | 139 |  |
| 1 (Yes) |  |  | 514 |  |  | 900 |  |

**[TABLE 4.b.1 – Summary Statistics]**

We initially went with making the number of 1’s and 0’s for GOOD in the test set equal, but because of the low number of 0’s available in the data set this option gave poor results. The decision was made to split the data 80/20 (80% in the training, 20% in the test set) to ensure that there was enough data for training. For that 80/20 split we made 80% of the 0’s be in the training set and 20% of them in the test set. This helped ensure that the model would consistently predict/interpret both the successful field goals and the misses.

1. **Assumptions and Influential Points**

*Linearity*

|  |  |
| --- | --- |
| **[CHART 4.c.1]** | We will check the linear relationship between continuous predictor variables and the logit of the outcome **[CHART 4.c.1]**. This is done by inspecting the scatter plot between each predictor and the logit values. We will first remove qualitative variables from the original data and bind the logit values to the data. ‘Distance’ and ‘timerem’ shows linear relationship with the response variable ‘GOOD’. Predictor variables ‘kickdiff’, ‘timeremqtr’ and ‘togo’ show the linear relationship after cubic transformation of these variables. |

*Influential points*

|  |  |
| --- | --- |
| **[CHART 4.c.2 – Cook’s D and Residuals vs Leverage]** | Two points have been observed with Cooks’d value **[CHART 4.c.2]** higher than other values, however, not all outliers are influential observations. Data points 853 is still under 0.04 (less than 4/n = 0.004807692), hence these points are not considered as outlier. We also could see that the data. We also see that in Residuals vs Leverage **[CHART 4.c.2]** show no leverage on this dataset. |

*Multicollinearity*

As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity. In our data there is no collinearity: all variables have a value of VIF well below 5.

1. **Model Selection – Forward, Stepwise, and 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 produces the following model.

7.06 – 0.12\*distance – 0.37965\*homekick1

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, the prescribed criteria was AIC. The stepwise model selection produce a model where the two predictors which were ‘distance’ and ‘homekick1’.

6.418 – 0.12\*distance – 0.37965\*homekick1

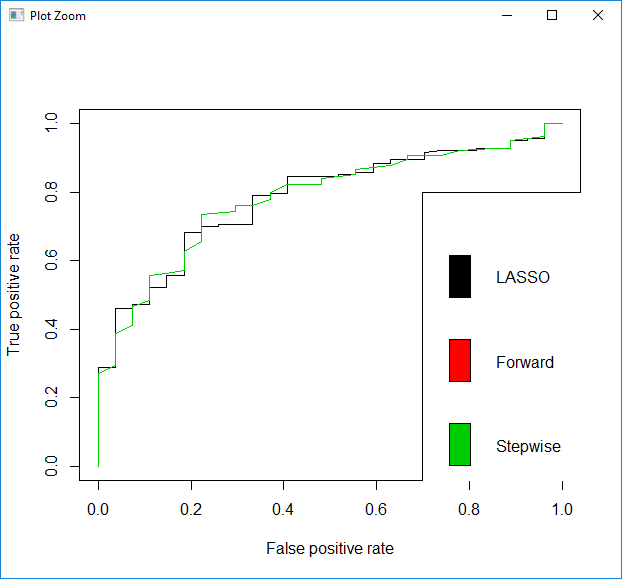
LASSO (least absolute shrinkage and selection operator) is a penalized regression model technique that is used to select a subset of variables. Using cross validation we computed a lambda of 0.009502046; the final LASSO model is:

6.406 – 0.111\*distance – 0.257\*homekick1 + 0.00029\*timeremqtr

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **AIC** | **AUC** | **Accuracy** | **Sensitivity** | **Specificity** |
| **Forward** | 554.48 | 0.782 | 0.8696 | 0.037 | 0.99 |
| **Stepwise** | 554.48 | 0.782 | 0.8696 | 0.037 | 0.99 |
| **LASSO** |  | 0.782 | 0.87 | 0.037 | 1.0 |

**[TABLE 4.d.1 – Initial Model Results]**

Diving into our models and looking at the confusion matrix, all of our models have great accuracy but poor sensitivity. Sensitivity is the true negative rate, when actually no, how often does it predict no? Our specificity is incredibly high. Specificity is the true positive rate, when it’s actually yes how often does it predict yes? To gain more insight, the ROC curves were investigated and the area under the curve for all models was above 75% (see **[CHART 4.d.1]** and AUC in **[TABLE 4.d.1]**). Note that the stepwise and forward selection’s ROC curves **[CHART 4.d.1]** are practically identical which is why you can’t see the ROC curve for forward selection.



**[CHART 4.d.1 – ROC Curves]**

The curves look very good but we need to choose a new cut point to fix the low sensitivity. After choosing a new cut point value, in **[TABLE 4.2]** below you will see that sensitivity and specificity are much better balanced. Although accuracy took a hit, it was necessary to have the model be decently good at predicting both the successful and missed field goals. AUC was used to be the deciding factor on which model was best. Based on this metric LASSO was chosen as the best model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **AIC** | **AUC** | **Accuracy** | **Sensitivity** | **Specificity** |
| **Forward** | 554.48 | 0.782 | 0.79 | 0.59 | 0.82 |
| **Stepwise** | 554.48 | 0.782 | 0.79 | 0.59 | 0.82 |
| **LASSO** |  | 0.787 | 0.8068 | 0.59 | 0.83889 |

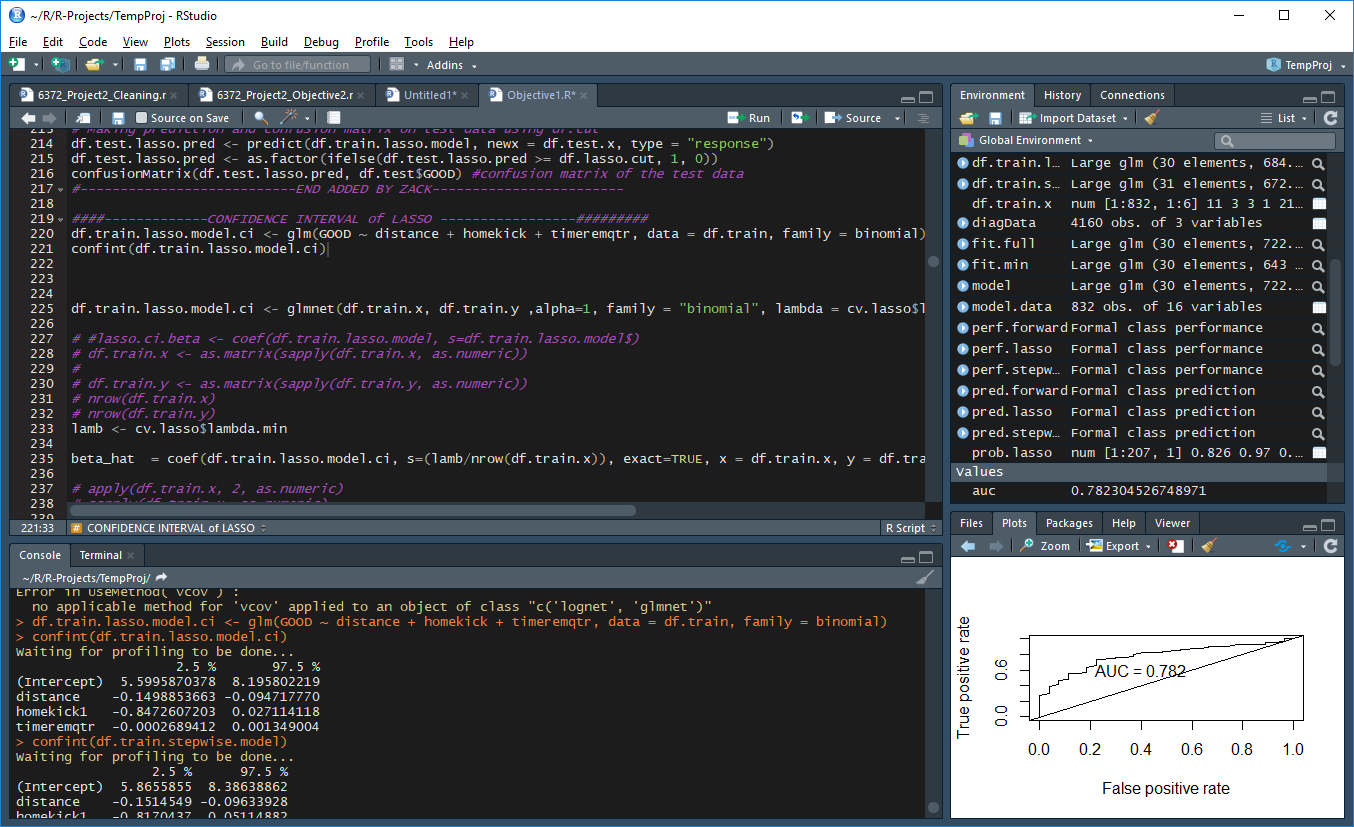
**[TABLE 4.d.2 – Model Results after cut point adjustment]**

1. **Interpretation**

Based on the above model metrics, the most accurately predicting model is LASSO. To generate the confidence intervals for interpretation the model was run using the R function glm with the variables from LASSO’s output. Because LASSO does not use a p-value for the decision of what parameter to select some of the confidence intervals from glm include zero.

The odds ratio for making a field goal for a 1 yard decrease in distance (with everything else held constant) is 1.117 with a 95% confidence interval of [1.099, 1.1617]. The odds ratio of making a field goal while away from home with everything else held constant is 1.293 times higher than when when at home (95% confidence interval [0.97, 2.33]). The odds ratio for making a field goal for a 1 second increase in time before the end of a quarter (with everything else held constant) is 1.00029 with a 95% confidence interval of [0.9997, 1.0013].

LASSO Model: 6.406 – 0.111\*distance – 0.257\*homekick1 + 0.00029\*timeremqtr

**Raw Confidence Intervals**

**Point estimates/CI**

Distance: e^(-0.111) = 0.8949 inverse is 1.117

e^(-0.149885) = 0.8608, e^(-0.0947) = 0.9096 inverse is [1.099, 1.1617]

Homekick1 e^(-0.257) = 0.7734 inverse is 1.293

e^(-0.84726) = 0.42858765, e^(0.0271) = 1.02747 inverse is [0.97, 2.33]

Timeremqtr e^(0.00029) = 1.00029

e^(-0.0002689) = 0.9997, e^(0.0013) = 1.0013

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

In objective 1, a simple logistic model was created to predict the success of NFL field goal kicks based on a set of predictors using 2008 NFL data set. Using this simple model as a baseline, we will create complex logistic model using interaction terms and will also investigate competing models such as LDA and Random Forest. To investigate of clean data set further we will use Principle Component Analysis to gleam insight into the explanatory variables of data set and too see if we have any separation of principle components by the response. If the responses separate out, then predictive such as LDA, logistic, and random forest typically will predict well.

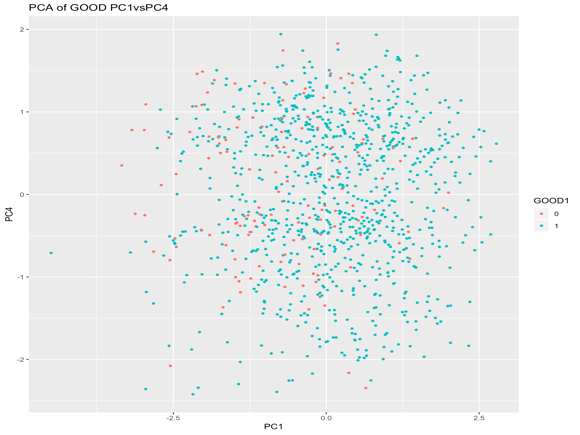
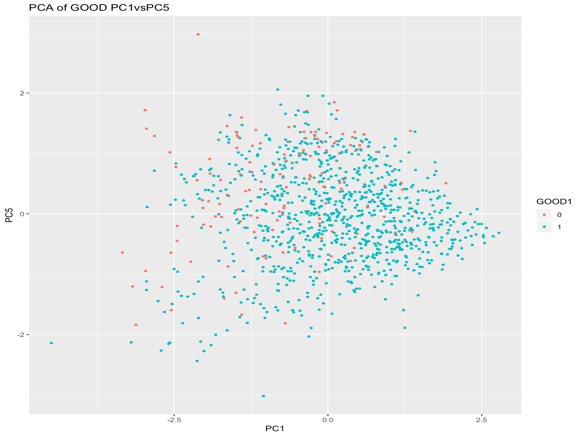
1. **Principle Component Analysis as Explanatory Analysis**

Principle Component Analysis was used as additional explanatory technique on our continuous variables of our clean data set. Why not us the entire data set? In objective 1 explanatory data analysis, we had already removed explanatory variables that we deem would not help us like the variable kicker. We wanted to make a truly predictive model, if kicker was included it would hamper our ability to predict the success of field goal’s because the model would not be able to be used for new kickers. Second, our initial data set also contained a lot categorical explanatory variables. Those are removed for this PCA because we only want to look at the continuous variables for Linear Discriminate Analysis. If you look at the table below all of our continuous variables are weighted differently through our principle components which seem promising of some separation.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **PC1** | **PC2** | **PC3** | **PC4** | **PC5** |
| **togo** | -0.64 | 0.3 | -0.04 | -0.07 | -0.7 |
| **distance** | -0.63 | 0.11 | -0.42 | 0.02 | 0.65 |
| **kickdiff** | -0.07 | -0.57 | -0.72 | 0.25 | -0.3 |
| **timerem** | 0.25 | 0.61 | -0.21 | 0.72 | -0.02 |
| **timeremqtr** | 0.36 | 0.43 | -0.51 | -0.64 | -0.03 |

**[TABLE 5.b.1 – PCA Results]**

Through the creation of scatter matrix of principle components (Appendix Fig #) coded by our response there are few of graphs that have a separation. PC1 vs PC5 and PC1 vs PC4 **[CHART 5.b.1]** shows the most separation of the response.



**[CHART 5.b.1 – PC1 vs PC5 and PC1 vs PC4]**

There are no distinct clouds but there seems to be some areas of concentration of the different responses that gives some confidence that LDA and random can give us a decent model that will predict well.

1. **Model Building Selection – Complex Logistic, LDA, and Random Forest**

For the complex logistic model we decided that because the three automatic selection techniques had very little difference between their outcomes, we would use stepwise selection with the prescribed criteria being AIC. In this model we included all parameters, all 2-way interactions, and the square of each parameter. As seen below this result is exactly the same as the ones produced in Objective 1 for the forward and stepwise selections.

7.0639 – 0.12\*distance – 0.37965\*homekick1

Linear discriminant analysis (LDA is a method used in machine learning to find a linear combination of features that separates two or more classes of objects or events. The resulting combination may be used as a linear classifier.

From our PCA explanatory work, the five continuous variables appear to create some separation between the responses of a good and bad field goal and LDA may be a good choice in prediction. Using LDA, the coefficients of linear discriminant output provides a linear combination our five continuous variables that will be used to form the LDA decision rule. In the table 4.5 are the coefficients for this data set.

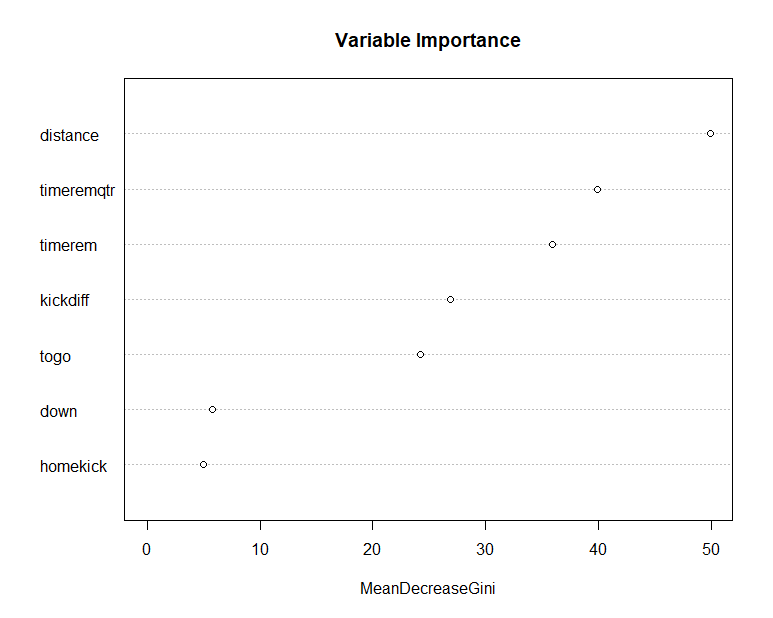
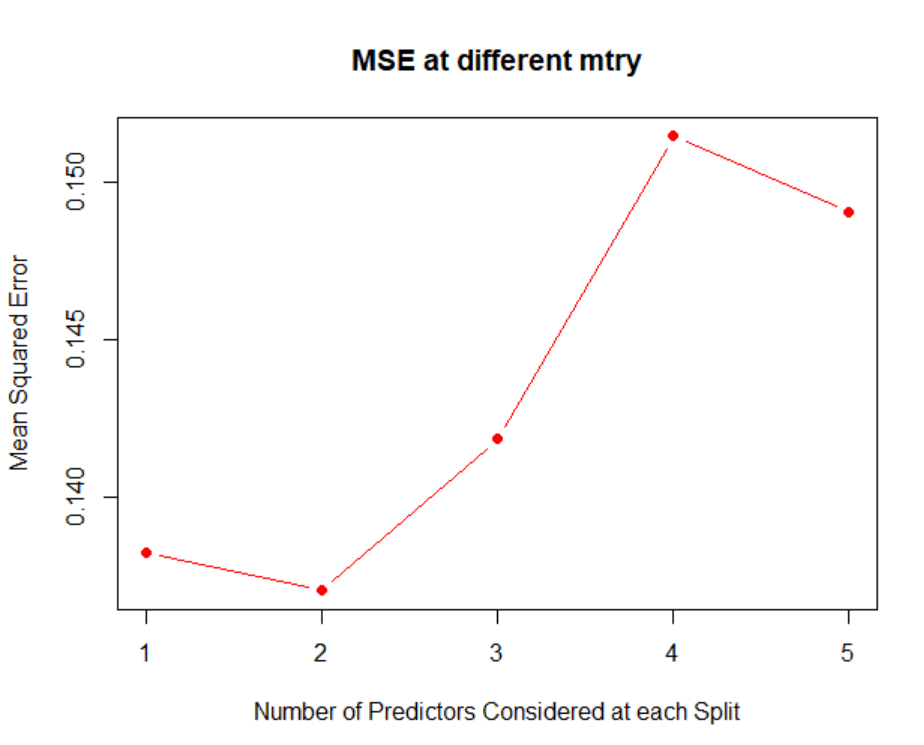
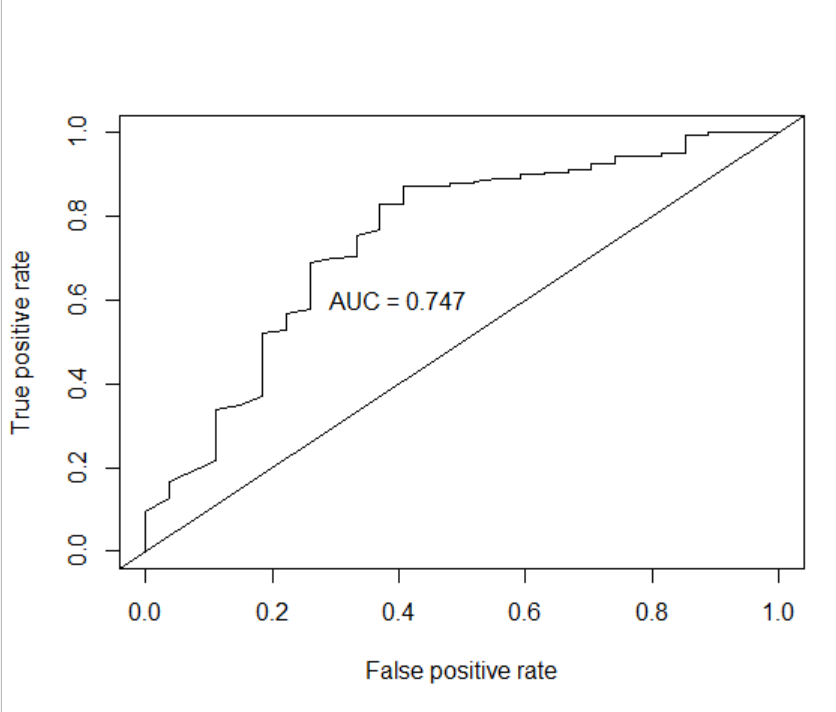
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variables** | **togo** | **distance** | **kickdiff** | **timerem** | **timeremqtr** |
| **LDA coeff** | 2.07e-2 | -1.07e-1 | -4.42e-3 | 2.35e-6 | 5.75e-4 |

**[TABLE 5.c.1 – Initial Model Results]**

If 2.07e-2\*togo -1.07e-1\*distance-4.42e-3\*kickdiff+2.35e-6\*timerem+5.75e-4\*timeremqtr is large then the LDA classifier will predict the field goal is good and if it is small then it will predict a missed field goal. When we look at our confusion matrix, the LDA classifier model has an accuracy of 78%. The specificity and sensitivity is 78% and 70%, respectively. Our LDA does not perform as well as objective 1 models in terms of accuracy but it does perform better in terms of specificity and sensitivity.

Random forest is a method for classification that operates by construction a multitude of decision trees. Random Forest attempts to decorrelate trees by random sampling from the predictors equally not dependent correlation weight of predictors.

Random forest was executed with all the variables to identify the importance of the variables, this model was executed with 400 trees, we found these variables distance, timeremqtr, timerem, kickdiff, togo to effectively reduce the Gini impurities for Variable Importance **[CHART 5.c.1]** to predict good and bad field goal response, we also tried with 500, 1000 trees and the variable importance remains the same. Now, random forest was again executed to find the OOB error at different number of variables available for splitting at each tree node (mtry), we found that the lowest mean square oob error to be 0.1394231 at mtry=2 [chart - OOB Error at different mtry]. ROC curve was ran important variables with 400 trees and probability predictions using the test data to find the best cut point from the ROC curve [chart - ROC] and found optimal cut point to be 0.88. With these optimal values random forest model was able to predict good and bad field goal with 69.57% accuracy with optimal sensitivity (0.74074) and specificity (0.68889).

**[CHART 5.c.1]**

**Model Results**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **AIC** | **AUC** | **Accuracy** | **Sensitivity** | **Specificity** |
| **LDA** |  | 0.775 | 0.7778 | 0.70370 | 0.78889 |
| **Complex** | 554.48 | 0.782 | 0.79 | 0.59 | 0.82 |
| **Random Forest** |  | 0.747 | 0.6957 | 0.74074 | 0.68889 |

**[TABLE 5.c.2 – Model Results after cut point adjustment]**

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)

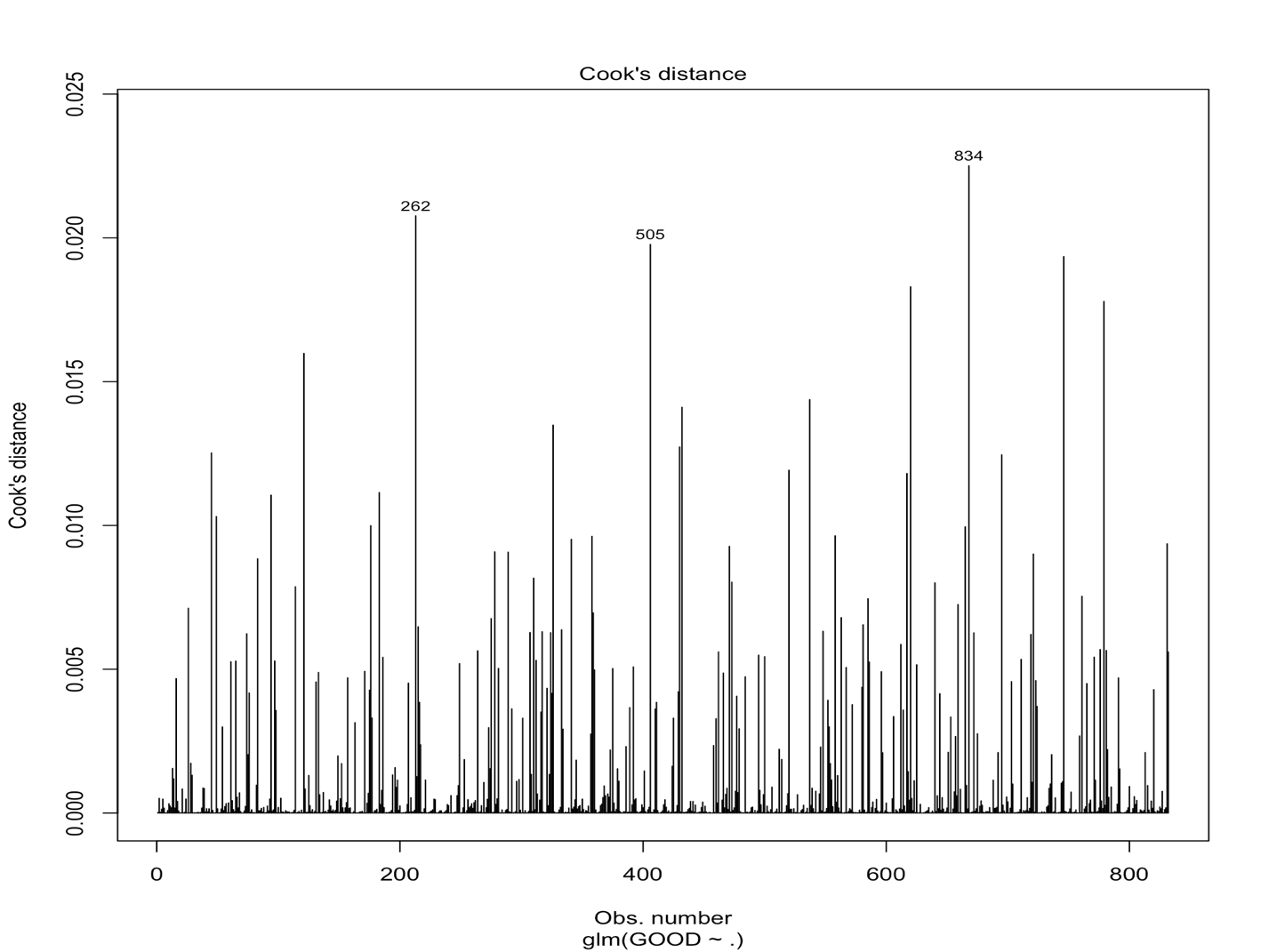
Blocked (1 if Blocked, 0 ow)

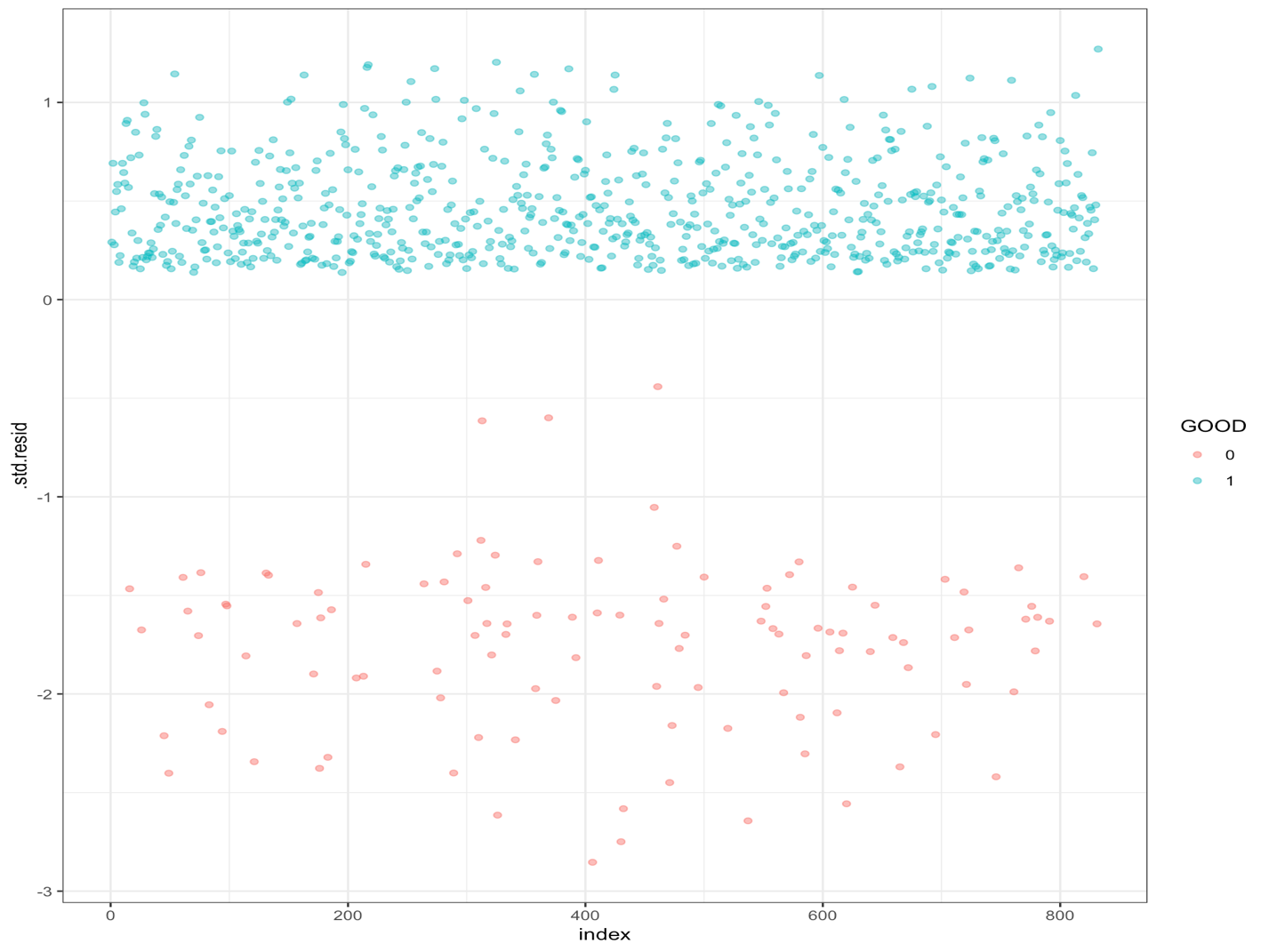
**Data Cleansing:** In our data set we found NA’s in our data set for two games. With us having the games date and the 20078 schedule, we were able to obtain the box score information and stats for each game.

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** |

