Predicting NFL Field Goal Outcome

Quang Nguyen

April 29, 2021

Abstract: This paper investigates different methods of predicting the success of field goals in the National Football League. In particular, we consider three statistical learning techniques: logistic regression, classification tree, and random forest. It turns out that random forest gave us the highest prediction accuracy for field goal outcome, followed by logistic regression and decision tree. Results indicate that field goal distance is the most important variable in predicting field goal success for NFL kickers.

1. Introduction

Football is certainly one of, if not the most popular sport in the United States; and it also plays an important role in the American culture. The National Football League (NFL) is the top-tier football league in the country, and it is home to some of the most financially valuable and famous sports franchises and athletes in the world. The most important goal for every team, coach, and players in the NFL is to win the game; and in order to win, it is clear and simple that you must score more than your opponent. There are multiple ways to score in football, with touchdown being the most commonly known form of scoring. Alongside with touchdown, most of the scoring in football is done by field goals. Scoring a field goal could significantly impact the outcome of a football game, as we often see a field goal end up clinching a game victory, or even more significant - a Super Bowl title, for an NFL team.

There are many factors affecting the outcome of a field goal attempt in football. The result of a field goal certainly depends on roster-related factors such as having a good ball snapper, holder, and most importantly, a reliable kicker. In addition, there are in-game aspects such as whether or not the team is leading, or the time of the game when a team attempts the kick (for instance, late in-game). Moreover, coaching decisions could have an impact on field goal success, as in tight late game situations, coaches of the opposing team often call a timeout and implement the strategy of "icing the kicker", which has proven to be successful quite often. There are also factors related to playing condition, such as the field surface, or how cold and windy the weather is. In this paper, we examine different statistical learning methods of predicting field goal success such as logistic regression, classification trees, and random forest. In particular, the response variable - field goal outcome - will be modeled using various explanatory variables.

2. Data and Methods

For this project, the data was obtained from a GitHub repository owned and maintained by Michael Lopez, who is the current Director of Football Data and Analytics at the National Football League. This publicly available dataset consists of all field goal attempts from 2005 to 2015, with the following attributes for each observation: the NFL team taking the field goal (`Team`); the season (`Year`); the game minute at which the field goal occurred (`GameMinute`), the name of the kicker (`Kicker`), the distance of the field goal from the opposing team's endzone (`Distance`); the difference between the scores of the offense and the defense (`ScoreDiff`); whether or not the surface is grass (`Grass`); the outside temperature when the game happened (`Temp`); and most importantly, whether or not the kicker made the field goal (`Success`). From the given information, we created the following new features: whether or not the kicking team is leading (`Leading`); the period in the game (first, second, third, fourth quarters, overtime) when the field goal took place (`Period`); the type of score (`ScoreType`) - one possession (when the score difference is 8 points or less), two possession (when the score difference is 16 points or less), and three or more possession (when the deficit is more than 16 points); and the dominant foot of the kicker (`Foot`).

Since our goal was to build predictive models for field goal success, we first created a holdout data set of observations from the 2015 season, then utilized the data from prior seasons from 2005 to 2014 to fit different models and examine how well they predict field goal scoring in 2015. Our first model for predicting the probability of a field goal success is a logistic regression model, and we determined the best model by performing stepwise backward selection using AIC to obtain the best subset of explanatory variables. We also looked into two other classification techniques to predict kicking success in the NFL - classification tree and random forest. For both methods, we performed cross-validation by splitting the

train data set (data from 2005 to 2014) using a 75-25 train-validation split. After training and validating the models, we fitted a final model for each method and made predictions on the 2015 field goals data set.

3. Results

From our first approach of logistic regression, we found that the best subset of predictors determined from stepwise backward selection using AIC consists of `Distance`, `Grass`. We started with nine variables: `GameMinute`, `Distance`, `ScoreDiff`, `Grass`, `Success`, `Leading`, `Period`, `ScoreType`, and `Foot`; and were able to narrow down the subset of predictors to just two using stepwise selection (Figure 1). The logistic regression model for predicting field goal outcome is $logit(\hat{\pi}) = 5.7776 - 0.1024\ Distance - 0.1624\ Grass$. According to this model's coefficients table (Table 1), every extra yard in distance multiplies the odds of successfully converting a field goal by 0.9027, after accounting for field surface. Hence, a longer distance is associated with lower chance of a field goal success, which is not surprising. Interestingly, the chance of success for kicks attempted on grass surface is lower than that for non-grass surfaces, given the kick distance, as the odds of a field goal success on grass is estimated to be 0.8501 times the odds of a field goal success on non-grass surfaces.

In terms of prediction, using the fitted logistic model, we obtained the predicted outcomes of the field goals in our holdout set - that is, for the field goals attempted in the 2015 season. The result is quite good, as we accurately predicted 84.90% of the field goal outcomes in 2015 (Table 2). For our classification tree model, Figure 2 shows the decision tree with information regarding the nodes and branches of the tree. We noticed that `Distance`, `Period`, and `GameMinute` were the only variables that got selected to be involved in our tree. We then used this tree model to predict outcomes for field goals in 2015, and get a 84.70% accuracy (Table 3), which is a little less accurate than what the logistic regression model gave us. Lastly, the random forest model gave us 85.29% accuracy (Table 4), which is the best prediction we got out of the three techniques. From the plots of variable importance measured by our random forest (Figure 3), it is obvious that `Distance` plays the most important role in contributing to the random forest model compared to other covariates.

4. Conclusion and Discussion

Overall, we utilized different statistical learning methods, namely, logistic regression, classification tree, and random forest, to predict what would happen to field goals attempted in the 2015 NFL season from prior data. Using a logistic regression model gave us a very good prediction accuracy (84.90%), and we were able to improve our prediction rate by using a random forest model (85.29%) but not a classification tree (84.70%). We notice that field goal distance is highly involved in all three of our methods, especially random forest, hence it has a huge influence on predicting field goal outcome in football.

In the future, it is totally possible for us to improve our prediction accuracy of field goal kicking outcome in the NFL. We could try to add more variables of that our dataset did not have, with features like the playing side of the kicking team (home or away), whether or not a timeout was called before the kick by the opponent, whether a penalty was called resulting in more yards being added, or more information on weather condition such as wind direction and wind speed at the time the kick was attempted. With more data available in the form of play-by-play, thanks to the creation of packages such as `nflscrapR` and `nflfastR`, and the establishment of the annual NFL Big Data Bowl competition, we could easily extract the desired information from the data provided by those excellent resources. In addition, instead of viewing our response - kick outcome - as a binary variable (success or failure), we could look at it as a multinomial outcome variable with levels like success, blocked kick, and non-blocked field goal miss. Another thing we should look into is to use other classification techniques such as k-nearest neighbors, support vector machine, or naive Bayes to predict scoring outcome for NFL kickers and compare how well they perform compared to the algorithms that we implemented in this paper.

References

Agresti, A. (2019). *An Introduction to Categorical Data Analysis.* Hoboken, NJ: John Wiley & Sons.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R.* New York: Springer.

Appendix

Table 1: Coefficients table for logistic regression fit obtained from stepwise selection.

	Estimate	OR	95% CI	Std. Error	z value	Pr(>|z|)
(Intercept)	5.7776	322.9859	(242.9107,432.2403)	0.1470	39.3052	0.0000
Distance	-0.1024	0.9027	(0.8969,0.9085)	0.0033	-31.2961	0.0000
GrassYes	-0.1624	0.8501	(0.7597,0.9508)	0.0572	-2.8385	0.0045

Table 2: Classification table for predictions of field goal kicking outcome using logistic regression. The diagonal elements of the table indicate correct predictions, while the off-diagonals represent incorrect predictions. The overall proportion of correct predictions was $(10 + 867)/1033 = 0.8490$.

	Actual	
Predicted	0	1
0	10	10
1	146	867

Table 3: Classification table for predictions of field goal kicking outcome using classification tree. The overall proportion of correct predictions was $(7 + 868)/1033 = 0.8470$.

	Actual	
Predicted	0	1
0	7	9
1	149	868

Table 4: Classification table for predictions of field goal kicking outcome using random forest. The overall proportion of correct predictions was $(10 + 871)/1033 = 0.8529$.

	Actual	
Predicted	0	1
0	10	6
1	146	871

```
Start:  AIC=7996.94
Success ~ (Team + Year + GameMinute + Kicker + Distance + ScoreDiff + Grass + Leading +
    Period + ScoreType + Foot) - Year - Team - Kicker
            Df Deviance    AIC
- Period     4   7977.1 7995.1
- Foot       1   7971.2 7995.2
- ScoreDiff  1   7971.6 7995.6
- GameMinute 1   7971.8 7995.8
- Leading    1   7971.9 7995.9
- ScoreType  2   7974.4 7996.4
<none>           7970.9 7996.9
- Grass      1   7979.1 8003.1
- Distance   1   9182.1 9206.1

Step:  AIC=7995.13
Success ~ GameMinute + Distance + ScoreDiff + Grass + Leading + ScoreType + Foot
            Df Deviance    AIC
- ScoreDiff  1   7977.3 7993.3
- Foot       1   7977.4 7993.4
- Leading    1   7977.5 7993.5
- GameMinute 1   7977.8 7993.8
- ScoreType  2   7980.4 7994.4
<none>           7977.1 7995.1
- Grass      1   7985.4 8001.4
- Distance   1   9209.1 9225.1

Step:  AIC=7993.27
Success ~ GameMinute + Distance + Grass + Leading + ScoreType + Foot
            Df Deviance    AIC
- Leading    1   7977.5 7991.5
- Foot       1   7977.6 7991.6
- GameMinute 1   7977.9 7991.9
- ScoreType  2   7980.8 7992.8
<none>           7977.3 7993.3
- Grass      1   7985.5 7999.5
- Distance   1   9209.2 9223.2

Step:  AIC=7991.47
Success ~ GameMinute + Distance + Grass + ScoreType + Foot
            Df Deviance    AIC
- Foot       1   7977.8 7989.8
- GameMinute 1   7978.2 7990.2
- ScoreType  2   7980.9 7990.9
<none>           7977.5 7991.5
- Grass      1   7985.7 7997.7
- Distance   1   9214.9 9226.9

Step:  AIC=7989.75
Success ~ GameMinute + Distance + Grass + ScoreType
            Df Deviance    AIC
- GameMinute 1   7978.5 7988.5
- ScoreType  2   7981.2 7989.2
<none>           7977.8 7989.8
```

```
- Grass       1   7985.7 7995.7
- Distance    1   9215.4 9225.4

Step:  AIC=7988.51
Success ~ Distance + Grass + ScoreType
            Df Deviance    AIC
- ScoreType  2   7982.0 7988.0
<none>           7978.5 7988.5
- Grass      1   7986.4 7994.4
- Distance   1   9215.4 9223.4

Step:  AIC=7988.02
Success ~ Distance + Grass
           Df Deviance    AIC
<none>          7982.0 7988.0
- Grass     1   7990.1 7994.1
- Distance  1   9218.9 9222.9
```

Figure 1: Stepwise backward selection using AIC output. Our model selection process begins with all of our potential explanatory variables being included as main effects. At each step, we remove the variable so that AIC decreases the most, until we get to the stage in which AIC increases if we remove any other variables.

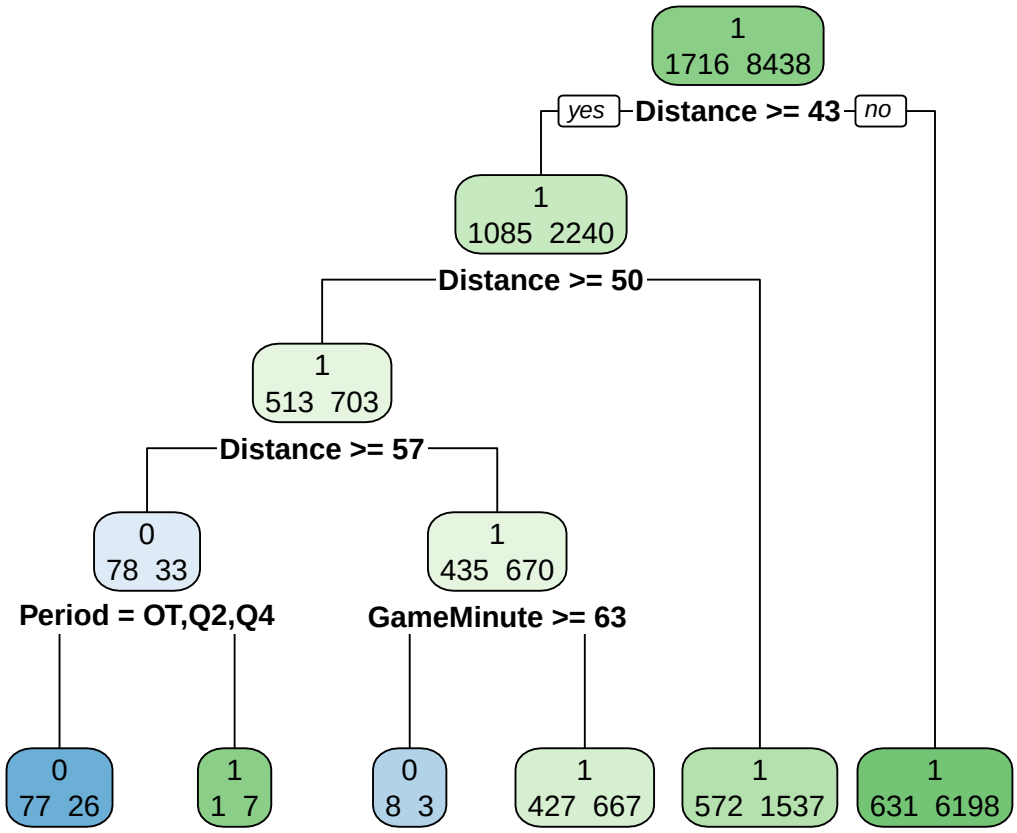

Figure 2: Classification tree for predicting field goal success, based on the field goal distance, game minute, and playing period. The classification tree summarizes responses to five questions with binary outcomes, with the "yes" response going to the left branch. Each node in the tree illustrates the decision question, together with the counts for each response.

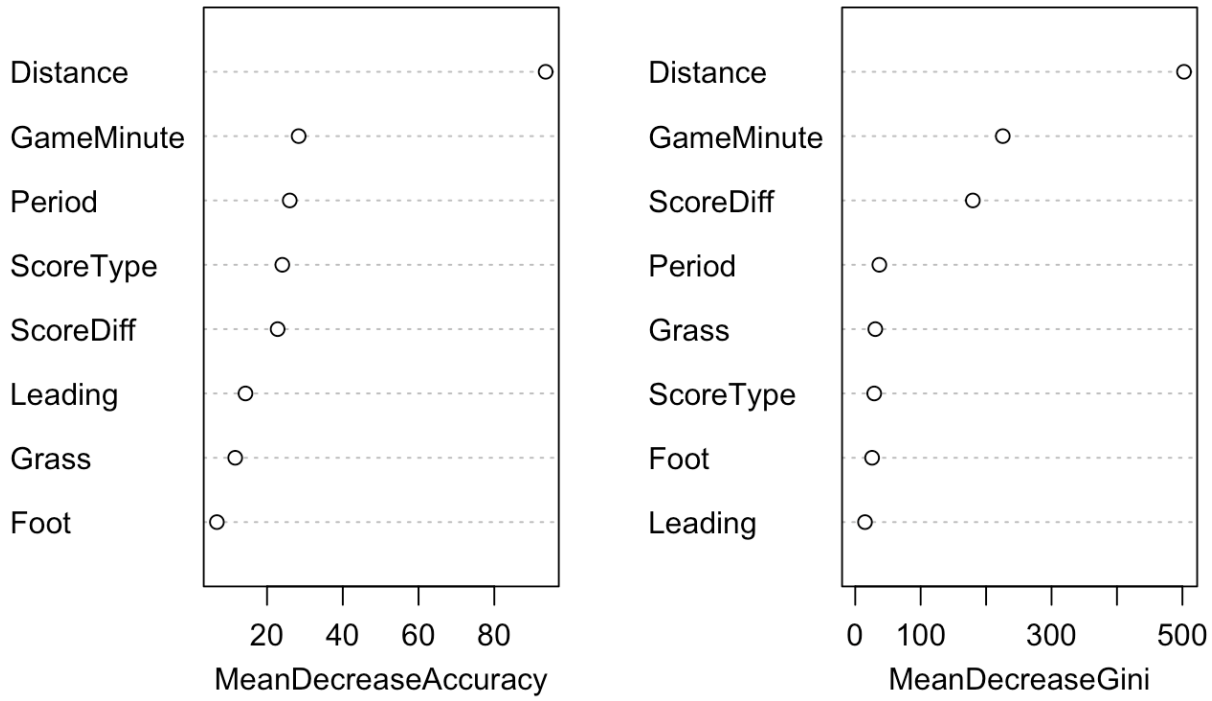

Figure 3: Variable importance plots for NFL kicking data. Variable importance is computed using the mean decrease accuracy, which expresses how much accuracy the model losses by excluding each variable, and the mean decrease Gini index, which measures how each variable contributes to the homogeneity of the nodes and leaves in the random forest. A higher mean decrease accuracy or mean decrease Gini score indicates higher importance of the variable in the model.