**Predictive Modeling for Success on Third Down**

Dan Goldhamer

December 5, 2018

**Executive Summary**

Logistic regression, decision trees and neural networks were all used to create predictive models exploring the variables that contribute to gaining a first down or scoring a touchdown when it is third down in a National Football League (NFL) game. The data set used contained every play from the 2009 through the 2017 NFL season. Decision trees performed the best when judged on misclassification rates while logistic regression provided some insight into the effect of the different variables. Overall, the yards needed to gain a first down or touchdown was the strongest predictor of success on third down. Improvements to the model could occur if plays on first and second down were also incorporated into the model as well as additional variables.

**Business problem/opportunity**

The National Football League (NFL) is extremely competitive both off and on the field. The NFL is also big buisness with the revenues surpassing $8 billion dollars in 2017 (Field Level Media 2018).

One differentiating characteristic of the NFL compared to the other American professional sports is the limited number of games played in the regular season. Each team in the NFL only plays 16 regular season games. While the National Hockey League (NHL) and the National Basketball Association (NBA) both play 82 regular season games and Major League Baseball (MLB) plays 162 regular season games. This means that each win for an NFL team contributes a greater percentage to a team’s overall record.

In the NFL there is a total of 32 teams. They each play 16 games each in the regular season and there are a total of 22 games played in the playoffs. That equals a total of 534 games played in a season (not including the post season). Assuming a revenue of $8 billion dollars, that means each game on average brings in $14,981,273 in revenue. If a team makes it to the playoffs they can play up to 4 additional games which would equal to almost $60 million dollars in revenue they would not have received if they did not make it to the playoffs. Clearly teams should pursue an analytical advantage to gain insight on what contributes to these precious wins.

**Specific business objective**

There is an adage in football that the games are won and lost on third down. The objective of this project is to use predictive modeling to explore factors that contribute to a team succeeding on third down as defined by gaining the necessary yards to earn a fresh set of downs or scoring a touchdown.

**Process followed for selecting and gathering data**

The data was provided through the Denver Broncos analytics competition. The original data set contained every play of every NFL game from the 2009 season through the 2017 season. This 215 MB set originally contained 407,689 cases and 102 variables. However, many of these variables would not be appropriate to include in constructing a predictive model. In fact, 21 of those variables were output variables that came from other predictive models.

In order to select the most appropriate variables, all variables that occurred during a play or were calculated as result of the play except for gaining a first down or a touchdown, were eliminated. Descriptive variables such as player name and player ID were also eliminated. In the end, 9 variables from the original data set were selected. Those variables were:

1. TimeSec- Time remaining in game in seconds
2. Yrdline100-Distance to opponents endzone, ranges from 1-99 situation
3. Ydstogo- Yards to go for a first down
4. posteam- The team on offense
5. HomeTeam- The home team
6. FirstDown - Binary: 1if the play resulted in a first down conversion else 0
7. Touchdown- Binary: 1 if the play resulted in a TD else 0
8. PlayType- The type of play that occurred, potential values are: Kickoff, Punt, Pass, Sack, Run, Field Goal, Extra Point, Quarter End, Two Minute Warning, Half End, End of Game, No Play, QB Kneel, Spike, Timeout
9. ScoreDiff- The difference in score between the offensive and defensive teams (offensive.score - def.score)

**Description of Data Reparation**

**Data preparation**

All of the data preparation was performed with Microsoft’s Excel. The first step in data preparation was to filter for only the third down plays using the “down” variable. This reduced the data set down to a total of 67,398 cases. The data was again filtered on the variable “PlayType” for the values of Run, Pass and Sack. This reduced the number of cases to 61,067. This was done to eliminate cases where plays did not occur such as penalties or should not be included in the model such as intentional kneel downs. These 61,067 cases are what was used for predictive modeling.

**Data Repairs**

The data set was checked for any missing values using SAS Enterprise Miner and the explore function. There was no missing data for any of the variables. So no data repairs were needed.

**Replacements**

No data replacements occurred before analysis began.

**Transformation/Reduction**

The first transformation performed was to change all cases with the value of PlayType equal to Sack to Pass for all cases. This was done because sacks only occur on pass plays and result in not achieving a first down or touchdown. There were 4,200 plays coded as sack that were changed to pass.

The second transformation performed was to create a new variable IsHomeTeam. The purpose of this transformation was to create a binary variable indicating if the offensive team was also the home team. This was performed by using the variables “posteam” and “HomeTeam”. If the two variables equaled each other the value of IsHomeTeam is 1 and if they were unequal the value of IsHomeTeam would be 0.

The third transformation performed was to create the TargetSuccess variable. The TargetSuccess variable indicates if the play results in a first down or touchdown with a value of 1. A value of 0 is assigned if the play did not result in a first down or touchdown.

The fourth transformation performed was to create a new categorical variable from “Yrdline100”. This new variable was named “SideOfField”. For this transformation, domain knowledge was used to divide the playing field into three sections. The first category was for values of 1-20 and received the category of “Own”. This group is represents playing close to your own endzone with a majority of the field in front of the offense. The values of 21-79 were received the category of “Middle”. These values represent the offense playing in the middle of the field where it can be easier to gain yards. The values of 79-99 received the category of “Red Zone”. This area of the field is typically known as the Red Zone where the offensive team is close to scoring but since the field is smaller, it can be more difficult to gain yards since the defensive team has to cover less of a playing area. Table 1 Has the breakdown of the percentage of cases that fit into each of the bins.

The fifth transformation performed was to create a new categorical variable from Ydstogo. This new variable was named “CatYdsToGo”. By looking at a histogram of Ydstogo, it was determined that the distribution of Ydstogo was skewed towards the right. The 5 bins for the new variable CatYdsToGo were created by trying to divide all the cases as evenly as possible into the 5 bins. The value for the short bin was if Ydstogo was less than or equal to 3. The value for the manageable bin was if Ydstogo was less than or equal to 5 but greater than 3. The value for the medium bin was if Ydstogo was less than or equal to 8 but greater than 5. The value for the long bin was if Ydstogo was less than or equal to 10 but greater than 8. The value of the extra long bin was if Ydstogo was greater than 10. Table 2 Has the breakdown of the percentage of cases that fit into each of the bins.

**Final Variables and Data Sets**

In the end, there were 6 predictor variables selected for the single target variable. The 6 predictor variables were:

1. TimeSec
2. PlayType
3. ScoreDiff
4. IsHomeTeam
5. Yrdline100 or SideOfField
6. Ydstogo or CatYdsToGo.

The single target variable was TargetSuccess.

In order to examine the effect of including SideOfField and CatYdsToGo into the model, 4 different data sets were used and all models were run on each of the data sets. The first data sets was called “No Modification” and used both Yrdline100 and Ydstogo. The second data set was called “CatYdsToGo” and used Yrdline100 and CatYdsToGo. The third data set was called “SideOfField” and used SideOfField and Ydstogo. The fourth data set was called “CatYdsToGo and SideOfField” and used both CatYdsToGo and SideOfField.

**Partitions**

All four of the data sets were partitioned the same way. They were partitioned with 60% of the data used for training and 40% of the data used for validation. The seed was set to 12345.

**Discussion of preliminary data exploration and findings**

The first data exploration task was to look at the TargetSuccess variable. In the entire data set, there were 36,776 cases where the value of TargetSuccess was 0 (60.22%) and 24,291 cases where the value was 1 (39.78%). When the data is further broken down by TargetSuccess and PlayType, the results are that when TargetSuccess is 0, 81.86% of those plays are passes while 18.04% of those plays are run. When TargetSuccess is 1, 75.29% of the plays are passes and 24.71% of the plays are runs. This seems to suggest that runs are more successful at gaining first downs or touchdowns. The issue is that it is more likely for a team to run the ball when the yards needed for a first down or touch down are small.

This led to the exploration of Ydstogo and the creation of the CatYdsToGo variable. The average Ydstogo when TargetSuccess was equal to zero was 8.26 with a standard deviation of 5.37 yards. The average Ydstogo when TargetSuccess was equal to one was 5.50 with a standard deviation of 3.94 yards. It was difficult to discern a precise pattern when Ydstogo was compared to TargetSuccess but there was a clear trend of lower percentage of plays achieving success as the Ydstogo variable increases. This trend is much clearer to see when the CatYdsToGo variable is used as a replacement. Table 3 shows the results of TargetSuccess when broken down by CatYdsToGo. From table 3 it can be inferred that the shorter the distance needed for success the higher percentage of plays fall into the success category. Likewise, as distance needed for success increases, the higher percentage of plays falls into the unsuccessful category.

Table 4 shows the results of when TargetSuccess is broken down both by PlayType and CatYdsToGo. What is interesting about the results of table 4 is that only on short distances are there relatively equal proportions of successful pass and run plays. At any greater distance, pass plays would appear more successful in achieving success.

This was the end of data exploration and predictive modeling was next employed to discover additional insight.

**Description of data modeling/analyses and assessments**

**Logistic regression**

The first model ran was a stepwise logistic regression. Stepwise logistic regression was chosen because it would allow variables to enter and exit and hopefully give highest probability of finding meaningful variables and a useful model.

First the stepwise logistic regression model was run with all the default settings intact except for selecting for a stepwise regression. A second model was run with the entry significance level to set to 1, the stay significance set to level to .5 and the maximum number of steps to set 10. Both model setups returned the exact same model.

When comparing the 4 different data sets, all of the logistic regressions used all 6 variables. The best performing model was the SideOfField data set. This data set had a misclassification rate of 0.362821. The worst performing model was the CatYdsToGo data set with a misclassification rate of 0.365359. This is not a large difference, but the results from the SideOfField data set will be used to describe the findings of the logistic regression.

All variables possessed a p-value under 0.04. P-values are displayed in table 5.

For the odds ratio estimates all variables besides TimeSec had a value not equal to 1. Odds ratio estimates are displayed in table 6.

There are some useful insights to be gained from these odds ratio estimates. First off, there does seem to be a home team advantage on third downs. Since the odds ratio estimate is below one for the away team, the chances of them achieving a first down or touchdown on third down are lower. This makes sense since often the home crowds will cheer loudly to disrupt the away team’s offence and will be quiet for the home team on third down to allow for effective communication. It also looks like passing the ball versus running the ball results in a lower probability of achieving success on third down. This could be due to the risk of sacks and/or the need to gain more yards thus necessitating the desire to pass the ball. This point is driven home by looking at Ydstogo. The higher the yards needed on third down, the lower the probability of success. This is particularly interesting because of the units of Ydstogo. With every increase in yard needed on third down, the probability of success decreases 88%. One can see how anything greater than 3 yards needed quickly becomes a low probability.

The only other variable with a greater absolute value than Ydstogo, is SideofField Own vs Red with being in your own end of the field having a lower probability of success than being in the Red Zone. This could be attributed to the higher probability of scoring a touch down in the Red Zone compared to the rest of the field. This is contradictory compared to SideofField Middle vs Red where being in the Red Zone means having a lower probability of success. One could guess that it is easier to get first downs in the middle of the field compared to the Red Zone because there is more playing field to use and more play call options. Football analysts often talk about how the offense gets compressed in the Red Zone and how it becomes more difficult to move the ball. The data seems to support this conclusion.

Lastly, as the ScoreDiff increases, the probability of getting a first down or touchdown decreases. It is unclear why this is but maybe teams with score advantage play more conservatively and do not take as many gambles to gain first downs or touchdowns. This could be especially true when ScoreDiff is large and the game is not in danger of being lost.

**Decision Trees**

For the decision trees, 2 3 and 5 branch trees were examined. Those branching levels were chosen to give the model the flexibility to select all the different levels of the different data sets. Misclassification rate as the selection criteria and set the maximum depth to 10 were selected for all the models.

The best performing tree was the 5 branch tree with the CatYdsToGo data set and is displayed in figure 1. This model had a misclassification rate of 0.355022. The worst performing model was the 2 branch tree with the CatYdsToGo and SideOfField data set with a misclassification rate of 0.36148.

Unfortunately, this tree does not provide significant valuable insight. The tree first breaks down the cases into the 5 categories of yards to go and then uses TimeSec for the remaining splits. What this tree can tell us is that the probability of achieving success on third down decreases as yards to go increases.

The 5 branch tree with the No Modification data set is the tree with the lowest misclassification rate that incorporates variables that are not Ydstogo and TimeSec. This tree has a misclassification rate of 0.356232. This tree incorporates PlayType and Yrdline100 at the third branch depth. This tree is displayed in figure 2.

**Neural Networks**

For neural networks, both an AutoNeural and standard Neural Network node were run on all 4 data sets. For the standards neural network, preliminary training was disabled and the maximum iterations was changed from 0 to 100. For the AutoNeural Network, the number of hidden units was changed to 1, Select Tolerance to was changed to low, Select Direct was changed to to no, Select Normal was changed to no and lastly, Select Sine was changed to no.

The neural network with the lowest misclassification rate was the standard Neural Network with the SideOfField data set. The misclassification rate for this model was 0.36147. The worst performing neural network was the standard neural network with the CatYdsToGo and SideOfField. The misclassification rate for this model was 0.365073.

The best performing neural network model was optimized at 34 iteration. The final weights are displayed in figure 3.

**Explanation of model comparisons and model selection**

All models were run through the SAS Model comparison node. The model with the lowest misclassification rate at 0.3550 was 5 Max Branch Misclassification Tree with the CatYdsToGo data set. This model was chosen as the champion model.

The model with the highest misclassification rate at 0.3654 was the stepwise logistic regression with the CatYdsToGo data set.

All model misclassification rates are shown in table 7. There is not a large spread between the different models in their misclassification rates. So all models performed relatively the same.

The cumulative lift chart show in figure 3 shows a similar trend in that there is not an easily observable separation between the models.

Overall, the massification trees with greater than 2 maximum branches seemed to perform the best and provide the most tangible insight on what a coach should do in different situations.

**Conclusions and recommendations**

Overall, this analysis showed the importance of having the lowest possible yards needed for success. Both the decision tree and logistic regressions models were highly influenced by Ydstogo or CatYdsToGo. This means that NFL teams should design their offences and play calling around the idea of keeping the distances needed for success on third down less than 3 yards.

There is room for further analysis and improvement to the model. Ideally, it would be beneficial to incorporate the plays on first and second into the model to help teams decide on the optimal play calls in different situations. One could also add in other variables such as temperature, wind speed, and precipitation to see if the model could be improved.

The NFL has been criticized for it slower adoptions of analytics compared to the other professional sports leagues in the United States (Breer 2017). However, there is a clear value in using analytics in American football as the Philadelphia Eagles demonstrated last year with their Super Bowl victory (McManus 2018).

**Appendix**

|  |  |
| --- | --- |
| Bin Name | Percentage |
| Own | 15.30 |
| Middle | 72.84 |
| Red Zone | 11.86 |
| Table 1. Percentage of cases per bin of SideOfField | |

|  |  |
| --- | --- |
| Bin Name | Percentage |
| Short | 26.72 |
| Manageable | 16.43 |
| Medium | 22.50 |
| Long | 15.40 |
| Extra Long | 18.95 |
| Table 2. Percentage of cases in each bin for CatYdsToGo | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| TargetSuccess | Short | Manageable | Medium | Long | ExtraLong |
| 0 | 11.74% | 8.97% | 13.69% | 10.62% | 15.20% |
| 1 | 14.98% | 7.74% | 8.81% | 4.78% | 3.74% |
| Table 3. Percent of all cases in each bin of CatYdsToGo by TargetSuccess | | | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| TargetSuccess and PlayType | Short | Manageable | Medium | Long | ExtraLong |
|  |  |  |  |  |  |
| TargetSuccess=0 |  |  |  |  |  |
| Pass | 7.16% | 7.70% | 12.19% | 9.55% | 12.76% |
| Run | 4.58% | 1.27% | 1.5% | 1.07% | 2.45% |
|  |  |  |  |  |  |
| TargetSuccess=1 |  |  |  |  |  |
| Pass | 7.63% | 6.58% | 7.96% | 4.37% | 3.40% |
| Run | 7.35% | 0.88% | 0.85% | 0.40% | 0.34% |
| Table 4. Total percent of cases per bin of CatYdsToGo by TargetSuccess and PlayType | | | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| Effect | DF | Chi-Squar | Pr > ChiSq |
| IsHomeTeam | 1 | 4.4197 | 0.0355 |
| PlayType | 1 | 7.1898 | 0.0073 |
| ScoreDiff | 1 | 15.0390 | 0.0001 |
| SideOfField | 2 | 58.5789 | <.0001 |
| TimeSec | 1 | 50.5058 | <.0001 |
| Ydstogo | 1 | 2096.3986 | <.0001 |
| Table 5. P-values stepwise logistic regression on the SideOfField data set | | | |

|  |  |
| --- | --- |
| Effect | Point Estimate |
| IsHomeTeam 0 vs 1 | 0.954 |
| PlayType Pass Vs Run | 0.926 |
| ScoreDiff | 0.996 |
| SideOfField Middle vs Red | 1.073 |
| SideOfField Own vs Red | 0.847 |
| TimeSec | 1 |
| Ydstogo | 0.88 |
| Table 6. Odds ratio estimates for the stepwise regression on the CatYdsToGo data set | |

|  |
| --- |
|  |
| Figure 1. Best performing 5 branching misclassification tree on the CatYdsToGo data set |

|  |
| --- |
|  |
| Figure 2. The 5 branching misclassification tree on No Modification data set |

|  |
| --- |
|  |
| Figure 3. Weights of the best performing neural network on the SideOfField data set |

|  |  |  |
| --- | --- | --- |
| Data Set | Model Description | Misclassification Rate |
| CatYdsToGo | 5 Max Branch Misclassification Tree | 0.355002456 |
| CatYdsToGo and SideOfField | 5 Max Branch Misclassification Tree | 0.355002456 |
| No Modification | 5 Max Branch Misclassification Tree | 0.356312428 |
| CatYdsToGo | 3 Max Branch Misclassification Tree | 0.356762731 |
| CatYdsToGo and SideOfField | 3 Max Branch Misclassification Tree | 0.357294907 |
| SideOfField | 3 Max Branch Misclassification Tree | 0.357540527 |
| No Modification | 3 Max Branch Misclassification Tree | 0.358563943 |
| SideOfField | 5 Max Branch Misclassification Tree | 0.359014246 |
| No Modification | Misclassification Tree | 0.360897331 |
| CatYdsToGo | Misclassification Tree | 0.360938268 |
| SideOfField | Misclassification Tree | 0.361061077 |
| CatYdsToGo and SideOfField | Misclassification Tree | 0.361347634 |
| SideOfField | Neural Network | 0.361470444 |
| SideOfField | AutoNeural | 0.361552317 |
| CatYdsToGo | AutoNeural | 0.362166366 |
| No Modification | AutoNeural | 0.362575733 |
| SideOfField | Stepwise | 0.362821353 |
| No Modification | Stepwise | 0.362985099 |
| CatYdsToGo and SideOfField | AutoNeural | 0.363312592 |
| No Modification | Neural Network | 0.363762895 |
| CatYdsToGo | Neural Network | 0.364049451 |
| CatYdsToGo and SideOfField | Stepwise | 0.364254135 |
| CatYdsToGo and SideOfField | Neural Network | 0.365072867 |
| CatYdsToGo | Stepwise | 0.365359424 |
| Table 7. Misclassification rate for all models with all data sets | | |

|  |
| --- |
|  |
| Figure 3. Cumulative lift chart for all models. |

Works Cited

Breer, Albert. 2017. “How Analytics Are Used in the NFL | SI.Com.” June 27, 2017. https://www.si.com/mmqb/2017/06/27/nfl-analytics-what-nfl-teams-use-pff-stats-llc-tendencies-player-tracking-injuries-chip-kelly.

Field Level Media. 2018. “Report: NFL Teams’ Revenue Share Topped $8 Billion in 2017 | Reuters.” July 17, 2018. https://www.reuters.com/article/football-nfl-gb-nfl-revenues/report-nfl-teams-revenue-share-topped-8-billion-in-2017-idUSKBN1K719F.

McManus, Tim. 2018. “Eagles’ Secret Weapon? An Analytics-Fueled Attack.” ESPN.Com. January 18, 2018. http://www.espn.com/blog/nflnation/post/\_/id/266982.