Project Proposal

Chris Fenton

DATA 698

**Context / Problem Statement**

One of the challenging things about doing data analysis of NFL Football is that since the sport is highly situational, it makes many of the basic game data hard to value without proper context. Certainly, a scoring touchdown is generally going to be a positive data point for a team, as it scores points for that team, and the object of the game is to score more than other team.

However, consider a player running for 4 yards (one of the more common plays of a football game). Is this considered a positive or negative play for that player’s team? The answer completely depends on the context. Is the player’s team facing a 3rd and 1, where four yards will bring a new set of downs? Than very likely it is a positive play. However, if the players team is facing a 3rd and 20, they will likely have to punt, thus giving the ball away, and it is a negative play. So the data that we have that a player ran for four yards is impossible to assign proper value to without additional, contextual data.

Despite this truth, the way the NFL community has typically analyzed game statistics ignores context, out of either a lack of availability, ignorance, or both. After all, it is much easier to calculate the total amount of yards a player has run for, than to go play by play and manually assign each play a value of “positive” or “negative”. And generally speaking, gaining yards is an indicator of success, since it states that the ball was moved toward the opponent’s goal, and scoring is the object of the game. But if we stop there, much of the true information about what happened is lost.

**Literature Review/Research Conducted**

In recent years solid progress has been made addressing this via a variety of efforts that attempt to encapsulate critical “game state” information that can be used to contextual game data. One of the first was Defense-adjusted Value Over Average (DVOA) by Football Outsiders (<http://www.footballoutsiders.com/info/methods)> , which manually encoded each play according to a proprietary scoring scale, and summed the results by player, game, or season.

A more recent attempt has been to incorporate the idea of Win Probability (WP), and its close companion, Win Probability Added (WPA). The idea of Win Probability goes back to mid 20th century baseball, but was widely introduced to NFL analysis by Brian Burke at his former site Advanced Football Analytics (<http://www.advancedfootballanalytics.com/index.php/home/stats/stats-explained/win-probability-and-wpa>). Since then, a variety of other approaches have emerged, summarized by Michael Lopez at his personal blog (<https://statsbylopez.com/2017/03/08/all-win-probability-models-are-wrong-some-are-useful/)>.

Although the models surveyed by Lopez include a variety of implementations with regard to both data and model selection, they all follow a similar approach. A sample of recent NFL plays are collected (typically no older than from the 2000 season, as that seems to be an inflection point where NFL data became widely available). These data sets will include play by play game logs that include game state data encoded, such as (but not limited to):

* Home Score
* Away Score
* Score Differential
* Total Points Scored
* Down
* Yards to go (for first Down)
* Time Remaining
* Las Vegas Point Spread
* Timeouts Remaining

Once play data is selected and encoded, a model was fit on the data. Varying levels of transparency have been provided, but among the approaches were Random Forest (Lock and Nettleton), ensemble methods (Burke), logistic regression, and others. The common denominator amongst all methods were they were classifiers. This is natural, since within this framework the outcome variable of interested is whether a team won or lost (ties are rare and were ignored).