ECON 8330 Term Paper

Adapting Connelly’s Five Factors to the NFL

Miles Russell

Professor: Dr. Ben Smith

Introduction

In contrast to most other popular sports in the United States, football has been notoriously resistant to advanced analytics. What were one or two-team experiments in both Major League Baseball and the National Basketball Association evolved into analytic breakthroughs that have completely changed the strategies teams employ. Moreover, analytically-derived concepts are accessible to the average baseball and basketball fan, generating analytics communities in each sport that are engaged with ongoing practice and research. In light of these developments, one might wonder why such a revolution never occurred in football. Some say that football's old-school, tough-guy ethos is incompatible with analytics: legendary coach Vince Lombardi is quoted "Some people try to find things in this game that don't exist, but football is only two things- blocking and tackling." Alternatively, some say that the structure of the game is too complicated, or that available data aren't good enough to produce the types of insights possible in baseball or basketball. Regardless of the reasons, few football analytics insights have seeped down from NFL front offices to the average fan, and publications that do football analytics research, such as ESPN, FiveThirtyEight, and Football Outsiders, generally guard their methodologies.

One notable piece of research that has come out recently can be found in a series of articles written in 2014 by Bill Connelly, a writer for *SBNation*. Connelly laid out what he had determined to be the "Five Factors" that cause college football teams to win games: *explosiveness* (the ability to generate highly successful plays), *efficiency* (the ability to generate successful plays at a high rate), *field position* (the ability to play closer to the opponent's end zone than your own end zone), *finishing drives* (the ability to maximize scoring when close to the opponent’s end zone), and *turnovers* (the ability to generate extra possessions). Connelly was not the first person to build a football win probability model, but his work is important because of its transparency; other researchers can use his blueprint to advance football analytics research.

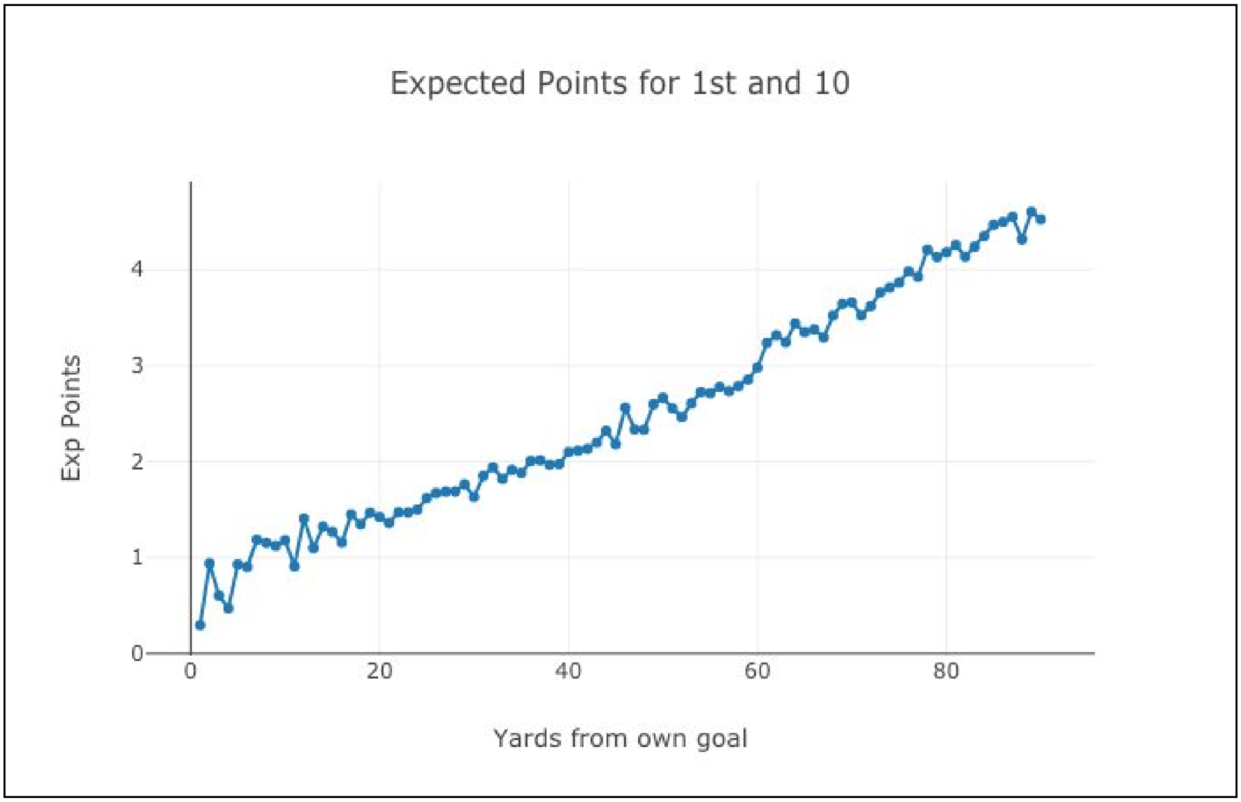
Connelly’s research informs us about the factors that cause college football teams to win, but it may not tell us so much about the factors that cause professional teams to win. It is possible that the differences between professional and college football would manifest themselves in different Factors. Alternatively, successful implementation of the same Factors with professional data may indicate that Connelly’s factors are the essential elements that decide football games across levels of competition. The sophistication of Connelly’s analysis is unclear, and deeper analysis could depart Connelly’s findings in interesting or unforeseen ways. In this analysis, I adapt Connelly’s Five Factors to professional football using 2010-2015 NFL game data. My analysis shows that Connell’s Five Factors are useful for generating game-level win probability estimates for NFL games. Moreover, by summing a team’s game-level win probabilities over an entire season, I generate estimates of a team’s second order wins, which can differ wildly from their actual number of wins. This has important implications for the NFL postseason and calls into question whether the NFL’s postseason structure truly rewards the best teams.

Data Collection and Cleaning

I retrieved my data from the College of Business Administration’s NFL database, which was purchased from ArmchairAnalysis.com. Data was available from the 2000 through 2015 season. Since the most recent years are the most relevant, I chose to query data only from the 2010 season and later. The majority of my querying was centered on retrieving play-by-play data for the time period I was interested in. In order to return only plays that I was interested in, I limited my dataset to only penalty-free offensive plays and field goal attempts. Since the tool I was using to query the data could only return about 35,000 rows for any query before breaking, and there are tens of thousands of plays in a football season, I had to query each season’s play-by-play data individually and union them together after the fact. I also queried data on turnovers, average starting field position, and game results.

Once I collected data, I needed to develop variables that could be fed into a win probability model. Since my goal was to adapt Connelly’s Five Factors to NFL data, I replicated his Factors as best as I could with the data I had. I started with *efficiency*, which can be thought of as the ability to generate successful plays at a high rate. Traditionally, a play is considered successful if it generates 50% of the remaining yardage on first down, 70% of the remaining yardage on second down, and 100% of the remaining yardage on third and fourth down. While this is a useful heuristic, we can devise a more sophisticated method using play-by-play data. In order to classify plays as successful or unsuccessful, I used play-by-play data to generate an expected point value for every down, yardline, and yards-to-go combination. The expected point value also accounts for the chance of a *defensive score* via a safety or a turnover returned for a touchdown.

Figure 1. Expected points for 1st and 10.



A play is considered successful if it results in an offensive score, or if it results in a situation with a higher expected point value than the previous play. Essentially, my method of classifying successful plays empirically verifies that a team is actually better off after the play than it was before it. We should expect that the closer an offense is to the opponent's end zone, the higher the expected points, all else equal. Figure 1 illustrates that expected point values generally increase as an offense gets closer to the opponent's goal. We should also expect that earlier downs and lower yards-to-go are associated with an increase in expected points, all else equal.

Next, I focused on *explosiveness*, or the ability to generate highly successful plays. To calculate a team's explosiveness, I averaged the expected point gain of a team's successful plays.

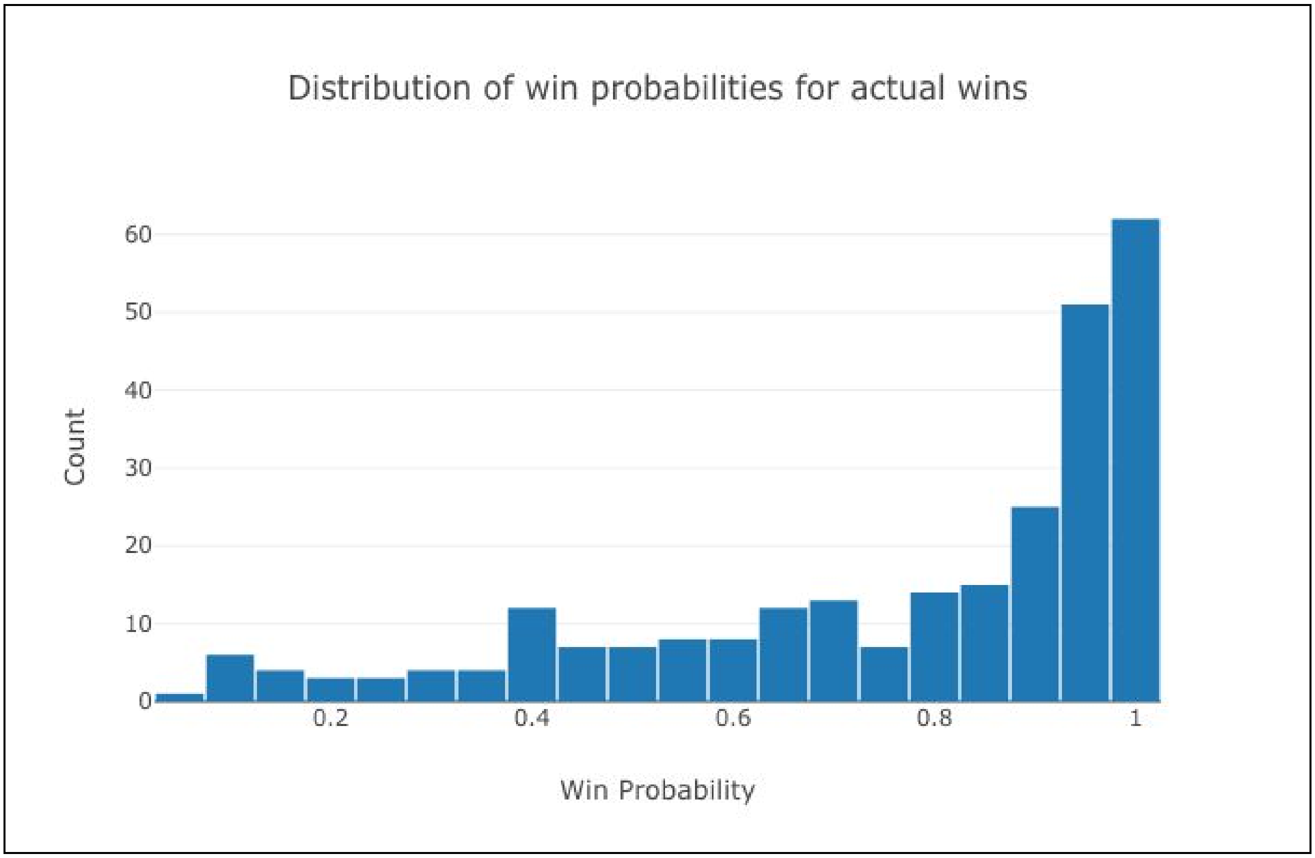
This conception of explosiveness helps us understand the magnitude of an offense's success when they were successful.

To account for a team's ability to *finish drives*, I calculated the average points scored for drives that entered what Connelly calls the 'scoring zone': the part of the field between the opponent's 40-yard line and the opponent's end zone. To account for *field position* and *turnovers*, I simply averaged the starting yardline of a team's possessions and summed the number of turnovers, respectively. Following Connelly's model, I calculated the margin between each team's five factors and their opponent's five factors. Finally, the dataset was ready for analysis.

Analysis and Discussion

To build a win probability model, I implemented a logistic regression on the dataset described above. I trained the model on data from the 2010-2014 seasons and tested it on data from the 2015 season. There are several methods by which we might judge the performance of a classification model, and many of them aren't useful in this context. Since we aren't actively trying to improve the model and it is designed for forecasting purposes, we shouldn't concern ourselves with various measures of R-square or the statistical significance of independent variables. The marginal effects of the independent variables are oftentimes interesting, but the independent variables in my model are measured very differently, making it difficult to compare their marginal effects. Rather, since the model is designed to estimate game-level win probability, we should focus on how successfully the model classifies wins. Figure 2 shows the distribution of win probabilities that were generated for actual wins in the 2015 season. Whereas a perfect model would generate a win probability of one for every actual win, a good model should generate win probabilities close to one (and far away from zero) for every actual win. At the very least, we should hope that the distribution of win probabilities for actual wins is skewed strongly to the left.

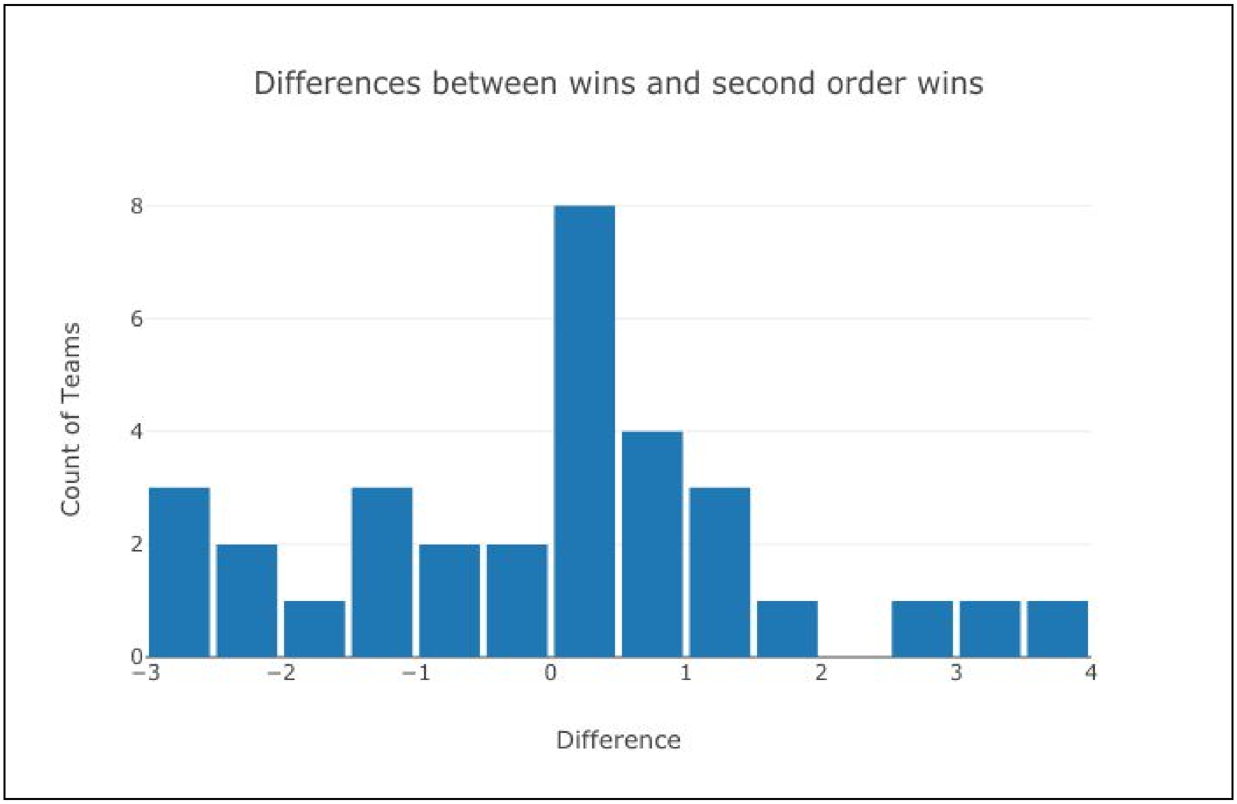
Figure 2. Distribution of win probabilities for actual wins.



We can see in Figure 2 that the distribution is strongly skewed to the left. A majority of actual wins were assigned a win probability of greater than 0.8, and exceedingly few actual wins were assigned a win probability of less than 0.2. Taking all of that into account, we should conclude that the model is reasonably good at game-level win projection.

One of Connelly's interesting extensions on his Five Factors model is to sum up a team's win probabilities across an entire season, which he calls a team's "second order wins." When we compare a team's actual win total to their second order win total, we can begin to get a sense of the teams for whom variance played in their favor. Since there are only 16 games in the NFL regular season, we could hypothesize that due to small sample size, second order wins would deviate substantially from actual wins. If a large relative difference between wins and second order wins exists, this would suggest that a team's record may be a noisy indicator of their actual capability. Figure 3 shows the distribution of differences between wins and second order wins.

Figure 3. Distribution of differences between wins and second order wins.



We can see from the histogram that second order wins can deviate substantially from actual wins. At the extremes, the Arizona Cardinals won 14 games but accumulated 10.5 second order wins, while the Seattle Seahawks won 10 games but accumulated 12.7 second order wins! We can safely say that variance plays a huge role in deciding a team's record at the end of the season, probably because there are so few regular season games in professional football. This is relevant because regular season record determines which teams get into the playoffs, obtain home field advantage, and get first-round byes. Since the relative difference between actual wins and second order wins can be so high, it is likely that teams that improbably won games are being admitted into the postseason over teams that improbably lost games. This potentially biases the playoff field towards teams that got lucky in the regular season.

Conclusion

In this analysis, I adapted Bill Connelly's Five Factors to professional football and concluded that the Five Factors are a useful analysis tool for NFL games. This finding suggests that the essential factors deciding football games are similar across different levels of competition. As a part of this analysis, I replaced a blunt but convenient heuristic with an empirically-driven method for identifying successful plays. By summing teams' win probabilities across the season and comparing that sum to teams' total wins, I found examples of large relative differences between actual wins and second order wins. These differences may indicate that NFL postseason entry is biased towards inferior teams that got lucky during the regular season. More analysis needs to be done to see if this is an extensive problem. If it is an extensive problem, a possible solution to is to expand the playoffs to include teams with high second order wins that would have otherwise not been included, or to reserve the wild card spots for teams with high second order wins.

References

Connelly, B. (2014, January 24). *The five factors: College football's most important stats*. Retrieved from www.footballstudyhall.com.

Connelly, B. (2014, January 27). *Five factors: Isolating explosiveness with IsoPPP*. Retrieved from www.footballstudyhall.com.

Connelly, B. (2014, January 30). *Five factors: What derives field position?* Retrieved from www.footballstudyhall.com.

Connelly, B. (2014, April 1). *Five factors: What matters when it comes to finishing drives?* Retrieved from www.footballstudyhall.com.

Source Code

Code can be found here.