Eric Wilson
Springboard
July 2020

Capstone Project Proposal

Professional football is one of the most popular entertainment options in the United States, with 16.7 million spectators attending National Football League (NFL) games during the 2019 regular season and an additional 180 million viewers watching on television. Owing to this popularity and spurred on by the emergence of fantasy sports competition, sportsbooks and other sports betting sites have experienced lucrative growth in their bookmaker business, taking advantage of the fact that reportedly more than 38 million Americans planned to bet on NFL games last season, according to the American Gaming Association.

Among the various betting options for NFL games offered by sportsbook outfits, the over/under (O/U) line is one of the most popular. The O/U line represents the predicted total score of the contest, taking into account all scoring including scoring in any overtime period. The bettor chooses to bet on the over or the under, meaning that the conjecture is that the final total score will be greater or less than the O/U line, respectively. If the selection of the bettor is correct, the bettor wins the amount of the bet placed; if not, the bettor loses the bet. In the event that the total final score matches the betting line, the bet is returned to the bettor. In this process, a vigorish is typically included, which is a percentage of the bet that is paid to the bookmaker, regardless of the outcome. In essence, the bookmaker is less interested in correctly predicting the total amount of points scored and more interested in promoting even action on both sides of the bet, maximizing profit for the bookmaker. In fact, an O/U line often changes during the week before the event in response to public demand on one side of the bet or the other. Thus, the bettor is competing not so much against the bookmaker but rather against public sentiment. With this in mind, this study seeks to determine whether a profitable business can be maintained employing a system that uses regression analysis of historical football team statistics to predict the total score of a contest and defeat sportsbook O/U odds on a sufficiently regular basis.

This study will utilize team statistics as well as game data to obtain the target feature, total points scored. The team statistics will include approximately 20 attributes for each participating team in the contest, such as total yards, league rank, and touchdown scoring efficiency, characterizing the abilities of the offensive, defensive, and special teams components of the squad over the previous five games of the respective season. In addition, event statistics such as the field type and temperature at kickoff will also be included. Contests over the past three seasons will be utilized in the analysis, with data used for training and testing beginning with week 6 of each season and

extending to the end of the respective regular season. These data will be obtained from https://www.pro-football-reference.com/, which provides historical data from decades of NFL contests. Through the application of conventional statistical metrics, an optimal model will be selected.

The criterion for success for this study will be a sufficient number of favorable outcomes against a betting line. As mentioned above, sportsbooks earn their profit through the implementation of a vigorish. A typical vigorish for a bookmaker is 10%, which means that for a bettor to win $100, $110 must be placed at risk, with $10 going to the bookmaker. Thus, the breakeven point for a bettor may be determined from the equation

$$100x - 110(1-x) = 0,$$

which solves to

$$x = 52.38\%.$$

For the purpose of this study, a success rate of at least 53% will be judged to be profitable and worthwhile. Betting lines of contests over the past three regular seasons will be obtained from https://sportsbookreviewsonline.com/.

Despite the inclusion of an extensive amount of information regarding recent team history, this study contains a couple of limitations that may affect the reliability of its results. An important factor in the outcome of a football contest is the realization of injury to the members of the respective teams. Although injuries during a given contest can be regarded as a random factor that contributes to the variance of the result of that contest, injuries that occur during the week before such a given contest or in the most recent contests before can affect the betting line of the given contest and, thus, the reliability of the analysis. This study will not give consideration to such injuries, except as reflected in the statistics of the previous contests considered in the study. Another important factor in the result of a football contest is weather. The temperature of the event will be considered, and the nature of the weather (e.g., sunny, rainy) will likely be included in the form of boolean variables. However, the magnitude of such weather, for which the forecast can affect betting lines, will be out of the scope of this study. Finally, the study will be limited by the amount of data taken into account. Three years of regular season contests from game 6 to game 16 will be considered for the training and evaluation of the model, amounting to 528 total contests. Whether this is a sufficient amount of data will need to be evaluated; however, extending the data further back to earlier contests would be problematic, owing to the evolution of the sport, which may render earlier contests less relevant.

A regression analysis of historical NFL team data will be performed, followed by thresholding on the prediction interval, to determine whether the total score of a football contest can be predicted with sufficiently low error for the model to serve as a betting tool. Stakeholders for this study would include anyone interested in evaluating whether

a business pertaining to betting the O/U lines at sportsbooks could be profitable. Interested parties in this study may also include coaches and managers that are interested in what aspects of team performance are most relevant to team scoring in football today. The solution of this study will be specifically focused on the effect of team performance on total score and will not directly cover individual player performance, nor will it address point-spread betting nor other margin-of-victory considerations.