Ford Knippa, Brennan Penafiel, Patrick Oliver, Walker Lee

LTC Smith, USA

Computational Methods II

12 September 2024

Final Project Summary

Sports data analysis is an important practice used in the predictions of NFL games. Being able to make an accurate prediction of the outcome of these games is a means of obtaining bragging rights for those who are successful. The aim of this project is to see how effective oddsmakers are at predicting these outcomes and comparing them against real NFL game data and characteristics.

To conduct our analysis, we used a dataset from Kaggle (an online repository of data sets) that contained information on NFL games from 1966-2023. Prior to visualizing the data and attempting to identify trends, we first needed to clean and procure the data into a usable format. We began by answering the question: what data should we use? We decided to use only data after the year 2002 for two primary reasons. First, the modernization of NFL offenses has changed the game with more of a pass-heavy focus as opposed to the run-heavy offenses of old. Second, in 2002 the NFL evolved into its modern format with 32 teams and 8 divisions upon the addition of the Houston Texans as an expansion franchise.

We then had to account for various team names, team locations, and stadium name changes. For teams that changed locations after 2002, we decided to drop games they played in before their move since the stadium and fan base would be different. The Washington Commanders (formerly the Washington Redskins and Washington Football Team) were a unique

case, that they only changed their team's name. For this reason, we considered the Washington Redskins, Washington Football Team, and Washington Commanders to be the same team in our analysis. Since there are only 32 home stadiums where games have been played since 2002, we were able to manually form a list of stadiums that we wanted to include in our analysis. We filtered NFL games since 2002 to exclude international games and games played in temporary stadiums. In the data, some of the values in the stadium column corresponded to the field name, while others corresponded to the stadium name. We identified columns where there were either the stadium names or field name was a possibility and ensured that both values were included in the list that was used to filter the data. In the final step of the data-cleaning process, we used the information from various columns to calculate if the spread was covered, if the game was over or under the point total, and if the favorite team won.

For our visualizations, we chose to focus on the teams themselves and some external factors that we thought may have affected the outcomes of the game. Using Dash, we were able to produce two visualizations, each with their drop-down menus. These drop-down menus allowed us to quickly switch the prediction method and analyze any potential trends. With the visualizations, we were able to quickly determine that Vegas's predictors are particularly good at predicting the outcome of a game. Additionally, we found that Vegas is adept (accurate within 70-80%) at adjusting its predictors to account for external factors (weather, stadium type, etc.).

Future work for this project could explore different factors affecting these spreads such as stadium location and away team travel distance, as well as expanding to other sports given the appropriate data. With our findings, we are confident that the algorithms and data analysis techniques that are employed by oddsmakers are effective tools in predicting the outcomes of NFL games.