Capstone 2 Milestone Report

I am going to create a model to predict winners of NFL games against the Las Vegas spread. My client will be an NFL franchise. The goal is for the model to pick winners against the spread and offer insight as to why the spread was wrong. By doing this it could reveal under or overvalued traits of an NFL team or an NFL game.

I am going to use a dataset from pro-football-reference.com. The data will be combined and broken down by play type, along with other features. By breaking down each team’s performance by play type (pass short left, pass short middle, pass short right, pass deep left, pass deep middle, pass deep right, run left end, run left tackle, run left guard, run center, run right guard, run right tackle, run right end, and middle plays) offensively, defensively, and their completion percentage (complete pass or run of 3 yards or greater), I hope to quantify advantages or disadvantages in this realm that others have not revealed. Additional win, loss, point, yard, and turnover data will be scraped from a different page of pro-football-reference.com and then converted into per game statistics when relevant as well.

Columns will be created that represent the difference between the offense’s completion percentages vs. the defense’s completion percentages by play type. This is done by subtraction in the relevant columns. Columns representing the home team’s time zone, the away team’s time zone, the time zone at the location of the game, the time after their previous game, and whether it is a division game or not will be added as well. This will be done utilizing the locations of the game and each team’s time zone. Using the type of game each team played in previously, the days between games was determined as well. Points given up, points scored, yards given up, yards gained, turnovers against, and turnovers for per game for were all determined by making a function and applying that function to the relevant games. After each season the per game numbers are reset to zero. There is no data for week 1 for any season since there are no per game averages until the second week. In total, there will be 199 columns and 956 rows for games from the 2014 to 2017 season.

Vegas spreads are notoriously accurate, and thus creating a model with approximately 60% accuracy could be considered extremely helpful. Hopefully using extensive play by play data along with other, more standard data, this model will be able to consistently outperform the Vegas spreads. I will test a random forest classifier, a logistic regression model, and a deep learning model to see which is most accurate.

So far, each model is slightly above 50%, but there is room for improvement. The best is the logistic regression model with slightly above 55% accuracy as of right now.

I could find more data to include on the previous year so that I could increase early season accuracy. In the first few weeks of predictions for the model there is not much information for the model to use and that information is particularly susceptible to problems of variance, likely making the predictions less accurate than they could be. Using previous years wins could help alleviate this problem.