Solving a Problem for Analytics Edge

Cameron Cubra, Brooke DiMartino, Evan Marrone, and Alexander Pinigis

Problem Statement: Every week millions of people sit down to watch NFL football games. Many of these fans bet on which team will win. We believe we can create a model to predict these win-loss outcomes in advance.

Objectives: We want to use the abundance of football related statistics available from previous and current years to:

1) Create models based on past years' statistics (see data section) and win-loss results for a given team (training set) to predict later season win-loss results (test set)
2) Use the final model to predict 2022 weekly win-loss outcomes between two teams based on 2022 team statistics for the games played this year
3) Create relevant data visualizations

The Data:

We will collect the following data:

- Weekly data on: offensive, defensive, special teams, win/loss, injury, home vs. away, current standing, weather, referee.
- Team vs team data on: trends on wins/losses between the teams, whether they are in the same division.

There are many different sources for NFL data, but we have researched the following possible source for data which gives us all the data we need:

- Sportsdata IO

Because each week adds to the dataset, we will create cumulative statistics for the entire season and potentially implement moving averages. For example, if we are trying to predict if the Eagles will win in week 10 and one of our features is rushing yards, then we will look at their cumulative rushing yards the entire season as well as potentially their moving average of rushing yards over the past X number of weeks. We plan to limit our data to 2020 and on since previous data is not readily available for all features (and if it is available, it costs money).

Analytic Techniques: Our primary analytical technique will be a logistic regression model utilizing random splits and stratified sampling and identifying our most statistically significant dependent variables. Our plan is to use the previous year's data in order to train and test our model. We also plan to implement regression trees, random forests, and any other relevant models learned from class (an ensemble of each best performing model type will be considered as our final model).

Impact and Goals: The overall goal of the project is to increase the winning prediction accuracy for a given week for a specific team. This will allow sports bettors to make more accurate predictions, and thus, more money. Additionally, by exploring the important features in our models we will be able to report and visualize the most important aspects of a specific team winning a game. This could be extended to discussion on which players or play styles are the most effective.