Introduction

Sports is rife with statistics. Girlfriends, boyfriends, housewives, and husbands have marveled at their significant others ability to recall the most minute mathematical oddities about players. In a vacuum these statistical numbers are curiosities, perhaps useful at analyzing the uniqueness of a player or the playstyle of a team, but they do not allow us to uncover anything deep about the sport. American football is one of the most statistically deep sports in the world. Nearly every action a player can take, from throwing a pass to committing a penalty is measured and recorded for posterity. Our analysis seeks to utilize a robust dataset of National Football League (NFL) teams from 2002 through 2019 to accomplish two things.

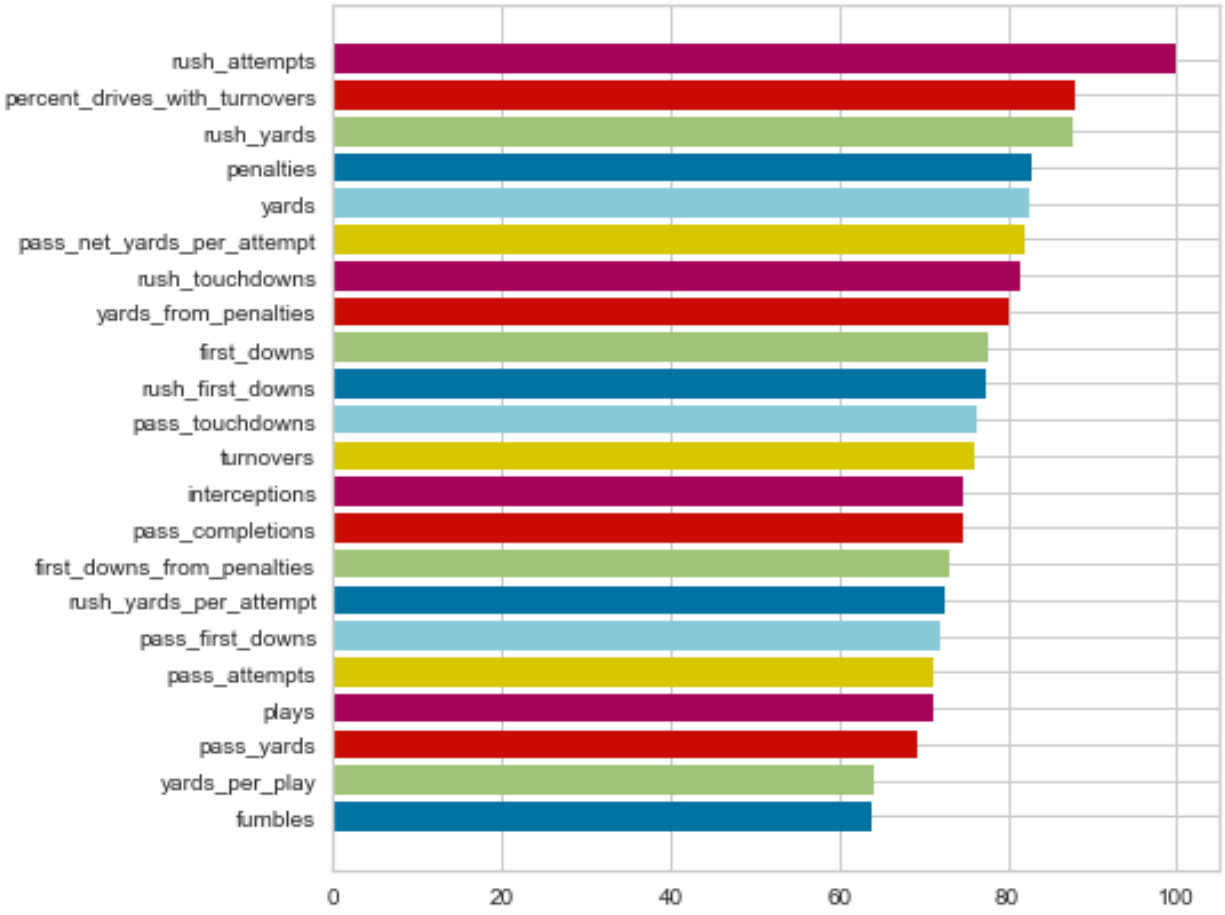
- Determine which summary statistics are most important at predicting success (determined as wins in this analysis)
- Create a model that can predict a team's wins based on their season statistics within at least 2 standard deviations

Methodology

Data was loaded into an SQL database hosted on AWS. Once loaded we began our analysis. We used two different machine learning methods to accomplish our goals. The first is a random forest model. This model was selected not primarily for its predictive capabilities, but for its ability to tell us the relative importance of features. For prediction we utilized a 3-layer neural network and fed in most of the databases statistics into it. While the networks predictive capability fell short of our hopes, it was still able to predict wins with a high degree, if not perfect, accuracy.

Findings

Our findings bucked many of our expectations. Firstly, we found that rushing was significantly more important than our initial expectation. Rushing statistics dominate the upper 50% of the chart while passing statistics tended towards the bottom. The reasons for this are numerous. Firstly, it could indicate that teams that are winning are more likely to run the ball in order to end the game sooner (the clock does not stop if the runner stays in bound leading to shorter games). Rushing may also just be less risky than passing, so it is possible that rushing teams simply commit fewer mistakes leading to a higher likelihood to win on average.



Our neural network also performed admirably. While it was only around 32% accurate at predicting the exact win total it was rarely off by very much when it missed. When comparing the predicted label value to actual label value in our test dataset we found that the prediction was only off on average by 0.12 games and had a standard deviation of 1.42. This is indicative of a highly predictive model that when fed a team's season statistics can, within less than a game, predict a team's overall wins.

Discussion

This analysis is incomplete. The data we collected provides an excellent starting point for a more rigorous analysis of the NFL. Additional datatypes such as offensive formations, defensive formations, weather, altitude, location, home vs away, and many more would aid our model's accuracy and provide clearer results. Additionally, creating models that looked at week to week statistics as opposed to annual statistics could allow us to predict a team's overall wins on the fly as a season progresses. A useful tool for enthusiasts, gamblers, and coaching staff alike. The methodology is sound, but an influx of more high-quality data could make things more clear.