

Predicting NFL Game Outcomes

Michael Nercessian, Rohan Lewis, and Matthew Armstrong

November 8, 2020

1 Introduction

Our project aims to predict the outcome and point differential of NFL games. We initially approached this problem by considering variety of different datasets including game-by-game and play-by-play data. Ultimately, we decided that accumulated game-by-game data would be a more accurate predictor of a game's outcome than play-by-play data, which we felt was too noisy. So far, we have used data gathered from Pro Football Reference (PFR) that consists of game-by-game team statistics for all 2019 regular season games.

2 Data Preparation

First, we gathered 32 CSV files, one for each NFL team, from the “Schedule Game Results” section of the 2019 season page on PFR (example link [here](#)). We then manually created a mapping Dataframe with a unique identifier for each team, the team's full name, abbreviation, conference, and division, which was a useful key for maneuvering our datasets. A preliminary investigation into the data led us to drop some unnecessary columns and rows. We removed ‘Expected Points’ columns which represented PFR's predictions for Offense, Defense, and Special Teams, as well as game start times. We also removed rows that corresponded to teams' bye weeks. In addition, we modified the home/away column, which contained missing values for home games, to a binary encoding in order to keep track of where a team played in a given week. We also found that entries in the turnovers column were missing if a team had no turnovers in a game, so we substituted missing values with zero in this column. Finally, we added in the unique identifiers for target team and opponent from our initial team mapping Dataframe.

After this initial cleaning of the data, we vertically concatenated the 32 team-by-team datasets into one dataset containing all of the games in the season. Notably, this dataset contained two copies of each game, since each game was counted in both of the datasets coming from the two teams playing in a game, so we removed all the duplicates. We then had a Dataframe with 256 rows, representing the 256 games in an NFL season, and 19 columns, representing identifying information such as week, and game statistics including points scored, first downs, rushing yards, and others for each team in that game. Our current dataset therefore contained the game-by-game data with columns representing what had happened in that game.

We seek to predict a point differential for each game, and in doing so we will also predict the winner. We envisioned the point differential as the output space, while the input space would contain relevant statistics for the two teams prior to that game. To facilitate this analysis, we needed to restructure our

data such that each row would still represent a game, but the columns would only represent data that would be available before that game began. In our preliminary analyses, we aimed to design a feature space containing the information available in our 256x19 dataset. We therefore transformed the data such that the features represented the same game-by-game statistics, but aggregated as a running cumulative average for the prior weeks of the season for each of these statistics. Specifically, we used 22 features, which were the following 11 statistics for both home and away team: win percentage, points per game, first downs per game, rushing yards per game, passing yards per game, turnovers per game, points allowed per game, first downs allowed per game, rushing yards allowed per game, passing yards allowed per game, and takeaways per game. We excluded the 16 week 1 games from our analysis as there is no prior data in the season to predict on, but we may alter our approach to incorporate the prior season's data in the future. This would allow prediction in week 1, and may provide important information for predicting other weeks as well. Our final dataset contained 240 examples and 22 features.

3 Preliminary Analyses

We first visualized the true point differentials across all regular season in a histogram, and noted that the majority of games were decided by fewer than 10 points (Figure 1). This distribution was a useful benchmark to consider as we examined the distributions of our predicted point differentials.

In each of our analyses, we shuffled and split the dataset 80:20 to training and test sets.

First, we used least squares regression on a simple feature space containing only home and away team points per game (plus bias). As we are aiming to predict point differential, we hypothesized historical points per game of each team might be a particularly strong predictor of point differential for a given game. The model achieved mean squared error (MSE) of 187 on the training set and 177 on the test set. We also found that the model correctly predicted the winner of games in the training and test sets with 58.8% and 64.6% accuracy, respectively. We plotted true point differentials against predicted point differentials in Figure 2a. While this scatter plot appears noisy, and the error values seem high, both of these are expected due to the high degree of inherent noise in predicting NFL games, stemming from reliance of game outcomes on individual performances which can fluctuate unpredictably, as well as other factors. The histogram shows the distribution of predicted point differentials on the test set (Figure 2b); most fell to ± 10 points which was consistent with the true distribution.

Our second analysis mirrored the first, but also included takeaways per game of each team as another pair of features. We hypothesized that this key statistic relating to ball control could strongly influence outcomes. Training (MSE) slightly decreased to 183, and test MSE slightly increased to 188. Accuracy in winner prediction slightly increased for both training and test sets. The same

plots are shown for this analysis in Figure 3.

Finally, we experimented with performing least squares regression with all of the data at our disposal as a 22-feature input space. This yielded a substantial decrease in training MSE to 157, and an increase in test MSE to 194, and a decrease in accuracy in winner prediction on the test set to 62.5%. Training winner prediction, however, jumped to 64%. This was indicative of overfitting. One potential explanation for this is the potential collinearity of many of our features. For example, a team's average number of 1st downs and average number of total yards are strongly correlated as every 10 yards gained within 4 offensive attempts equates to a 1st down. This could lead to our model over-learning on the same noise captured across two collinear features, causing our model to overfit the training set. The corresponding plots are shown in Figure 4.

4 Next Steps

Going forward, we aim to work towards the following goals in order to improve our accuracy in predicting the outcome and point differential of NFL games:

- Add data from prior seasons to increase the size of the dataset and perhaps incorporate useful features regarding prior season statistics.
- Modify our feature design approach to limit collinearity of features and thereby limit overfitting.
- Experiment with moving windows for statistics instead of only using season cumulative averages. For example, in addition to using cumulative averages, also add features for the same statistics averaged over the prior 3 weeks as the most recent games might have a larger impact on outcomes.
- Account for variables that can have extreme impacts on a game such as key injuries, suspensions, or other miscellaneous factors that impact a team's performance.
- Explore the LowRankModels and ScikitLearn packages to experiment with different loss functions, regularizers, and models to construct an approach that may be superior to least squares regression.
- Use k-folds cross-validation rather than a single split for training and testing.

5 Appendix

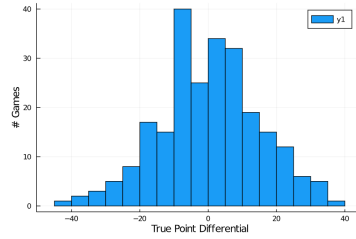


Figure 1: True point differentials distribution.

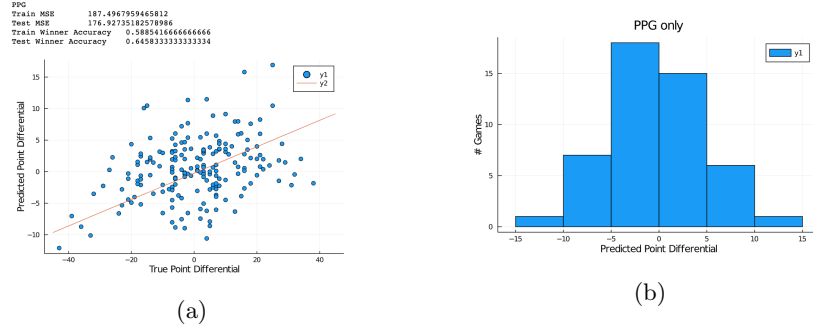


Figure 2: Scatter plot and histogram for predictions from points per game.

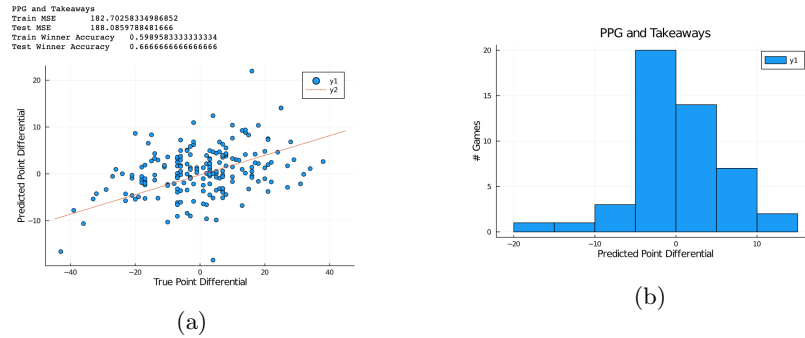
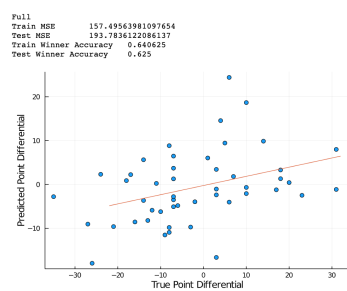
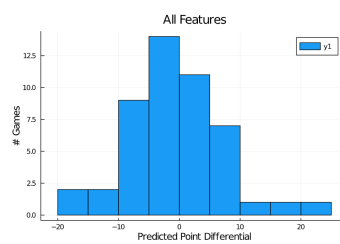


Figure 3: Scatter plot and histogram for predictions from points and takeaways per game.



(a)



(b)

Figure 4: Scatter plot and histogram for predictions from all features.