Project Midterm Report (Dream Team)

Darren Chang, Jack Vaughan, Zach Schaffer

1 Dataset Description

There are two primary datasets used for this project. The first is weekly fantasy football data from 1999 to 2019 sourced from Fantasy Football Data Pros, which contains fantasy football points and game statistics for offensive players each week. About 200-300 players are represented in each dataset for the 17 weeks of the regular NFL season. The three measures of fantasy football points for each player include the standard calculation (based on one point for ten yards of offense generated, and adds bonus points for touchdowns and subtracts points for interceptions and fumbles), the points per reception (PPR) calculation (the standard rules with the modification that each reception is an extra point), and the half PPR calculation (the PPR rules except that each reception is half a point). Offensive statistics for each player is included, such as passing yards, rushing yards, and reception yards, completions, interceptions, and touchdowns. The overall quality of the data is great, as no player is missing information; however, players vary from week to week and there is little information at different levels of granularity.

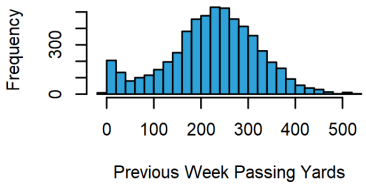
The second dataset is game-level data from 2009 to 2019 from the `nflfastR` package, which is available from CRAN. This dataset contains a `game_id` field, which can be linked to other game-level stats, scores, and the teams involved in each game. These two datasets are joined together so that each player entry for each week has game data. There are $n = 52,744$ player-weeks represented in the data, with nominal, discrete, and continuous values.

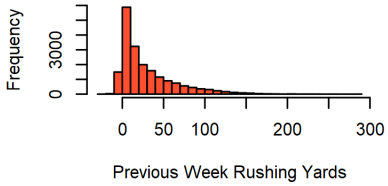
2 Feature Engineering

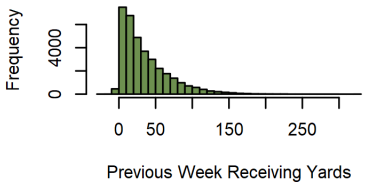
For fantasy football projections, feature engineering is a critical part of transforming the data to be used in models. Projections must be created based on data from the previous weeks; for the first week, there is essentially no data. Two sets of features are engineered for projection purposes:

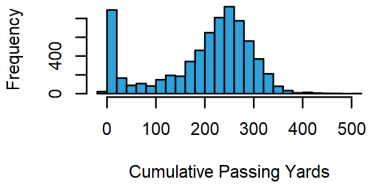
1. **One week lag** data, which are game statistics from the previous week and are denoted with the suffix `_prev`. For the feature set $X_{i,t}$, $X_{i,t,prev} = X_{i,t-1}$.
2. **Season cumulative** data, which are game statistics averaged over the season and are denoted with the suffix `_cum`. For the feature set $X_{i,t}$, $X_{i,t,cum} = \frac{\sum X_{i,t}}{t}$.

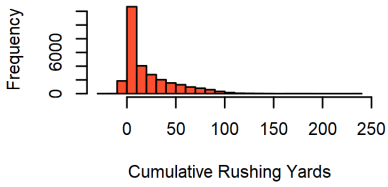
We think that these are useful features for predicting player fantasy performance in a specific week because the lag represents a "momentum" factor for how well a player is performing at a specific point in the season and the cumulative data represents a "baseline" factor for how well a player is performing throughout the season. Thus, there are 30 features for player statistics and one-hot encoded features for player team, opposing team, and player position.

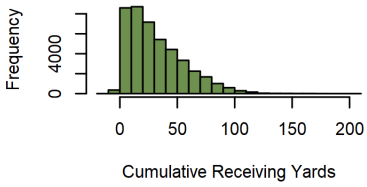

Frequency
0
300
0
100
200
300
400
500
Previous Week Passing Yards

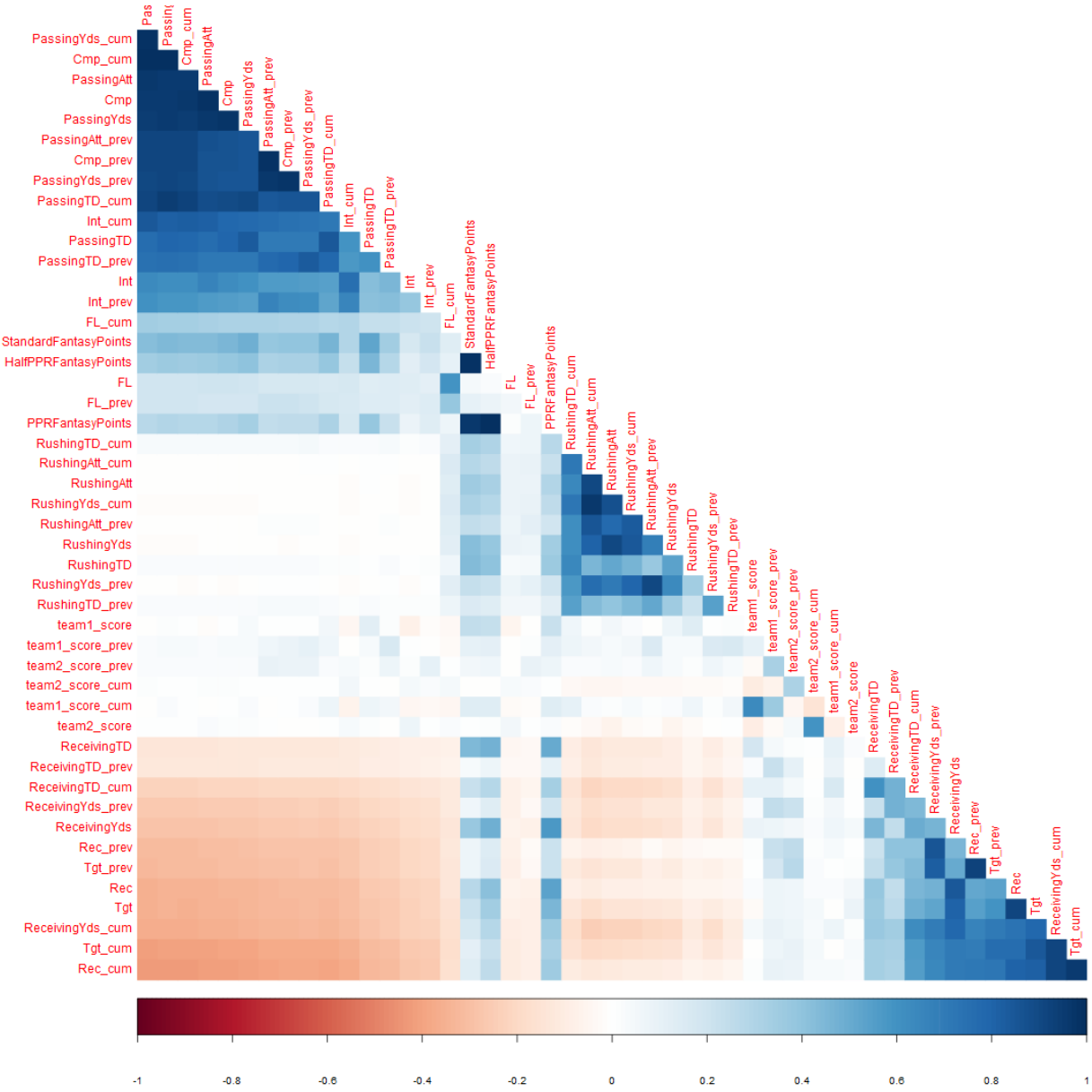

Frequency
0
3000
0
50
100
200
300
Previous Week Rushing Yards


Frequency
0
4000
0
50
150
250
Previous Week Receiving Yards


Frequency
0
400
0
100
200
300
400
500
Cumulative Passing Yards


Frequency
0
6000
0
50
100
150
200
250
Cumulative Rushing Yards


Frequency
0
4000
0
50
100
150
200
Cumulative Receiving Yards



3 Preliminary Analysis

4 Plan

1. loss function
2. projections
3. defense
4. WAR