Project Progress Report
November 7, 2024

1. Names of all group members

Carter Larsen, Seth Remer, Michael Laswell

2. A brief description of the problem

In fantasy sports, predicting player scores is crucial for making the best decisions but is challenging due to the variable nature of player performance in individual games. We will use current projections to help us to look at expected outcomes but these lack depth in that they don't account for historical performance trends for the player, in game context, and other situational factors. We will also use player historical data this data will give us the depth we need but would be lacking if used on its own as the current projections take into account factors that are more recent or in the future like who is being played, against any recent information about the player (performance trends, health, etc), the weather predicted for that match day etc. The biggest problem will be effectively combining short-term projections and long-term performance data in a way where our model will be able to more accurately predict fantasy scores. We hope by using both current projections and historical data we will be able to overcome the disadvantages of each form of data and develop a predictive model with improved accuracy and reliability when predicting fantasy player scores thus allowing users to make more informed lineup decisions.

3. How and from where is the data being gathered

We found all the data we need on the https://github.com/hvpkod/NFL-Data github repo available as csv files that we will download. All data on the repo was scraped off of fantasy.nfl.com then put into files by the author of the repo. The repo has the mit license and thus should be fine for us to use.

4. A description of the data set including:
    - Actual example instances, including a reasonable representation (continuous, nominal, etc.) and values for each feature
    - How many instances and features you plan to have in your final data set

For our data set, we plan to start with the following features:

1. Fantasy points projected for the week in question
    a. Using archived weekly predictions from NFL Fantasy, this will be the points projected by experts for the player on the given week. This will provide a grounding on our projection for the player.
2. Fantasy points per game, projected at the beginning of the season
    a. Using archived seasonal projections from NFL Fantasy, this will be the fantasy points per game projected by experts at the beginning of the season (i.e. before

any games are played and projections shift). This will link the player to their preseason expectations.

3. Fantasy points per game in the previous season
    a. Using NFL box scores, this will be an average of the fantasy points scored per game played by the player in the previous season. The previous season's performance is generally a strong predictor for the current season's performance.
4. Touchdowns per game, averaged over the last 3 weeks
    a. Using NFL box scores, this will be an average of the touchdowns scored by the player over the past 3 weeks. Players that score more touchdowns will be more likely to score points on any given week, and a player with a higher rate of touchdowns is more stable on any given week.
5. Position of the player
    a. For our project, we will likely focus on WR/RB/TE. WRs and RBs may be more likely to score higher than TEs, or TEs may be more volatile in their scoring.
6. Game home/away
    a. This will encode whether the player is playing the game in question at home or away. This affects performance for some players due to playing in a hostile/friendly environment.
7. Fantasy points scored in the previous week
8. Fantasy points scored, on average, over the last 3 weeks
    a. These two features will be calculated from NFL box scores. Recent performance is perhaps the best predictor for future performance in fantasy football. Including two levels of recency allows the model to focus either on most recent trends or repeated performances over a few games.
9. Performance of the opposing defense against the player's position, averaged over the past few weeks
    a. From NFL box scores - defenses that consistently give up big points to WRs are more likely to continue that trend
10. Targets (for WRs/TEs) and carries (RBs)
11. Red zone targets/carries
    a. From NFL Advanced Stats - Targets and carries represent the opportunity and workload of a player in a given offense. Red zone targets/carries are especially valuable because they show that the player is being given more scoring opportunities. These would be normalized by position.

These features will be refined as we test them in the model, and we will likely finish with some subset of these features. We plan to use about 3,500 instances, using the top 30 WRs, top 30 RBs, and top 15 TEs from each season from 2021-2023.

Here are some example instances:

PtsProj	SeasonPPGProj	PrevSeasonPPG	TDPG (P3)	Pos	H/A	Pts (P1)	Pts (P3)	DefPA (P3)	Tgt/Crr (P3)	RZTgt/Tch (P3)	Target: PtsScored
14.76	11.24	12.74	0.67	WR	H	12.20	9.87	10.07	0.92	1.00	15.40
17.27	14.24	15.55	0.67	WR	A	20.90	18.33	17.23	1.38	1.33	29.80
12.92	13.53	NaN	1.00	RB	A	25.80	15.93	16.27	2.42	2.00	9.40
13.64	10.26	5.70	0.67	RB	A	23.40	13.20	13.03	2.22	4.00	19.80
8.90	7.03	7.01	0.00	TE	H	7.40	5.17	9.63	0.88	0.67	18.70

Note that for some players, the previous season points per game will be an unknown value. This is because some players will be rookies and will not have data from the previous season. We plan to treat these unknown values as their own unique value.

5. What machine learning models are you initially trying to learn with

Initially we want to try learning on a variety of different models, MLP, Decision trees, K-NN. We are curious to see which models perform the best and which ones we can get better performance out of by adjusting their hyperparameters. By applying what we have learned in the labs, we believe we can adjust their hyperparameters to reduce overfit and get a better generalization.

We also want to use ensembles to increase accuracy by using the results of the different models to increase accuracy together. Random Forests is one that intrigues us, especially with the software that increases the ease of use.

6. Brief discussion of plans and schedule to finish the project

In the month before this project is due, we believe there are 3 major stages for us to accomplish:

1. Data collection and cleaning
2. Model training
3. Final report writing

We hope to complete the data collection and cleaning in the first week. Collectively we are going to create a script to pull the data that is conveniently in a csv format from the github repo. It will then format it and drop the columns that we are not using and calculate any of the ones we are interested in. This will also include feature selection and reduction through methods like PCA.

For the next 2 weeks after that, we will train the models and adjust them. We will initially train them with their default parameters and tweak the hyperparameters to get better accuracy while reducing overfit. Each of us will take a subset of the models to focus on and share the data amongst each other.

For the final week, we will review and analyze our data. We will put it into a cohesive format for our final presentation.