**Project Proposal - Fantasy Football Success Predictor**

**Team Name** - Cloudy with a Chance of Football

**Team Member, GitHub Username**

Griffin Taub, griffin-taub

Aidan O’Connor (Nominated project coordinator), oconnorac

Ermina Mujan, ermina-mujan

Yaphet Tewahade, ytewahade

**Domain** - NFL Fantasy Football

**Hypothesis/Project Topic** - On any given week, will an NFL player score above or below their projected score? What are the factors that have influenced performance the most so that we can best predict future performance?

**Data Sources**

We plan on using at least two of the below data sources (Fantasydata.com and Nflsavant.com). We are mostly interested in weather data from Nflsavant.com and player/fantasy stats from fantasydata.com. To merge the two datasets, we will probably need to match on the date of the game and home team. If we use player stats from more than just Fantasydata.com like pro-football-reference.com then to merge the different data sources we will likely consider matching on team name, jersey number, and season year.

**Fantasydata.com** - This is where we will get all the 2019 and 2020 season stats as well as fantasy projections and actuals by player by week. This is also where we plan to get red-zone and third down stats by player by week to use as a “clutch” feature.

**Nflsavant.com** - This is where we plan to get weather data for each game. We hope to use weather as a feature.

**Pro-football-reference.com** - This is a backup data source for player stats. If for some reason fantasydata.com doesn’t work out, this is where we would go next. There also may be some features this data has that fantasydata.com doesn’t in which case we will end up using this data source as well, we need to further explore this.

**Description**

In fantasy football, each player has a projected score for their upcoming game. It is not uncommon for a player to have a high projected score only to produce a score below projection. On the flip side, it is not uncommon for a player to have a low projected score only to produce a score above projection. We are pursuing this project to see what influences a player's actual output score the most so that we can make data-based decisions when building a team or starting lineup week to week during the NFL season. We plan on using 2019 and 2020 NFL season data for this project.

Our first step is to get together as a group and start thinking about the data we are interested in and where we will get it from. Since we are interested in individual player performance it will be important that our data contains every NFL player, every game they played, and all the stats they accrued during the 2019 and 2020 season. This data will also need to contain projected and actual fantasy points by player for each game. Once we bring all the data in, we’ll need to start thinking about what specific variables we will want to use from the data. Since we are interested in which variables will influence a players fantasy score most, some of the key variables we already identified that we hope to have in our data are: weather conditions, performance in high stakes situations, receiving touchdowns, rushing touchdowns, touchdown passes, receiving yards, rushing yards, injury history, number of targets, number of receptions, number of snaps played, pass attempts, pass completions, completion percentage, interceptions thrown, fumbles committed, and opponent faced. We know that any single one of these variables or a unique combination of these variables affects player performance, the question is, to what degree? Our next step will be to include all these variables in a correlation matrix to see which variables are most correlated with whether a player performed above or below their projected fantasy score. In this model our dependent/target variable is whether the player scored above or below what was projected.

We intend to use a classification model with an 80/20 train/test split and/or a 12-fold k-folds method to determine whether a player will perform above or below the projected fantasy points. Critical to this model is engineering binary features to complement the numeric features we have found, and fine-tuning parameters to make the model account for complexities in the selected features, but flexible enough to take in 2021 fantasy projections and help fantasy managers to decide who to draft, start, trade for, and pick up off the waiver wire.

Our intent is to deploy our trained model to a web application that requests basic user input and takes certain features into account to provide a prediction (which could be taken as a recommendation) about a player’s performance in the coming week.

**Questions or avenues of exploration required for the project**

When seeing a projected fantasy score, how much can I trust it?

What can we learn about specific players using score projections/actuals from the past?

Did the player perform above or below projected score? What are the factors that caused that?

Will we be able to get all our data from one source or will we have to merge data sets?

How will our team successfully navigate this project using different operating systems?

Are there players that perform better at certain points in the season?

How will we define time and instances?