**Data Selection/Data Engineering Write Up**

The goal of a fantasy football draft for each manager is to end the draft having created the strongest team they possibly could. This strength comes in the form of scoring points. The team in fantasy football that scores the most has the highest chance every week to win. To achieve this goal managers will each approach the draft with a personal plan. This plan is commonly built by weighing previous performance (passing, receiving, and rushing statistics), health statistics (age, build, injury history), news analysis, and opinion (whether they like a player/team) against each other. To create an agent to initially mimic and eventually defeat standard human picking (represented as average draft position (ADP)) we collected data spanning back to 2016 on players performance history, health statistics, and news articles.

For player performance we stuck to box score statistics. Things like receiving yards, passing yards, touchdowns, fumbles, interceptions, etc. Our assumption was that these would be the biggest driver in player selection. These statistics either directly affect how many points a player has at the end of a game or can drastically affect their ability to get more points. To get the actual data we found a dataset that contained offensive statistics dating back to 2001. We limited this to 2016 due to data limitations further down the line.

For health statistics we considered a player’s age, weight, height, and injury history. This was the category I expected affect the selection the least, since usually a player’s height and weight don’t change too much over an NFL career, and injuries from an outside perspective seem random. This data was collected from a dataset that spanned back to about 2012, but again was limited to 2016 due to data limitations.

For article selection we used a BERT text analyzer to assess whether an article written about a player was positive or negative. I expected this to be our differentiating factor from other models. This something I’ve only ever heard spoken about on a broad and general basis, and never as a quantifiable fact. I was expecting the model to use a negative article score to be a tie breaker in between players. To collect the article a team member wrote a web scraping script for a website called [The Huddle](https://tools.thehuddle.com/nfl-fantasy-football-player-news/) that held articles dating back to 2001. Again, we limited this data to 2016 and on.

Now that we had built out the portions of our database that would be used for prediction, we had to create a way for the prediction’s accuracy to be checked. The metric we chose for prediction accuracy was total points scored by a team. This meant that we needed to record how many points each player scored in each season. Once an agent selected a team, it could total up all the points scored by every player and that would be the agents score.

The final portion we needed was a way to predict regular people’s picks. To do this we used ADP. This was the big limiting portion for our data. Not only is getting ADP from previous seasons difficult, choosing a single ADP to use is incredibly challenging. With this being such a large industry, many people have their own ADPs. To get around this we used an average ADP dataset dating back to 2017. This was the most accurate data that we could find that was usable. We could use this data to stop the agent from taking the top players at every pick, since this would not reflect how a real fantasy football draft would go.

These different datasets saved us invaluable time to work on the algorithm and other portions of the project but have proven to be a frustrating group to deal with. Many of the datasets we found had obvious holes, issues with formatting (Yahoo ADP value csv files are unusable), or didn’t contain the data we were looking for. I think if given more time to work on this project we would have scraped the data ourselves.