Christian Thieme

DATA 698

DRAFT – Project Proposal

09.12.21

**DRAFT – PROJECT PROPOSAL**

**PURPOSE**

Sports analytics has grown exponentially in the last decade. The success of “Moneyball” type analytics in baseball has led to many other sports attempting to find an edge using analytics. Sports teams around the country, both professional and collegiate, employ full time analysts to try and squeeze out any ounce of improvement for their teams that they can. Another area in the sports arena that has grown tremendously over the last 10 years are fans involvement in “fantasy” leagues. These leagues allow the individual to “draft” players on to their team and then are given points each week based on the performance of the players in the team. The fantasy football following is the largest of all the fantasy sports. While these leagues may sound silly to some, there are millions of people who play and there is **big** money involved. The NFL’s biggest prize is $1,000,000 (Subramanian, 2014).

More and more people are turning from using their gut, to really diving into the analytics of players and teams. In a special kind of league called a dynasty league, participants are allowed to keep their players year after year. That means that a major way to improve your standings each year is to draft the right rookies on to your team as players transition from college to the NFL. One of the positions that can have a huge impact on a dynasty team as a rookie is that of the wide receiver.

Each year, dozens of sports analysts make predictions about who the “break out” wide receivers will be based on having watched their college careers. Often the analysts get the obvious players right, but each year there are also surprises and disappointments. Is there a more quantitative way to determine who these breakout players will be – more specifically, using a wide receiver’s collegiate and NFL combine statistics, can we accurately predict 80% of the wide receivers that will average 10 fantasy points per game (the breakout threshold)? If we can predict these players at the specified accuracy, what are the most predictive factors in the model? Understanding these factors can help analysts understand where they should focus their attention to find standout players.

**RELEVANT RESEARCH**

This is a foggy area. While there are certainly models in production to answer these questions, I have not seen anyone mention more than a “model” and their predictions in articles because there is so much at stake for models that actually work. Additionally, it appears that many models tend to focus on the “breakout” age for wide receivers as opposed to predicting whether a player WILL BE a breakout. There are certainly numerous articles containing predictions from year to year, but none I have seen willing to share more than an r-squared about how their model is working (Hampton, 2020). Additionally, from what I’ve seen, many of these articles focus on simple regression type models (multiple regression) as opposed to using more advanced algorithms. The approach to build a classification model to determine if a player will “break out” appears to be novel.

**DATA**

There is a host of available data for college player’s historical data as well as the NFL combine, however, there are few datasets that combine data from all of the sources and then also compute many of the well-known and lesser well-known metrics. The [data source](https://docs.google.com/spreadsheets/d/1uVcLovsFMHJvHVqqFAGTW2hTBuiCMI5qHE7RlxOkntk/edit#gid=1166725717) I found that is the most complete and contains hundreds of variables for each player is maintained by Dave Wright who is a well-known author at DynastyLeagueFootball.com (Wright, n.d.). He provides this dataset free of charge. It contains player data from 2018 forward and is broken down specifically to wide receivers (see WR tab). There are 729 rows (players) in the dataset and 587 columns. The dataset contains both numeric and categorical data.

**METHODS**

To determine which receivers are “breakout” wide receivers, I’ll need to build a classification model. This model will be trained on the college and NFL combine stats of current and past NFL players who were both “breakouts” and non-breakouts (these will be labeled based on my criteria). I’ll withhold a portion of this data as the testing set and run the final model against this test set to see if we can achieve 80% accuracy. Many different classification algorithms will be attempted to determine the most appropriate/accurate algorithm to use and ensembling will also be explored. Finally, I’ll use the most accurate model (before ensembling) to view which factors are most important to the predictions of the model.

**POTENTIAL FEATURE ENGINEERING**

One variable that is difficult to measure but is important that I don’t see in this dataset is *Opportunity*. A great player may get drafted on to a team that already has a star wide receiver. If that player has little opportunity, they may not have much potential to break out. However, if a player comes in and gets to start, they will have a lot more opportunity to break out. If time allows, I’d like to dig in to how to quantify this and add it to the model to see if it’s a meaningful factor.

**REFERNCES**

Hampton, B. (2020, May 24). *Blake's Predictive 2020 Wide Receiver Model*. Retrieved from The Undroppables: https://www.theundroppables.com/blakes-predictive-2020-wide-receiver-model/

Subramanian, P. (2014, August 29). *The 3 biggest prizes in fantasy football*. Retrieved from Yahoo Finance: https://www.yahoo.com/entertainment/news/the-3-biggest-prizes-in-fantasy-football-174712081.html

Wright, D. (n.d.). *DAVID WRIGHT*. Retrieved from https://dynastyleaguefootball.com/author/spacemnspiff207/: https://dynastyleaguefootball.com/author/spacemnspiff207/