Aaron Drexler

DSC 680

Term Project Proposal 3

**Topic: NFL Combine**

I want to find the correlation between NFL combine data and that of their NFL draft positioning. In other words, how well does the NFL combine data predict draft position?

**Business Problem**

The focus of this research is on the NFL combine. For those who do not know, the NFL combine is a showcase of skills and measurables for incoming rookies coming to play in the NFL. At this week-long event, players go through measurements, common workouts, and positional drills to showcase talent and skills to the NFL teams looking to draft them. This is a prestigious event that most draft prospects take part.

Now, some take big stock in how players perform in the combine while others choose to rely on game data and previous scouting. For my project, I want to investigate what impact the combine truly has on the draft. What level of relationship does the combine performance data have on a prospects draft position? How much emphasis should we put into how players perform in the combine?

**Data Sets**

For my data, I am going to use ta Kaggle database with the combine results and the draft performance of those prospects from 2009-2019. I plan to use the data from the following variables in the dataset: player, year, age, school, height, weight, draft position, 40-yard dash time, Vertical Jump, bench press, broad jump, 3-cone drill, Shuttle, BMI, position, and position type.

**Methods**

To start off, I plan to use random forest to model my data and determine the accuracy of the data. I will start off with a summary of the data and look at each variable and their correlation to one another. I will then run a series of random forest models, looking at overall relation, then by player type then by position, looking how the measurables relate to the drill scores, and finally to the draft position.

Now, I will not just run the random forest once. I plan to look at several different categories. I plan to look at the whole group of players, then I plan to group by position type and then by position itself. I feel like the more specific the position, the more accurate the results will become.

**Ethical Considerations**

The biggest consideration for the ethics of the data is the lack of non-quantitative data that is proprietary and collected by teams. On drills that are position specific, there is no way to rate the players in the proper context. For example, QB throws to receivers would be something that I cannot quantify in this study, or the footwork of the cornerbacks as they cover the receiver. This data that is being compared is only half (if that) of the data teams collect at the combine, not to mention medicals and interviews that are conducted at the combine.

**Challenges/Issues**

One of the biggest issues will be parsing up and cleaning the data. Several players skip certain drills, either because they are the wrong position, they are injured, or they choose not to show that skill. There will be several null data as a result. In addition, because of all the different models I want to run, the data set will be split and shaped and modeled a bunch of different ways, which leaves open the chance of data corruption or wrong data set use.

**References**

https://www.kaggle.com/datasets/redlineracer/nfl-combine-performance-data-2009-2019