Group: Godzilla Buttkickers

Summary

Our Raw Data is going to consist of NCAA Player career stats along with seasons played (1-3) and total Approximate Value over the players' first three pro seasons. The Approximate Value is a figure from 1-20 per season generated by [www.pro-football-reference.com](http://www.pro-football-reference.com/) based on the player's performance that NFL season. We plan to analyze **which NCAA stats (more likely weighted combination of stats) correlate best to Approximate Value** (overall performance) in the NFL in their first three seasons for Quarterbacks, Running Backs, and Wide Receivers. We're going to do a LEFT JOIN, meaning we will keep NCAA players who did not make it to the NFL for the purposes of comparison, but we will discard NFL players who are not represented in the NCAA set (older NFL players, or came from outside the USA). We will discard players who did not finish their NCAA career and did not join the NFL.

This experiment is involving an entirely **supervised** model. That is we expect positive correlation between stats and performance in the NFL, and we are looking to see which correlation is strongest. We are not fishing for correlations that were not set out in the first place.

Why just the first 3 years? This is to control for career length. It becomes difficult to compare a 10 year QB against a 1 year QB; at that point it isn't feasible to take the rookies' AV for his first year and extrapolate the same performance to a 10 year season.

What is Approximate Value and why are we using it? We need a concise and sufficiently accurate indicator of player value in the NFL. It controls for team strength, so we can compare players man to man. Here's a detailed explanation of the Approximate Value:

<http://www.pro-football-reference.com/blog/?p=466>

What will our work enable us to do? For example, we'll be able to address something like “For a Wide Receiver, does NCAA Yards Received correlate strongly to player performance?” If not, we must examine things like Conference strength, players' college team strength, etc. If it does, we should weight it heavily when we make an aggregate linear model of all the stats to player performance.

Data Pruning Techniques Used

There are multiple Excel files to minimize the size of each file. During our data collection, Excel calculations crashed if we put all of our data in a single spreadsheet. The NCAA raw data came in csv files for each year. The file had one row per player per game. In order to get combined career statistics I had to first combine the statistics for each player’s season and then combine those season totals to get the career totals. I had to do this in several steps due to the limitation of computing power from Excel. After downloading the 12 csv files from ncaa.org, I made each one a worksheet in the NCAA Raw Excel file.

For each year, I removed the columns that weren’t necessary. These included the special teams statistics (e.g. punt and kick return stats), as well as total offense yards because they were redundant. Since the NCAA files didn’t come with a primary key to identify each unique player over several years, I created one by concatenating in order and delimiting with carats (“^”) the player’s last name, first name, uniform number, school ID number, and year. I added this new ID to a new column and then used Excel’s subtotal function to add the stats together for each player to get a season total. I then copied this new information into the NCAA All Year Total file, where each worksheet is a year from 2000 to 2011.

Then I combined all of the season totals for each player by removing the year from the name identifier and using Excel’s subtotal function to generate career stats. Before moving forward with this step, I spot checked some names in the data set to verify that players would be grouped together properly. The NCAA data doesn’t account for if a player transferred schools between years. In that course, our method would identify that player as two unique players. Our team is working on how to fix this issue.

Then I combined the two large data sets by using VLOOKUP in Excel to match player names in the NCAA data set to those in the NFL data set. I added the AV and position columns in the combined final set in the Final Data Set File.