Analyzing 2015 NFL Quarterback Prospects

Dan Loman

May 6, 2015

1 Introduction

As NFL draft season rolls around once again, I can't help but wonder why, for as long as I can remember, my Buffalo Bills just can't seem to find the right quarterback to lead them to the playoffs. Every year quarterbacks are the most hyped prospects in the draft - they're scrutinized, picked apart and evaluated more closely than players at any other position, yet the biggest challenge for NFL teams is finding one who can consistently lead them to victory. This year I decided to dig more closely into quarterback prospect data to see if I could gather any insight on such an unpredictable system.

In the past I've done projects where I use supervised machine learning to make predictions, and my original goal for this project was to project how successful each of the newly drafted quarterbacks would be. However, I believe in most cases a quarterback's success is only partly explained by their individual talent and ability, and is also widely dependent on the situation they land in. Thus, instead of attempting to tackle the impossible problem of forecasting these prospects' career success, I turned to focusing on how they stack up to previous draftees, and who they are most similar to.

I've summarized my results on my blog: *www.greenandgoldanalytics.blogspot.com.*

2 Data

In this study I use a quarterback's combine and college stats as a basis for comparison. I ended up using 15 raw stats:

Height - QB height (in)
Weight - QB weight (lbs)
Wonderlic - QB Wonderlic score
Forty - QB 40 yard dash time (sec)
Pass Attempts - Number of pass attempts in college
Completion Percentage - Completions per attempt
Pass Yards per Attempt - Passing yards per attempt
Pass Touchdown Rate - Pass TD per pass attempt
Interception Rate - Interceptions per attempt
Passer Rating - College passer rating
Rush Attempts - Number of rush attempts in college

Rush Yards per Attempt - Rushing yards per attempt
Rush Touchdown Rate - Rush TD per rush attempt
Pass/Rush Ratio - Pass attempts per rush attempt
BCS - 2 for BCS school, 1 for FBS (non-BCS), 0 for FCS

3 Data Acquisition

I scraped data from three places to get what I needed:

nflcombineresults.com - For combine data
http://www.nfl.com/draft/history - For all drafted quarterbacks
http://www.sports-reference.com/cfb/players - For college stats

All scraping was done in Python using BeautifulSoup and urllib2. My methodology was this: Find all quarterback combine data, which gave me a list of all quarterbacks who participated in the combine since 1999. Then, use only those who were drafted and find their college stats, and merge the two tables together.

Along the way I had to inspect for players who didn't make it onto the list because of a name mismatch issue (i.e. 'AJ McCarron' vs 'A.J. McCarron') or a stats issue (sports-reference only has FBS data). This was corrected in the data preprocessing stage.

4 Data Preprocessing

Probably the most difficult of this project was dealing with missing data and getting the data to a usable format for analysis. In the previous section I noted how some players didn't smoothly make it to my final stats table. To correct this, I removed the 'bad names' from my names list and then manually inserted their stats afterward. There were around 20 quarterbacks I had this issue with.

I also converted raw college stats to per-attempt rates. For example, I used Yards/Att, TD/Att and Int/Att instead of total yards, touchdowns and interceptions.

I also had to deal the the issue of missing combine data. Since many players don't participate in certain combine events, or only have unofficial scores, there was a good amount of missing data here. I was able to fill in most of the lost 40 yard dash times by researching unofficial times or estimating values for injured players (like Zach Mettenberger), and I replaced the 20 or so missing Wonderlic scores with an average score of 25. However, other combine stats such as vertical jump, broad jump and bench press were missing in around 25% of players, and I didn't feel comfortable simply assigning those players an arbitrary value due to high diversity among players. If a statistic was missing too much data, I ended up not including it in my final analysis.

Finally, before analysis, I added a few columns (BCS and Pass/Rush Ratio) and normalized the data.

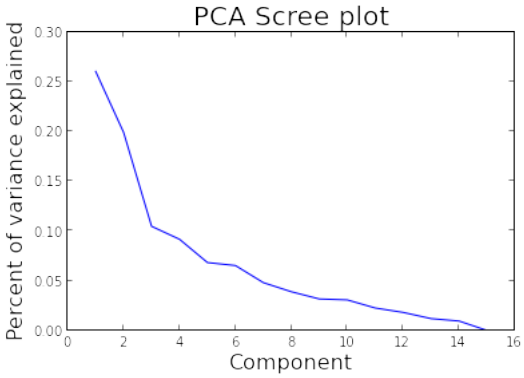
5 Player similarity

To compute player similarity for 2015 prospects I used a euclidean distance metric on all of their stats, which looked like the following:

$$distance = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2} \tag{1}$$

The problem with simply doing this on the normalized data, however, is that it puts equal weight on each stat, so stats that are correlated (like passer rating and yards per attempt) throw off the study because you're accounting for essentially the same thing twice. To solve for this issue, and to remove multicollinearity, I ran a orthogonal factor analysis on my data.

To see how many factors I should use, I ran a Principal Component Analysis and examined the scree plot:



The first four components explain almost 2/3 of the variance in the data and there is a sharp dropoff after the fourth component in my scree plot, so I felt comfortable using four factors to represent my data. It turns out that these four factors can be interpreted well with underlying trends in the data. They look as follows:

Factor 1: Highly correlated with Passer Rating, Yards per Attempt, Completion Percentage and TD rate, negatively correlated with INT rate. I call this the "Passing Efficiency Factor".

Factor 2: Highly correlated with Rushing Attempts and Rush Yards per Attempt and low 40 Time, negatively correlated with Pass/Rush ratio. I call this the "Rushing Factor".

Factor 3: Highly correlated with Passing Attempts and Rushing Attempts. I call this the "Experience Factor".

Factor 4: Highly correlated with Height and Weight. I call this the "Measurables Factor".

To achieve my final results I re-ran my distance metric on each player's factor score instead of their normalized raw data.

6 Data Issues

While I'm proud of what I accomplished with this data that was available to me there is one huge, massive problem with this study which makes it impossible to draw any solid conclusions from this project: the data has no context. There's a reason Bryce Petty was drafted in the 4th round despite having a similar statistical profile to the top prospects, and why Tim Tebow was seen as a first round reach despite being one of the most productive college QBs of all time. That's where scouts, people who watch the actual games come in, to provide context to the data and give additional information that isn't captured or is misrepresented in the stats. The underlying assumption behind college stats is that they correspond to natural ability that should translate to the next level, but that assumption is weak. My data doesn't know if stats were compiled because of or in spite of the system the quarterback comes from, what kind of teammates they were playing with, or what kind of intangibles the quarterback may bring to the table. So if I ever decide to revisit this study and improve it, I'd have to get my hands on some quantifiable scouting data to separate the Pettys from the Winstons.

In my blog I took some of this into account by reporting only similar players who were drafted in similar rounds. For example, Marcus Mariota's top comparison was actually Josh Johnson out of San Diego, but I didn't report it because in reality, Mariota is a far superior prospect. Same goes for Jameis Winston and Ingle Martin. Overall though, players were compared most favorably to players drafted in similar positions.

Another more minor data issue is inconsistent combine stats. I mentioned previously that I didn't even include a lot of combine stats in my model because there were too many missing values, but that doesn't even take into account the amount of players that skip the combine all-together (or at least never made it onto my data source). I was able to get just about every 1st, 2nd and 3rd round pick into my study but there are plenty of late-round QB's that aren't included because I don't have data on them. I believe this skews my results by making the truly spectacular quarterbacks look less spectacular, since they show up as being less rare in my data.

7 Clustering/Visualization

To visualize my results I use a scatterplot in D3. I show the Passing Efficiency Factor on the x-axis and the Rushing Factor on the y-axis. Since I use 2 additional underlying factors that aren't shown in this plot, I ran a K-Means cluster analysis with 5 clusters on the factor scores. Players with similar profiles are clustered in the same group and are represented by the same color. Size of player bubbles represent how high they were drafted.