Predicting Collegiate Quarterback Success

Ted Henson

Ryan Thornburg

MEJO 570

2 May 2019

Narrative Anecdote

Quarterback has always been the most important position in football. College coaches are increasingly being evaluated on how they develop and recruit their quarterbacks; moreover, they are now having to recruit quarterbacks knowing they may transfer after one or two seasons (Justin Fields, Tate Martell). Projections have not improved despite the increased emphasis on the position: three stars and walk-ons (Baker Mayfield, Gardner Minshew) are still exploding onto the scene in the NCAA. At the MIT Sloan Sports Analytics conference, I heard offensive guru Mike Leach, coach of Garnder Minshew, state: “I’ve never taken a quarterback that was inaccurate and made him accurate.” This is in stark contrast to what many people believe: that you cannot teach arm strength. Data analysis could evaluate the validity of these types of statements. Most people would take the stance that predicting the success of a quarterback is a crapshoot; however, using data gathering and analysis, a collegiate coach could grab the next Baker Mayfield.

Data gathering was both straightforward and taxing at times. R code has already been written to scrape 247 recruiting rankings for quarterbacks and for MaxPreps passing statistics. 247 is a recruiting service that evaluates all high school football players and gives them a rating. MaxPreps collects high school football statistics. The 247 dataset includes information regarding a prospect’s rating, height, weight, and hometown. This 247 scraping code could be altered, or additional code could be written to scrape each player’s individual 247 profile. This would obtain additional information including scout’s detailed breakdown of a prospect’s strength and weaknesses on a

scale of 1-10 and a player's combine data (40-yard dash time, 20-yard shuttle, etc.). Using IMPORTHTML in Google Sheets, this profile information can be obtained. Here is a sample of Trevor Lawrence's high school statistics and scouting breakdown from 247:

Year	PaAtt	PaCmp	PaYd	PaTD	Int	RuAtt	RuYd	RuTD
2017	227	161	3148	40	1	30	200	2
2016	372	236	3900	51	9	-	-	-
2015	364	233	3655	43	4	-	-	-
2014	310	187	3042	26	7	-	-	-
TOTAL	1273	817	13745	160	21	30	200	2
- Size *10*								
- Intangibles *9*								
- Pocket Presence *9*								
- Accuracy *9*								
- Reactive Quickness *8*								
- Feet *8*								
- Delivery *8*								
- Arm Strength *8*								

His statistics have already been gathered so that portion can be ignored, but the scouting breakdown could be valuable in the analysis if time permits. This example could be easily replicated by substituting Trevor Lawrence's URL with other players' using google sheets, another program, or both.

There are a lot of unknowns as far as what can and cannot be predicted for a given player. The largest issue with predicting collegiate success is the variation of competition in college. Finding a relationship between a player's recruiting ranking and his completion percentage, number of touchdowns, or yards may be difficult if quality of collegiate opponent is not considered. One possible approach would be to use ESPN's FPI to scale a player's numbers up or down based on his conference. The more

encompassing and robust option would be to consider predicting a player's ESPN QBR, which heavily incorporates quality of opponent. Gathering this information was a copy and paste job. Here are last season's leaders in ESPN QBR:

RK	PLAYER	PASS EPA	RUN EPA	SACK EPA	PEN EPA	TOTAL EPA	ACT PLAYS	RAW QBR	TOTAL QBR
1	Kyler Murray, OKLA	109.6	42.5	-9.6	0.6	143.1	553	93.6	**95.4**
2	Tua Tagovailoa, ALA	87.5	10.8	-9.7	-0.1	88.5	438	89.6	**93.1**
3	Jake Fromm, UGA	64.3	-4.2	-10.9	1.4	50.6	374	76.9	**85.1**
4	D'Eriq King, HOU	65.9	35.5	-9.9	4.4	96.1	504	87.6	**84.8**
5	Dwayne Haskins, OSU	106.4	-2.7	-11.4	3.3	95.3	673	77.5	**84.8**
6	K.J. Costello, STAN	83.4	0.8	-16.3	8.4	76.4	522	78.0	**82.6**
7	Drew Lock, MIZ	63.2	11.2	-11.1	7.8	70.8	552	75.8	**82.0**
8	Shea Patterson, MICH	49.3	9.4	-13.3	3.8	49.0	434	72.7	**81.5**
	Trevor Lawrence, CLEM	62.4	-0.4	-6.6	0.5	55.9	486	74.2	**81.5**
10	Will Grier, WVU	89.1	-1.1	-16.1	-0.1	71.8	485	79.1	**81.0**

ESPN has calculated this metric for every Quarterback in division 1-A since 2004. This large amount of quality data will allow for seamless integration and modeling with the 247 and MaxPreps data. This process has already been accomplished in PostgreSQL for quarterbacks dating back to the 2014 season. Below are the first four rows of the combined dataset, with the primary variables of interest shown:

Player	Total QBR	247 Rating	MaxPreps yds/g
Kyler Murray	95.4	0.9855	294.7
Tua Tagovailoa	93.1	0.9843	258.3
Lamar Jackson	85.4	0.8788	142.7
Jake Fromm	85.1	0.9794	391

There are some quarterbacks who played in our sample seasons that are not in our overall dataset; this is due to missing information. For the preliminary analysis, omitting

these players will the best approach. In the future one could explore using imputing methods such as k-nearest neighbor imputation. This would allow the models to keep players with only either recruiting data or MaxPreps data. Delving into why these players have missing information may be another avenue to explore. A wide variety of regression methods could be considered to model this data. The CARET package in R provides an easy way to incorporate multiple types of models and generate predictions.

So What

The skeptic would wonder how college football has not tried every method to predict quarterback success, and why this project would be any different. There have been several papers and articles attempting to do this in the NFL; however, none of them have attempted to consider using high school data to predict performance, collegiate or otherwise. FiveThirtyEight recently published an article attempting to predict NFL quarterback success with mixed results. If their audience is interested in NFL quarterback success, then they would certainly be interested in predicting collegiate success as the concepts behind the projections are analogous. This analysis could hold value on its own at the collegiate level, and potentially improve upon their previous methodology.

If collegiate performance has predictive value for the NFL, then there is reason to believe that a player's high school numbers have bearing on his collegiate career. Starting this analysis in college will allow for a larger sample size; there are over 100 FBS 1-A programs, many of which are multi-million-dollar programs that would kill to get every possible advantage over their rivals. Smaller programs may even be more interested in this analysis as they are not able to easily grab the concensous top

prospects. Many teams have won national titles largely due to tremendous quarterback play (Cam Newton, Jameis Winston, Vince Young). Even the smallest increasing of the odds of grabbing a better quarterback has instant ramifications for a program and coach's future. Interested football organizations would be willing to pay for our models and data if they proved value. Creating a story with direct financial implications for certain individuals will generate loads of curiosity. Diehard football fans will also enjoy the analysis to evaluate their own team's decision making.

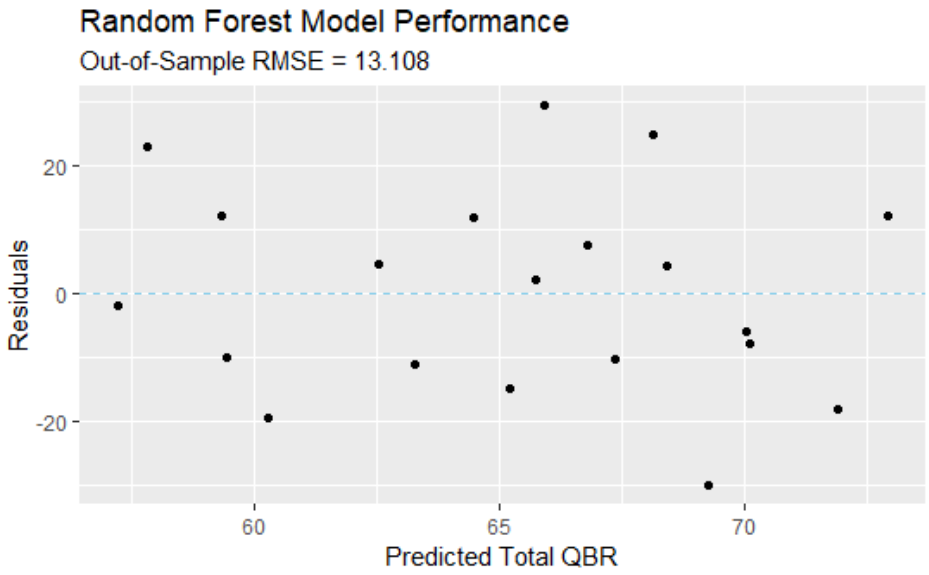
Data Analysis

The data was joined in PostgreSQL using a left join, keeping all observations in the ESPN dataset, and adding in data from MaxPreps and 247. See the attached SQL file for details. Only a player's final high school season was used in the joining of the MaxPreps data. A common problem with football is small sample size. This will be mitigated due to the large number of seasons and players in all three datasets. A lot of the time consumption that comes from modeling high school data is gathering it and cleaning it. Most of that legwork has already been completed; code would simply need to be executed for each requested season.

Modeling in R was the most practical choice as the CARET package streamlines the regression training process. Four model types were trained: ridge regression, lasso regression, neural network, and random forest. Additionally, a linear stacked model was built using predictions from the previous four. The variables used to predict a player's college total QBR were his 247 recruiting rating and high school statistics: yards per game, total touchdowns, total interceptions, completion percentage, passer rating, total completions, total yds, and total games played.

Below is a table of each model's out-of-sample root mean squared error and the plot of residuals versus the fitted values for the best model:

Model Type	Out-Of-Sample RMSE
Stacked Model	14.46
Bayesian Ridge	13.20
Random Forest	13.11
Neural Network	14.15
Bayesian Lasso	13.27



On average our best model predicted a player's QBR within about 13 points as shown by the RMSE. This performance is somewhat mediocre, but this will improve with considering more seasons of data, including more inputs, training more models, and applying transformations to the inputs. At the very least, the models show that there is a relationship between the high school data and a quarterback's college performance.

See the attached R scripts for details regarding the modeling and web scraping; Modeling.R models the data, 247 football.R scrapes one season of quarterback recruiting data from 247, and maxpreps football scraper.R scrapes one season of high school statistics for quarterbacks from MaxPreps.

Key Documents

The key documents are the data from MaxPreps, 247, and ESPN that have been collected. It may be beneficial to consider other works of research to decide what variables to focus on. FiveThirtyEight wrote an article predicting NFL quarterback success which found that completion percentage, especially when adjusted, was the most predictive variable. Brian Burke at ESPN also found something similar in his analysis of current NFL quarterbacks. This may lead one to consider variables more related to accuracy in the analysis.

Key Sources

In order to fairly evaluate how collegiate coaches currently make scholarship decisions, speaking with someone in the industry would go a long way. Ben Weiss, founder of Zcruit, a college football recruiting service, may have some insight into how quarterbacks are evaluated. Ben's email is ben@zcruit.com. A lot of people believe coaches ignore a player's recruiting ranking or high school statistics; however, after speaking with Ben a few times, he informed me that college coaches pay thousands of dollars to get rankings and evaluations from special scouting services. His company even started paying a contractor to gather player's combine data. College coaches and scouts are already using high school data in their decision process; this project would

simply use that data in a more rigorous and mathematical sense, which could potentially improve decision making. As Ben has already been a helpful resource in preparing for this project, he will certainly be helpful in the future.

Bibliography

ESPN, ESPN Internet Ventures, www.espn.com/college-football/qbr/_/year/2015.

"2015 Top Quarterback Recruits." *247Sports*, CBS Broadcasting Inc., 247sports.com/Season/2015-Football/CompositeRecruitRankings/?InstitutionGroup=highschool&PositionGroup=QB.

Burke, Brian. "DeepQB: Deep Learning with Player Tracking to Quantify Quarterback Decision-Making & Performance." www.sloansportsconference.com/wp-content/uploads/2019/02/DeepQB.pdf.

Friscojosh. "The NFL Is Drafting Quarterbacks All Wrong." *FiveThirtyEight*, FiveThirtyEight, 27 Feb. 2019, fivethirtyeight.com/features/the-nfl-is-drafting-quarterbacks-all-wrong/.

"National Football Stat Leaders." *MaxPreps*, CBS Broadcasting Inc., 1 Jan. 2016, www.maxpreps.com/leaders/football-fall-15/offense,passing/stat-leaders.htm