Who's on Draft?

Predicting NFL Quarterback Selections

Master's in Big Data Analytics | University of Michigan, School of Information

Harrison Hall

Background

- Most college QBs enter draft day uncertain about if they will be selected.
- Predicting if and when a QB will be drafted into the NFL can help players understand their potential and decide if they are certain about a career in American Football

Data

- Player performance statistics of 4,300 QBs from 2004-2021
- Historical draft records from 2008-2021
- All records are focused on Division-1 colleges and universities in the Football Bowl Subdivision.

Player Stats

- Touchdowns
- Pass attempts
- Pass completions
- Yards per pass attempt
- Rushing yards
- Interceptions
- Number of years played

Citations - Data Sources

- sportsdata.io
- collegefootballdata.com
- sports-reference.com

Research Question

Can Machine Learning models be used to predict whether any given college QB will be drafted?

Results

- Of the Big-10 QBs, only Sean Clifford (currently, Penn State) would be selected in the 2022 NFL draft.
- **Currently, Cade McNamara is predicted to not be drafted**
- Note on limitations: the results will change with more years of data
  - Number of years played is a factor
  - 2021 is only Cade's 2nd year recording QB stats.
- **List of College QBs who are predicted to be drafted in 2022:**
  - Kenny Pickett (Pittsburgh)
  - Bailey Zappe (Western Kentucky)
  - Carson Strong (Nevada)
  - Sam Hartman (Wake Forest)
  - Brennan Armstrong (Virginia)
  - C.J. Stroud (Ohio State)*
  - Bryce Young (Alabama)*

*Not eligible until 2022

Future Work

- Account for current NFL team needs
- Include data on players from FCS schools
- Analyze separately rush-oriented versus pass-oriented QBs
- Dynamic weighting based on teammate skills

Modeling

- Supervised Classification models
  - Random Forest
  - K-Nearest Neighbors
  - Naïve Bayes
- Random Forest is best
  - Accuracy: 93.3%
  - Mean absolute error: .067
  - Cross-validation score: 0.9173

Insights

- Predicted whether each of the starting QBs in the Big 10 this year will be drafted
- Surprisingly unimportant features:
  - School/conference
  - Player height

Contact Information

- **Email**: HarrisonJosephHall@gmail.com
- **LinkedIn**: https://www.linkedin.com/in/harrison-joseph-hall

Who's on Draft?

Predicting College Football Quarterback Drafts

Harrison Hall

Abstract

No quarterback (QB) in college football is truly confident in whether they will be drafted to play in the National Football League (NFL). Historical draft patterns should be utilized to evaluate the likelihood of any given college QB to be selected to play in the NFL. This project serves to shift the narrative of data-driven football analysis away from determining the success of a player in the NFL towards identifying the threshold of entry into NFL in the first place.

1 Introduction

While modern data scientists and statisticians provide ample informational support for predicting the success of QBs in the NFL, there is a serious lack of similar research conducted on the barrier to entry into the NFL.

For my project, I developed a classification model to predict whether any given college QB will be drafted into the NFL. The model draws from historical college QB passing statistics, player demographics and the number of years in which the QB has competed in college.

I structured the development of my project following a step-by-step approach, following a series of milestones described in more detail in the Work Plan section near the end of this document. First, I gathered all the necessary data for my project, which is detailed in the Data section. Second, I expanded upon the data from various sources by drawing connections based on unique player ID values, before conducting correlation tests to determine the most relevant variables to predicting whether the QB would be drafted. Third, I tested 8 different models to find the most appropriate classification model for the project. Finally, I applied the model to a series of relevant data to predict whether each of the starting QBs in the Big 10 division will be drafted next year.

Football fanatics would likely care about this model most, as it can help fans feel more, or less, confident in the futures of their favorite players. While the model would not provide significant utility for the fans, it could at least have an emotional impact on the fans. For example, if the model found a QB from Michigan is likely to be drafted next year, then Michigan fans may be more inclined to purchase apparel with that QB's name on it, or to more actively and passionately support that QB.

Additionally, individual college football QBs and their coaches would likely care about this model because they can gain a better sense of the QB's likelihood of being given a chance to make a career in professional football.This knowledge is important for college QBs to decide whether they should expand their skill-set beyond football or prepare for a career in the sport. Additionally, they can better understand which performance indicators they should focus on improving to increase their chances of being selected in the upcoming NFL draft. Finally, the model might help certain QBs feel more confident in their likelihood of being drafted because the model finds that certain features, which are more difficult for a QB to alter than simply performing better on the field, are less important than indicators on the field. For example, a QB from a small school without a historically strong football program may feel more hopeful about their prospects for entering the NFL after interacting with this model because the model would not consider the school in which the QB attends.

My primary goal for building this model is to use as a talking-point in interviews with sports-related businesses. I would be incredibly excited for an opportunity to work as a data scientist for an American football organization, and I believe this project can help me get my foot in that door by

applying advanced machine learning techniques to the football industry. While I would like to work for a specific NFL team, I have seen many interesting predictions and statistics provided by Amazon's AWS in the last few years. I would also hope to share my experience working on this project with the sports department within the AWS organization.

2 Problem Definition

My research question is as follows, "Can Machine Learning models be used to predict whether any given college QB will be drafted?"

This project predicts whether a particular QB will be drafted into the NFL based on historical draft patterns and player demographics. While collecting data and interpreting football statistics is simple enough, there is a serious gap in the ability to apply that knowledge to questions about post-college life. Most years see two or three stellar QBs, who are nearly guaranteed to be selected for the NFL professional draft. However, the vast majority of QBs do not have this privilege.

This project is specifically focused on players at the QB position from Division-1 colleges and universities with American football programs in the "Football Bowl Subdivision". This project helps answer the question of, "Will this QB be drafted next year?".

This project does not attempt to predict which team a particular QB may be drafted to. Additionally, the algorithm is agnostic about whether any given NFL team actually needs a QB or not.

Success of the project has been determined based on the algorithm's ability to accurately predict true-positives (whereby the algorithm predicts the QB will be drafted, and the QB was actually drafted) and true-negatives (whereby the algorithm predicts the QB will NOT be drafted, and the QB was actually NOT drafted). This metric will also be modified by the false-positive and false-negative rate; i.e., how frequently the algorithm makes an incorrect prediction.

3 Data

This project uses data on QB statistics since 2004. This data came from the CFBD API, which was generated by the Swagger Codegen project [1]. This data includes information about college football QBs from colleges and universities in the "Football Bowl Subdivision" between the years 2004 and 2021. Additionally, the data from the CFBD API is combined with historical NFL draft records of QBs, between the years 2008 and 2020. The first year is set at 2008 so that all 4 years of a QB's college football career are included in the analysis. The source for the historical draft rankings is public knowledge, but this project scraped the information from DraftHistory [2]. The third and final data source came from SportsReference, which I used to collect player demographics [3]. Specifically, I used the SportsReference API (free trial) to gather the height (in feet and inches) and weight (in pounds) for each individual QB.

Some extensive pre-processing of this data was required. For example, the data from the CFBD API only included information on QBs from the "Football Bowl Subdivison" and not the "Football Championship Subdivision". Accordingly, ten NFL QBs are not included in the analysis: Trey Lance (North Dakota State), Ben DiNucci (James Madison), Easton Stick (North Dakota State), Kyle Lauletta (Richmond), Carson Wentz (North Dakota State), Jimmy Garoppolo (Eastern Illinois), Brad Sorensen (Southern Utah), John Skelton (Fordham), Keith Null (West Texas A&M) and Josh Johnson (University of San Diego). Furthermore, Terrelle Pryor has been excluded in the dataset because he was drafted in the 2011 "Supplemental draft" instead of the 2011 annual "NFL draft".

Furthermore, without access to the dictionary used by SportsDataIO to connect player names with their unique "player ID" values, I needed to manually recreate the links to lookup each QB in their associated database. This time-consuming process required extensive fine-tuning of a combination of basic string concatenation techniques with RegEx for manipulating regular expressions. Upon inquiring to SportsDataIO about receiving access to this dictionary, they responded that I need to purchase a premium subscription for such access, which was not an option for the purpose of this project. Given the manual URL creation process - which was foundational to the gathering of player demographic information - the data needed to be reduced in size and future research should develop a more sophisticated system for ensuring

[1] collegefootballdata.com

[2] drafthistory.com

[3] https://sportsdata.io/developers/api-documentation/ncaa-football

integrity when querying for player demographic information.

There are a total of 1,317 QBs in the dataset, with 126 QBs having been successfully drafted into the NFL, ranging from the years 2004 to 2020. I intentionally excluded QBs in 2021 because the 2021 season has not yet concluded, so their stats are not finalized nor have they even been given an opportunity to be drafted. However, I saved the stats on the current QBs in 2021 to provide insightful and relevant predictions for how current QBs may fare if an NFL draft were to happen tomorrow.

For each player, the data initially reported on the following variables for each year of play in college football: touchdowns, interceptions, passing yards, attempted passes, completed passes, pass completion percentage, yards per pass attempted, and the name of the QB's college football team, the QB's height (feet and inches), and the QB's weight (pounds). Every QB is assigned a unique player ID value.

I expanded on the initial dataset into the following features for each individual QB: number of touchdowns, interceptions, passing yards, attempted passes, completed passes, pass completion percentage, yards per pass attempted, the QB's weight (pounds), the QB's height (feet and inches), the number of students enrolled at their college, the number of years in which they appear in the dataset - which translates to the number of years in which they have actively competed in college football, with more than 5 pass attempts - and a series of Dummy variables representing the conference in which their college competes. Some examples of conferences are, "Big Ten", "Pac-12", "SEC" and "C-USA".

Here I will provide a brief synopsis of how I gathered the new features. I explained above how I extracted the QB's height and weight features with the help of the SportsDataIO API. Because the dataset included statistics on each QB for every year, there were many players with multiple years worth of stats; 4,566 players had at least 1 year of stats recorded within the dataset. Eventually, this fact will prove useful as an important correlative feature within the predictive model. I manually inputted the number of students enrolled at a college and the conference in which the college competes. While the manual input process for college enrollment numbers and the conferences in which each college plays is not ideal and risks misreported information, this is not an issue because the data was eventually removed from consideration in the feature selection phase.

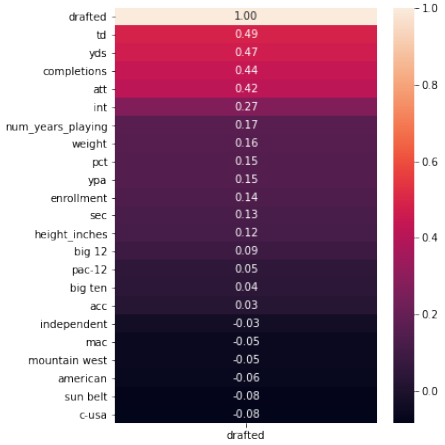

Figure 1: Heatmap of correlation coefficients between features and 'drafted', sorted.

To narrow my large (26) set of features before developing any models, I analyzed the correlations between each feature independently with the dependent variable, which was the boolean of whether the QB was drafted into the NFL. I used the built-in Pandas feature, file.corr() to craft a correlation matrix, and visualized the correlations as a Heatmap with Seaborn. Figure 1 demonstrates the outputted correlation matrix sorted by the correlation between the feature and the 'drafted' values.

Based on the correlation coefficients, I decided to limit my features to only those with a correlation coefficient greater than 0.15. As such, the relevant features for my model are the number of touchdowns ("td"), the number of passing yards ("yds"), the number of pass completions ("completions"), the number of pass attempts ("att"), the number of interceptions thrown ("int"), the weight of the QB ("weight") and the number of years in which the QB has competed in college football ("num_years_playing"). Although the dependent variable, "drafted" is still included in the dataset at this time, it was be removed from the training set before any modeling commences.

4 Related Work

While professional football predictions are a hot topic, especially amid the recent visibility of AWS "Next Gen Stats", most predictive models currently published and circulating in academia

are focused on answering the following question: "How likely is this college QB to succeed in the NFL?". While this is certainly an important question to ask, there are innumerable X-factors which make crafting an accurate prediction nearly impossible. How can these models account for a QB's risk of falling to injury? How does the model incorporate the effectiveness of other players on the team, such as linemen designed to protect the QB, or receivers who need to complete the passes?

In contrast, my model is specifically designed to cease making predictions once the NFL draft has occurred. My model purely intends to predict whether a particular QB will be drafted in the upcoming NFL draft. My project does not extend into the success of any given QB once they have entered the NFL.

Furthermore, most currently-published predictions are informed by expert opinions and subjective evaluation metrics, not by the historical correlations between the statistics included in my data and their respective draft results.

In light of the aforementioned significant difference between my work and other published works, there is one particular study with another purpose which should be referenced here. This relevant study proposes utilizing Euclidean distance, Spearman rank-correlations and a custom "dollar misallocation score" in evaluations of NFL mock draft predictions (Caudill et al., 2014). Any or all of these metrics could be used to help test the accuracy of my model. However, in contrast to my project, this study is limited in scope to the evaluation of NFL mock draft predictions; it is not intended to create new draft predictions nor is it focused on QBs. While this study evaluates the work done by experts and commentators, my project is intended to automate that work which is being evaluated. However, the technical details of their model-accuracy-scoring system can inform future research into the fine-tuning of my approach.

Another relevant study applies more traditional statistical approaches to the success of drafted QBs once they have entered the NFL. Drawing from data on QB performance between 1997 and 2009, this study built a model based on logistic regression and negative binomial regression (Wolfson et al., 2011). However, this study is intended to predict the success of a particular QB once they enter the NFL. In contrast, my project will only serve to predict whether a college QB will be drafted into the NFL.

Additionally, there is a third relevant study which is more closely related to my project than the others. However, this study is focused on predicting the professional draft for the Women's National Basketball Association (WNBA) (Harris and Berri, 2015). While the relationships explored in the study between performance and non-performance factors may prove useful in evaluating my model, the fundamental differences between basketball and football are too great to draw deep connections between our works. Additionally, the information learned from utilizing the Poisson Distribution model and the Negative Binomial model in the study may ensure the study maintains relevance to my project (Harris and Berri, 2015).

Finally, the above Harris and Berri study draws from a previous study, conducted by a shared author. This study is specifically intended to analyze how the performance factors implicate the decisions by NFL teams on whether or not to select a specific QB for the NFL draft (Berri and Simmons, 2011). This study proved extremely valuable in the research phase of my project to help me understand the dataset I worked with because the research draws from player performance metrics, such as passing statistics. However, this study also evaluates the QB's performance at the NFL Scouting Combine, a session of multiple physical and psychological tests used by NFL decision-makers when selecting their draft picks (Berri and Simmons, 2011). Including the data gathered from the NFL Scouting Combine likely blurs the results due to unrealistic data where the QB is not facing the typical pressures of a real game, such as interacting with opposing defensive players. Additionally, this study includes considerations about the given QB's likelihood of success while competing in the NFL, which is beyond the scope of my project.

5 Methodology

All of the data for my project has been collected and combined into a singular data set. Figure 2 provides a small illustration of my dataset after the appropriate features have been selected. It is useful to note that at this point, I had already removed the name of the players from the dataset. Out of personal interest, I would like to report that the two players who are noted as drafted in Figure 2 are Tom Brandstater (2009) and Stephen McGee

player_id	td	yds	completions	att	int	weight	num_years_playing	drafted
133550	0.000000	52.0	4.500000	8.000000	0.500000	213.107268	2.0	0.0
136354	9.000000	1263.0	114.000000	203.000000	4.000000	213.107268	1.0	0.0
146421	0.666667	47.0	3.666667	8.333333	0.333333	213.107268	3.0	0.0
147211	11.250000	1596.5	137.000000	234.000000	7.750000	213.107268	4.0	1.0
156336	6.750000	1283.5	112.250000	189.500000	3.000000	213.107268	4.0	1.0

Figure 2: Snippet of the DataFrame before model creation.

(2009).

I conducted a basic train_test_split using Sci-kit Learn's model selection function (Pedregosa et al., 2011). I conducted a 75-25 split, using the random state value of 671, out of respect for the popular course at the University of Michigan. Next, I applied a MinMaxScaler - also from Sci-kit Learn - to fit and transform the training data, while only transforming the testing data.

I built 8 different machine learning models for this supervised classification problem. The models I tested are as follows: Gaussian Naïve Bayes, Bernoulli Naïve Bayes, K-Neighbors, Support Vector Machine, Logistic Regression, Decision Tree, Random Forest and Gradient Boost. I heavily relied on Sci-kit Learn's GridSearchCV() function, which allowed me to efficiently test various hyper-parameters to fine-tune my models.

For each model, I ensured to utilize a 5-fold cross-validation metric to prevent over-fitting or selection bias within the training of my models.

6 Evaluation and Results

I compared my model against the true results of drafted QBs based on history.

By comparing the Mean Absolute Error, Accuracy and Cross-Validation scores, I determined that the RandomForest classifier is the best predictor for QB draft patterns. The RandomForest classification model has a Mean Average Error of 0.067, an Accuracy score of 0.9333, and a Cross-Validation (5-fold) score of 0.9173.

The RandomForest classification model uses the following hyper-parameters: 50 estimators (50 trees in the forest), a criterion of 'gini' (which is used to measure the quality of the split, where 'gini' represents Gini impurity), a maximum depth of the tree at 10, a minimum number of samples for each leaf node at 5, and 1 job to be run in parallel (Pedregosa et al., 2011).

7 Discussion

The data collection and organization steps were certainly the most difficult and time-consuming components of this project. Combining information from a wide array of sources, including permission-restricted APIs and manual input, required significant attention to detail and may have resulted in some inaccurate player information in my final model. I made sure to comb through my data multiple times before modeling, but such a level of detail should not be the norm when it comes to building a predictive model. Purchasing a premium subscription service to a useful API, such as the API hosted by SportsDataIO, would prove greatly beneficial to reducing the time and effort required to connect various sources of data.

NFL draft experts and enthusiasts as well as analysts may be surprised to learn that, according to the feature selection phase of my project, the college and conference for which a QB competed at does not significantly effect their likelihood to be drafted into the NFL (correlation coefficient of ¡0.1). Although the Future Work section will describe how I believe the model could be improved to better account for the college and conferences, the current results show that such factors are not worthwhile considerations. Additionally, such stakeholders may be surprised to learn that the height of a QB is similarly not important, with a correlation coefficient of 0.12. Of course, there are limitations to this claim: the average height for QBs in the dataset is 6 feet and 2 inches, which is roughly 4 inches larger than the average height for all males in the United States (WorldPopulationReview, 2021). Height likely is an important consideration for selecting QBs, but such a feature may already be applied to weed-out potential college football QBs before they are considered NFL Draft prospects. Thus, relative to other QBs which could be drafted next year, height for any given QB in the dataset does not matter.

It is important to note that the predictions in my model do not extend beyond the current year. Namely, the model can only predict whether a given QB will be drafted this year, not in any of the next years. Thus, while observing the model's prediction that Cade McNamara - the star QB of the Michigan Wolverines - will not be drafted may come as a disheartening surprise to some. However, noting this limitation can help revive the hopes of die-hard McNamara fans, because he has

only logged two years of QB statistics and thus still has at least another two years left of eligibility left to play football at Michigan and record more impressive QB stats.

The model predicts that the following seven QBs will be drafted into the NFL in the upcoming draft: Kenny Pickett (Pittsburgh), Bailey Zappe (Western Kentucky), Carson Strong (Nevada), Sam Hartman (Wake Forest), Brennan Armstrong (Virginia), C.J. Stroud (Ohio State) and Bryce Young (Alabama). Although C.J. Stroud and the recent Heisman winner, Bryce Young, may be incredibly enticing to NFL programs after their stellar performances this year, they are not eligible to be drafted into the NFL until their years of college eligibility are finished. Even still, to the naked eye of a college football fan such as myself, all seven of these QBs would likely be drafted into the NFL next year, if they could. Thus, my model passes the "gut-check".

8 Future Work

This project should not be considered the end all be all of predicting draft likelihoods. There is ample more data which could be included in future iterations of a similar predictive model, such as information on which city the QB is from.

The data on specific QB statistics should also be scaled in future developments, especially to consider the colleges of each QB. For example, a QB at a historically high-performing college, such as Alabama, should be evaluated differently than a QB at a less dominant school, such as Boston College. This is an important consideration because, unlike in the NFL, colleges gain better players each year based on achieved success in the previous year, so a QB from Alabama likely has better receivers - and thus is predisposed to log better passing statistics - than a QB from Boston College may be.

Additionally, future improvements of this predictive model should incorporate time factors. In recent years, the NFL has demonstrated an increase in the desirability of mobile QBs instead of traditional pocket-passers. This may impact the value of passing statistics and require future models also include rushing statistics of QBs, such as the number of yards they run and the number of touchdowns they run for in a season.

One important limitation to my project which should be accounted for when building a future, similar model is the needs and desires of current NFL teams. If an NFL team drafted a new QB who is expected to become a successful franchise QB, then that particular NFL team will likely not need to draft another QB for another few years. Considering the extraordinary QBs in the 2021 NFL draft class along with the dwindling number of franchise QBs who are approaching a retirement-ready age, I would anticipate the probability of a QB being drafted to change drastically moving forward.

References

David J. Berri and Rob Simmons. 2011. Catching a draft: on the process of selecting quarterbacks in the national football league amateur draft. *Journal of Productivity Analysis* 35(1):37–49. https://doi.org/10.1007/s11123-009-0154-6.

Steven B. Caudill, Franklin G. Mixon JR., and C. Paul Mixon. 2014. Nfl draftnikology: Euclidean metrics and other approaches to scoring ranking predictions. *Communications in Statistics - Simulation and Computation* 43(2):237–248. https://doi.org/10.1080/03610918.2012.700363.

Jill Harris and David J. Berri. 2015. Predicting the wnba draft: what matters most from college performance? *International Journal of Sport Finance* 10:299+. https://link.gale.com/apps/doc/A435542104/AONE?u=umuser&sid=bookmark-AONE&xid=adbbdbee.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Julian Wolfson, Vittorio Addona, and Robert H Schmicker. 2011. The quarterback prediction problem: Forecasting the performance of college quarterbacks selected in the nfl draft. *Journal of Quantitative Analysis in Sports* 7(3). https://doi.org/doi:10.2202/1559-0410.1302.

WorldPopulationReview. 2021. Average height by country 2021. https://worldpopulationreview.com/country-rankings/average-height-by-country.