Jacob Zimmer

Getting to the Pros:
Predicting NFL Success from College Football Statistics

Motivation, Data Sources, and Focus

In this project, I seek to analyze the college and NFL statistics of recently drafted players to explore the relationship between college and professional success. The data is pulled from two different sources: the NFL csv was found on kaggle, while the college data is from a dropbox of csv's - found on Reddit's r/CFBAnalysis - containing single game statistics for all players from 2005 to 2013. For this reason, I limit my analysis to that decade. Furthermore, since football positions have wildly different statistics for measuring performance, I further limit the scope to a single position: Running Back. As a huge Detroit Lions fan, this position is of particular interest, as many fans want the team to draft a running back in this year's NFL draft (2018).

Research Questions, Methods & Context

1. Does having good college stats imply a higher likelihood of success in the NFL? Can this relationship be used to predict future NFL success for college players?
    a. Does this relationship differ for players who played against 'top' competition vs. second-tier and third-tier competition?
    b. Do better stats always lead to better draft positions?
    c. **Methods:** Correlation, Regression, Classification, Dimensionality Reduction, Clustering
2. What college statistics are better predictors of making a successful transition to the NFL?
    a. **Later changed to**: What college statistics are the most important predictors of whether a player will be drafted to the NFL?
    b. **Methods:** Classification, Logistic Regression

In defining competition level for college football, I split the field into three levels: Power 5 conferences (top-tier), Group of 5 conferences (second-tier), and FCS (third-tier). The Power 5 are generally the most financially and athletically successful programs in the nation (left).

ACC	Big 12	Big Ten	Pac-12	SEC
Boston College	Baylor	Illinois	Arizona	Alabama
Clemson	Iowa State	Indiana	Arizona State	Arkansas
Duke	Kansas	Iowa	California	Auburn
Florida State	Kansas State	Maryland	UCLA	Florida
Georgia Tech	Oklahoma	Michigan	Colorado	Georgia
Louisville	Oklahoma State	Michigan State	Oregon	Kentucky
Miami	TCU	Minnesota	Oregon State	LSU
North Carolina	Texas	Nebraska	USC	Mississippi
North Carolina State	Texas Tech	Northwestern	Stanford	Mississippi State
Pittsburgh	West Virginia	Ohio State	Utah	Missouri
Syracuse		Penn State	Washington	South Carolina
Virginia		Purdue	Washington State	Tennessee
Virginia Tech		Rutgers		Texas A&M
Wake Forest		Wisconsin		Vanderbilt

*Notre Dame, technically an independent, is affiliated with the ACC by scheduling 5 games a year against ACC teams. For this research, they will be considered an ACC team.

The Group of 5 conferences form the rest of FBS Division 1's field, and are generally viewed as lower quality programs than the Power 5, as they have less funding and no National Championships in the modern era (below).

AAC	Conference USA	Mid-American	Mountain West	Sun Belt
UCF	Florida Atlantic	Toledo	Boise State	Troy
South Florida	Florida International	Central Michigan	Wyoming	Appalachian State
Temple	Marshall	Northern Illinois	Colorado State	Arkansas State
Cincinnati	Middle Tennessee	Western Michigan	Utah State	Georgia State
Connecticut	Western Kentucky	Eastern Michigan	Air Force	New Mexico State
East Carolina	Old Dominion	Ball State	New Mexico	Louisiana-Lafayette
Memphis	Charlotte	Akron	Fresno State	Louisiana-Monroe
Houston	North Texas	Ohio	San Diego State	Idaho
Navy	Southern Miss	Buffalo	UNLV	South Alabama
Southern Methodist	UAB	Miami (OH)	Nevada	Coastal Carolina
Tulane	Louisiana Tech	Bowling Green	Hawaii	Georgia Southern
Tulsa	UTSA	Kent State	San Jose State	Texas State
	Rice			
	UTEP			

**Several teams exist in D1 as independents: for this analysis, these teams (except the aforementioned Notre Dame) will be treated as Group of 5 teams:

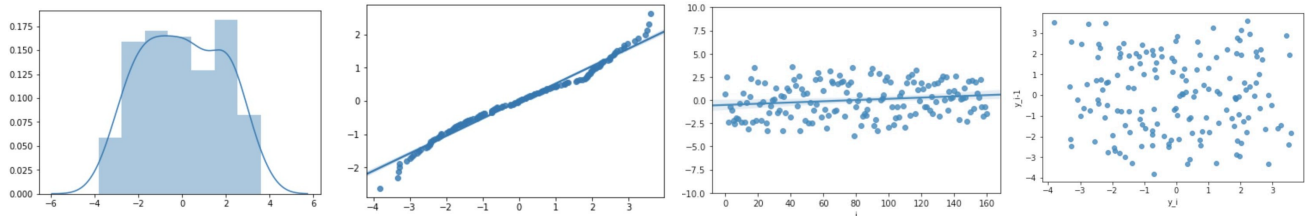
- US Military Academy (Army)
- Liberty University
- New Mexico State University
- University of Massachusetts Amherst (UMass)
- Brigham Young University (BYU)

Analysis, Results & Implications

Q1: Does having good college stats imply a higher likelihood of success in the NFL? Does this relationship differ for players competing against 'top' college competition? Do better stats always lead to better draft position?

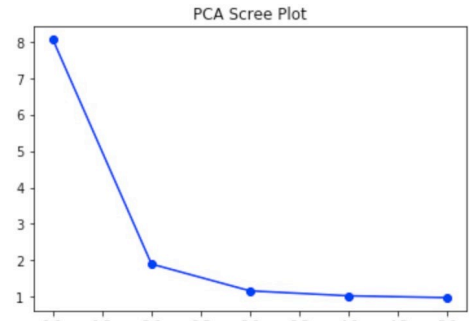
To get an idea of the overall probability of having an opportunity to be 'successful' in the NFL (i.e. get drafted), I calculated the percentage of players in the dataset that were drafted - 29.19%. After normalization and further exploratory analysis, I chose spearman's correlation because my data is non-parametric. I then selected college statistics with high correlations to NFL performance statistics as the dependent variables in OLS regression models aimed at predicting those NFL statistics, specifically: DrAV, or total AV (Approximate Value) a player accumulates for their drafting team, to explore whether good college stats imply NFL success; and draft round, to determine if better stats led to better draft position. My best model for DrAV utilized dependent variables rushing touchdowns (positive coefficient), rushing yards (positive), and rushing fumbles (negative), but yielded a disappointing R-squared of 0.117 and non-normal and

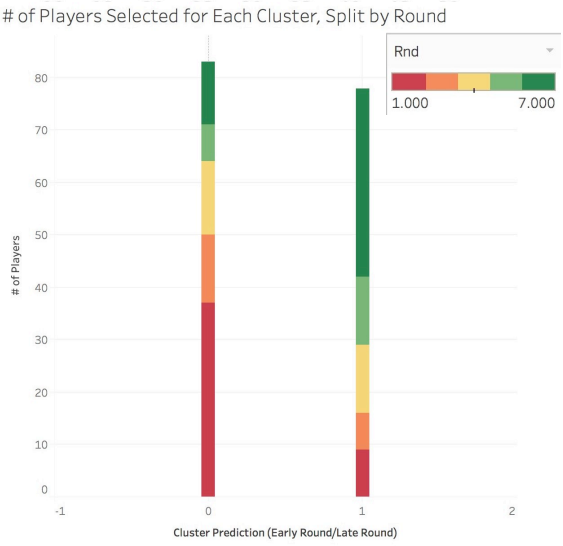
non-independent residuals. My best model for Round utilized height, weight, and rushing touchdowns, and had normal and independent residuals (below), but yielded an R-squared of only 0.116.



Finding that my regression models were not accounting for enough of the variance in the to answer my questions, I decided to build classification models to predict draft round from college statistics. I created both a Random Forest and a Naive-Bayes classifier, and checked their accuracies against each other: The Random Forest classifier was roughly 17% accurate, while the Naive-Bayes was only about 15% accurate.

Given these terrible accuracy scores, I reasoned that the inaccuracy was likely due to the large number of labels and small number of observation points. As a result, I chose to pursue clustering analysis, as it would allow me to alter the number of labels and see if clusters of players with similar draft rounds and/or professional statistics arose. First, I performed Principal Component Analysis to reduce the dimensionality of my data and increase the power of my analysis, ultimately selecting 2 principal components based on the resulting scree plot (top right). Using these 2 components, I performed K-means clustering analysis using several different numbers of clusters. The most effective number of clusters in this analysis was 2, splitting the players into the 'early round pick' group (0) and 'late round pick' group (1) with only mild success (bottom right).





Question 1 Implications

With the poor findings of my correlation & regression analysis, classification models, and PCA & clustering analysis, I am forced to conclude that there is simply not a strong enough relationship between college and professional statistics to draw meaningful conclusions - or make useful predictions - about a player's NFL potential from their collegiate performance. With more observation points

to perform analysis on, a slight relationship may have revealed itself, but the limited availability of the necessary college football data online placed a severe limiting factor on this analysis. Due to this conclusion, I elected to adjust my second research question to the alternative question specified below.

Q2: What college statistics are the most important predictors of whether a player will be drafted to the NFL?

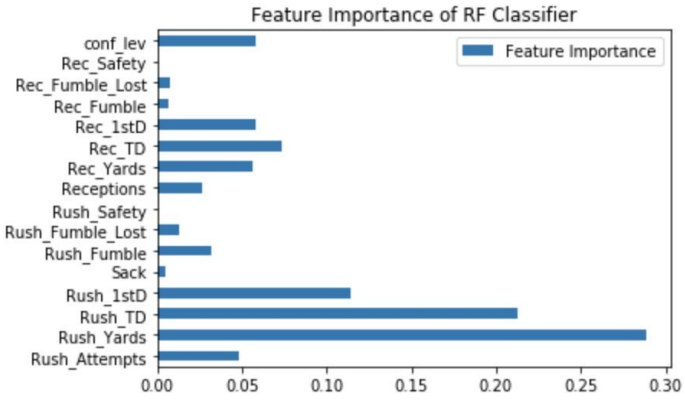
To explore which college performance statistics have the largest influence on a player's draft probability, I began by building a classifier to predict whether or not a player was drafted. I constructed both a Random Forest and a Naive-Bayes classifier in order to check their accuracies against each other, with the Random Forest achieving about 94% accuracy and the Naive-Bayes achieving roughly 88%. The confusion matrices (left top & middle) of the two classifiers revealed that these results need to be viewed cautiously, as their high performances mainly resulted from the vast majority of players in the dataset not being drafted. However, since the null accuracy for the classifiers was 70%, their performances are still notable; particularly that of the Random Forest classifier, as its accuracy score was roughly 25% above the null. Given the reasonable reliability of the Random Forest classifier, its feature importance (left bottom) can paint a solid picture of the statistics most indicative of a high draft likelihood: rushing yards, rushing touchdowns, rushing first downs, receiving touchdowns, and conference (competition) level form the top 5.

Pro Prediction: Random Forest Classifier Performance

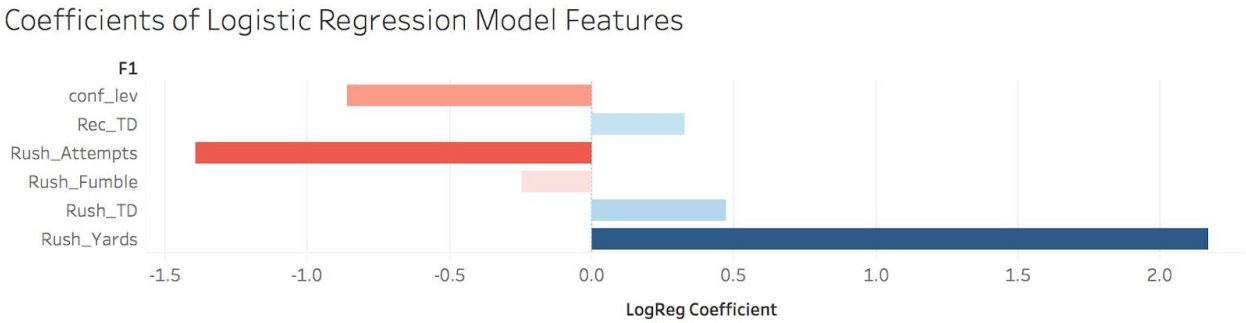
	Predicted Rf Tree	
Was Drafted	No	Pro
No	839	6
Pro	39	6

Pro Prediction: Naive-Bayes Classifier Performance

	predicted nb AllFeats	
Was Drafted	No	Pro
No	765	80
Pro	17	28



Next, I built a logistic regression, since I wanted to predict a dummy dependent variable. Using sklearn libraries, I selected 7 features as independent variables in my model: rushing yards, rushing touchdowns, rushing attempts, rushing fumbles, receptions, receiving touchdowns, and conference level. The p-values of each of these features was below the 0.05 threshold, with the exception of receptions (p-value = 0.051) - meaning that 6 features represented meaningful inclusions in the model. I checked the prediction accuracy of the model, 0.94, against an average of the 10-fold cross validation scores, 0.948, and found that the model holds. Examining the coefficients of the independent variables provides valuable insight into the impact of each key feature on draft probability.



Question 2 Implications

Notably, several important features identified by the classifier match features highlighted by the Sklearn feature selection method (subsequently selected by me) for use in logistic regression - lending credibility to the importance of these features. Focusing on the regression model and the above graph of its independent variable coefficients: three influential features - rushing yards, rushing touchdowns, and receiving touchdowns - have positive coefficients, meaning that as the feature value increases, so too does draft probability. This follows contextual intuition, as good players (i.e. those inclined to be drafted) likely produce higher rushing yards and score more often on rushes and receptions. Negative coefficients, which signify that increases in the statistic reflect a drop in draft probability, were associated with the other 3 features - rushing attempts, rushing fumbles, and conference level - and all likewise fit the context: NFL teams will be wary of players with many rushing attempts, as more carries means taking more hits and having more 'wear and tear' (i.e. the player is more susceptible to future injuries); similarly, NFL teams avoid players with numerous fumbles, as those prospect are bad at protecting the ball. Finally, since the numeric values for conference level are 1 for best (P5), 2 for middle (G5), and 3 for low (FCS), the negative coefficient makes sense: it implies that, as competition level decreases (meaning the dummy value increases), the draft probability falls as well. In conclusion, prospects that are more likely to be drafted play in the Power 5, have high rushing yards and touchdowns on a low number of attempts, few fumbles, and high numbers of receiving touchdowns.

Future Work

Currently, a website called Pro Football Focus has been creating unique statistics for NFL players since 2007 based on context and their performance on every single play; in 2015, the site launched a Division 1 college football version. In the future, these context-based statistics could clarify the relationship between college and pro statistics by quantifying some of the many currently-unquantifiable football intangibles. Ultimately, however, potential researchers must wait until at least 2023 for a dataset matching the size of that used in this analysis, or even longer for additional observation points.