Predicting NFL Rookie Offensive Performance from Combine and Collegiate Data

Ted Henson

Hunter Finger

Jalen McClain

It is ok to use an anonymized version of this report as an example of a great project for future classes

Summary

This project used NFL combine data and college rushing/receiving stats to predict overall NFL rookie success, measured by their scrimmage yards per game and total scrimmage output over the average. Many methods were trained over 5 fold cross validation to try to predict both of these outcomes. Overall, these methods performed reasonably well despite having a low number of observations: achieving below a 17.2 RMSE for all our positions for the scrimmage yards per game, and achieving an accuracy rate of over 62% for both positions when predicting over/under a successful total scrimmage output. Prior research and prevailing wisdom concluded that combine data was useless, but the models presented in this project showed the 40 yard dash, bench press, broad jump, and shuttle run to be highly predictive.

Introduction

There are many different hypotheses as to what should be emphasized when making NFL draft decisions. There are so many different sources of information (scouts, college statistics, NFL Combine, draft interviews) that decision makers differ drastically in how they approach the NFL draft. Since the organizations making these decisions all have their own criteria, fans and the media constantly debate what to look for in NFL prospects. There has been a fair amount of research into the NFL draft as a whole; particularly into the NFL combine [1], [2], [3], and in predicting quarterbacks [4]; however, preliminary searches proved that no analyses combined both collegiate and combine data in predicting NFL success. Combine data may be useless on its own, but it may further reduce prediction error after accounting for a player’s football talent measured by collegiate stats.

Data

The NFL data used to predict the response were rushing and receiving statistics from sports reference [5]. Since there are a finite number of NFL players, every NFL player who logged a rush or a reception from 2014-2018 was included in the data. For this analysis, only rookie seasons were considered for simplicity. Scrimmage yards was the response variable for regression and classification. This was chosen to measure total offensive output. By in large, a rushing yard is equal in value to a receiving yard. There are many variables that play into the yards gained on a player that are out of a player’s control, such as the quality of players around him, strength of schedule, and number of opportunities given to him. A complete analysis would incorporate all these factors, but for simplicity, they will be largely ignored in these models.

The college data was also gathered from sports reference. Unlike the NFL data, not every Division I college player that logged a given statistic was in the dataset. For a given season, a player had to log 6.25 rushes per game in a season to be in the rushing dataset and 1.875 receptions per game in a season to be in the receiving dataset. Although this was not ideal, players that are legitimate draft prospects would not fall below these given thresholds for most seasons. Collegiate seasons from 2009-2017 were considered in order to have 5 seasons of college data for each player. The college statistics used in prediction were as follows: games played, collegiate conference, rushing attempts, receptions, yards per rush, yards per reception, rushing touchdowns, receiving touchdowns, scrimmage attempts, scrimmage yards, scrimmage yards per attempt, and scrimmage touchdowns. For the numeric variables, transformed variables were created for each player across their seasons: the sum of each statistic, and the mean of each statistic. For the sum, NAs were replaced with 0s, and for the mean, the NAs were left out of the calculation. This was done to get an idea of total production and yearly production respectively.

Combine data was also gathered from sports reference. NFL combines from 2013-2017 were considered to get the required seasons for the NFL players. The NFL combine is an event in which NFL draft prospects go and have their physical attributes measured. These include height, weight, 40-yard dash, vertical leap, bench press reps, broad jump, 3 cone drill, and the 20-yard shuttle. Although unrelated to the combine, every player’s age was gathered from this dataset. The combine data was the primary limiting factor for the observations as many players only had partial combine data, and some had none. In order to determine the true value of the combine data relative to collegiate statistics, only players with complete combine data were used in prediction. This criterion could be revisited, as it certainly limited the number of observations, but in order to avoid using imputation techniques (such as k nearest neighbor or bag imputation), players with partial combine data were excluded.

Regressions on Scrimmage Yards Per Game

As discussed above, scrimmage yards was chosen to get an overall estimate of total offensive output. Yards per game played was chosen to attempt to control for injuries and a player’s opportunity to play. Despite wanting to predict the same response variable, the differences in player’s positions presented modeling issues. Rushing statistics were more robust and predictive for running backs than receivers and tight ends. Rather than create a model with hundreds of interaction terms, the players were split up into two buckets: running backs and wide receivers/tight ends. Ideally one would split the tight ends into their own bucket, but there were not enough players in this data to do so without hindering predictions.

Many different models were trained to try to predict the response. Some of the models that were trained but predicted particularly horribly were neural networks and lasso regression. For both buckets, the number of features exceeded the number of observations, so perhaps that contributed to these models predicting poorly. Other models trained that predicted decently, but will not be discussed were ridge, stepwise, support vector machines, Bayesian additive regression trees, multivariate adaptive regression splines, and gaussian processes. The best models for the running backs were the best for the catcher bucket (receivers and tight ends) so the same models were trained for both through 5-fold cross validation. The models trained were extreme gradient boosting, partial least squares, principal component regression, random forest, and a stacked model, which runs a linear regression over 5 folds on the predictions of the other four on the training set. Models were trained on both normalized and unnormalized data. The normalization did not improve predictions, so the data was left unnormalized. All variables were numeric except for the player's position (only included in the catcher regression; indicating wide receiver or tight end) and a player's conference (if a player did not play in the SEC, ACC, Big 12, Big 10, or Pac 12, their conference was revalued as "Other" due to a lack of observations). Although the principal component techniques do not technically handle these dummy variables as categorical, they were included in these techniques as this is important variation to include and weight accordingly. Below is the table of out of sample root mean squared errors trained on 21 running backs and predicted on 11:

Model	Out of Sample RMSE
Extreme Gradient Boosting	21.4722210071323
Partial Least Squares	**17.1976923856682**
Principal Component Regression	18.5701360157812
Random Forest	19.7571688401092
Stacked	20.1105066525871

Despite the simplicity, decision not to scale the data, and poor handling of dummy variables, principal component techniques performed the best. This is most likely due to having an extremely high number of variables and low number of observations. To gain insight, out of bag importance measures scaled from 0 to 100 for the random forest model were calculated.

Predictor Variable	Importance
cfb_Sum__RushingAvg	**100**
cfb_Junior_RushingAvg	98.58
cfb_Mean__ScrimmageYds	95.17
cfb_Sum__Rec	**92.27**
combine_40YD	**89.5**
cfb_Sum__ScrimmageYds	85.84
`combine_Broad Jump`	**84.93**
cfb_Sophomore_ScrimmageYds	82.88
cfb_Sophomore_RushingAvg	81.52
cfb_Mean__ScrimmageTD	79.64

Despite some of the prior mentioned research deeming combine data irrelevant, the 40-yard dash was highly important in prediction, as well as the broad jump. The sum of a player’s rushing yards per attempt across all his college seasons was the most important, and a player’s conference was not listed at all. Many decision makers have been scared to take highly productive player’s in weaker defensive conferences (such as the Big 12 or Pac 12). These basic results give support that decisions to pass on highly productive and athletic running backs in these conferences could be flawed. Also of note, the number of receptions a running back gained throughout his career was highly important. Running backs are expected to catch a lot of passes in the NFL so this makes sense intuitively; however, running backs do not catch many passes in college. It is intriguing that even with a low number of running backs catching the ball in college, the few that are involved in the passing game are productive in the NFL.

For the wide receivers and tight ends, predictions were slightly more accurate across the board, but not by much. The same process was applied here as well. Below are the results of the models trained on 54 players and predicted on 29 test players

Model	Out of Sample RMSE
Extreme Gradient Boosting	19.2575352006551
Partial Least Squares	18.7407308890609
Principal Component Regression	**16.6429955928222**
Random Forest	17.0317685623057
Stacked	19.3310375421897

As with the running backs, principal component techniques performed very well. Of note, here, random forest predicted better than partial least squares. Below are the important metrics calculated by the random forest:

Predictor Variable	Importance
cfb_Senior_Rec	100
cfb_Senior_ReceivingYds	90.89
combine_BenchReps	**81.98**
cfb_Senior_ScrimmageYds	67.36
combine_Shuttle	**53.9**
cfb_Senior_ReceivingTD	48.47

cfb_Senior_ScrimmagePlays	45.92
cfb_Sophomore_Rec	42.36
cfb_Sophomore_ScrimmageAvg	40.65
cfb_Sophomore_ReceivingTD	40.39

Interestingly, the number of bench press reps was highly predictive, as was the shuttle run (to a lesser degree) for wide receivers and tight ends. Conventional wisdom would propose that strength would be more important for running backs than pass catchers, but the results here support the alternative. Perhaps bench press reps signify a player's ability to fight through defenders contact with their upper body and get open, whereas a running back's ability to break tackles comes from their lower body strength, captured by the 40-yard dash and broad jump to some degree. The shuttle run could signify a player's ability to change directions quickly and get open to receive passes.

Although a player's receiving production was the most important feature here, like the running backs, these results show that the combine does have at least some value, contrary to prevailing wisdom and existing research. As with the running backs, receiving yards and receptions were the most important variables as one might expect. In this case, senior receiving production was more important than total production. This could signal that catchers should be evaluated later in their careers, but it could have happened due to random chance, as our data is very wide. Of note here, the player's conference was not highly important, further refuting this idea that collegiate production cannot be compared across conferences.

Classifying Successful Scrimmage Yardage Output

For classifying a successful season of scrimmage output, a cut off of 480 scrimmage yards was made. This cut off point was determined as 30 yards per game was close to the mean output per game, and 30 times 16 (the number of games in an NFL season) would equal 480 scrimmage yards per season. As in the regression analysis, models were trained separately for our running back and catcher buckets. Each model was tuned over 5 fold cross validation, and predicted on a test set. There were 29 and 11 players in the test sets for catchers and running backs respectively. The same seed was used to ensure the same split of players used in regression.

As with regression, many methods were trained that will not be discussed in detail. Many of these methods, such as logistic regression, extreme gradient boosting, Bayesian additive regression trees, and quadratic discriminant analysis, were unable to converge or create meaningful predictions. This was not too surprising for logistic regression as our linear methods performed poorly in regression; however, the other methods generating errors were puzzling. For this project, these methods were ignored, but it may signal that our data is too sparse for certain techniques. Below are the out of sample accuracy rates for our top models predicting on the catcher bucket.

Model	Out of Sample Accuracy
Adaboost	**72.41%**
Random Forest	**72.41%**
Naive Bayes	51.72%
Support Vector Machine (Radial Kernel)	**72.41%**
LDA	65.5%

The tree based methods performed the best along with the support vector machine with a radial kernel. A support vector machine with a linear kernel was also trained, but the radial support vector machine predicted better. To gain insight, importance measures for the random forest were computed in the same fashion as in regression.

Predictor Variable	Importance
cfb_Sophomore_ScrimmageAvg	100
cfb_Sophomore_ReceivingYds	67.08
cfb_Senior_ReceivingYds	59.12
cfb_Senior_ScrimmageYds	56.06

cfb_Senior_ReceivingTD	55.19
combine_BenchReps	**50.28**
cfb_Sum__G	46.07
cfb_Junior_RushingYds	38.69
cfb_Senior_Rec	37.66
cfb_Sum__ScrimmageTD	36.11

As with the regression analysis, the number of bench press reps held predictive power. Oddly, sophomore receiving and scrimmage production was highly predictive. As with some of the variables in regressions, this probably happened due to chance. For the running backs, the same models were trained, and out of sample metrics computed.

Model	Out of Sample Accuracy
Adaboost	36.36%
Random Forest	45.45%
Naive Bayes	45.45%
Support Vector Machine (Radial Kernel)	**62.07%**
LDA	55.17%

The support vector machine was the best for our running backs, followed by the LDA model. The other methods were worse than one would expect from randomly guessing. The random forest importance metrics explained some of the poor performance

Predictor Variable	Importance
cfb_Junior_RushingTD	100
cfb_Junior_RushingYds	95.14
combine_Height	**88.56**
cfb_Mean__RushingYds	87.88
cfb_Mean__ScrimmageTD	87.51
cfb_Sum__RushingYds	**50.28**
cfb_Mean__ReceivingTD	46.07
cfb_Sum__RushingAvg	38.69
cfb_Mean__ScrimmageAvg	37.66
cfb_Junior_ScrimmageTD	36.11

Height may signify a running backs ability to catch difficult balls, but based on it not being significant in any of the other prior models, this presents evidence of overfitting.

Conclusion

All prior inferences should be taken with some suspicion as the sample size for the training and test sets were relatively small, particularly for the running back bucket; however, throughout all the different buckets, response variables, and positions, combine data proved to possess predictive power, as well as collegiate data. Prevailing wisdom and research in football suggests that collegiate players should only be evaluated by experienced scouts due to large differences in competition and so many variables to control for. The methods presented successfully showed the insignificance of some of these factors (such as a player’s conference, height, and weight) and provides support that NFL draft projection systems can assist in minimizing drafting error as they have been used more heavily in the MLB and NBA.

References

[1] "Does the NFL Combine Really Matter", Undergraduate, University of California at Berkeley, 2019.

[2] F. Kuzmits and A. Adams, "The NFL Combine: Does It Predict Performance in the National Football League?", *Journal of Strength and Conditioning Research*, vol. 22, no. 6, pp. 1721-1727, 2008. Available: 10.1519/jsc.0b013e318185f09d.

[3] S. Jenkins, "The NFL combine tells us nothing", *The Washington Post*, 2019. [Online]. Available: https://www.washingtonpost.com/sports/redskins/the-nfl-combine-tells-us-nothing/2019/03/04/888721ca-3e8a-11e9-922c-64d6b7840b82_story.html. [Accessed: 03- Dec- 2019].

[4] L. Jones, "Modeling NFL quarterback success with college data", Undergraduate, University of Georgia, 2019.

[5] "Sports Reference | Sports Stats, fast, easy, and up-to-date | Sports-Reference.com", *Sports-Reference.com*, 2019. [Online]. Available: https://www.sports-reference.com/. [Accessed: 01- Dec- 2019].

Breakdown of Team Member Contribution

- Data Collection:
  - The collection of combine data, collegiate data, and NFL data was largely split up into equal thirds.
- Data Merging.R:
  - Ted and Hunter jointly wrote this script to clean and merge our datasets.
- Data Prepping for Modeling.R, Data Prepping for Modeling (Classifer).R:
  - Ted and Hunter wrote this script jointly to prepare data for regression and classification.
- Regression Analysis Code (Predicting RB Scrimmage Yards Per Game.R,Predicting Catcher Scrimmage Yards Per Game.R):
  - Ted wrote this code to perform regressions.
- Classifier Code (Hunter's Models).R:
  - Hunter wrote this code to run KNN, lda, linear svm, radial svm, and other methods that did not converge (qda, logistic regression).
- Tree and other methods classify catchers.R, Tree and other methods classify running backs.R:
  - Ted wrote these to run adaboost, naive bayes and random forest algorithms.
- Presentation:
  - The slides in the presentation were done by committee.
- Write Up:
  - Ted and Hunter jointly wrote the write up, Ted focused on rough draft, and Hunter focused on revisions.