Boom or Bust: Examining the Relationship between High School Recruiting Rankings and the NFL Draft

Nicholas Tice

University of South Carolina

Table of Contents

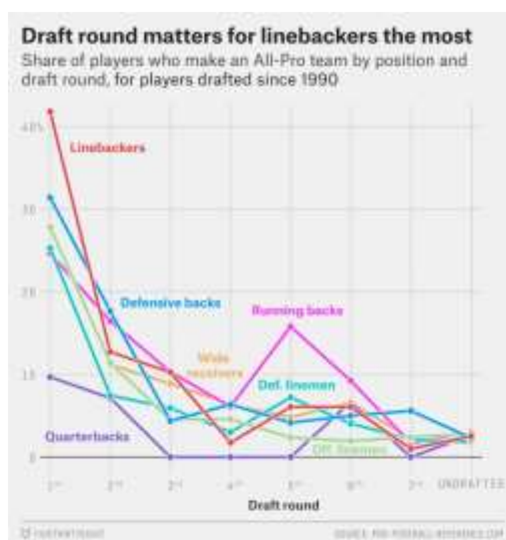
Thesis Summary

The goal of this thesis is to model the probability of a high school football player's chance of being drafted based on information from their recruiting profile. The response variable is a binary and defined as drafted (1) or undrafted (0). The independent variables come from data scraped from the recruiting websites including height, weight, position, hometown, and recruiting grade and other socioeconomic data based on the player's high school. 247Sports and ESPN were the two recruiting services used and compared in this study. Because of the binary nature of the dependent variable, logistic regression and decision trees were chosen as the methods to analyze and model the data. All analysis was conducted using the statistics program RStudio. Once the data were cleaned, they were separated into two sets: one including all public-school players and another including public school players from the south region. Logistic models were chosen based on AIC, BIC, ROC, and misclassification error. The decision trees were pruned to reduce overfitting and increase the power of the test. Ultimately, the best model for both sets was achieved by using logistic regression from the 247Sports data.

Introduction

As a statistics and sport and entertainment management major, my thesis seeks to combine these two interests. Ever since reading *Moneyball* by Michael Lewis and *Outliers* by Malcolm Gladwell in my high school freshman English class, I have always been intrigued by the use of numbers to quantify real-world problems. Fans of almost every sport today have seen an influx of data and analytics incorporated into the decision making of the teams they love, despite occasional reluctance and disdain from some front office staffs, media members, and broadcasting crews. With the amount of money involved in these leagues and invested by the owners, organizations would be foolish not to take any advantage that is offered to them. One area I always believed could benefit from incorporating more analytics was the drafting process. Every year I would watch these drafts and witness so many draftees go on to bust in the future years, especially in the MLB and NFL. Many of the best college players get drafted and never amount to anything in the professional leagues, and most comparisons and expectations the draft "experts" set never even come close to being realized.

Among the four major sports in the United States, the NFL creates the most fanfare for its draft. The fans are constantly inundated with versions of mock drafts leading up to the actual event, and all three days of the draft are televised on ESPN and the NFL Network. This graph created by Rob Arthur and Zach Binney of FiveThirtyEight in 2016 shows the percentage of players who were named to an All-Pro team by draft round and position. Being named an All-Pro indicates that the player was one of the best at their position in that season. Even in the first



round, where the most talented players are still available, the only position with more than a 40% chance of becoming a really good player is linebacker. The randomness and inefficiency of the process made me want to investigate it more. In order to best prepare myself for a career in sports analytics, I have chosen to look at the relationship between high school football recruits and the NFL draft because of the extensive amount of information available from recruiting websites and the historical data from past drafts.

Every time a high school football recruit commits to a college, fans immediately begin searching for how many stars the player has and where they are ranked in comparison to other high school prospects. While numerous services offer these recruiting comparisons, the most popular and available to the public are 247Sports, ESPN, and Rivals. 247Sports also offers a composite ranking that combines the grades from all three systems into one weighted grade. This composite ranking previously included Scout.com until it was purchased by 247Sports in February of 2017. 247Sports was created in 2010 and subsequently began their own evaluation system of the Top 247 players in each high school class. ESPN began publishing their rankings in 2009 with their Top 150 list and switched to a Top 300 list in 2013.

For this study, the prospect rankings from 2009 to 2014 were used. The players in the 2014 high school class would have been redshirt seniors by the time of the 2019 draft. The 247Sports sample has 247 players per year from 2010 to 2014 for a total of 1235 players. The ESPN sample took the top 150 players per year from 2009 to 2014 for a total of 900 players. The players from 247Sports and ESPN samples were then compared to a database of NFL draftees from 2010 to 2019 (about 2,560 players) and assigned a "1" if drafted and a "0" if undrafted. The response variable of drafted or undrafted was then modeled with the use of logistic regression and decision trees based on independent variables taken from each player's background.

All data were collected with the use of RStudio. Relevant code has been included on a public github website[1] found in the appendix. The majority of the data were scraped from public websites, but some of the high school information was found through personal research and will not be easily replicated. Once scraped, the data were exported into excel, filled in with the appropriate high school data, and imported back into RStudio to conduct the analysis.

Data Collection

Most of the data for this project was scraped from 247Sports, ESPN, and Pro Football Reference. Examples of all code used have been included as downloadable R files. The first step of this thesis project was finding an effective way to scrape large amounts of data to reduce the amount of manual imputation needed. I used the code outlined in “Beginner’s Guide on Web Scraping in R” by Saurav Kaushik and adapted it to the websites I was interested in. I began by creating a database of drafted players and used Pro Football Reference as my source to store the drafted players from 2010 to 2019 in a data frame. This created a list of 2,552 drafted players over that ten-year span.

Next, I began looking at the scraping process for the three recruiting websites: 247Sports, ESPN, and Rivals. Rivals uses a dynamic html format to build their website which made it difficult to scrape. Therefore, I focused on the other two. I wrote a similar loop to the one used for the draft but had to adjust for some areas that caused some trouble. This included trimming the white space, separating the high school and state variables, converting height from feet and inches to total inches, and removing some duplicates in the data. The first step was to specify the year range for each set to create a large enough sample size. As explained above, the 247Sports sample was from 2010 to 2014 and the ESPN sample was from 2009 to 2014. Some of the information that was provided on each player from the scrape was the grade, name, position, state, high school, height, and weight. Since it did not include whether or not a player had been drafted, an additional loop was written that compared the name of the player in the drafted data frame to the name of the player in the 247Sports or ESPN data frame. If a match occurred, a new column was appended in the recruiting services data frame noting that the player was drafted.

An area of interest of this study was the impact socioeconomic factors had on a player's chances of being drafted. These were clearly not included on the recruiting websites and had to be individually researched based on the high schools that the players attended. Once the appropriate year range was selected and a data frame was created, the data were saved to an excel file. I then searched every player's high school on USNews Education to find if they attended a private school, what the enrollment was, the percentage of minority students, the percentage of students receiving free or reduced lunch, the graduation rate. The USNews Education consolidates this information from the National Center for Education Statistics which "fulfills a Congressional mandate to collect, collate, analyze, and report complete statistics on the condition of American education" (nces.ed.gov/about/). After this was completed, the data were imported back into R and began to be divided into different independent variables. The recruits attending private schools were missing much of the socioeconomic data since they were not required to report, so the created models only focus on students that attended public schools. Using only public-school students eliminated 18.4% of the 247Sports sample and 16.9% of the ESPN sample.

The two public samples were further broken down into two data frames with the 15 variables listed below. The first table includes all public-school data and the second includes public school data in the south region. All binary variables were made by assigning a dummy variable.

Variable	Description
Grade	Discrete quantitative. Based on assessment by scouts at each website.
Height	Continuous quantitative. Height of the recruit.
Weight	Discrete quantitative. Weight of the recruit.
Enrollment	Discrete quantitative. High school enrollment based on USNews Education.
Minority	Continuous quantitative. Percentage of minority students in high school of recruit.
EconomicDis	Continuous quantitative. Percentage of students in high school of recruit receiving free or reduced lunch.
Graduation	Continuous quantitative. Graduation rate of students in recruit's high school.
Northeast	Binary, 1=Yes, 0=No. States include ME, NH, MA, RI, CT, VT, NY, PA, NJ, DC, DE, and MD
South	Binary, 1=Yes, 0=No. States include WV, VA, KY, TN, NC, SC, GA, AL, MS, AR, FL, and LA
Southwest	Binary, 1=Yes, 0=No. States include TX, OK, NM, and AZ
Midwest	Binary, 1=Yes, 0=No. States include OH, IN, MI, IL, MO, WI, MN, IA, KS, NE, SD, and ND
West	Binary, 1=Yes, 0=No. States include CO, WY, MT, ID, WA, OR, CA, AK, HI, UT, and NV
OFF	Binary, 1=Yes, 0=No. Positions include DUAL, PRO, OC, OG, OT, RB, APB, WR, TE, for 247 and QB, QB-DT, QB-PP, OC, OG, OT, RB, WR, TE, TE-H, and TE-Y for ESPN
DEF	Binary, 1=Yes, 0=No. Positions include SDE, WDE, DT, ILB, OLB, S, and CB for 247 and DE, DT, ILB, OLB, S, and CB for ESPN
Drafted	Binary, 1=Drafted, 0=Undrafted

Since the largest region of players was the South region, we conducted additional analysis of recruits from these states. New independent variables were used because the region was now given. These are outlined below:

Variable	Description
Grade	Discrete quantitative. Based on assessment by scouts at each website.
Height	Continuous quantitative. Height of the recruit.
Weight	Discrete quantitative. Weight of the recruit.
Enrollment	Discrete quantitative. High school enrollment based on USNews Education.
Minority	Continuous quantitative. Percentage of minority students in high school of recruit.
EconomicDis	Continuous quantitative. Percentage of students in high school of recruit receiving free or reduced lunch.
Graduation	Continuous quantitative. Graduation rate of students in recruit's high school.
QB	Binary, 1=Yes, 0=No. Positions include DUAL and PRO for 247 and QB, QB-DT, and QB-PP for ESPN
OL	Binary, 1=Yes, 0=No. Positions include OC, OG, and OT
RB	Binary, 1=Yes, 0=No. Positions include RB and APB
REC	Binary, 1=Yes, 0=No. Positions include WR and TE for 247 and WR, TE, TE-H, and TE-Y for ESPN
DL	Binary, 1=Yes, 0=No. Positions include SDE, WDE, and DT for 247 and DE and DT for ESPN
LB	Binary, 1=Yes, 0=No. Positions include ILB and OLB
DB	Binary, 1=Yes, 0=No. Positions include S CB
Drafted	Binary, 1=Drafted, 0=Undrafted

For both samples, the Enrollment, Minority, EconomicDis, and Graduation variables all came from manual imputation from the government provided education statistics on United States' high schools. All other variables were directly scraped from 247Sports, ESPN or Pro Football Reference. The data have been cleaned and the appropriate dummy variables have been assigned to make the independent variables binary. One issue that arose was names being listed differently in the draft database vs. the recruiting database. For example, sometimes the Jr. suffix was omitted in one and not the other or someone like Jalon Tabor would be listed as his nickname "Teez" Tabor. There were also some duplicate names in the data such as 2010 Texas wide receiver Chris Jones and 2013 Mississippi State defensive tackle Chris Jones. However, I was able to write code to identify and fix these errors or manually change the mistake in the excel file.

Fitting the Model

After the data were collected and cleaned, two techniques were used to fit a model: logistic regression and decision trees. The primary factor that led to using these methods was that the response variable was binary (drafted vs. undrafted). The analysis was conducted on four major groupings of the data: 247Sports Public, ESPN Public, 247Sports South, and ESPN South. As mentioned above, no recruits from private schools are used.

Logistic regression is one of the more commonly used regression methods with binary data. The generalized linear form is shown below:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_p x_{p,i}$$

In logistic regression, the independent variables can be quantitative or categorical. For this model, the categorical variables were represented using binary dummy variables with “1” indicating the presence of the variable and “0” indicated the absence. The response variable is drafted (1) or undrafted (0), and the output of the prediction model will be a probability between 0 and 1.

For each set of data, various techniques and strategies were employed to pick the best model. The variables to include in potential logistic models were first chosen using the “bestglm” package in RStudio. This package contains a function that will produce a specified number of models based on the lowest AIC or BIC criteria. For each set of data, the ten best models based on AIC were generated. From here, other factors used to determine the best model were the ones with the lowest BIC and fewest predictors, the significance of the predictors, conducting likelihood ratio tests between the models, and comparing model diagnostics like sensitivity, specificity, precision, confusion matrices, and ROC Curves[2]. The cutoff point for classification as drafted or undrafted was chosen with the "OptimalCutoff" function. This function minimizes the misclassification error and was between 0.52 and 0.54 for each model.

For creating decision trees, the "tree" package was used in RStudio. This created a decision tree from the same predictors as the logistic regression model. In order to protect against overfitting, each tree was pruned using K-fold cross-validation to identify the minimum index necessary. Since both 247Sports and ESPN are trying to model the recruits from public schools and the south, they can be compared against each other to see which recruiting service best models the data. I have chosen the best models for each and will spend the next section coming to this conclusion.

Public

In each of the logistic models for the public data, the variables of Grade, Height, and the Southwest Region (TX, OK, NM, AZ) were used. Grade and Height both had a positive impact on the probability, but Southwest had a strong negative effect. The 247Sports logistic model performs slightly better than the ESPN logistic model in every comparative measure except for specificity. Overall, the two models perform fairly similarly. For the 247 logistic model, the top 22 players in terms of draft probability were all drafted. The model is strong at predicting a drafted player will be drafted (precision) but misclassifies a lot of drafted players as being undrafted (sensitivity). The specificity is excellent with 95.6% of undrafted players being correctly classified. Either of the 247 models would be good to use, but I would choose the logistic because it correctly classified the most drafted players of any of the four models.

	247 Logistic	*247 Tree*	*ESPN Logistic*	*ESPN Tree*
Misclass. Error	**0.2431**	**0.2431**	0.2634	0.2660
Sensitivity	**0.2837**	0.2457	0.1909	0.1227
Specificity	0.9563	0.9723	0.9640	**0.9886**
Precision	0.7321	0.7889	0.6885	**0.8181**
AUC-ROC	0.6993	N/A	0.6819	N/A

South

Both 247 models are much better classifiers in this set of data. Similar to the last models, Grade and Height are still included. For these models, the independent variables included position groups. Positions with a negative coefficient could indicate a deficiency in the scouting ability of the recruiting services. For the 247 model, REC has a negative impact whereas for ESPN, OL has a negative effect and RB has a positive effect. The 247 south logistic model is the best out of any of the logistic models. It is a slight improvement over the 247 public model. While it looks like the 247 tree performs better than the 247 logistic, this can be slightly deceiving. The tree is very conservative in its classification of drafted players. The logistic correctly classified (55) almost as many players as the tree predicted would be drafted (56).

	247 Logistic	*247 Tree*	*ESPN Logistic*	*ESPN Tree*
Misclass. Error	**0.2573**	**0.2573**	0.2889	0.2727
Sensitivity	0.3594	0.3072	0.2481	**0.3609**
Specificity	0.9422	**0.9694**	0.9343	0.9051
Precision	0.7639	**0.8393**	0.6471	0.6486
AUC-ROC	0.7323	NA	0.708	NA

Conclusion

The results of this study would be of interest to the recruiting services. They would like to improve the accuracy of their grading systems, and this can potentially identify regions or position groups that they are deficient in scouting. Additionally, it could be of interest to the high school players and colleges in pinpointing which traits translate and have the most impact on success. In general, the 247Sports data seems to be a better predictor than the ESPN data. This result was somewhat expected going in, since 247Sports is completely dedicated to recruiting, and ESPN’s interests are spread out over a number of projects.

Future research may expand the scope of the project to identify characteristics of players outside the top 300 as potential high-level players. This study was constrained to the top 300 players from each recruiting service and making conclusions outside of that should be avoided because of extrapolation. Another application of this study is that it can be translated to other sports. Basketball has a more similar recruiting system to football than baseball or hockey, and the code used to scrape and analyze this example could be easily modified.

Since the logistic models were chosen for the public and south region data, the probability of each player being drafted can be found. Listed below is are the top twenty prospects based on this probability. These are the high school recruits from 2010-2014 that the model predicted would have the best chance of being drafted. All twenty players were drafted in the public model, and seventeen of the twenty were drafted were drafted in the south model.

Top Public Prospects

Rank	Name	Prob	Drafted
1	Jadeveon Clowney	0.950253	Yes
2	Robert Nkemdiche	0.822887	Yes
3	Arik Armstead	0.811705	Yes
4	Chris Jones	0.779803	Yes
5	Ronald Powell	0.768344	Yes
6	Da'Shawn Hand	0.768344	Yes
7	Cam Robinson	0.756862	Yes
8	Sharrif Floyd	0.756475	Yes
9	Dominique Easley	0.756475	Yes
10	Timmy Jernigan	0.756475	Yes
11	Dorial Green Beckham	0.7446	Yes
12	Laremy Tunsil	0.7446	Yes
13	Eddie Vanderdoes	0.7442	Yes
14	Dorian Johnson	0.731939	Yes
15	Myles Garrett	0.730794	Yes
16	Aaron Lynch	0.729681	Yes
17	Landon Collins	0.718461	Yes
18	Eddie Goldman	0.716562	Yes
19	Vernon Hargreaves	0.705017	Yes
20	Montravius Adams	0.703065	Yes

Top South Prospects

Rank	Name	Prob	Drafted
1	Jadeveon Clowney	0.973236	Yes
2	Chris Jones	0.84311	Yes
3	Robert Nkemdiche	0.84074	Yes
4	Laremy Tunsil	0.821528	Yes
5	Aaron Lynch	0.821085	Yes
6	Da'Shawn Hand	0.804526	Yes
7	Cam Robinson	0.8033	Yes
8	Matthew Thomas	0.791029	No
9	Christian Miller	0.790427	Yes
10	Eddie Williams	0.780428	No
11	Jeff Driskel	0.774896	Yes
12	Stephone Anthony	0.764986	Yes
13	Rashaan Evans	0.764986	Yes
14	Marlon Humphrey	0.763812	Yes
15	Vernon Hargreaves	0.763179	Yes
16	Quin Blanding	0.761222	No
17	Landon Collins	0.760584	Yes
18	Timmy Jernigan	0.759164	Yes
19	T.J. Yeldon	0.754778	Yes
20	D.J. Humphries	0.753335	Yes

Appendix

1. All R code can be found at: https://github.com/NickTice/thesis

2.

Misclassification Error: number of incorrect predictions divided by total number of observations. This should be minimized to 0.

Sensitivity: number of players predicted to be drafted and actually drafted divided by total number of actually drafted. This should be maximized to 1.

Specificity: number of players predicted to be undrafted and actually undrafted divided by total number of actually undrafted. This should be maximized to 1.

Precision: number of players predicted to be drafted and actually drafted divided by total number of predicted drafted. This should be maximized to 1.

AUC-ROC: tells how well a model distinguishes between classes. When there is an AUC-ROC of 0.5 the model has no separative ability and when it is 1 it perfectly classifies.

2. Public

247Sports Public Logistic Model

Drafted ~ Grade + Height + Southwest + OFF

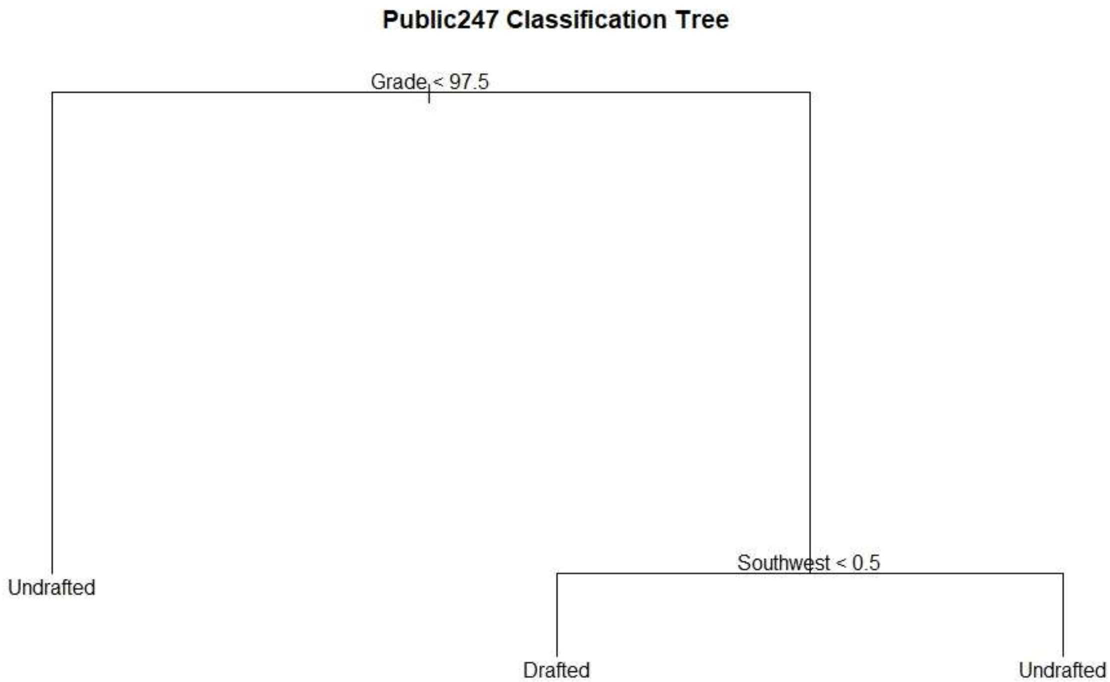
Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-37.48670	4.12454	-9.089	< 2e-16 ***
Grade	0.33705	0.03553	9.486	< 2e-16 ***
Height	0.06554	0.03055	2.145	0.03195 *
Southwest	-0.53738	0.20765	-2.588	0.00966 **
OFF	-0.26004	0.15176	-1.714	0.08662.

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

Null deviance: 1185.2 on 974 degrees of freedom
Residual deviance: 1064.1 on 970 degrees of freedom
AIC: 1074.1

247 Public Logistic	***Undrafted***	***Drafted***	***Total***
Predicted Undrafted	656	207	863
Predicted Drafted	30	82	112
Total	686	289	975



247 Public Tree	***Undrafted***	***Drafted***	***Total***
Predicted Undrafted	667	218	885
Predicted Drafted	19	71	90
Total	686	289	975

ESPN Public Logistic Model
Drafted ~ Grade + Height + Minority + Graduation + Southwest

Coefficients:

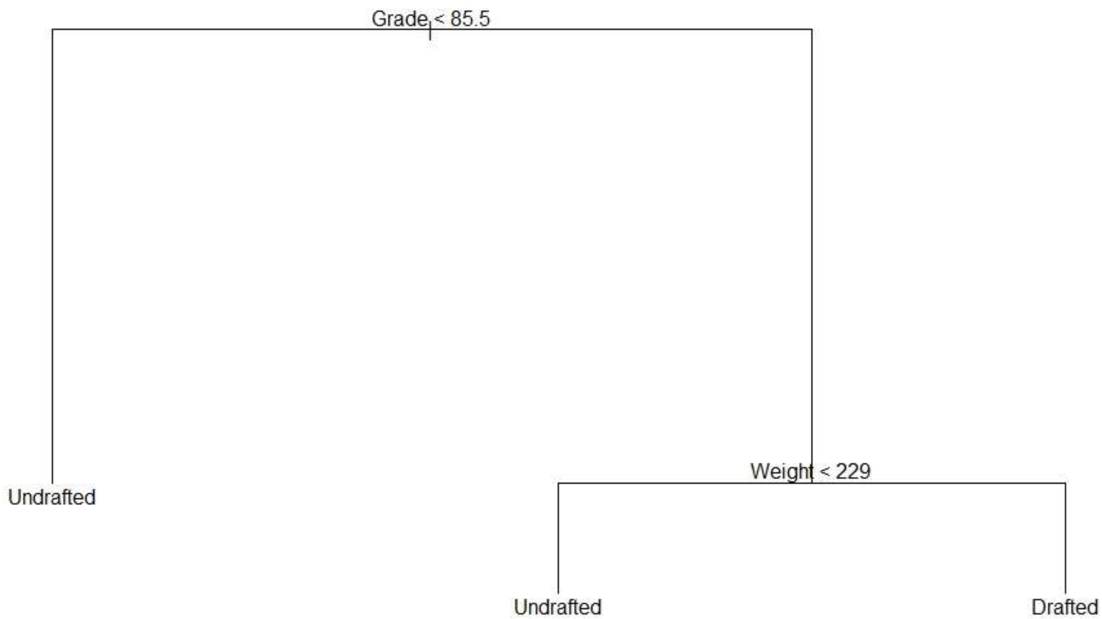
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-24.27808	4.07606	-5.956	2.58e-09 ***
Grade	0.22264	0.03263	6.824	8.85e-12 ***
Height	0.09510	0.03682	2.583	0.00981 **
Minority	-0.61572	0.34946	-1.762	0.07808.
Graduation	-1.87037	1.10155	-1.698	0.08952.
Southwest	-0.71678	0.24660	-2.907	0.00365 **

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

Null deviance: 906.27 on 747 degrees of freedom
Residual deviance: 827.19 on 742 degrees of freedom
AIC: 839.19

ESPN Public Logistic	***Undrafted***	***Drafted***	***Total***
Predicted Undrafted	509	178	687
Predicted Drafted	19	42	61
Total	528	220	748

PublicESPN Classification Tree



ESPN Public Tree	***Undrafted***	***Drafted***	***Total***
Predicted Undrafted	522	193	715
Predicted Drafted	6	27	33
Total	528	220	748

3. South Region

247Sports South Logistic Model
Drafted ~ Grade + Height + Weight + REC

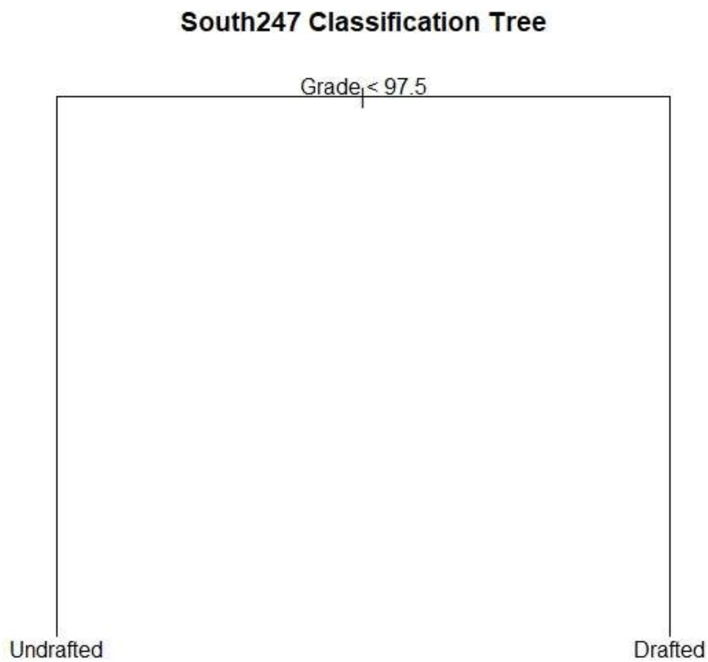
Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-47.928085	6.676266	-7.179	7.03e-13 ***
Grade	0.389430	0.053043	7.342	2.11e-13 ***
Height	0.161339	0.059634	2.705	0.00682 **
Weight	-0.007026	0.003397	-2.068	0.03864 *
REC	-0.727773	0.331118	-2.198	0.02795 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

Null deviance: 574.43 on 446 degrees of freedom
Residual deviance: 496.68 on 442 degrees of freedom
AIC: 506.68

247 South Logistic	***Undrafted***	***Drafted***	***Total***
Predicted Undrafted	277	98	375
Predicted Drafted	17	55	72
Total	294	153	447



247 South Tree	*Undrafted*	*Drafted*	*Total*
Predicted Undrafted	285	106	391
Predicted Drafted	9	47	56
Total	294	153	447

ESPN South Logistic Model
Drafted ~ Grade + Height + OL + RB

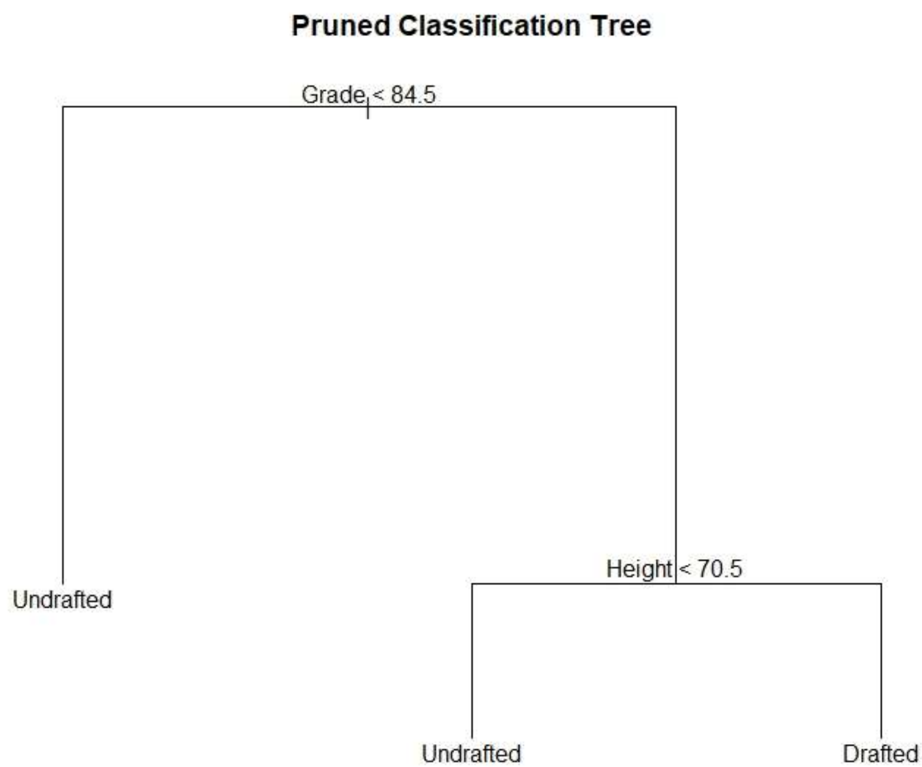
Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-36.82862	5.82991	-6.317	2.66e-10 ***
Grade	0.21580	0.04240	5.089	3.59e-07 ***
Height	0.24555	0.06213	3.952	7.74e-05 ***
OL	-1.04915	0.42226	-2.485	0.01297 *
RB	1.07694	0.39945	2.696	0.00702 **

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

Null deviance: 514.35 on 406 degrees of freedom
Residual deviance: 460.76 on 402 degrees of freedom
AIC: 470.76

ESPN South Logistic	***Undrafted***	***Drafted***	***Total***
Predicted Undrafted	256	100	358
Predicted Drafted	18	33	51
Total	274	133	407



ESPN South Tree	***Undrafted***	***Drafted***	***Total***
Predicted Undrafted	248	85	333
Predicted Drafted	26	48	74
Total	274	133	407

Bibliography

2010:2019 NFL Draft Listing. (n.d.). Retrieved from https://www.pro-football-reference.com/years/2010:2019/draft.htm

2010:2014 Top Football Recruits. (n.d.). Retrieved from https://247sports.com/Season/2010:2014-Football/RecruitRankings/? InstitutionGroup= highschool

Arthur, R., & Binney, Z. (2016, April 28). It's Hard To Tell How Good NFL Teams Are At The Draft. Retrieved from https://fivethirtyeight.com/features/its-hard-to-tell-how-good-nfl-teams-are-at-the-draft/

ESPN Football Recruiting - 300 Player Rankings. (n.d.). Retrieved from http://www.espn.com/college-sports/football/recruiting/playerrankings/_/view/rn300/sort/rank/class/2009:2014

Kaushik, S. (2019, June 24). Beginner's Guide on Web Scraping in R (using rvest) with example. Retrieved from https://www.analyticsvidhya.com/blog/2017/03/beginners-guide-on-web-scraping-in-r-using-rvest-with-hands-on-knowledge/

Prabhakaran, S. (n.d.). Logistic Regression. Retrieved from http://r-statistics.co/Logistic-Regression-With-R.html

The Best High Schools in America. (n.d.). Retrieved from https://www.usnews.com/education/best-high-schools