Author: Winston Graves

Accuracy of High School Football Recruiting Ratings

Executive Summary:

The prevailing belief among people in & around the college football industry is that high school recruiting rankings are great predictors of individual success. Other research has even gone so far to say they've proven the relationship between recruit ranking and NFL draft status. Most say the higher rated a recruit the more likely they are to get drafted and therefore recruiting rankings are good predictors of success and draft status. After analyzing the relationship between draft status and recruit metrics (rating and stars) for the top 500 players from 2010-2017, the validity of these claims is highly questionable. No evidence exists to confidently say that high school football recruiting rankings are predictive of NFL draft status. Analyzing draft rates by star ranking does show higher stars tend to get drafted at higher rates. However, this is likely because 5-stars are so rare it only takes a few hits to balloon the draft rate. The number of draft spots is the same each year and thus only so many players can be drafted each year, 5-stars have a much lower hurdle to get their draft rates up. Adjusting for available draft spots, shows a different story. Additionally, when training machine learning models on this data all models struggle to outperform the accuracy of blind guessing who gets drafted (78%). Blindly guessing had an accuracy of 78%, while machine learning models only had an accuracy of ~79%, on the data collected. After several rounds of validation testing, the best model still has around a 12% chance of being less accurate than blindly guessing who gets drafted. High school recruiting rankings should be taken with a grain of salt by agents who rely on them to predict future NFL prospects and by proxy college success, as they are not strong indicators of prospects success. Do not use recruiting rankings as the only data point for client identification and be wary of using high school recruiting rankings in NIL or other contract negotiations. Find other predictors of success and/or conduct thorough interviews/background checks on potential clients. This applies for underclassmen & upperclassmen alike.

Introduction:

Recruit rankings are widely trusted as measures of player talent and thus used in contract negotiations to determine the value of players. However, sometimes we see players garner a lot of hype going into college and end up flopping. They attract a lot of attention and big-time contracts because their high school rating was so high, but this investment can easily backfire. Vice versa, we sometimes miss out on signing a big-time client because they were low rated in high school. The purpose of this paper is to gauge the predictive accuracy of high school football recruiting rankings on their own. We want to understand if using these recruiting rankings to help value players is a wise decision. Are they truly reflective of a player's talent/success in the long run? To answer this question we will investigate the relationship between recruiting ranking and NFL draft status. Can we predict who gets drafted based on their high school recruiting ranking? Do using recruiting rankings give us a better understanding of who ends up being good? Getting drafted is a good catch-all gauge for talent, because only the best college football players get drafted and NFL scouts are the best in the world at identifying talent. We believe draft status is the best measure of college talent because NFL scouts take crucial context surrounding a player's performance into account that traditional stats cannot. Additionally, positions on the offensive line don't have widely tracked stats to being with. Primarily, we will investigate the problem using statistics and machine learning models.

Context of the Data:

The data used in this paper is oriented around the top 500 recruits each year from 2010-2017. This means the data does not include every player that was drafted from 2010-2017, just the top 500 recruits in those years. Variables used included: recruiting rank from 3 major services, draft pick, draft round, draft position, etc. The main data source for this paper is On3.com; other NFL draft data was joined to this On3 data from collegefootballdata.com.

Data Dictionary (only for variables used in this report):

247_Rating: numerical rating given to a high school football player by 247sports.com.
- Theoretical range between 0-110, prospects below a certain star get no rating.

247_Star: the star rating given to a high school football player by 247sports.com. Ranges from 0-5.
Delta: The percentage difference between actual drafted rate and the max possible draft rate.
- Max draft rate – draft rate.

Drafted: whether a player eventually got drafted or not.
- undrafted player = 0, drafted player = 1

Draft Rate: proportion of players drafted for that service for that star rank.
ESPN_Rating: numerical rating given to a high school football player by espn.com.
- Theoretical range between 0-99, prospects below a certain star get no rating.

ESPN_Star: the star rating given to a high school football player by espn.com. Ranges from 0-5.
Max Draft Rate: if ALL players of that star rating were good enough, what percentage could be drafted.
- Number of available draft spots / number of players in that star level, with maximum of 1.

Rivals_Rating: numerical rating given to a high school football player by rivals.com.
- Theoretical range between 0-6.1, prospects below a certain star get no rating.

Rivals_Star: the star rating given to a high school football player by rivals.com. Ranges from 0-5.

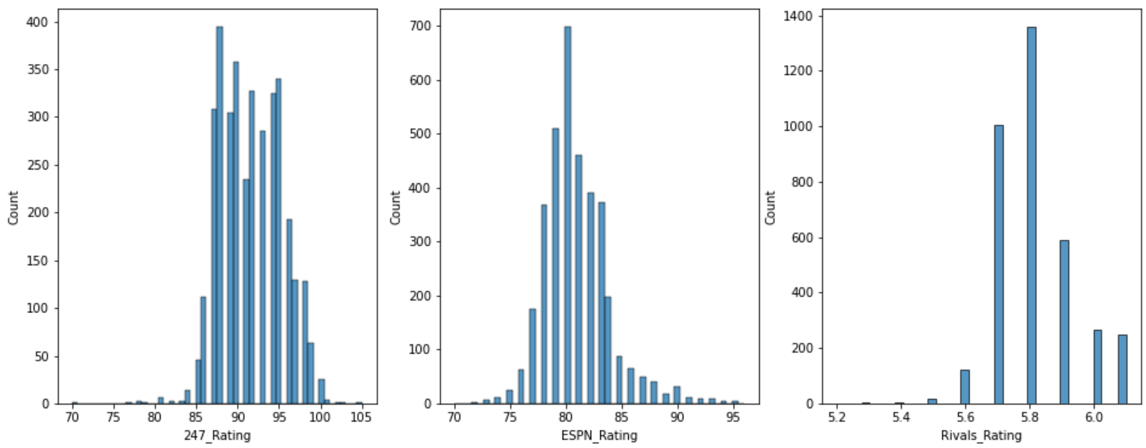
Initial Data Exploration:

Figure 1 (below) shows the proportion of players drafted based on their Star rating via each service from 2010-2017. Players who graduated high school after 2017 were dropped, because they could still be playing in college with covid year and redshirt year. 71 (69%) of all ESPN 5-stars between high school classes of 2010-2017 were drafted. Sample size & draft rate for 3-stars (4900) has an asterisk because not all 3-stars were captured in this analysis. We gathered the top 500 players of each year, so we only end up getting around the top 33% of 3-stars. Most years, around 700 players are rated as 3-stars; 4900 is 700 players for 7 years. If we assume 3-star draft rate of our top ~33% subset is representative of the entire group, the results are pictured below. A smaller delta is better, as the delta basically represents how far the draft rate was from its ceiling. The 3-star ceiling (max draft rate) is so low because there are more 3-stars than there are draft spots to begin with. Even if all 3-stars were good enough to get drafted only 32% of them could be, because the draft is only ~224 picks long each year. Comparing with this caveat in mind, we see 3-stars are getting drafted at a much higher rate over expected compared to 5 and 4 stars. **Note that extrapolating the draft rates of the top 33% of 3-stars to the remaining 67% is a major assumption. Although even if you cut the 3-star draft rate in half, their delta is still the most favorable**

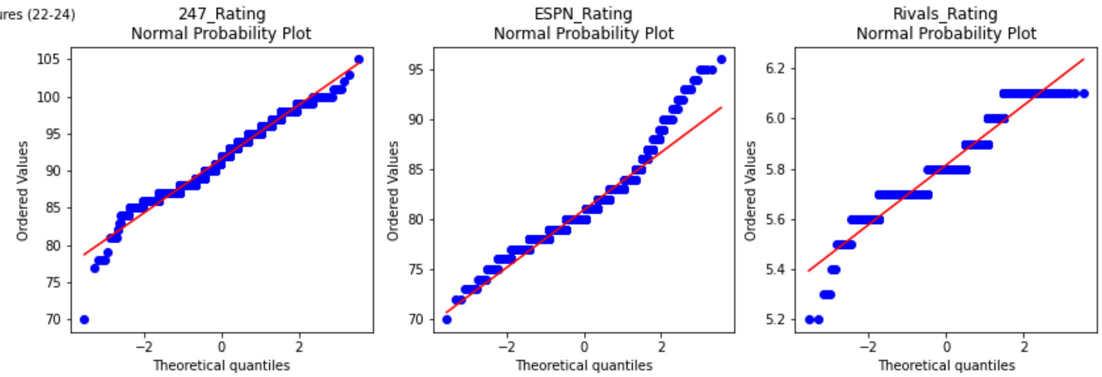
(Figure 1)	N	Draft Rate	Max Draft Rate	Delta
247 3-Star	4900*	11.5%*	32%	20.50%
ESPN 3-Star	4900*	15.5%*	32%	16.50%
Rivals 3-Star	4900*	12.8%*	32%	19.20%
247 4-Star	2214	23%	71%	47.30%
ESPN 4-Star	2438	23%	64%	40.90%
Rivals 4-Star	2352	22%	67%	44.30%
247 5-Star	225	63%	100%	36.70%
ESPN 5-Star	71	69%	100%	30.60%
Rivals 5-Star	244	58%	100%	41.70%

Figure 21 (below) shows the distributions of our continuous predictors with NAs removed. The predictors are close to a normal distribution based on histogram visual inspection and the normal probability plots in Figures 22-24 (below). This means we should be relatively immune to outliers skewing our results and making it hard on models to detect patterns.

Continuous Predictors (NAs removed)
(Figure 21)



(Figures (22-24)



Figures 15-17 (below) show correlation amongst our continuous predictors. The correlations are moderately positive and range from 0.55 – 0.66. A moderate correlation between these variables is expected and they will all be kept in to see if any unexpected patterns emerge.

(Figures 15-17)

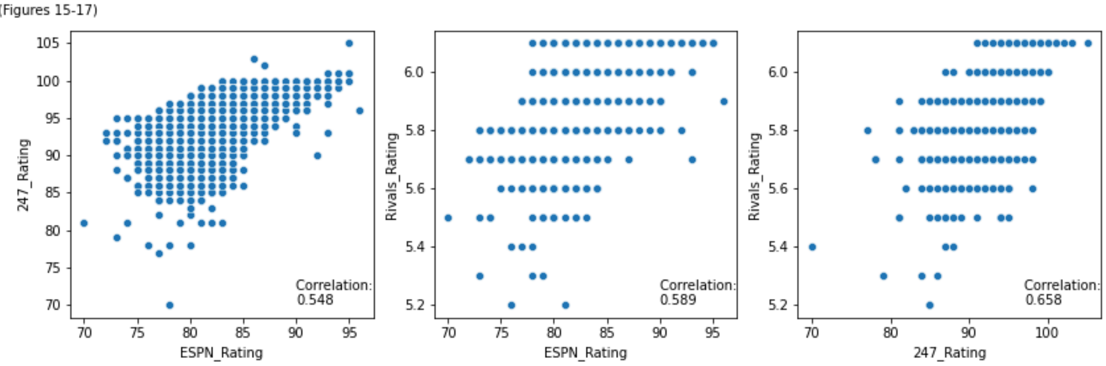
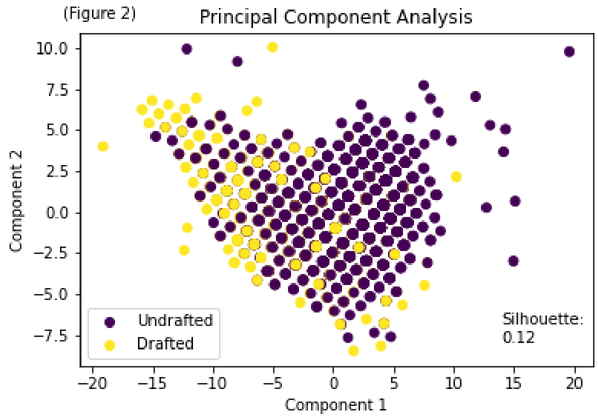
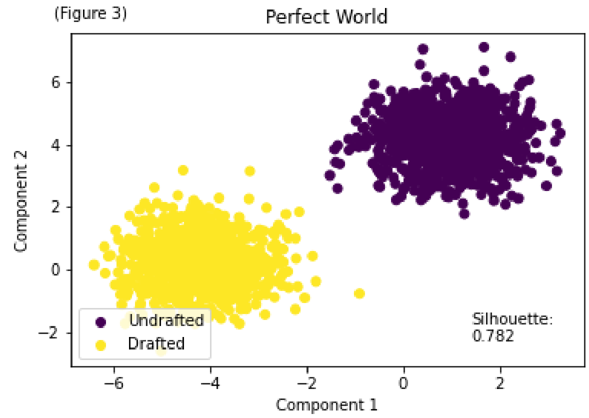
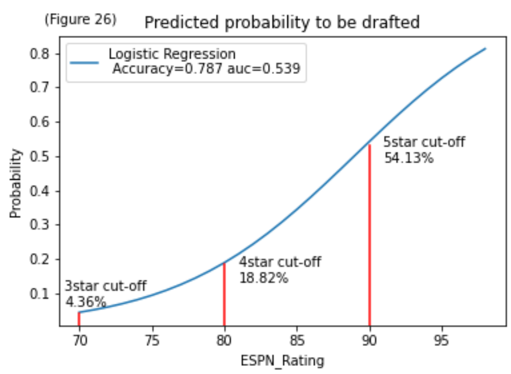


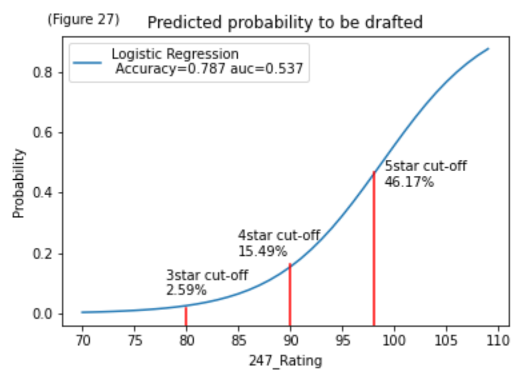
Figure 2 (below) shows our recruit rating space condensed down to two dimensions using principal component analysis. This was done to show our multidimensional data in a 2D scatter plot, so we can easily visualize how "similar" are the groups (drafted/undrafted) we're trying to predict. Points closer to each either on this 2D plot are closer to each other in our actual prediction space and vice versa. In layman's terms, two points right on top of each other in Figure 2 are players with the exact same ratings from ESPN, 247, and Rivals. Figure 2 shows we have significant overlap between groups, drafted players and undrafted players have very little discernable difference based on recruit ratings alone. The silhouette score being close to 0 confirms this. Figure 3 (below) shows what having good predictor variables can do. The fake variables have separated our two groups into distinct clusters.

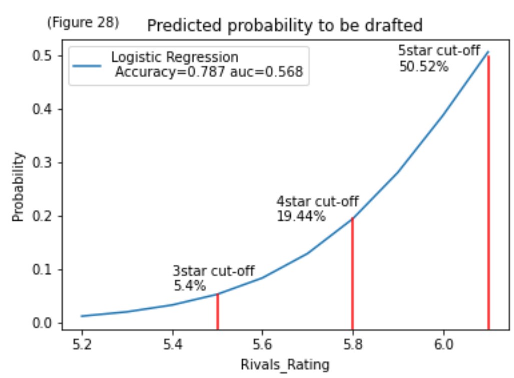




Figures 26-28 (below) show the predicted probabilities of getting drafted at different rating thresholds. Each figure is created based on a univariate logistic regression model using only one variable (x-axis) to predict a player's draft status. All 3 models were also fit using unbalanced (78/22) data. So, accuracy of 78.7% is not amazing, because we could be 78% accurate just by guessing everyone goes undrafted. In other words, it's not much better than blindly guessing. Area-under-curve (AUC) & accuracy will be explained in more detail later in this report. **Please note Figures 26-28 do not represent the strength of the relationship, it simply shows how a weak model interprets the relationship**







Predictive Modeling

So far, it looks like there might be a slight positive relationship between recruit rating and chance of being drafted. A way to test this hypothesis is by training various machine learning models to see if they validate the same pattern. If these patterns truly are strong enough, then models should be capable of accurately predicting a player's future draft status.

Data Preparation:

These preparation steps were done for all models regardless of situation.

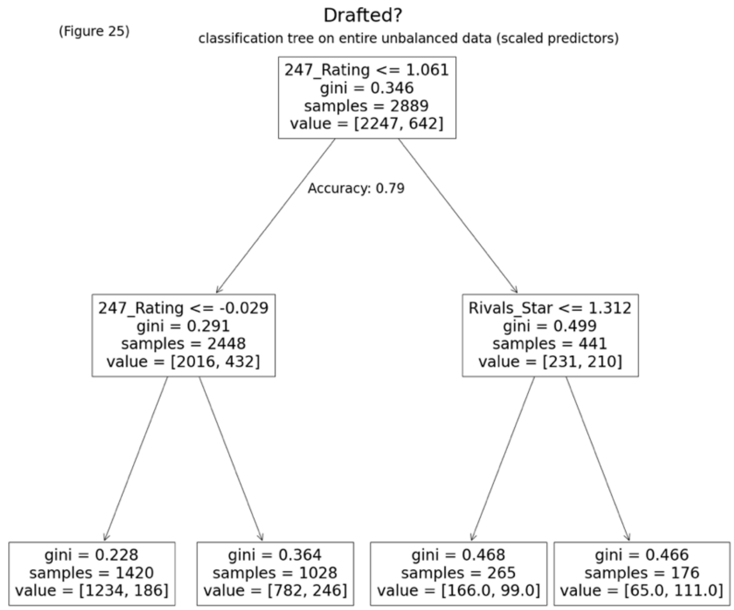
1. Only keep players who have been rated by all 3 existing services (ESPN, 247, Rivals) and graduated before 2018.
    a. Players who graduated high school after 2017 could still be in college currently due to covid years and a redshirt year given to every player. We want to make sure everyone who hasn't been drafted in our data set is because they were not good enough, not because they just were not old enough yet.
    b. Occasionally, a player is rated by one service but not rated by another. Imputing NA values for service ratings was attempted, however this greatly skewed the distributions of our predictor variables and hampered model learning. So, rows with any missing values in service ratings columns were dropped.
    c. Both these actions resulted in ~1457 rows being cut. (28% of observations deleted). Only 8% of this comes from dropping players with any NA rating.
2. Drop columns (variables) that are not of interest. We drop any columns that are not a player rating or a star rating out of high school and their draft status.
3. Scale each predictor (independent) variable into a z-score. This helps ensure the machine learning models do not favor predictors that are on a larger scale (ie: tens vs billions).

Dependent Variable: drafted (boolean if drafted n/y).

Independent Variables: Rivals_Rating, Rivals_Star, ESPN_Rating, ESPN_Star, 247_Rating, 247_Star

Initial Model:

To start we will try to predict a players draft status (n/y) on the entire unbalanced training data set (78/22). 78% of our observations after cleaning are undrafted players. Figure 25 (below) shows a visual of a very basic classification tree with scaled predictors. This is only included to give a visual of what is going on behind some of these models. Classification trees make decisions to split the data based on values of certain predictor variables. With the goal of getting the purest leaves possible: pure meaning all observations left in the bottom leaves are all the same class (all drafted or all undrafted).



Next, we will use a basic logistic regression & a random forest model on the entire unbalanced dataset. Figure 18 (below) show the results of those models. This graph shows two important metrics for evaluating our model's performance: accuracy and area-under-curve (AUC). Accuracy is just the percentage of predictions our model got correct. We see our model was right 79% of the time! Pretty good right? No, not when you consider the naïve model. The naïve model represents what you would predict without having any independent (predictor) variables. Imagine a task to predict whether someone gets drafted into the NFL or not. We have no other information about the person. One fact we know is a very small portion of humans make the NFL, so the best guess for any random person is they go undrafted. This is exactly what the naïve model does, it predicts the majority class for every observation in a classification problem. The naïve model is a great gauge to tell us how helpful our model is. Figure 18 shows the models are better at correctly sorting classes (increased AUC by ~0.19). In layman terms this AUC improvement means our model can better identify draftees correctly (versus the naïve getting none) but is now less accurate at identifying undrafted players. This results in a better AUC but similar accuracy. However, overall each model's accuracy rate is essentially the same as just blindly guessing nobody gets drafted. **Note using other models with hyper parameter tuning was done to no avail**

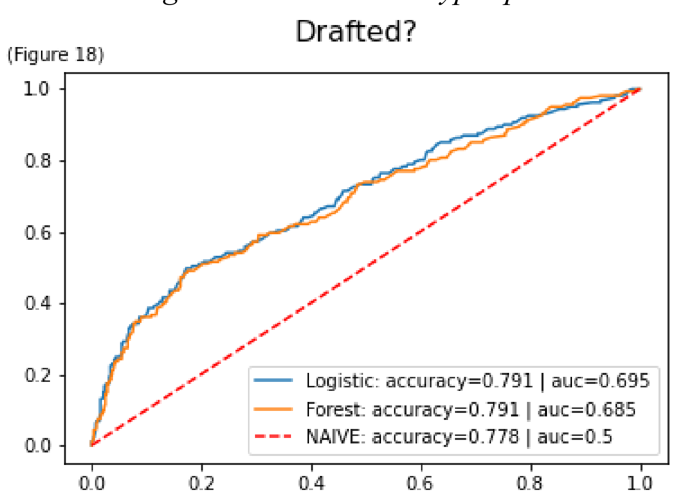
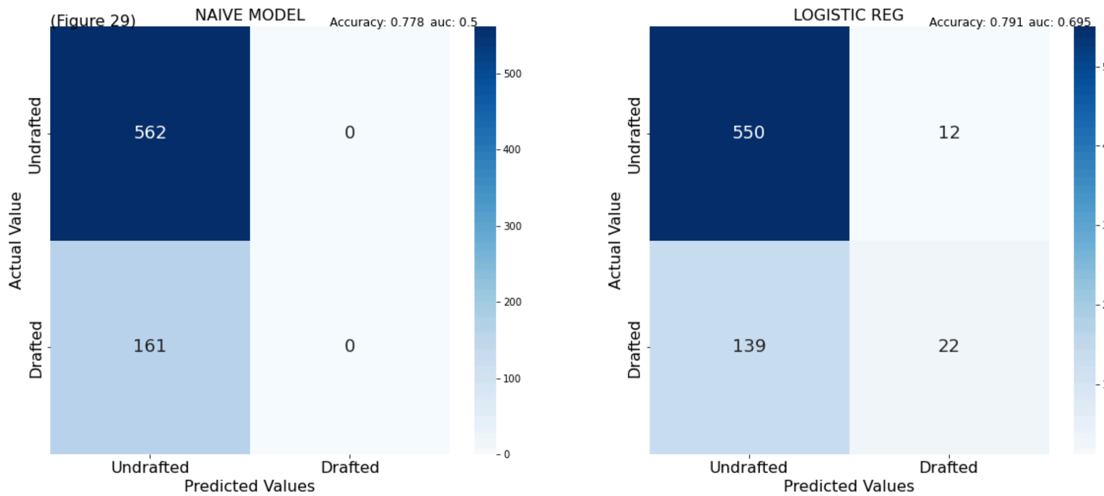


Figure 29 (below) compares the confusion matrices of the naïve model and logistic regression model shown above in Figure 18. On the left, we see the naïve is predicting 100% of test samples to go undrafted, resulting in 77.8% accuracy and 50% AUC. On the right, we see one of our models attempt at sifting through all those undrafted players and correctly selecting the drafted ones. The logistic regression model pinpoints 34 players it thinks will be drafted based on their recruit ratings, but only 22 of those got drafted. This should help illustrate how the models AUC improved but its overall accuracy stayed relatively the same. Class for class we are better at predicting, but overall we only went from 161 wrong predictions to 151.

(Figure 29)



Meta Ensemble:

One thing that enables the naïve model to be so powerful, is if classes are really unbalanced. In our case we have moderately unbalanced classes (78/22). To circumvent this, we will build several models on random balanced subsets of the data, and ensemble them together as a meta classifier.

To start, we will build several models all trained on different 50/50 subsets of our data. Each subset will contain all drafted players (yes class) and an equal number of randomly selected undrafted players (no class). Figure 19 (below) shows each individual model's performance on its random balanced subset. The performance over naïve baseline have improved in both general accuracy and AUC compared to the unbalanced total dataset. However, a model that is only 60% accurate on an unrealistically balanced dataset has questionable value. **Note the naïve changes here because the classes are no longer imbalanced**

(Figure 19)

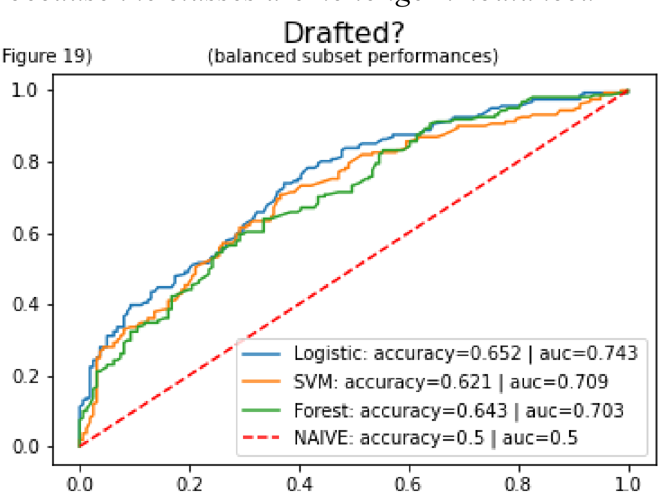
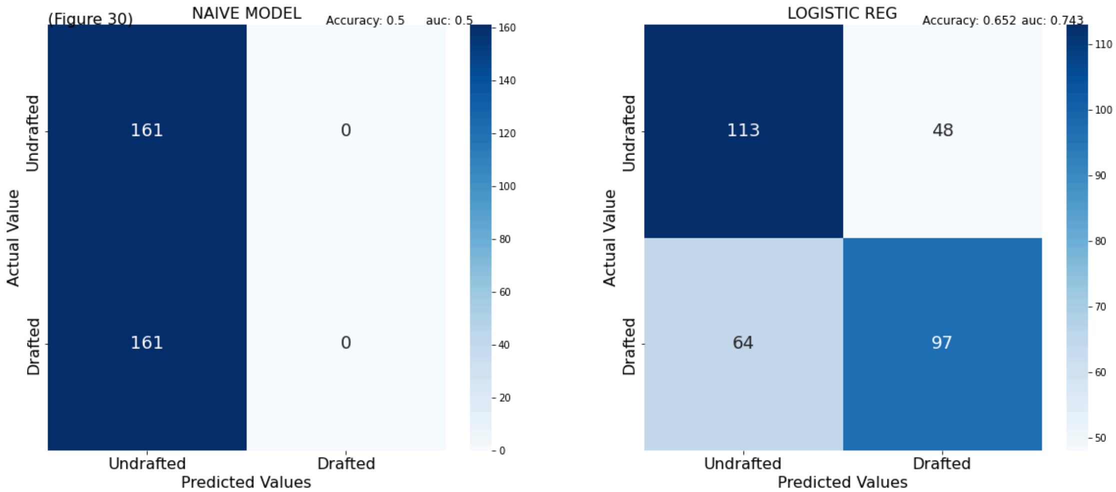


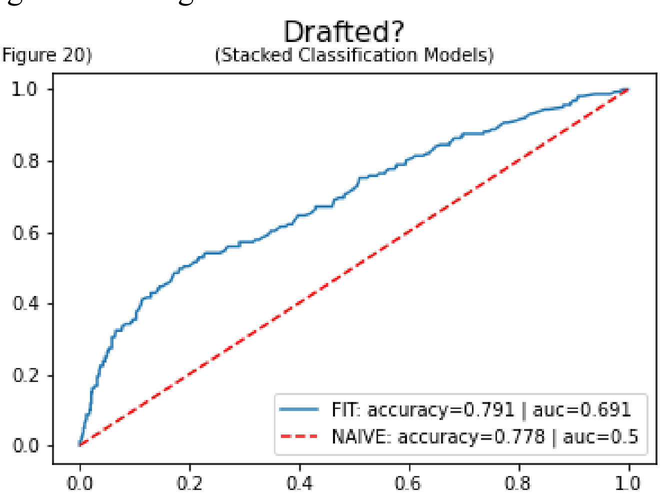
Figure 30 (below) compares the confusion matrices of the naïve model and logistic regression model shown above in Figure 19. The interpretation is the same. We see improvements in correctly sorting observations into their respective class. However, these improvements are modest and are unfortunately on an unrealistic perfectly balanced dataset. Modeling on this unrealistic balanced subset may help models understand what distinguishes classes. However, we would never observe this in nature and to get a balanced dataset we'd need to know the draft status of our players before it happened. This almost completely defeats the purpose of predicting new individuals whose draft status is unknown but of interest to us.

(Figure 30)



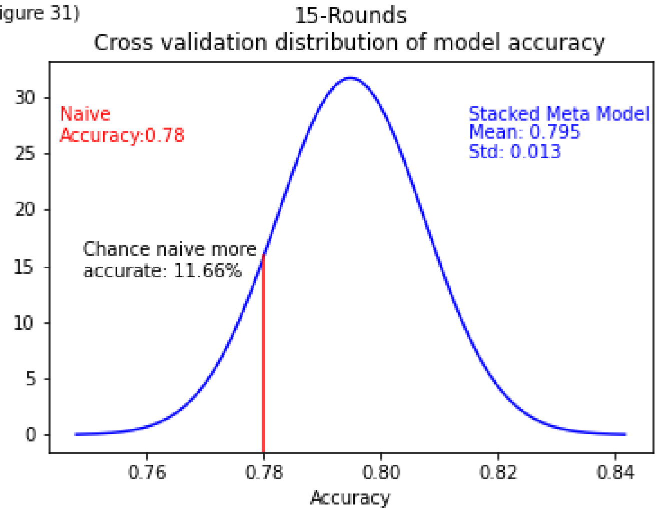
A potential way to improve our metrics (accuracy & auc) is to stack all the machine learning classifiers together. Stacking the models means we have each model make its own prediction on the dataset, then feed all those predictions into another model. This final (meta) model predicts outcomes based on the predictions of the sub models in the stack. Each of our sub models was fit on slightly different balanced subsets of the data; the goal of this is to make them learn more about the class of interest (drafted players). Since each model was fit on different balanced subsets, we should evaluate the stacked meta model on the entire dataset. We want to ensure the stacked model is generalizable to a much larger unbalanced dataset (more reflective of reality). Figure 20 (below) shows the results of stacking our 3 subset models in Figure 19 and predicting on the entire dataset. Figure 20 shows that even after all the effort to help the models overcome class imbalance and even let them "team up" to make predictions, they still fail to significantly improve upon the general accuracy of the naïve model. The AUC is slightly improved; the machine learning models are better able to sort players into the correct classes, but not by a significant margin.

(Figure 20)



The stacked model in Figure 20 (above) was put through 15 rounds of predicting. This gives us reasonably enough samples to infer about the true (population) distribution of this model's accuracy in the long run. Figure 31 (below) shows the estimated distribution of our model's accuracy. Based on the distribution parameters there is a ~12% chance that our model is equal to or worse than the naïve's accuracy. In layman's terms, this means there's a 12% chance our model is no more accurate than just blindly guessing. Additionally, there's a 99.7% chance our model is at best only 5% more accurate than guessing. Alternatively, there's only a 0.3% chance the model is 5% more accurate than blind guessing.

(Figure 31)



Conclusion & Recommendation:

The results are not looking good. We see there is a positive relationship between rating/star-level and the probability of being drafted, however this relationship is so weak it has little-to-no predictive value. At this point there are a few options. Option 1: we could collect data on other variables that are unaccounted for. In the context of our problem, these would be things that would derail the normal development/career path of a college player: injuries, death, suspension, etc. These are things that would completely prevent a player from getting drafted that are not related to how good he is. Option 2: we can reframe our question to predict NFL/college success on things other than strictly high school recruiting rankings or attempt to study the relationship of recruit rankings on a different measure of future success. Option 3: we can collect more than just the top 500 players in each recruiting class; however this is likely to exacerbate the imbalance problem. Option 4: we can tentatively conclude that recruiting rankings are not accurate predictors of making the NFL and by proxy not accurate predictors of successful/good college players either.

Based on the analysis done in this paper, under the assumptions and framing we have, there is sufficient evidence to conclude that high school recruiting rankings are not of significant use for predicting future success. No matter what perspective you take or how you shape it, recruiting rankings are not valuable in predicting if a player will make the NFL or not. In other words, there is very little relationship between a player rating out of high school and their success on the field (being drafted is a proxy for college success as well). Rankings can be useful in laying the foundation for deeper analysis, providing a starting point. However, these rankings alone are not good predictors of success. This makes sense because we know things like coaching, injuries, work-ethic, etc. are all important in the development of a player, not just how recruiting services viewed their talent.

Agents should be wary of anyone leveraging a player's recruit ranking in negotiations for sponsorship deals. Use these findings to your advantage in negotiations. Do not allow someone to use a client's high school recruit rating as leverage against you. Do not allow someone to devalue your client based on their high school rating, as we now know these to be more inaccurate than not. However, you should leverage your clients' ratings against others. If you have an unproven player or an underperforming player, upsell how highly regarded they were in high school. This analysis clearly shows that things outside the players high school rating are more relevant to determine how successful they become. Furthermore, it is not recommended agents make representation decisions based on recruiting rankings. If you're looking to make an early investment into the next great football star as an agent using recruiting rankings alone, you still invite a significant chance of being wrong. Find other predictors of success or delay signing clients will little on-field experience or freshmen.