Predictive Analytics for University of Virginia Football Recruiting

Kunrui Peng, Jonathan Cooke, Annie Crockett, Dongmin Shin, Austin Foster, Jacob Rue, Roby Williams, John Valeiras, William Scherer, Chris Tuttle, Stephen Adams, and Mark Rhodes
kp9eb@virginia.edu, jwc2ka@virginia.edu, aec2mc@virginia.edu, djs5ve@virginia.edu, adf3dv@virginia.edu, jbr3um@virginia.edu, rrw5am@virginia.edu, jdv3hn@virginia.edu, wts@virginia.edu, chris@augmenteddatascience.com, sca2c@eservices.virginia.edu, markrho@mindspring.com

Abstract - **College football programs rely on recruiting to attract high-quality talent, which helps build a team's foundation and ensure success year after year. By conducting systems analysis of the current University of Virginia recruiting process and building upon Walter et al.'s work [1], the Virginia recruiting staff can gain a competitive advantage in the recruiting landscape.**

Analyzing Virginia's football recruiting and utilizing data analytics could provide the coaching staff with powerful tools to gain such a competitive edge. This study uses a database encompassing over 53,000 football recruits and over 200 predictive attributes to model the four aspects of collegiate football recruiting, as defined by Virginia's football coaches. Specifically, a desirable athlete is defined as one who 1) would succeed on the field at the collegiate level, 2) will meet Virginia's strict academic standards to achieve four years of playing eligibility, 3) fit Virginia football's "gritty" team culture, which is characterized by players who are resilient and able to overcome challenges, and 4) would commit to Virginia if given an offer.

Index Terms - Data Analytics, Football Recruiting, Regression Modeling, Sports Analytics

Introduction

College athletic departments not only bring in millions of dollars in revenue, but are relied upon by the university to subsidize other collegiate expenses. In 2017, the University of Texas posted athletic department revenues and expenses of at least $200 million [2]. Most of the time, athletic department profits are then managed by the university board. The gap between average Division I football revenue and the other sports are staggering. While the average school records nearly $30 million in football revenue, the next 25 sports generate less than $25 million combined [3].

Thus, universities rely on college football to subsidize athletic scholarships for all other sports. Studies have shown that in college football, team performance is correlated with stadium attendance, with the direction of causation going from win rate to revenues [4]. Winning just one more football game in a season could lift athletic department revenues by as much as $3 million for a university in a Power 5 conference (ACC, Big 10, Big 12, Pac 12, SEC) [5]. As our interviews with UVA's coaching staff have determined, recruiting "good" players is the first step towards improving the team's win rate and building a sustainable program moving into the future. Conversely, one underperforming player has the potential to propagate mistakes throughout the team, due to the team-centered nature of football.

With so many high school athletes playing football, recruiting is an expensive and labor-intensive process. Much of recruiting is still predicated on watching film and visiting players face to face, both of which consume a lot of time and energy. As of 2015, the average ACC school spent $426,000 on college football recruiting [6]. In this digital age, companies such as 247Sports, MaxPreps, and ESPN have begun to address this problem by creating "star" ratings based on numerous performance metrics and hours of film watching. These "star" metrics have come to be the industry standard for predicting the future value of a player, but leave out metrics valued by UVa's coaches, such as academic standing and personality. In addition, since every school has access to this rating system, these ratings provide our team no unique insight. In order to provide additional valuable insight into a recruit, our models need to incorporate additional attributes not included in the publicly available "star" ratings.

In college football, winning attracts better recruits and yields higher revenues. Success on the football field also garners the university national recognition [7]. Thus, recruiting solid high school players is the lifeblood of any successful college football team. When recruiting a student-athlete, some factors to consider include on-field performance, academic performance, cultural fit, mental toughness, proximity to home, and other notable scholarship offers. UVA's football staff uses a software platform, WarRoom [8], to organize thousands of recruits per year. This platform manages a plethora of recruiting data, enabling the staff to view attributes such as height, weight, 40 time, GPA, contact information, etc.

This paper details the development of predictive performance models to supplement the information currently in WarRoom. These models will provide the

978-1-5386-6343-1/18/$31.00 ©2018 IEEE

football staff with more data to make data informed decisions on which recruits to pursue based on their potential for success on the college field. Our findings build off of the likelihood model constructed by Walter et al. [1].

Literature Review

Data analytics is becoming a major part of decision-making in sports [9]. One of the most well-known examples of this was *Moneyball*. In 2002, the Oakland Athletics baseball team used data in order to craft a roster that far outperformed its payroll [10]. Inspired by *Moneyball*, sports franchises are beginning to explore statistical analysis as an alternative way to improve the efficiency and effectiveness of their recruiting processes [11]. Rather than judging a recruit subjectively by coaches' opinions, analytics offers an objective method for identifying and acquiring desirable recruits [12]. Analytics is especially useful in this case, and like *Moneyball* has done for Major League Baseball, has the potential to find players who are "diamonds in the rough" [13]. In college football, Stanford has lead the way by finding success on the field even without top rated recruits, prompting our team to attempt to find the same success for UVA football [14].

Prior work on analytics in sports guides our work. PhD candidate Kristina Bigsby's commitment likelihood model uses social media data to predict where a high school recruit will play in college with over 70% accuracy [15]. Furthermore, Walter et al.'s work on academic models, likelihood models, and grit models for the UVA football team set the foundation for our team's work [1]. They created models that help quantify three of the four components of college football recruiting: academics, likelihood of attending UVA, grit, and performance. Our team validated their models and sought to complement them with a performance model.

Background

The goal of this project is to improve the recruiting staff's efficiency in choosing which prospects to recruit through data analytics. With regards to recruiting, UVA's coaches are primarily concerned with four aspects - football performance, academic performance, mental toughness, known as "grit", and likelihood of the recruit committing to UVA. In previous years, the recruiting staff has invested hundreds of hours on high school athletes who were never going to attend UVA and moving forward, the coaches want to know that their time is going towards players that actually want to be a part of their program.

I. Football Performance

When looking at potential recruits, UVA's coaching staff desires a player who can make an impact on the field. Provided by 247Sports, ESPN, and other recruiting services, the star rating is widely used to predict the talent level of a recruit and how much impact he can make in a football program. These companies attempt to assign higher star ratings to high school athletes they believe to be more talented, through various methodologies [16]. UVA Football, as of right now, is unable to compete for the top talent (four and five star recruits) to the extent that top programs do, and thus relies primarily on the two and three star recruits to fill its recruit class. There is great potential for finding "diamonds in the rough" using analytics - while under 10% of 2 and 3 star recruits receive accolades (All-American or All-Conference honors or being drafted by the NFL) in college, 88% of rated players and the overwhelming majority of accolade-receiving players fall within these two star ratings [17].

II. Academic Performance

While a recruit may be athletically talented, he or she must still meet NCAA's strict academic standards, encompassing grade point average (GPA) and course minimum requirements, in order to play on the field [18]. This is especially important in context of UVA's rigorous academic culture - the university admits just 30% of students, averaging 4.23 in high school GPA and 1410 in the SAT composite score (Math and Critical Reading) [19].

Statistical models may be useful in 1) predicting whether a recruit will be eligible for four years of play, and 2) assessing quantitatively the risk and uncertainty associated with admitting an athlete to UVA. Walter et al. developed linear regression models, incorporating several factors, but the models hold wide confidence intervals [1], resulting in low levels of precision. It may be possible to improve the model's performance by using richer data such as census data, which includes such factors as population density and percent of single-parent household. However, no additional models were created in this study.

III. Grit

The UVA coaching staff believes that grit, defined as mental toughness and resilience, is an important personality factor when evaluating a recruit. However, grit is difficult to quantify and predict because it does not show up on the statistics sheet. Walter et al analyzed recruits' Twitter accounts using IBM Watson's Personality Insights tool to pinpoint 25 unique personality attributes, of a recruit, including grit [1]. The predictive strength of their model was validated by comparing predictions with this year's 2017 recruit class.

IV. Likelihood of Committing to UVA

The football staff believes that the ideal recruit is not only a strong fit athletically and academically, but also has a sincere interest in attending UVA. There exists a limited number of elite football recruits and demand for these players is fierce. Furthermore, the recruiting process incurs high costs, including not only salary and travel expenses for

a staff dedicated entirely to talent acquisition, but also three weeks per year traveling to recruits' hometowns by the head coach. Due to these immense investments in time and resources, college football programs expect high acceptance rates by recruits who are extended an offer.

Walter et al. built a logistic model that calculates the probability of a recruit committing to each school from which he has received an offer [1]. This tool aids the recruiting staff in evaluating whether more time or resources should be invested in a recruit. Moving forward, it is possible to improve the model's prediction accuracy using publicly available census attributes, such as the number of single-parent households and average transit time to school.

Approach & Methods

Continuing to enrich and quantify insights into high school football players, this research focuses primarily on the football performance aspect of recruiting, building statistical models that predict success on the field in college using high school statistics as predictor variables. As a proof-of-concept, classification models for seven football positions - quarterbacks, wide receivers, running backs, tight ends, linebackers, defensive backs, and linemen - were developed, outputting the probability that a recruit will receive an accolade in college. Accolades considered include All-American distinction, All-Conference distinction, and being drafted by the NFL. According to the UVA coaches, any player who receives one of these accolades will impact the team in some significant fashion.

As mentioned earlier, our data shows that about 10% of rated players (2 star and above) receive an accolade in college. Thus, the objective of these models is to identify players who beat that average and deserve a second look. A majority of these two and three star recruits do not have major D1 offers, so there is a high likelihood that they would choose to attend UVA. According to the coaches, five good prospects per year (out of a class of 25) would generate significant lifts on the team's win rate.

I. Data Collection

To forecast college success, college statistics available on CoachesByTheNumbers (CBTN) [20] were used as response variables, and high school statistics available on 247sports [21], MaxPreps [22], publicly available census data [23], and social media data from Twitter [24] were used as predictor variables. Furthermore, in selecting predictor variables, it was necessary to ensure that the models can be implemented into the future. That is, only variables that will be available at the time that UVA's coaching staff must decide to invest time and resources in a high school recruit were collected.

Our database for college statistics encompasses all recorded statistics for all Division 1 college football players since 2005 (e.g. rushing yards, passing yards, touchdowns, interceptions, etc.), as well as the accolades received and various star ratings (e.g. ESPN, 247sports, etc.). Overall, this includes over 53,000 college players from the past 13 years (Figure I). This college data was paired with the players' high school statistics using Rivanna, UVA's high-performance computing system [25]. It is worthy to note that high school statistics from MaxPreps are self-reported by high school players and their coaches, resulting in many sparse data fields, especially for less-talented athletes.

To expand the selection of predictor variables beyond traditional performance statistics, census and Twitter data was also collected. Personal demographic data was retrieved from UVA's WarRoom database, and then used to pair recruits with their census data. Twitter handles were also gathered from UVA's WarRoom database, allowing for text analysis that evaluates 25 unique character traits using IBM Watson software. Although WarRoom data is still in its infancy (i.e. most recruits listed have not reached college yet, and thus have no recorded college statistics) and is thus unusable in our performance models, these predictor variables may be informative moving into the future as data cleanliness improves.

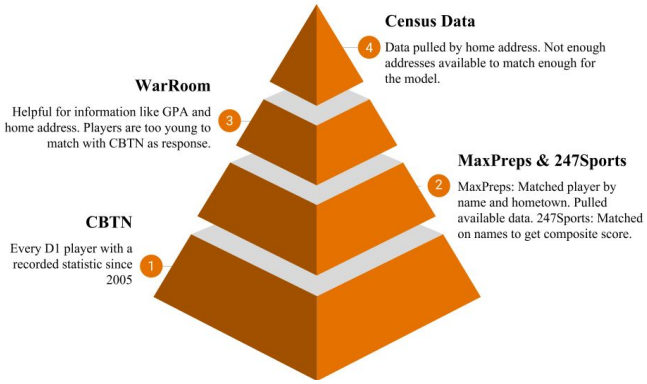

FIGURE I
VISUAL BREAKDOWN OF DATABASE AND PAIRING PROCESS.

II. Modeling Methods

Classification models were built with the following objectives:

- Identifying "hidden gem" recruits, defined as 2 and 3 star recruits who will receive an accolade in college
- Maximizing accuracy, defined as the proportion of true positives and true negatives, out of all predictions
- Maximizing precision, defined as the proportion of true positives, out of all positive predictions
- Outperforming publicly available star rating systems (e.g. ESPN, 247sports, etc.)

Due to missing and incomplete data, selecting predictor variables must be done with caution. For the classification

models built in our research, every data entry must have valid values for every predictor used in the model. Given the scarcity and self-reporting nature of MaxPreps data, using too many predictor variables in a model will yield sample sizes too small to prove statistical significance.

The simplest models built were generalized linear models (GLM's), predicting the aforementioned binary response variable - the probability of receiving an accolade. For each of the seven GLM models, high school performance predictors were selected with the intention of maintaining a stable sample size, and stepwise regression was used to select only the statistically significant predictors. 10-fold leave-one-out-cross-validation (LOOCV) and 10% random holdout samples (RHOS) were used to compensate for limited sample sizes. Finally, to address the last objective listed, each position's GLM was compared to a "null model" that used the 247Sports Composite Star Rating as the sole predictor. However, diagnostic plots show that the dataset lacks normality, violating a key assumption of linear modeling. Interestingly, the QQ plots, one shown in Figure II, in each model show that poor to average players clustered along the line of normality, while exceptional players essentially appeared as a separate segment.

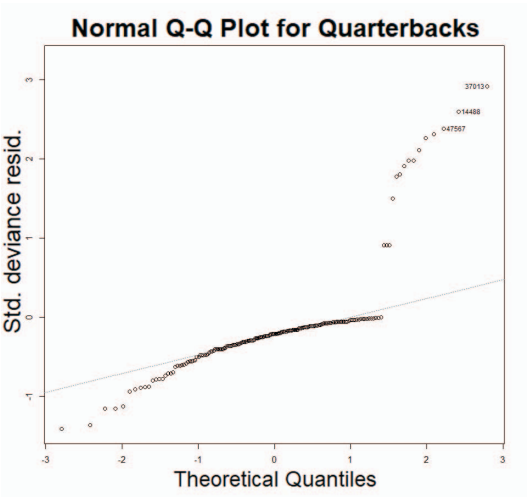

FIGURE II
QQ PLOT OF THE GLM FOR QUARTERBACKS.

Tree-based modeling was used to address these non-linearities. Initial classification and regression trees were built using similar predictor variables as in the GLMs. In each position model, predictor variables were then selected using R's predictor informativeness metric, which evaluates predictors' impact on determining whether a recruit will receive an accolade. Once again, LOOCV and RHOS were used to compensate for limited sample sizes. To further improve results, random forest models were built for each position in a similar fashion, benefitting from the additional power in an ensemble method [26]. Given limited testable sample sizes and nonlinearity in the data, these tree-based and ensemble models yielded the greatest level of success based on accuracy and precision.

RESULTS

Precision was chosen as the primary metric for comparison because false positives are highly undesirable in football recruiting. As mentioned earlier, our interviews with UVA's coaching staff determined that if the program invests thousands of dollars in pursuing one high school player, then he should show high levels of on-field success in college. Furthermore, UVA's coaches believe that one underperforming player has the potential to mitigate mistakes throughout the team.

Within each model, the decision threshold (i.e. probability of receiving an accolade) can be adjusted to skew models towards precision at the tradeoff of the number of "successful" recruits identified. With higher decision thresholds, there would be few to no false positives, but only a small handful of recruits identified, increasing the false negative rate. Conversely, lower decision thresholds yield larger lists of recruits identified and higher false positive rates. This decision threshold was adjusted to the degree of certainty desired by UVA's coaching staff, yet still allowing the model to identify a handful of 2 and 3 star recruits with above-average accolade potential.

While each modeling technique still resulted in low accuracy due to a high number of false negatives, with models missing over 60% of accolade players, some models still yielded precision above 80%. Skill-player models resulted in the best precision, outperforming the publicly available star ratings by far, especially in tree-based and random forest ensemble models. Furthermore, when these models classified a 2 or 3 star recruit as an accolade receiver, he was truly a "diamond in the rough" in almost every instance. One potential explanation is that skill players, such as quarterbacks and wide receivers have the most recorded statistics and their success on a given play is the most easily quantified. Conversely, non-skill positions such as offensive linemen and defensive backs do not have play-by-play statistics or metrics for how well they performed or whether they were successful. Without these performance statistics and without other potential predictors (census data, reliable personality analysis, etc.), most non-skill position models were based on simple statistics, such as height and weight, that are less informative than star-ratings, which are partly based on intensive scouting and film research.

DISCUSSION

Adding to Walter et al.'s work on likelihood models and personality analysis, integrating performance metrics into UVA's Warroom recruiting database can further enrich insights into recruits for the coaching staff. Football

performance is currently evaluated by the coaches, who decide which recruits to pursue primarily based on their star rating and by watching their high school film. These performance models will not replace human intuition and decision making, but will aid UVA's coaches by generating a list of players to consider that the coaching staff would have otherwise missed. Figure III below visualizes how the model's precision can provide the value add by finding high potential recruits who are both identified by the model and under ranked by the star system, giving UVA an advantage in recruiting them.

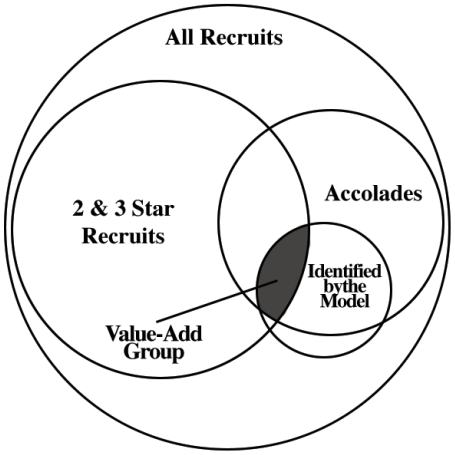

FIGURE III

VENN DIAGRAM SHOWING THE SMALL SUBSET OF RECRUITS WHO WOULD BE RECOMMENDED FOR THE RECRUITING STAFF TO PURSUE.

The biggest limitation in building accurate performance models is the lack of data. While our dataset holds records for over 50,000 athletes, only 200-300 entries are usable for each model due to the modeling techniques selected. However, this study shows that machine learning methods, cross validation, and random holdout samples can in fact compensate for limited data in some cases. For example, when testing our quarterback random forest with a holdout sample the model yielded over 90% true positive precision, and identifies some of the biggest "recruiting steals" in the past decade. In other words, when the model identifies a player as likely to receive an accolade, that prediction is almost always correct. Furthermore, the ability to identify lower-star players reinforces the idea that our models outperform the star rating system.

Each of these models outputs the probability of a player receiving an accolade in college: All-Conference, All-American, or NFL Drafted. As mentioned above, about 10% of players receive an accolade in college. Therefore, a random player selected has a 10% chance of being one of the receivers of an accolade. Thus, if the model predicted someone as having a 30% chance of receiving an accolade, he would be three times more likely to receive an accolade than his peers.

Models for other non-skill positions did not produce results reliable enough for the recruiting staff to use. Moving into the future, further data collection and the development of quantifiable metrics for non-skill players can improve these models. Companies such as Pro-Football Focus provide extra insight on college athletes' performance using game-by-game performance scores, but these statistics are not currently available at a high school level.

Used in conjunction with Walter et al.'s likelihood and academic performance models [1], our performance model will provide value to the staff by by identifying "diamonds in the rough," thus illustrating a more-complete picture of high school football recruits. Figure IV shows a mockup of the potential decision making power behind the two models in conjunction. These tools together can identify a set of recruits who will both improve the football program and be likely to accept an offer if pursued. Together these tools can help the recruiting team allocate resources like time and money in a data informed way.

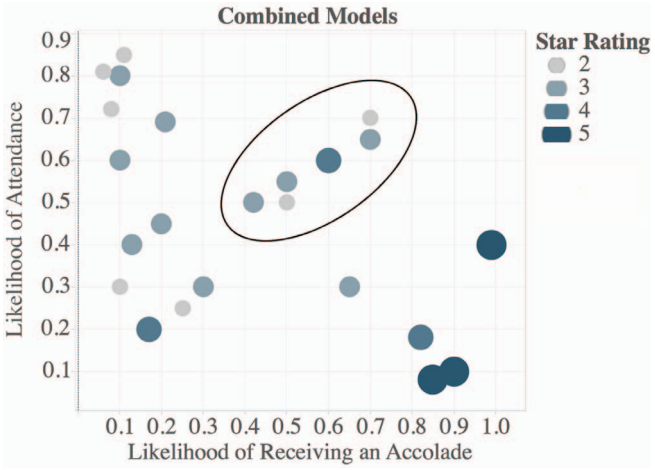

FIGURE IV

GRAPH SHOWS WALTER ET AL'S LIKELIHOOD PREDICTION ON THE Y AXIS AND THE PERFORMANCE MODEL'S PREDICTION ON THE X AXIS. KEY INSIGHTS WILL BE FOUND IN THE TOP RIGHT WHERE PLAYERS ARE PREDICTED TO BOTH PLAY WELL AND COMMIT TO UVA IF PURSUED.

CONCLUSION

The performance model estimates the probability a recruit will earn an accolade during his college football career. Initial results show the model is able to select accolade-receiving players with high precision. Even so, it should not be used to tell the UVA recruiting staff who to give offers to, but to help the recruiting staff not miss out on a potentially great player. Despite relevant data analytics improving the recruiting department's efficiency, choosing which university to attend is a fully emotional decision. Predictive models also have no way to account for unethical recruiting practices (such as cash payments) or underlying familial components that affect a recruit's decision. While it

is hard to believe that recruiting will ever become a fully ethical process, these models increase the efficiency of what the coaching staff can control. Moving into the future, this performance model should be validated on current UVA players, and more data should be collected to improve the model. In the long term, similar analytics strategies can be used to model recruiting in other college sports as well.

Acknowledgment

This work was supported by the University of Virginia Football Team.

References

[1] Walter, L., Citera, A., Knowles, K., et al. 2017. Implementation of a Recruit Visualization Tool for UVA Football. 2017 Systems and Information Engineering Design Symposium, 1-6.

[2] DiNitto, M. (2018). Texas athletics department continues to post staggering financial numbers. *Sporting News*. Retrieved from http://www.sportingnews.com/ncaa-football/news/ncaa-texas-athletics-department-finances-revenues-expenses-football-basketball/33vdxfzfvtvvzkn6ezzor7wo

[3] Gaines, C. (2016). The average college football team makes more than the next 25 college sports combined. *Business Insider*. Retrieved from http://www.businessinsider.com/college-sports-revenue-2016-10.

[4] Milrabile, M. (2014). The Determinants of Attendance at Neutral Site College Football Games. *Managerial and Decision Economics, 36*:191-204.

[5] Holland, R. (2015). Just how much is one win worth in college football and basketball? *Forbes*. Retrieved from https://www.forbes.com/sites/hbsworkingknowledge/2015/10/26/just-how-much-is-one-win-worth-in-college-football-and-basketball/

[6] Brady, E., Kelly, J., & Berkowitz, S. (2015). Schools in power conferences spending more on recruiting. *USA Today*. Retrieved from https://www.usatoday.com/story/sports/ncaaf/recruiting/2015/02/03/college-football-recruiting-signing-day-sec-power-conferences/22813887/

[7] Clotfelter, C. (2011). *Big-time Sports in American Universities*. Cambridge University Press.

[8] WarRoom (n.d.). Retrieved from https://www.collegewarroom.com/

[9] Steinberg, L. (2015). Changing the Game: The Rise of Sports Analytics. *Forbes*. Retrieved from https://www.forbes.com/sites/leighsteinberg/2015/08/18/changing-the-game-the-rise-of-sports-analytics/

[10] Lewis, M. (2004). *Moneyball: The Art of Winning an Unfair Game*. New York: WW Norton & Company.

[11] Cassilo, D. & Sanderson, J. (2017). "They Hired a Baseball Guy": Media Framing and Its Influence on the Isomorphic Tendencies of Organizational Management in Professional Football. *International Journal of Sport Communication*, *10*(3), 290-306.

[12] Porreca, R. P. (2016). General Managers and the Importance of Using Analytics. *Sport Journal*, 19.

[13] Stewart, M., Mitchell, H., & Stavros, C. (2007). Moneyball Applied: Econometrics and the Identification and Recruitment of Elite Australian Footballers. *International Journal of Sport Finance*, *2*(4), 231-248.

[14] Eckert-Fong, T. (2016). Stanford Football Recruiting: Success with Second-Tier Recruits. *SBNation*. Retrieved from https://www.ruleoftree.com/2016/2/10/10914618/stanford-football-recruiting-success-with-second-tier-recruits

[15] Bigsby, K., Ohimann, J., & Zhao, K. (2017). Online and Off the Field: Predicting School Choice in College Football Recruiting from Social Media Data. *Decision Analysis*, 14(4):261-273. Retrieved from https://doi.org/10.1287/deca.2017.0353

[16] Nusser, J. (2016, Jan 31). Rival, Scout, ESPN, 247: Star Ratings Explained. *SBNation*. Retrieved from https://www.cougcenter.com/wsu-football-recruiting/2013/2/5/3956800/rivals-scout-espn-247-star-rating-system-national-signing-day

[17] Kirshner, A. (2018, Jan 26). College Football Recruiting Star Ratings: How Rare Is It to Be a 5-Star. *SBNation*. Retrieved from: https://www.sbnation.com/college-football-recruiting/2018/1/26/16936186/recruiting-stars-rankings-high-school-football

[18] Division I Academic Eligibility (n.d.). Retrieved from http://www.ncaa.org/about/division-i-academic-eligibility

[19] UVA Requirements for Admission (n.d.). Retrieved from http://www.prepscholar.com/sat/s/colleges/UVA-admission-requirements

[20] Coaches By The Numbers (n.d.). Retrieved from http://coachesbythenumbers.com/

[21] 2018 Top Football Recruits (n.d.). Retrieved from https://247sports.com/Season/2018-Football/RecruitRankings?InstitutionGroup=highschool

[22] Top 100 Recruits (n.d.). Retrieved from http://www.maxpreps.com/signingday/football/top_athletes.aspx

[23] Census Data (n.d.). Retrieved from https://www.census.gov/data.html

[24] Twitter (n.d.). Retrieved from https://twitter.com/

[25] Advanced Research Computing Services (n.d.). Retrieved from https://arcs.virginia.edu/rivanna/

[26] Kaushik, S. (2017, February 15). How to Build Ensemble Models in Machine Learning. Retrieved from https://www.analyticsvidhya.com/blog/2017/02/introduction-to-ensembling-along-with-implementation-in-r/

Author Information

Kunrui Peng, Undergraduate Student, Department of Systems and Information Engineering, University of Virginia

Jonathan Cooke, Undergraduate Student, Department of Systems and Information Engineering, University of Virginia

Annie Crockett, Undergraduate Student, Department of Systems and Information Engineering, University of Virginia

Dongmin Shin, Undergraduate Student, Department of Systems and Information Engineering, University of Virginia

Austin Foster, Undergraduate Student, McIntire School of Commerce, University of Virginia

Jacob Rue, Undergraduate Student, Department of Systems and Information Engineering, University of Virginia

Roby Williams, Undergraduate Student, Department of Systems and Information Engineering, University of Virginia

John Valeiras, Undergraduate Student, Department of Systems and Information Engineering, University of Virginia

William Scherer, Professor, Department of Systems and Information Engineering, University of Virginia

Chris Tuttle, Principal, Augmented

Stephen Adams, Senior Scientist, Department of Systems and Information Engineering, University of Virginia

Mark Rhodes, Consultant, Strategy by Design