William Fleischer, Alec Shimell, William Starr, & Krishna Subramanian
Professor Valdes
April 20th, 2018
Data Mining (BUSSCM 1760) Final Project - College Basketball Analysis

---

Motivation

Collegiate Basketball has become a massive industry in the United States, with some of the top coaches making nearly $9 million annually and the top programs valued over $300 million. With the enormous sums of money invested into college basketball it crucial for Universities to either sustain their program's success or to grow their program into one that will have sustained success. (Appendix, Section 1)

There are numerous benefits that can arise through a successful College Basketball program. First, teams who perform well enough in their regular seasons to earn a bid into the March Madness Tournament earn, *at a minimum*, $1.7 million and receive an additional $1.7 million for each additional round they advance. Second, many potential undergraduate students consider school athletics when deciding a University to attend and having a successful basketball program can help to draw many bright students their school. (Appendix, Section 1) The questions we sought out to answer are:

1. How can we best predict Team Wins for a given season?
2. What variables are most important for schools to consider when recruiting the top high school recruits?
3. What revenue can teams expect for any given season based on their projected tournament performance?

Introduction to Data Set

We used numerous data sources for our data analysis. For our question determining variables that best predict Total Team Wins for a given season, we used data for the 2017-2018 NCAA Division 1 Basketball season for all 351 teams. For our data to answer the question of how many 4 or 5 star recruits a team will get we manually created a database containing information on Class of 2021 and 2022 4 and 5 recruits committed to each "Power Five" (Appendix, Section 3) school, Money Invested into the Athletic Program for each school, School Enrollment at each "Power Five" school, and Average ACT Score at each "Power Five" school. For the ordinal Regression we used historical data on Probabilities of NCAA Tournament Wins by Seed. (Appendix, Section 2)

Predicting Team Wins with Linear Regression (Unabridged R Code in Appendix, Section 5)

The first method we used to predict teams wins using past data was basic linear regression. Our data contains multiple variables that are highly correlated with others, so when creating our linear regression model the ones removed were Field Goals Made, Free Throws

Made, Three Pointers Made, and Offensive Rebounds. (Appendix, Section 4) The shot type made and shot type percentage variables for shooting stats convey similar information, so selecting either as the an independent variable is appropriate for a predictive model. Offensive rebounds were removed from the linear regression because it is already factored into the variable Total Rebounds.

Prior to creating our linear regression model we had preconceived notions on the importance of each variable. The amount of each shot type may provide useful insight in that it may be valuable for a team to have a player who is not afraid to attempt shots frequently and wants to have the basketball in their hands. Free throw attempts may be especially useful as it can give an idea of how well a team draw fouls from the opposition. The percentage of each shot made is useful in that shows how efficient a player is with the shots he does attempt.

After removing the variables listed above, the variables found to be significant were: Minutes Played, Field Goal Percentage, Field Goals Attempted, Three Point Percentage, Three Points Attempted, Total Rebounds, Steals, Turnovers, Personal Fouls, and Assists. The non-significant variables were: Free Throws Attempted, Free Throw Percentage, and Blocks. This linear regression did very well when predicting wins on new data, as evidenced by an Out of Sample R-Squared value of 86.11%.

After removing numerous combinations of variables from the dataset to find the best model, the optimal combination of removed variables (as measured by Out of Sample R-Squared) is removing Free Throw Percentage and Free Throw Attempts in addition to removing the redundant variables mentioned previously. In this model, the only insignificant variable was Blocks. The top three significant variables with positive beta values are: Field Goal Percentage, Three Point Percentage, and Steals. The only two significant variables with a negative beta value are: Turnovers and Field Goal Attempts. The Out of Sample R-Squared Value for the optimal model is 87.37%, an improvement of 1.26% from the model where no non-redundant variables were removed.

Some valuable insights can be taken away from the optimal linear regression model. First, college recruiters should look for players who are very efficient on offense. The two variables with the largest positive beta value both measure how efficient a player is at making shots and therefore it is in a college basketball program's best interest to recruit players who perform well in this measure. The model also places negative value on players who turn the ball over to the opposition and attempt too many shots. Therefore, the model tells recruiters it is best to field a team of players who are extremely efficient on offense, those who do not voluntarily turnover the ball to the opposing team and make a large percentage of the shots they attempt. The model also implies Free Throws are not important factors to consider when determining the number of games, a team will win. This is contrary to popular opinion, as many executives and fans believe Free Throws are crucial to winning games, especially those that are highly competitive in the final minutes of regulation. The possible downsides of linear regression are mentioned in the Appendix, Section 6.

Predicting Team Wins with Regression Tree (Unabridged R Code in Appendix, Section 7)

The second analytical method utilized to predict Team Wins was a Regression Tree. Using a *complexity parameter* of 0.01, a series of five regression trees were made to compare size, structure, included independent variables, and model simplicity. The difference in the trees were its respective *minbucket* values, or the minimum number of observations to be in each terminal node of the tree. The training set of the dataset contained 263 observations and therefore the tested *minbucket* values were 5 (Tree 1), 10 (Tree 2), 20 (Tree 3), 35 (Tree 4), and 50 (Tree 5) - (Appendix, Section 8). When comparing the multiple trees it is important to find an appropriate balance between predictive power and usefulness, as well as simplicity and complexity. Trees 4 and 5, though very simple to interpret, were based predictions from one or two variables with four and three terminal nodes, respectively. Although it would be easier to explain these predictive models for predicting wins, the value of tree would be very low as wins cannot typically be grouped into three or four categories. However, decreasing the *minbucket* value consequently increases both the size complexity of the tree. Tree 1 yielded a twelve terminal-noded tree that, though may offer greater insight into the predictor variables of Team Wins, is rather complex and uninterpretable. Trees 2 and 3 appeared to present a effective balance between simplicity and applicability, and Tree 3 (*minbucket* = 20) was ultimately selected as the best primarily due to its size. Tree 3 resulted in six terminal nodes, an appropriate number of different Team Win groupings.

Proceeding with the *minbucket* parameter of Tree 3, cross-validation was utilized to determine the optimal *complexity parameter* (Appendix, Section 8). Through this process the greatest *complexity parameter* value within one standard deviation of the standard-mean error was 0.021782. Then, using these two parameters a final regression tree was produced. The Final Tree (Appendix, Section 8) incorporated the predictor variables: Total Field Goals Made, Total Minutes Played, and Total Blocks. Total Blocks is an unexpected independent variable when originally reviewing the dataset. Although this model does not imply causation, it was interesting to figure a high correlation exists between blocks, common with aggressive and efficient defensive teams, and Team Wins. Furthermore, the regression tree contained five total splits, and three of those splits were based on Field Goals Made. Using the model as a predictive tool the calculated out-of-sample R^2 was 59.52% and the Mean Absolute Error was 3.45. In terms of the Mean Absolute Error, though may appear high on the surface, needs to be factored into the context. One of the core entertainment factors of college basketball is its unpredictability, and therefore 3.45 is a very reasonable that confirms value in the model. In sum, the regression model, compared using out-of-sample R^2, performed significantly worse than the linear regression model, and though it provided valuable insight is not the best method to predict Team Wins.

Predicting Team Wins with Artificial Neural Network (Unabridged R Code in Appendix, Section 10)

The final method we used to predict wins was an Artificial Neural Network(ANN). The first step in any neural network is data standardization, which was accomplished by scaling the data using min and max values. After doing that, we need to create a benchmark win "predictor" which is just using the mean of the test set to predict each win.

Finally, we get into our neural network prediction. Using the neuralnet function derived from the "neuralnet" library, we place all independent variables into it, with linear output set to true, and ran different hidden layer variations. Interestingly, the best hidden layer combination was 4 nodes in first layer, and 2 nodes in second, which when we compute our out-of-square regression, gives us a predictability of 85.7%. While this is a very high value, it is slightly lower than our regression line, which combined with neural network's "black box" nature, makes it a much less optimal predictor to go with. (Appendix, Section 11)

Team Seeding Ordinal Logistic Regression (Unabridged R Code in Appendix, Section 12)

When considering the managerial problem of financial allocations for college basketball programs, forecasting expected monetary gains from the NCAA Tournament is imperative. To better increase one's statistical likelihood of winning the tournament, the team would prefer to be a lower seed on the 1-16 scale. Using historical statistical probabilities (Appendix, Section 14), the lower a team is seeded, it has a much greater chance of advancing to the next round and thus gaining an additional $1.7 million for each tournament win. One of the most highly correlated variables with seeding is Team Wins, however this cannot be used as the sole predictor variables. Based on the shear size of certain schools and conferences, different programs face a different caliber of total oppositional talent throughout the year. Therefore, a historically successful team, such as Duke or Kansas, may accumulate a large number of wins against top-tier talent. While a smaller school, such as Loyola (Chi.) may accumulate the same number against objectively inferior talent. This is not only an issue with the models performed in this analysis, but in college basketball as a whole -- appropriately weighting the "quality" of a Team's Wins against another. For the ordinal logistic model, the imperfect weighting metric used as Strength-of-Schedule (SOS), a mathematical measure published by ESPN to rate the overall difficulty of the opponent's a team has faced that year.

The ordinal logistic regression model was created using only teams in the 2018 NCAA Tournament, based on Team Wins and SOS. Then, the Linear Regression model predicted a win total for each team in 2018 NCAA Tournament, and the ordinal logistic regression model made a seeding projection based on the predicted win total. With the projected seed value any team could utilize historical probability to estimate an expected revenue through its NCAA Tournament Performance (Appendix, Section 15). The logistic regression model provides immense value to financial and recruiting divisions of collegiate athletics. Often, teams beginning recruiting seasons in advance and having an approximate budget for expenses, as

well as an analytical prediction of seasonal performance, will truly assistant management in many aspects of the process.

Predicting 4- and 5-Star Recruits with Classification Tree (Unabridged R Code in Appendix, Section 17)

When further exploring the financial impact in college athletics, universities can save millions in their recruiting expenses, while simultaneously growing their revenue, by attracting top-tier talent. The first, of many, issues we ran into was the lack of an available dataset with all of the variables of interest or relevance. In the end, we compiled a manually-created dataset with numerous variables we thought would be pertinent to the question. Taking all Power-5 conference schools, as well as adding a few additional powerhouse-schools. Our dataset ended up with a total of 72 observations. The independent variables included were: school enrollment (undergraduate and graduate students), recruiting expenses, head coach salary, and the average ACT score for admitted students. The last step for cleaning this data was removing the outliers, Duke, Kentucky and UCLA for having a total of ten 4- and 5-star recruits for 2017 and 2018.

The original goal was to predict 4- and 5-star recruits separately from one another with a regression tree. Almost instantly, the size and lack of diversity in the dataset proved it was going to be a larger than anticipated hindrance. Giving us a negative out-of-sample R^2, guessing the average every time being more accurate, it was clear this route was just not possible with our dataset. The next step was molding (broadening) our question into a form our data would have a fair shot at answering. Our end result, was attempting to predict whether a school will have above, or below, the average ($\overline{X}$ = 3.116) total for 4- and 5-star recruits. Setting the *minbucket* at 3 and the complexity parameter at 0, I was able to find the optimal complexity parameter of 0.052632. The final tree (Appendix, Section 19) ended up with two splits, the first being whether the previous year's recruiting expenses were less than $1.2 million. The next split (to the right side) was whether the head coach's salary was more than $984 thousand dollars. The accuracy was 61.11%, specificity 80%, and sensitivity 37.5%. It was a lot better at predicting teams to be under the average, as opposed to above. We were unable to think of a correct way to implement a loss-matrix. There were too many moving-parts to try and come up with an arbitrary "loss" value for a false-positive. For example, a 5-star recruit this year ended up only playing four games the entire season. Injury and other factors often make highly-ranked recruits underperform. The intuition taken away from the classification tree was, each athletic program has a budget they must abide by, so already allocating over $1.2 million to recruiting expenses, and paying their coach under $984,000, it gives the school more resources to put towards wining/dining and impressing their recruits.

Conclusion/Summary

College basketball, one of the largest branches of collegiate athletics, stands as one of the most profitable and monetary-driven programs in all of sports. The conducted analysis encapsulates a program's notable expenditures into three subcategories: personnel, performance, and recruitment. Personnel focuses on the unique skill sets and attributes programs should seek to acquire in prospective players. As the goal of any school is to maximize Team Wins, three models were produced to best understand the imperative variables correlated to more Wins. Of the Linear Regression, Regression Tree, and Artificial Neural Network models, the Linear Regression was best predicting Teams Wins on external datasets. In the Linear Regression model, while many of the variables such as Field Goal, Rebound, and Assist related measures were expected, some such as Minutes Played and Personal Fouls were surprises. Using this information we can accurately deduce that coaches may opt to seek out multi-sport athletes or those who are associated with track and cross-country. Finding athletes that also excel in these different activities can give teams an advantage in terms of depth, stamina, and endurance, all relating to Minutes Played. Similarly, while fouling another player is typically considered a negative, the Linear Regression output expresses this variable as positively correlated with Team Wins. For interpretation, a coach may find immense value in aggressive defensive players, that sometimes make mistakes, but in the end reward their team with that aggression.

Next, the performance of each team was analyzed using ordinal logistic regression. Teams that recruit based on the most pertinent variables for increasing Team Wins count, can now use their statistics to make financial projections for university gains and potential future program expenditures. From a managerial perspective, being able to effectively and accurately propose an annual budget, and present scouts and coaches with the tools to properly being recruiting efforts is crucial in identifying prospective talent and staying ahead of the competition.

Finally, the factors associated with attracting and signing recruits based on school demographics allows for Universities to see how they can improve upon their current recruiting efforts, and pinpoint necessary changes they must ultimately make. Attracting premier talent to their respective University will likely allow for the program to become more successful and thus result in higher revenues. All in all, these methods will assist teams in targeting certain numbers and types of 4- and 5- star recruits, and better allocate program funds and create the most attractive collegiate atmosphere.

In one of the most lucrative sectors of athletics, figuring the best tactics for researching and acquiring talent, properly allocating financial resources, and establishing and achieving long-term program objectives can provide institutes with an unparalleled advantage. Through the evaluation of the college basketball space, variables and statistics that are greatly related to improving these coveted goals were analyzed, in order to provide collegiate athletic leaders with a better understanding of the strategies in which teams can be constructed and managed for future growth and sustained success.

Appendix

Appendix, Section 1: Financial Information
Data on team valuations from: College Team Valuations

Data on revenue per NCAA Tournament win from: Tournament Win Revenues

Appendix, Section 2: Data Sources
Team statistics data source: Team Statistics

Data on 4 and 5 star recruits by school from: Recruit Data

Athletic Program Investment and Coach Salary data from: Athletic Investment

Average ACT Score and School Enrollment is from each individual school's website

Historical Probabilities of NCAA Tournament wins by seed is from: Win Probabilities

Appendix, Section 3: College Basketball Key Concepts Explanation
Power-Five schools refers to all Universities that are a members of one of the five most prolific athletic conferences. This includes the ACC (Atlantic Coast Conference), SEC (Southeastern Conference), Big 10, Pac 12, Big 12. A list of all the schools include can be found in the following link: Power Five Conferences. Also, there are some wildly successful Basketball programs that are not a part of the Power-Five that were included in our data set. The schools included are: Villanova University, University of Cincinnati, Saint Mary's College, Butler University, Wichita State University, Xavier University, and Gonzaga University.

The NCAA Tournament is the National Championship tournament for the college basketball season. The format is a field of 68 teams, with four "play-in" games. The winner of the four games reduce the field to 64 teams, divided into four geographic regions. Each region has 16 teams, with the 1-seed playing the 16-seed, 2-seed playing the 15-seed, etc. The winner of each region is invited to the Final Four, the biggest stage for college athletics.
The seeding is determined subjectively by a committee of collegiate athletic professionals and analysts. While the criterion of how the teams are seeded is both unknown and varies each season, there are strong correlations between the seeding and the intra-season polls, which rank the teams throughout the year. These polls are typically made based on Team Wins and multiple weighting variables to measure oppositional talent.

Appendix, Section 4: Correlation Between All Variables in Team Statistics Data