Using Machine Learning to Predict the Tournament Outcomes of the Power 5 and Big East Conferences for Men’s NCAA Basketball

Riley Greene

Connor Finn

MECE4520: Data Science for Mechanical Systems

Professor Joshua Browne

Columbia University

Abstract

The goal of this project was to predict the tournament outcomes of the Power Five and Big East conferences for men's NCAA Division I college basketball. Statistical data for teams in all six conferences for regular season play and tournament play was collected from sports-reference.com. Data was collected for the previous 5 seasons for every team in every conference. Once the data was scraped and cleaned, it was put into databases where further analyses were performed. Jupyter notebook was used in conjunction with Pandas and Beautiful Soup, for data collection, analysis and extraction. The current deliverable yields every team's probability of beating every team in their conference with a testing accuracy of ~0.70. Further analyses were done to predict the outcome of each conference tournament. The models were found to select the correct winner of the Big East Tournament with 36% accuracy, the Big 10 Tournament with 65% accuracy, the Big 12 Tournament with 58% accuracy, the Pac 12 Tournament with 40%, the ACC Tournament with 24% accuracy, and the SEC Tournament with 27% accuracy. These results were compared with those of models based on random selection of a winner, and tournament simulations where each game's odds were indicated to be even. The machine learning model outperformed both other methods.

Keywords: machine learning, basketball, python, college sports

Using Machine Learning to Predict the Tournament Outcomes of the Power Five and Big East Conferences for Men's NCAA Division I Basketball

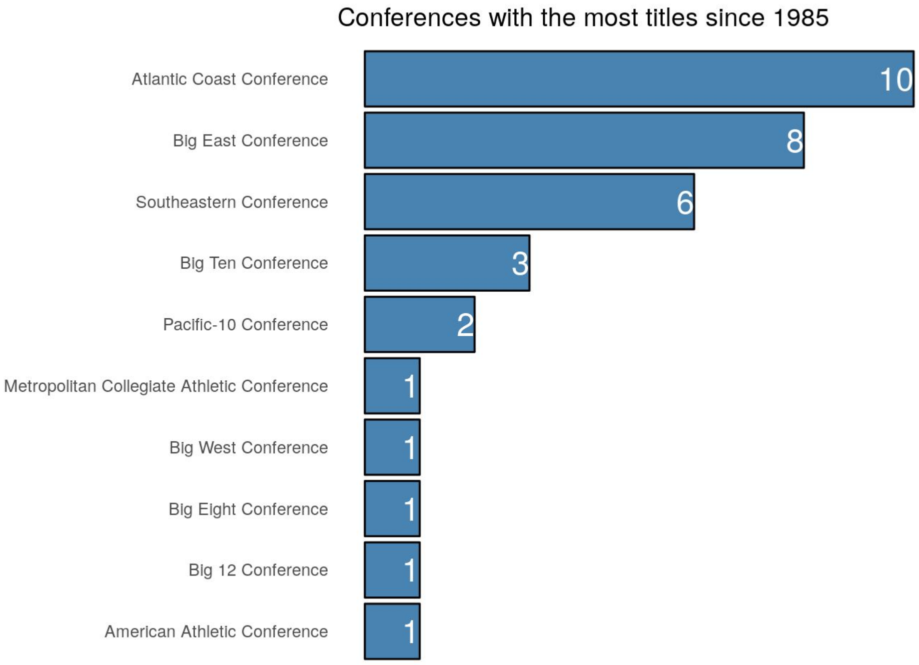
Our semester began with a lecture by Dr. Erik Allen. Professor Browne introduced us to Dr. Allen by telling us that, "This is what a data scientist looks like." He continued to show us data scientists throughout the year during their lectures that guided us in the direction of the class project. Data scientists look like: Dr. Allen, Austin Sendek, Alejandro Mesa, and Dr. Cassie Borish to name a few. Instantly our team was infatuated with this class after Dr. Allen's tales of Moneyball during the very first lecture. Both of our lives have been heavily influenced by athletics so naturally a project that combined our passions with the course material made sense. When given the opportunity to write a paper on the subject of our choice versus actually getting into the weeds of the data - we wanted to get our feet wet. How often do you have the opportunity to email a handful of Ph.D's on a weekly basis with coding and architecture questions relevant to your project?

The project we chose to pursue was to predict the tournament outcomes of the Power Five and Big East conferences in men's college basketball. Basketball was chosen because of the sheer amount of statistical data available to us freely online. As a collegiate skier and soccer player - we would have loved doing a project involving either of these sports but the amount of data available is scarce and difficult to collect. Additionally, using this model to beat our friends who claim to be, basketball inclined, in the March Madness tournament in the future would be a lot of fun.

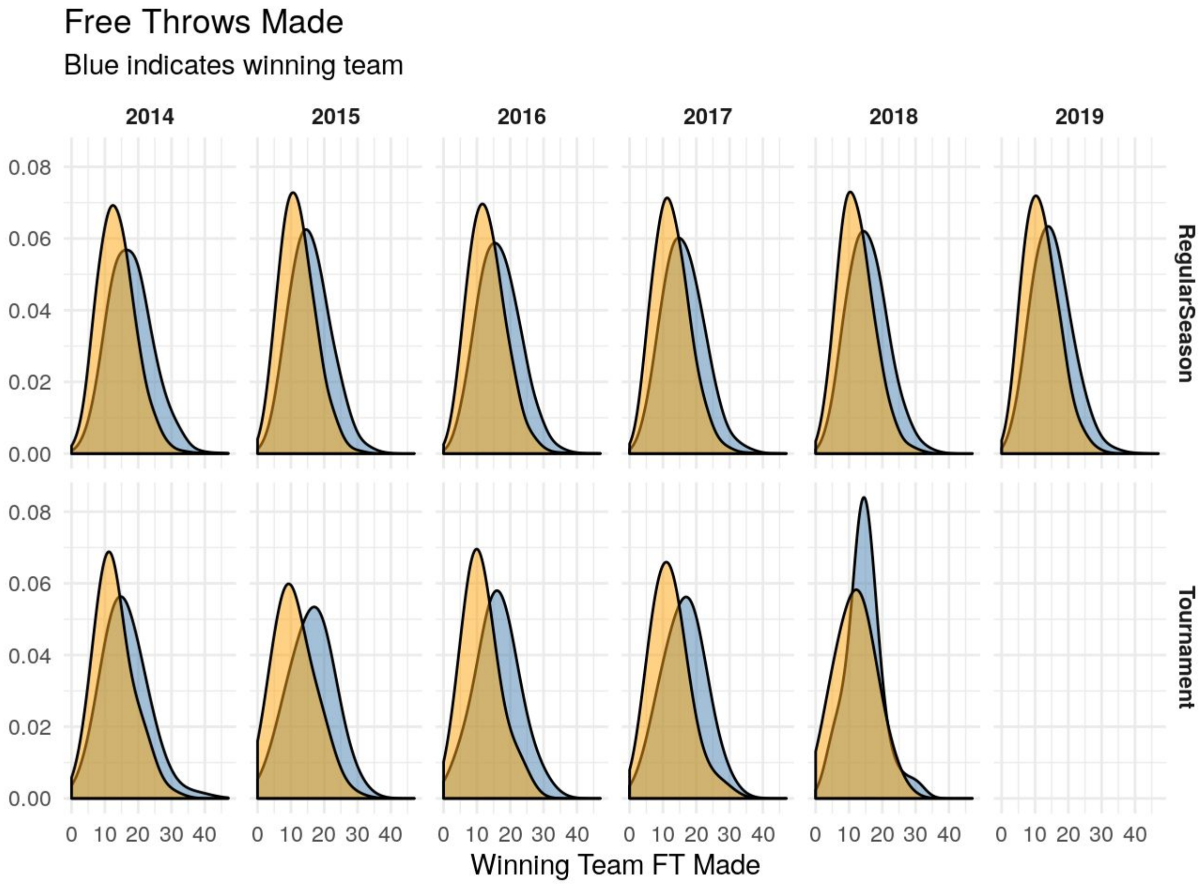
Literature Review

It was no surprise to us that when we looked into the details of work that had previously been done on this subject - there were plenty of extensive and complex projects that achieved similar goals. Some of the most fascinating papers encountered included, an extensive statistical model written by Jason Zivkovic on Kaggle, an article written by Daniel Oberhaus for WIRED magazine on the subject, and an academic paper titled, “A Machine Learning Approach to March Madness” by Forsythe and Wilde.

Some of the most interesting statistics relevant to the model we built encountered in Zivkovic’s analyses were: the most dominant conferences since 1985, and various team statistics both in the regular season and tournament play. Below are a couple visualizations provided by Zivkovic, J. (2019).



As is shown here, the conferences we are analyzing are credited as being the most dominant in basketball. The Power Five consist of: the SEC, the Pac-12 (formerly Pac-10), the Big 12, the ACC, and the Big Ten. We will also be looking at the Big East Conference.



The diagram above shows data regarding Free Throws Made in the regular season versus in tournament play. There were many of these charts shown in Zivkovic's analysis. We chose to show Free Throws Made as they are generally considered a very important statistic, highly correlated to winning games in college basketball, and this confirms that common knowledge.

Oberhaus, D. (2019) for WIRED indicated that this year an estimated $8.5 billion was spent betting on the outcome of the NCAA tournament. Sports may seem like fun and games but data scientists, statisticians, and athletes alike are doing everything they can to win sums of money like that by correctly picking winning teams in brackets. As the NCAA has become more popular throughout the years, the odds of correctly filling out a bracket have fallen off the face of the planet. In 1939, with 8 teams in the tournament your odds were 1 in 128, now the odds of correctly picking all 64 teams is 1 in 9.2 quintillion (9.2 followed by 18 zeros) - if each team's winning probability was 50%, but we know this isn't the case. If the 1 seed versus 16 seed games are treated as automatic wins (which isn't always a given, as Virginia showed us last year), the odds of picking a perfect bracket substantially increase to 1 in 2.4 trillion. Oberhaus continued in the article by explaining that March Madness has a very good dataset available for people trying to build machine learning models for tournament outcome predictions, but this data has realistically only become accessible in the past ten years or so. The training data is often as important as the statistics themselves when building these models. The more you can train the model you build, the better the model learns and the accuracy will increase accordingly.

Forsythe and Wilde, who studied computer science at Brigham Young University (BYU), considered what methods would give them best prediction odds for classification accuracy for March Madness. Their report served to inform and inspire certain aspects of the model we built. This included, using feature reduction, random data accuracies as a comparison method, and what machine learning model we chose to use. They found that, "...random ordering used in conjunction with k-nearest neighbor, Manhattan distance, and random forest as the feature reduction algorithm provide the best classification accuracy." From Forsythe, J. & Wilde, A. (2014), Page 6. Using those algorithms, provided the pair with a top classification accuracy of 0.7362. Forsythe and Wilde also discussed being able to implement what they called "Style

Theory” into their model. If there is a very fast team playing a very tall team - the quick team will look to speed up pace of play during the game to increase their chance of winning, if it’s a reasonable probability that they can do this against the tall team - it needs to add weight to the model to predict better outcomes. They credit much of the success of their model to feature combination. This means that their model looks at steals, rebounds and free throws simultaneously, and yields the outcome that a team who has high steals and free throw percentages is more likely to beat a team with high rebound and free throw percentages.

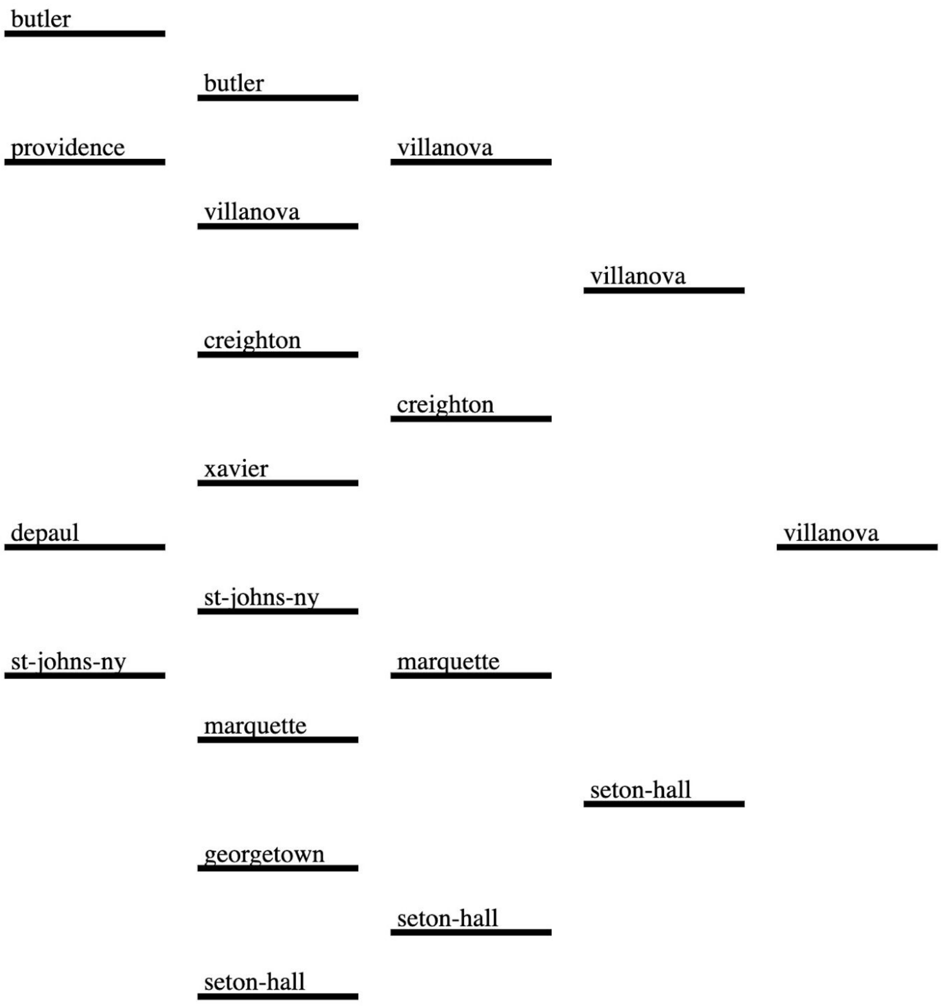
Our Model

The built program to predict conference tournament winners was done in three parts. The first file created was used to collect the necessary data. We took the season long data for all six conferences we were interested in, and organized it into a dataframe. For the 75 teams of interest, the scrypt scrapes statistics from 450 different webpages, from sports-reference.com. These statistics included like individual game scores, three point percentages, and win-loss percentages, etc. Once the data is collected, it is organized such that for each of the conference tournament games in the past five years, each tournament game and its result is written into a pandas dataframe, along with the season long stats of each team. Many of the challenges encountered during this portion of the project involved cleaning the data so it behaved the way we wanted to in the dataframe. Some of the cleaning issues encountered were weird breaks in statistics, not unlike the N/A cells discussed in class lectures. Additionally, because we were pulling so many different statistics from different web pages we had a lot of smaller dataframes which we had to compile into one complete dataframe that we could then use for analysis. Because we sought to look at individual games being played, we needed the game score, the year, and the two team’s statistics all in a single dataframe. Once the data was compiled into the primary data frame we could then move on to the second step in the model, generating the odds that a team will win a game.

A model was built for every possible game in a given conference tournament. The first step in this process was to generate a list of every possible game for all six conferences; for example for Villanova, in the Big East, this would be: Nova vs. Xavier, Nova vs. Providence, Nova vs. St. John's, Nova vs. Seton Hall, Nova vs. Marquette, Nova vs. Georgetown, Nova vs. DePaul, Nova vs. Creighton, and Nova vs. Butler. An important step in this part of the scrypt was making the distinction that no game is repeated - where Nova vs. Providence is a possible game, Providence vs. Nova should not be included. Next, the 2019 data was taken out of the primary dataframe as we used that data as case study of the model. Taking machine learning techniques we learned from class we used a machine learning model to generate the odds for every possible conference tournament game for the six conferences under study. According to Ben Allison on one of our team's new favorite resources, stackoverflow, there are two competing concerns when deciding how to size the train/test split: with less training data, your parameter estimates have greater variance; with less testing data, your performance statistic will have greater variance. We chose to use a train/test split of 80/20, following the Pareto principle. On average, the decision tree models had an accuracy of ~70%. Decision tree models were a great option for us because we talked about them a lot in class so we were most familiar with how they worked, and they don't require feature selection or normalization. We were at risk of overfitting however, so the tree was pruned to not include, FG, 2PT, 3PT, MP (field goals, 2 pointers, 3 pointers, and minutes played). These categories were chosen because they were the results of other statistics, where FGA (field goal attempts) * FG% (field goal % made) = FG, 2PTA (2 point attempts) * 2PT% (2 point % made) = 2PT, and 3PTA (3 point attempts) * 3PT% (3 point % made) = 3PT. MP was removed as overtime considerations were not being made.

The final part of the model was to test the performance of the built machine learning model on the 2019 tournaments, using our generated odds, as a case study. A scrypt was made to use a weighted random selection to choose a winning team, based off of the odds generated by the built model. The scrypt relies on Object Oriented Programming to develop the different models. Because each tournament has a different structure, a different class is generated for a game, as well as each individual tournament. Finally, the tournaments themselves are invoked. There are classes made for a single game, and all tournaments: Big East, Big 12, Big 10, Pac 12, SEC, and the ACC. Reason being is, different tournaments have different layouts. For example, in the SEC tournament, there are only two games in the first round, whereas in the ACC, there are three. The tournaments are set up such that teams are entered in increasing seed. The games are played with the generated odds and weighted random selection. The results are output for every game played, including the championship - giving us our predicted winners for all six tournaments.

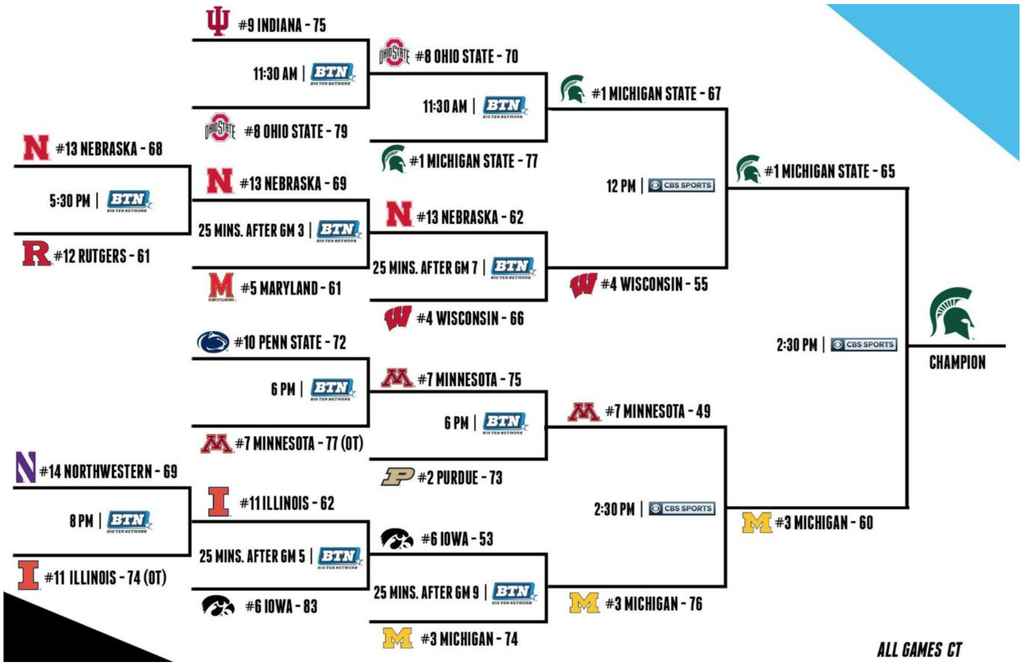
The figure below shows an example of the model predicting the Big East Conference Tournament.

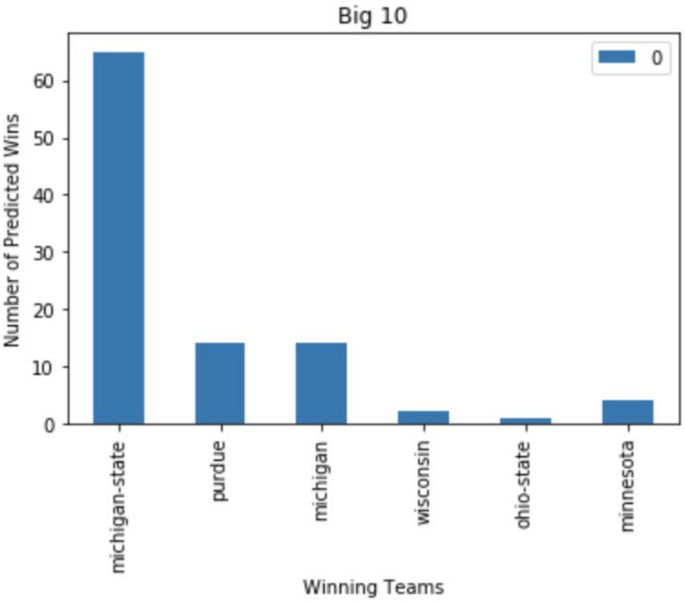


Here we can see that Villanova is the predicted winner of the 2019 Big East tournament based on our model. This was a great confirmation of our work as Villanova *was* the champion of the 2019 Big East Tournament, defeating Seton Hall.

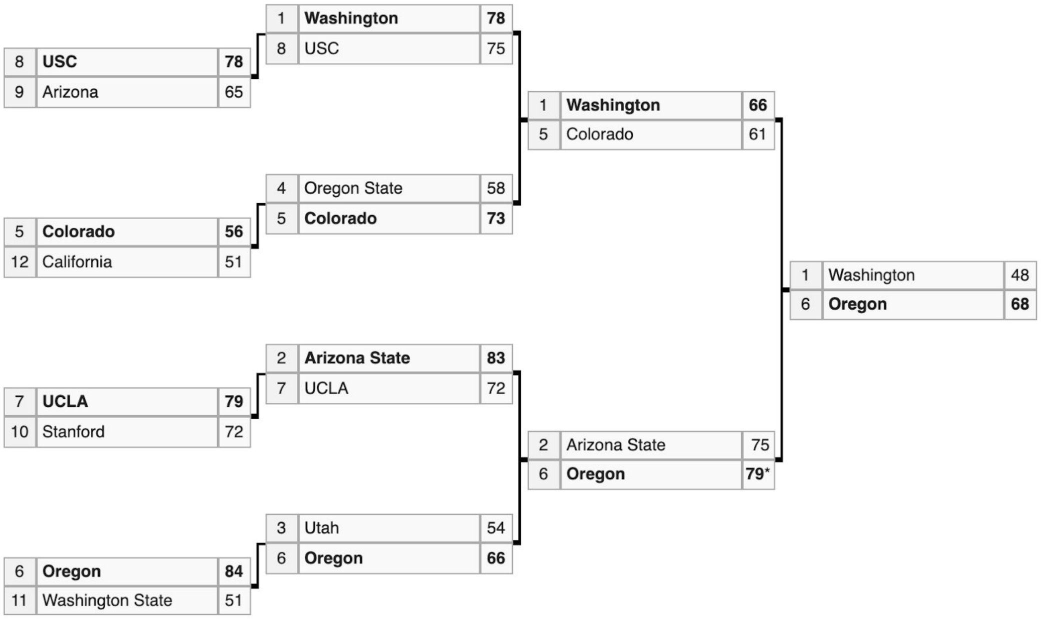
Discussion

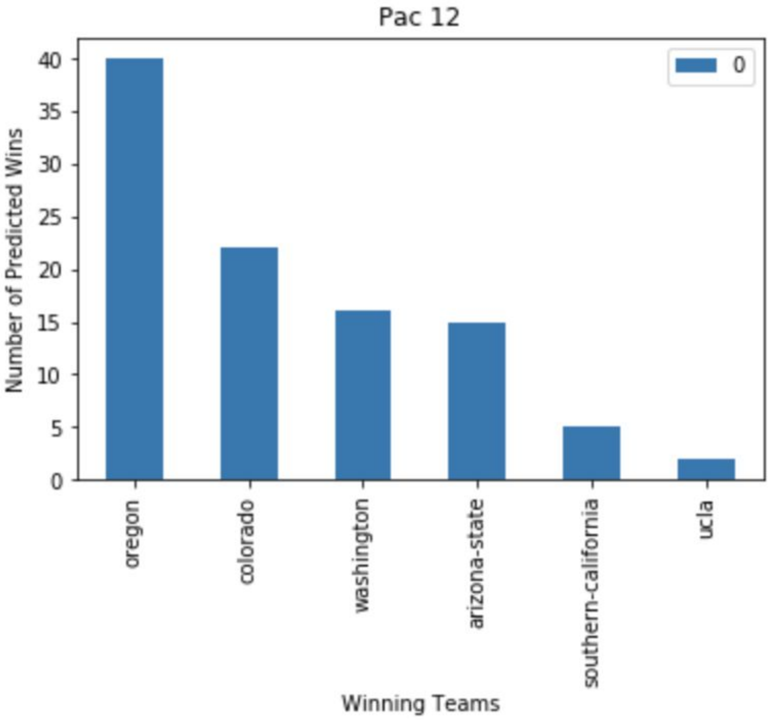
Below are all 2019 tournament results for the six conferences studied. The first images are the results of the real tournaments, and the following bar graphs are who the model predicted as the winner.



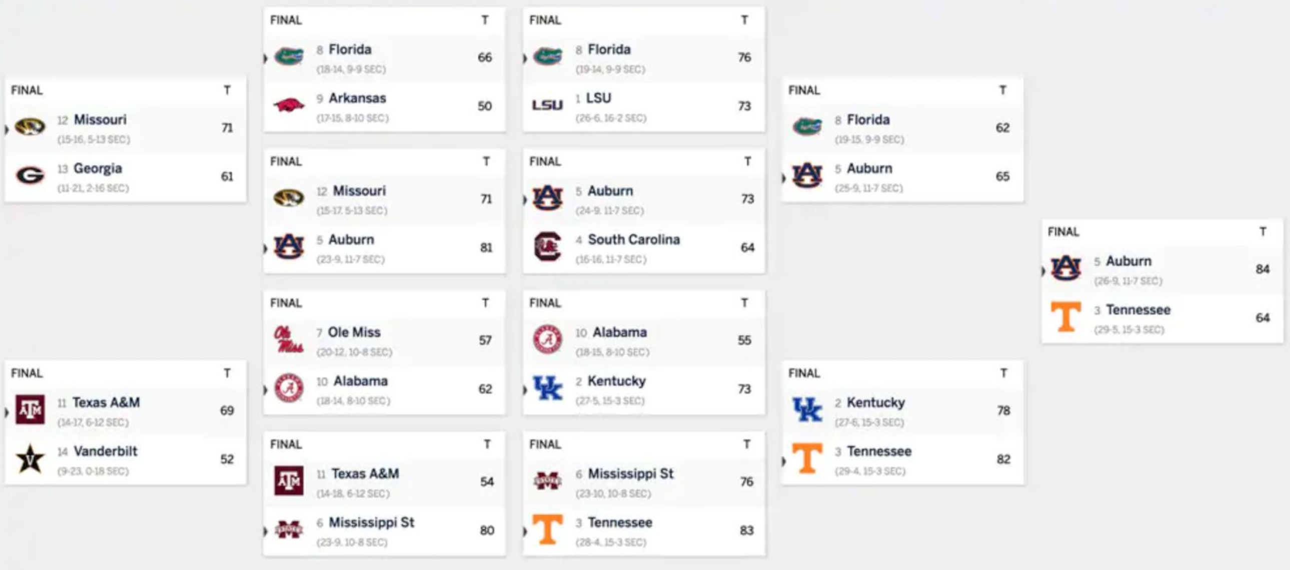


Here we see a commanding performance of our predicted winner Michigan State. This result was exciting because of how dominant Michigan State was in the NCAA tournament as well, making it all the way to the Final Four before being eliminated by Texas Tech.





The results from the Pac 12 conference were also very good. Our model accurately predicted Oregon winning their Conference championship as a six seed. As they had to play more games than the top seeded teams they were likely more exhausted but they prevailed nonetheless.



FINAL	T
12 Missouri (15-16, 5-13 SEC)	71
13 Georgia (11-21, 2-16 SEC)	61

FINAL	T
11 Texas A&M (14-17, 6-12 SEC)	69
14 Vanderbilt (9-23, 0-18 SEC)	52

FINAL	T
8 Florida (18-14, 9-9 SEC)	66
9 Arkansas (17-15, 8-10 SEC)	50

FINAL	T
12 Missouri (15-17, 5-13 SEC)	71
5 Auburn (23-9, 11-7 SEC)	81

FINAL	T
7 Ole Miss (20-12, 10-8 SEC)	57
10 Alabama (18-14, 8-10 SEC)	62

FINAL	T
11 Texas A&M (14-18, 6-12 SEC)	54
6 Mississippi St (23-9, 10-8 SEC)	80

FINAL	T
8 Florida (19-14, 9-9 SEC)	76
1 LSU (26-6, 16-2 SEC)	73

FINAL	T
5 Auburn (24-9, 11-7 SEC)	73
4 South Carolina (16-16, 11-7 SEC)	64

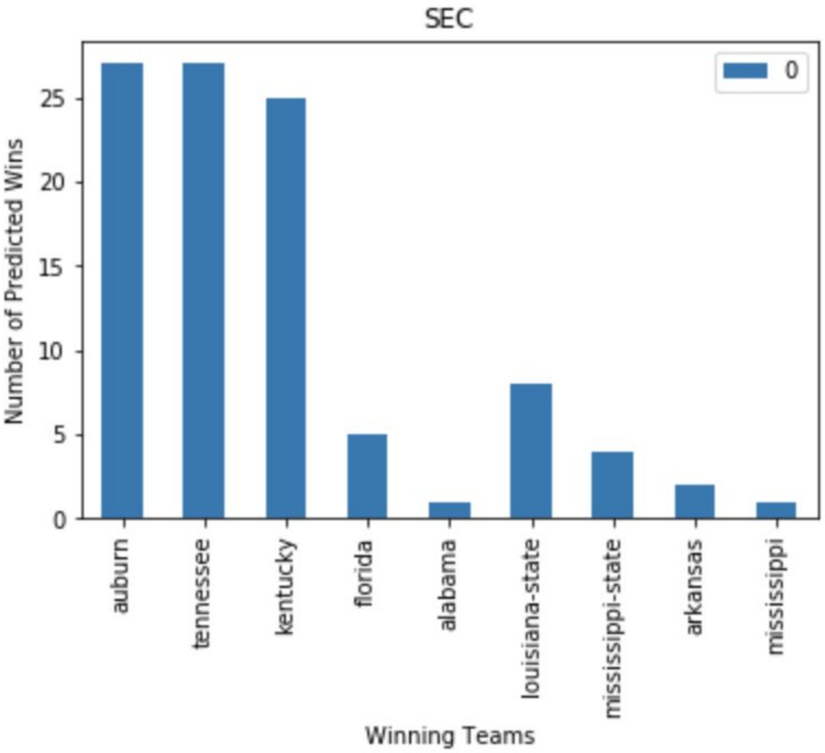
FINAL	T
10 Alabama (18-15, 8-10 SEC)	55
2 Kentucky (27-5, 15-3 SEC)	73

FINAL	T
6 Mississippi St (23-10, 10-8 SEC)	76
3 Tennessee (28-4, 15-3 SEC)	83

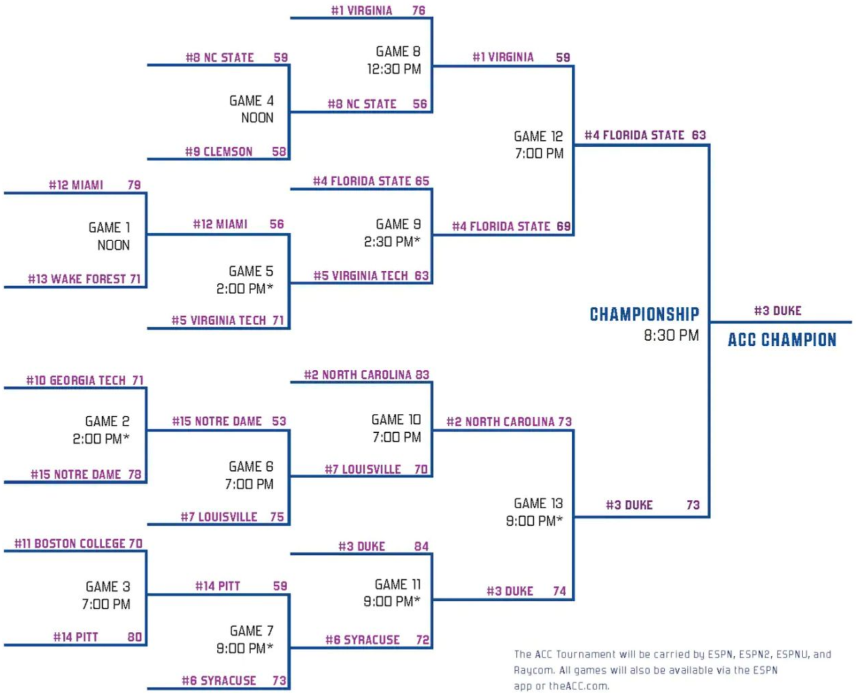
FINAL	T
8 Florida (19-15, 9-9 SEC)	62
5 Auburn (25-9, 11-7 SEC)	65

FINAL	T
2 Kentucky (27-6, 15-3 SEC)	78
3 Tennessee (29-4, 15-3 SEC)	82

FINAL	T
5 Auburn (26-9, 11-7 SEC)	84
3 Tennessee (29-5, 15-3 SEC)	64

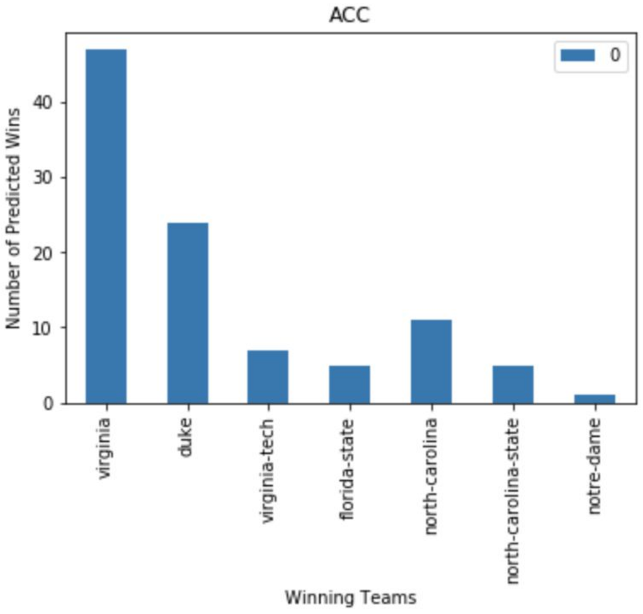


Here we see another correct prediction based on the real data versus model output. These results were fun to see because the model not only correctly predicted the winner but it also had the runner up and three out of the four teams correctly in the semi-final.

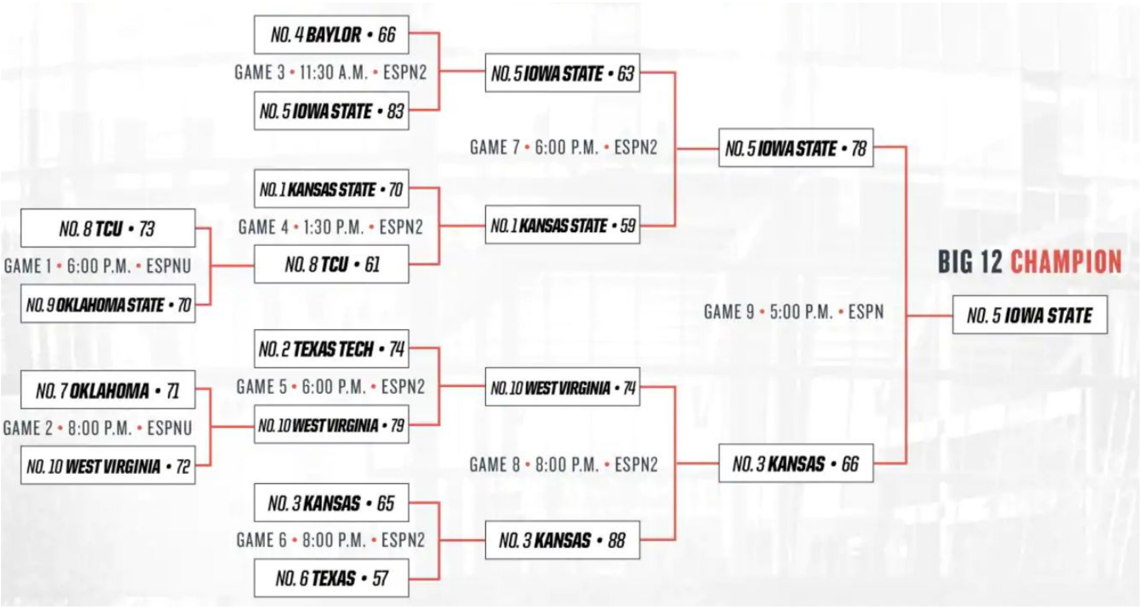


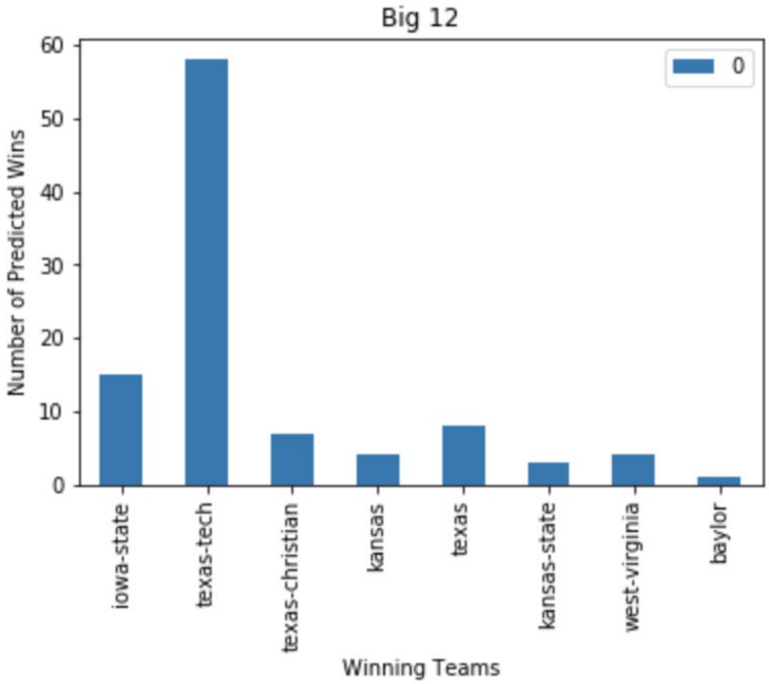
The ACC Tournament will be carried by ESPN, ESPN2, ESPNU, and Raycom. All games will also be available via the ESPN app or theACC.com.

*Denotes approximate time

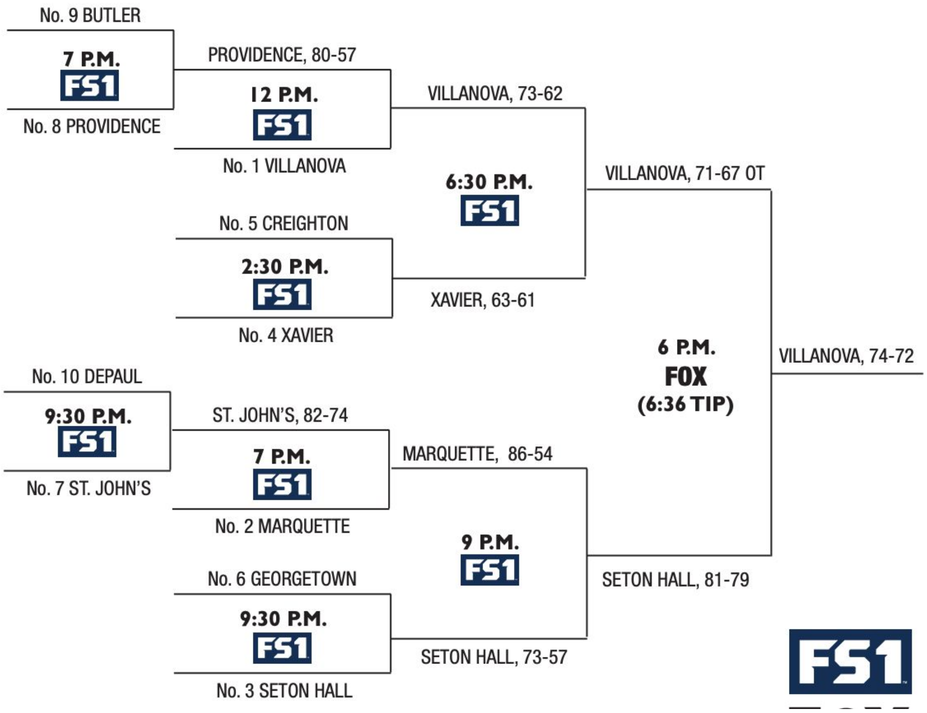


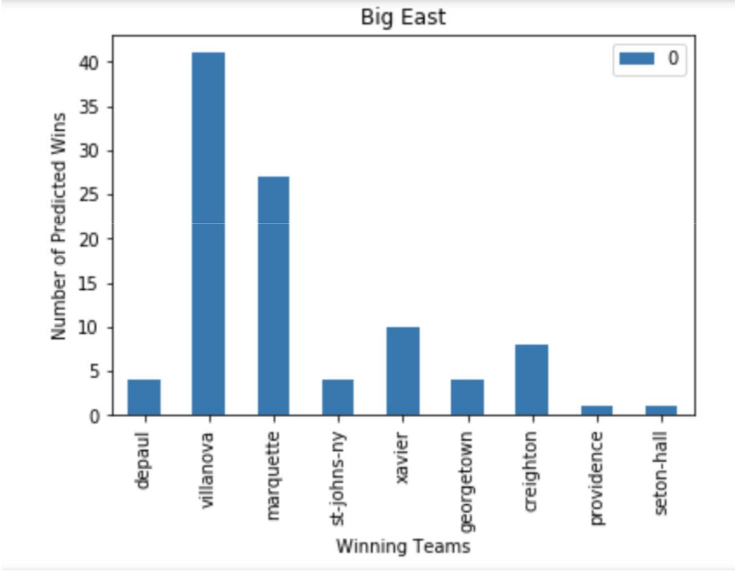
In the case of the ACC tournament our model unfortunately did not accurately predict the winner. However, it is important to note that the built model predicted Virginia winning - the team who went on to become the 2019 NCAA Men’s DI Basketball champions.


NO. 4 BAYLOR • 66
GAME 3 • 11:30 A.M. • ESPN2
NO. 5 IOWA STATE • 83
NO. 5 IOWA STATE • 63
GAME 7 • 6:00 P.M. • ESPN2
NO. 5 IOWA STATE • 78
NO. 1 KANSAS STATE • 70
NO. 8 TCU • 73
GAME 4 • 1:30 P.M. • ESPN2
NO. 1 KANSAS STATE • 59
GAME 1 • 6:00 P.M. • ESPNU
NO. 8 TCU • 61
NO. 9 OKLAHOMA STATE • 70
BIG 12 CHAMPION
GAME 9 • 5:00 P.M. • ESPN
NO. 5 IOWA STATE
NO. 2 TEXAS TECH • 74
NO. 7 OKLAHOMA • 71
GAME 5 • 6:00 P.M. • ESPN2
NO. 10 WEST VIRGINIA • 74
GAME 2 • 8:00 P.M. • ESPNU
NO. 10 WEST VIRGINIA • 79
NO. 10 WEST VIRGINIA • 72
GAME 8 • 8:00 P.M. • ESPN2
NO. 3 KANSAS • 66
NO. 3 KANSAS • 65
GAME 6 • 8:00 P.M. • ESPN2
NO. 3 KANSAS • 88
NO. 6 TEXAS • 57



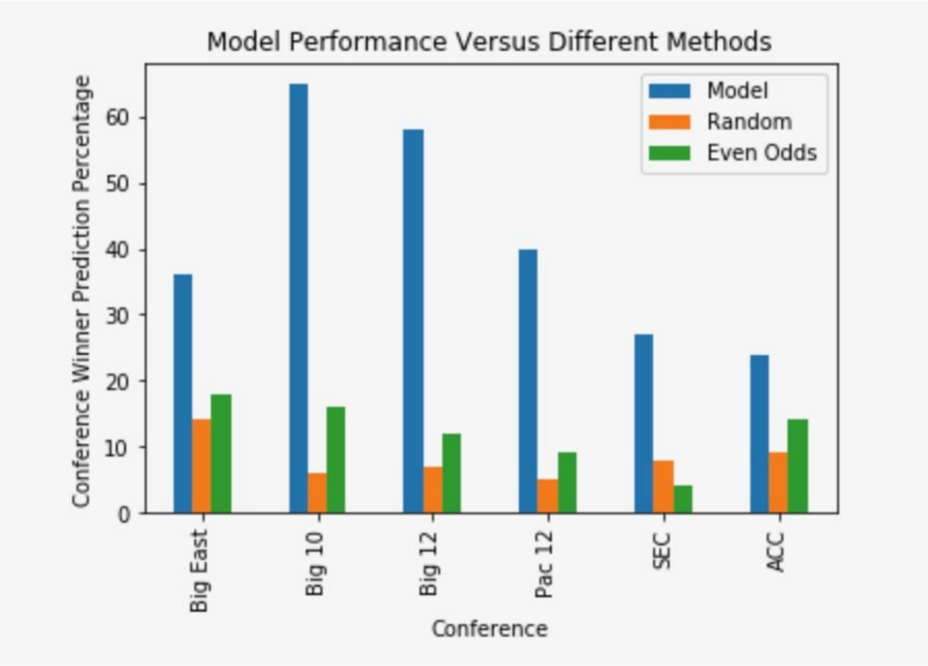
Similar to the results of the ACC tournament, the Big 12 prediction made by the built model was unfortunately incorrect. Texas Tech was eliminated by West Virginia, a ten seed, in a huge upset in their conference tournament. Texas Tech, however, moved on to play in the championship game in March Madness - which meant the model was still choosing arguably the best teams.





Lastly, the Big East tournament was accurately predicted as the model had Villanova winning it all. Interestingly, before win/loss % was included - the model had DePaul winning. This may be insignificant but, they may be a team to look for in the coming seasons, as they seem to have the skills but cannot make it to the final buzzer.

The final piece of our case study that we wanted to look at was seeing how the model performed against random selection as well as the scenario where each team was considered equal; thus each team's probability of winning a game was 50%. The graph below provides a visualization for each of the six conferences studied.



These results were interesting to see as random, and even odds remained relatively constant independent of conference. However, the built model fluctuated a bit more from conference to conference with correct prediction as high as 65% for the Big 10 and as low as 24% for the ACC.

Conclusions and Future Study

Instant replay and data science alike have propelled sports to even greater heights than ever thought achievable. Both developments have come for better and for worse however though - replays take away points and goals that many argue should have stayed on the scoreboard. Data science in conjunction with performance has changed the landscape of many sports for athletes. Baseball, basketball, hockey, and many other sports have undergone huge changes due to the data being manipulated in association with their respective games. Sports will remain imperfect however, thankfully, as it is why we watch in the first place. No amount of data or instant replay can see everything going on inside a player's mind, and thus perfect predictions or results are unachievable. As basketball will continue to change in the coming years, the model built will need to adapt along with the sport. Additional factors that may improve overall accuracy include: time between games, on court chemistry (think D. Wade to LeBron), and weight change depending on rivalries. If we were to continue on with this project we would like to build a March Madness predictive model in conjunction with the respective conference tournaments.

References

Oberhaus, D. (2019). Machine Learning for March Madness is a Competition in Itself. *Wired.* https://www.wired.com/story/machine-learning-march-madness

Zivkovic, J. (2019). A Deep Dive Into Men's NCAA College Basketball. *Kaggle.* https://www.kaggle.com/jaseziv83/a-recent-deep-look-at-the-men-s-ncaab

Forsythe, J. & Wilde, A. (2014). A Machine Learning Approach to March Madness. *Department of Computer Science, Brigham Young University.* http://axon.cs.byu.edu/~martinez/classes/478/stuff/Sample_Group_Project3.pdf