# BASKETBALL TOURNAMENT PREDICTION USING REGRESSION

A PREPRINT

**Soukarya Ghosh**
Department of Computer Science
University of Virginia
Charlottesville, VA 22903
sg4fz@virginia.edu

**Alex Wassel**
Department of Computer Science
University of Virginia
Charlottesville, VA 22903
aw7re@virginia.edu

**James W. Yun**
Department of Computer Science
University of Virginia
Charlottesville, VA 22903
jy2gm@virginia.edu

March 28, 2019

**ABSTRACT**

Every year, the men's National Collegiate Athletic Association (NCAA) Division I basketball season concludes with a single elimination, 68-team tournament to determine the national championship, commonly refered to as "March Madness". Played mostly during March, it has become one of the most famous annual sporting events in the United States. In 2018, the Virginia Cavaliers basketball team, representing the University of Virginia (UVA), entered the tournament as the No. 1 seed overall but much to the dismay of its students and the country, suffered a historic upset in the first round to UMBC and became the first No. 1 seed to lose to a No. 16 seed in the era of the new tournament format established in 1985. It is proposed that the UVA's performance in the 2019 tournament can be predicted using a variety of machine learning regression algorithms on NCAA's historical basketball data. The findings of this project will answer whether UVA students should expect to win the "Big Dance" for the first time in program history.

## 1 Motivation

UVA was under the heat of the national spotlight after last year's loss to UMBC. This season, UVA (currently No. 2 at the time of writing) is on track to make a strong showing at this year's tournament, under three-time Henry Iba Award winner for national coach of the year, Tony Bennett. There have been millions of attempts to predict the outcomes in the tournament, but there has not been a single confirmed perfect bracket for all of recorded history. In fact, the odds of correctly predicting all of the 63 games in the NCAA tournament is one in $2^{63}$ or $9,223,372,036,854,775,808$ [1], an astronomically low probability. The strength of the model can be assessed by comparing the prediction model against the predictions of basketball analysts, as well as the outcome of the actual tournament. The results of this project can have implications in basketball analysis, sport news coverage, and the billion-dollar sport betting industry. The setting being considered throughout this project include individual team performance, player performance, tournament seeds, and analyst rankings.

This project is inspired by UVA's current success in the regular season as well as its contributors' affinity for the team.

## 2 Method

Machine Learning techniques are great to use for this project because it is possible to gain insights from large amount of data, and since this is a complex problem for which there is no good traditional solution as the environment changes

and it is not always the same teams that win, requiring a lot of hand-tuning. Regression, a form of supervised learning, was used to make predictions was used for this project.

Throughout our experiments, a few distinct machine learning techniques were used, including some that have failed. At first, a custom regression model was created by using three dimensional analysis techniques and graphing tools, in order to minimize the log loss score and determine whether team 1 will beat team 2. We tried to do our own because we thought a unique and specific model to our data would be better than using the already-provided regression models. Additionally, we used the built-in SVM model, linear regression, multivariate regression, and logistic regression on the dataset.

Another machine learning technique we performed is called attribute combination, for which we divided field goals made by field goals attempted, giving us the percent of shots that have been made into the basket.

## 3 Dataset

The datasets being used for this project may be found at the following website:

`https://www.kaggle.com/c/mens-machine-learning-competition-2019/data`

This collection is a compilation of 6 distinct, but related datasets:

1. **Dataset 1** contains team ID's, team names, tournament seeds, final scores of all regular season conference tournament games, NCAA tournament games since the 1984-85 season, and season-level details.
2. **Dataset 2** holds game-by-game stats at a team level (free throws attempted, defensive rebounds, turnovers, etc.) since the 2002-03 season.
3. **Dataset 3** provides city locations of all regular season, conference tournament, and NCAA tournament games since the 2009-10 season.
4. **Dataset 4** contains weekly team rankings for dozens of top rating systems - Pomeroy, Sagarin, RPI, ESPN, etc., since the 2002-2003 season.
5. **Dataset 5** contains play-by-play event logs for 99% of regular season, conference tournament, and NCAA tournament games since the 2009-10 season - including plays by individual players.
6. **Dataset 6** contains additional supporting information, including coaches, conference affiliations, alternative team name spellings, bracket structure, game results for NIT and other postseason tournaments.

Additional datasets may be used at the authors' discretion.

## 4 Related Work

March Madness is a prime target for machine learning enthusiasts, with countless tutorials for beginners appearing on a simple Google search. Many of the methodologies used in these prediction models use Random Forest Regressors, Decisions Trees, and Convolutional Neural Networks.

One particularly impressive project originates from a pair of students from Brigham Young University [2]. The team scraped their data from ESPN and cleaned for a total of ninety-four features. Conclusively, the results show 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 of 0.7362. The group advised those who pursue this project in the future to consider individual player statistics and injury reports, win/loss record of the past ten games, and a few more features that could potentially affect the result of an individual match.

## 5 Preliminary Experiments

First, a custom regression model experiment was conducted by calculating and visualizing a unique and specific regression model that would fit the data and minimize the log loss score. This equation was: Pr(Primary Win) = $\frac{1}{\eta}(y^{\alpha} - x^{\beta} + 16)$, where $\eta \geq 32$ and $\beta \leq 1$, where x and y are the ranks of the teams and $\eta, \beta, \alpha$ are the hyperparameters. This was done by running many trials of changing variables and numbers until reaching the least log loss score. However, this method did not turn out to work as well as expected, because it assumed the distribution was linear. Thus, it was decided to turn to other built-in models such as the SVM model.

Individual weights were assigned that describe the significance of each feature in a basketball game such as steals, offensive rebounds, defensive rebounds, dead rebounds, assists, blocks, personal fouls, technical fouls, turnovers, free throws, dunks, layups, tips, and jump throws. These features were calculated using a Python program to get individual player ratings out of 100, outputting the result into a separate CSV file.

The individual team stats and player stats were aggregated for training against the test set. This was done by training on regular season games and testing on that same year's NCAA tournament.

## 6 Next Steps

The next steps for the project will be to finalize the regression model and finish data cleaning before the March Madness tournament starts, in order to have predictions available to use for the tournament. And in the meantime, brainstorming as a team about the content of the video to be created and presented, as well as starting some of the editing once the predictor is 'ready to use. We are also planning on training our model on different algorithms (boosting and stacking) and using classifiers like SVM, Gradient Descent, and Random Forest Regressor. After identifying best model, we will find the optimal hyperparameters that reduce the RMSE in the validation set. Another achievement we are planning on getting done this week is making a binary classifier of winner (1 and 0) between team match-ups, and split our data into train and test sets with which we will apply our regression on. The machine learning algorithm will be evaluated on the test set by following the college basketball games happening before the NCAA tournament. The model's correctness during this valuation set will be used to further to the model ahead of the actual tournament. Performance measures such as RMSE (L2 Norm) and log loss will also be used to measure the distance between the predicted values and targeted values. After getting our regression model as well as our video content, we will start working on our final report.

## 7 Member Contribution

We all worked on data cleaning and getting the sparse CSV files aggregated into a form we could run tests on. In addition we brain-stormed how to combine a few elements and account for players by team. We came up with a model that allows us to rate players based on season performance where Alex helped adjust feature weights, Soukarya created the algorithm for the ratings and James worked on integrating the player data with the team data. We spoke with Professor Nguyen for assistance and we all contributed equally towards the checkpoint report.

## References

[1] Daniel Wilco. The absurd odds of a perfect NCAA tournament bracket put into perspective, `https://www.ncaa.com/news/basketball-men/article/2019-02-13/absurd-odds-perfect-ncaa-tournament-bracket-put-perspective`

[2] Jared Forsyth and Andrew Wilde. A Machine Learning Approach to March Madness. In 2014, IEEE, 2014. `http://axon.cs.byu.edu/ martinez/classes/478/stuff/Sample_Group_Project3.pdf`