# What Goes Into a March Madness Upset?

## An Analysis of the NCAA March Madness Tournament

## Introduction

In 2019, the historic NCAA Men's Basketball Tournament saw its first 16 seed, the University of Baltimore County, beat a number 1 seed, the University of Virginia, in the Round of 64. Touted as an "improbable" accomplishment by the media, we wanted to find out if this upset really was improbable. **Can we successfully predict upsets before they happen using historical data? What factors come into play the most when trying to predict upsets?** We set out to explore these questions by analyzing the regular season play of teams who have made the "March Madness" tournament.

## Analysis and Methods

**Data Sources**

Our data were sourced from the hoopR[1] R package (counting statistics) and kenpom.com[2] (advanced statistics). These data encompass all men's NCAA seasons from 2007 to 2019.

**Data Cleaning and Preparation**

Both the hoopR and kenpom data sets contained data regarding all NCAA teams, so we subsetted to only tournament teams and aggregated regular season statistics to compare to tournament performance.

**Logistic Regression**

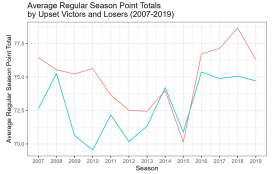
Logistic Regression is a statistical method that is used for estimating the probability of a binary outcome. The purpose of our project is to predict upsets, which are a binary outcome (either an upset happens or it does not), so employing logistic regression is a logical choice.

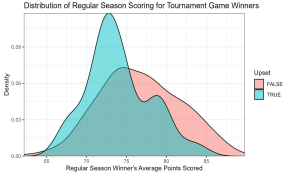
**Support Vector Machines (SVMs)**

SVM are supervised machine learning methods that are used for both classification regression. In the context of classification, SVMs attempts to find a hyperplane (delimiter) that best separates each class. SVMs are versatile and can be used with both linear and nonlinear data, making it a valuable method to explore.

**Artificial Neural Networks (ANNs)**

ANNs are composed of interconnected nodes and layers that can be used for a wide variety of machine learning problems. These connections are quantified by weights which are iteratively calculated during the model training process. ANNs are renowned for their accuracy and versatility across various domains and problems, making it a useful tool for this task.

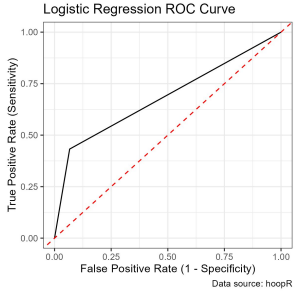



## Results

### Logistic Regression

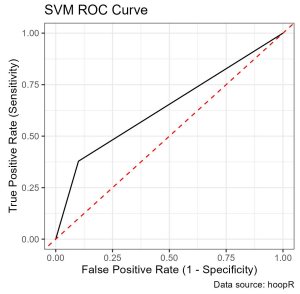
The parameters of interest for this model were various regular season counting statistics of the winners and losers of each tournament matchup(e.g., points scored, assists, rebounds, etc.). 10-fold cross validation was employed. The efficacy and relevant metrics accompanying this model are shown below in the ROC curve plot and table.




| Metric | Value |
|---|---|
| Training Accuracy | 0.8301 |
| Validation Accuracy | 0.8118 |
| Test Accuracy | 0.7874 |
| Test Precision | 0.7272 |
| AUC | 0.6830 |

### Support Vector Machines

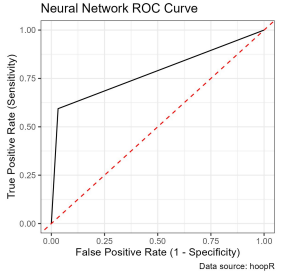
Using the same predictors used in the logistic model, an SVM model was fit to the data. Through iterative hyperparameter tuning, the optimal value for the "C" parameter was 10. This model exhibits poorer performance across the board compared to the logistic model, as evidenced by the ROC curve plot and model performance metrics shown below.




| Metric | Value |
|---|---|
| Training Accuracy | 0.8177 |
| Validation Accuracy | 0.8020 |
| Test Accuracy | 0.7480 |
| Test Precision | 0.6087 |
| AUC | 0.6392 |

### Artificial Neural Networks

A neural network with one hidden layer of 16 nodes using the sigmoid activation function (both hidden and output) minimizing the binary cross entropy loss function. The predictors used in this model were identical to both the logistic and SVM models. The training accuracy exceeds the test and validation accuracies by a considerable amount, indicating that the model is overfitting.




| Metric | Value |
|---|---|
| Training Accuracy | 0.9507 |
| Validation Accuracy | 0.7524 |
| Test Accuracy | 0.7402 |
| Test Precision | 0.5714 |
| AUC | 0.7810 |

### Best Model

Overall, our **logistic regression model best predicts upsets in the March Madness tournament** with a 79% test accuracy and 72% precision in correctly predicting successful upsets. Unlike our ANN, our logistic model does not show significant signs of overfitting, The most significant predictors indicative of an upset are how many points, on average, a lower seed has allowed and the amount scored in the regular season.

## Conclusions

Based on the results generated previously, we found that **a lower seed team that recorded more offensive rebounds and points scored throughout the regular season facing a team that averaged lower steals and blocks per game is a prime indicator of an upset**. Predicting upsets based on regular season statistics will enable analysts and coaches to more accurately assess the chances of and the statistics behind March Madness tournament upsets.

For future research, we would like to not only pin down what these point, offensive rebounds, steal, and block "thresholds" are, but also explore more types of models and improve the existing models. Additionally, we would like to apply these models to more contemporary March Madness matchups.

## References


1. Gilani, Salem et al. (2024). hoopR: The SportsDataverse's R Package for Men's Basketball Data. https://hoopr.sportsdataverse.org/.
2. Pomeroy, Kenneth (2024). kenpom.com: Advanced Analysis of College Basketball. kenpom.com.



Lan Cao (lc1311) Alex Pattarini (amp419)
Matt Carswell (msc301) Xingrui Huo (xh231)