# Predicting March Madness Upsets Using Machine Learning Techniques

Lan Cao, Matt Carswell, Xingrui Froome Huo, Alex Pattarini

Graduate School of Arts and Sciences, Georgetown University

DSAN-5300: Statistical Learning

Project Group 29

April 30, 2024

# Table of Contents

# List of Figures

# Abstract

The 2023 NCAA Men’s Division I Basketball Tournament witnessed one of the most significant upsets in March Madness history when the 16-seed Fairleigh Dickinson Knights defeated the 1-seed Purdue Boilermakers. This event sparked curiosity about the predictability of upsets in the tournament, leading to questions about their occurrence and prediction metrics. In this paper, we explore these questions by applying various statistical methods and machine learning models to understand the indicators and predictability of March Madness upsets. We utilized recent NCAA men's basketball data from 2007 to the present, focusing on regular season statistics and tournament outcomes. Data from hoopR, Ken Pomeroy's website, and a dataset containing tournament matchup results were employed. Each matchup's regular season statistics were aggregated and joined with the tournament outcomes to analyze predictors of upsets.

Through visualizations and statistical analyses, we examined the relationship between team performance metrics, seeding, and historical outcomes. We investigated the impact of factors such as team power, path difficulty, field goal percentages, and pre-conference performance on the likelihood of upsets. Logistic regression, linear discriminant analysis (LDA), support vector machines (SVM), and artificial neural networks (ANNs) were employed to predict upsets. These methods were chosen for their suitability in binary classification tasks and ability to handle high-dimensional data. Our LDA model outperforms others, boasting a test accuracy exceeding 80% and an impressive precision of 83.5%. While its AUC slightly trails logistic regression, precision is paramount for our goal of predicting upsets accurately. With an 83.5% test precision, our model correctly identifies upsets in the NCAA tournament 83.5% of the time, a crucial asset. Unlike the ANN, both logistic regression and LDA models show

minimal signs of overfitting. Key predictors of upsets include the average points allowed by lower-seeded teams and regular season scoring.

Our findings suggest that upsets in March Madness can be predicted based on regular season statistics, particularly focusing on defensive metrics and offensive rebounds. Understanding these predictors can aid analysts and coaches in assessing the likelihood of upsets in future tournaments. Future research will explore threshold values for predictive metrics and investigate additional modeling approaches.

# Chapter 1: Introduction

## 1.1 Overview

On March 17, 2023, the 16-seed Fairleigh Dickinson Knights defeated the 1-seed Purdue Boilermakers 63-58 in the first round of the 2023 NCAA Men's Division I Basketball Tournament, more commonly known as March Madness. The result of this matchup was one of the biggest and most surprising upsets in the history of March Madness, marking only the second time in history a bottom seeded team defeated a top seeded team. The Purdue team, favored by 23 points (CBS Sports, 2023) and fielding the nation's best player, Zach Edey, fell to a team that were only granted a spot at the tournament based on a technicality (Gregory, 2023) and was ranked over 250 spots below them (out of 363 Division I teams), according to advanced analytics guru Ken Pomeroy's rankings (Pomeroy, 2023).

While the Fairleigh Dickinson victory over Purdue may be the most notable upset in March Madness history, upsets are and have been a regular occurrence in every iteration of the March Madness tournament. Since 1985, in the first round alone, there have been 315 upsets based on seeding (out of 1,216 total first round matchups), and many more in the following rounds of the tournament (NCAA, 2024). Upsets are typically significant surprises, despite their regular occurrence. The ever-present popularity of the annual tradition of March Madness bracket building only increases the spotlight on upsets, as fans and analysts attempt to predict whether each tournament matchup will end in an upset or the seeding-determined expected outcome.

This brings up several questions, such as: Why do upsets occur? Are they merely random anomalies and outliers or can they be accurately predicted? If they can be accurately predicted,

what metrics or statistics are most pertinent? What methods can best predict these outcomes? In this paper, we explore these questions by applying various statistical methods and machine learning models to attempt to understand what metrics are indicative of a March Madness upset, and can these outcomes be predicted.

## 1.2 Data Sources and Cleaning

College basketball is a constantly evolving sport, with trends in scoring and other statistics varying over time. Our project is intended to pertain to the contemporary era of NCAA collegiate basketball, so we decided to focus on more recent seasons from 2007 to the present. 2007 was chosen as a cut-off year due to the implementation of the “one and done” rule, where men’s players can be eligible to play in the NBA after one year of collegiate play (Lancaster, 2022). This leads to team rosters and performances to vary significantly year to year as the top players usually try to join the NBA as early as they can.

The data we employed in this project were acquired through R’s “hoopR” package, Ken Pomeroy’s website for advanced analytics for collegiate basketball, and a data set containing tournament matchup results from data.world. We used the hoopR package to access a repository of men’s college basketball statistics for both regular season and tournament games. These data are primarily “counting” statistics, which include the likes of team scoring, assists, rebounds, turnovers, and any other simple statistics that can be “counted” in a game. For each individual tournament matchup in each individual season, the regular season counting statistics acquired via hoopR were aggregated and averaged for both the individual winner and loser to be used as potential predictors indicative of upsets.

Basketball statistician Ken Pomeroy’s website for advanced college basketball analytics is a premier destination for basketball enthusiasts seeking in-depth statistical analysis and

insights into the world of college basketball. Renowned for its advanced metrics and predictive models, the site offers a treasure trove of data that allows fans, coaches, and analysts to delve deep into the nuances of the game. It is often used as a subsidiary resource by basketball analysts in their "bracketology" predictions for the March Madness tournament. From offensive and defensive efficiency ratings to tempo-based metrics and player-specific statistics, KenPom.com provides a comprehensive understanding of team performance and player contributions. The website even goes as far to construct a statistic that attempts to quantify how "lucky" a team is. While much of the kenpom statistics are simply variations of aggregations of the hoopR data, posing immediate concerns of multicollinearity in our dataset, we primarily wanted to use kenpom data as another avenue for EDA and forming some of our early baseline assumptions about predicting winners and upsets in the NCAA tournament.

The data extracted from data.world (Lancaster, 2023) details the outcomes of all March Madness matchups from 1985 to 2019. These data include the round of the matchup, and the winning and losing teams' name, score, and seeding. Keeping in line with the aforementioned 2007 cutoff year, these data were subsetted to only include matchups that occurred between 2007 and 2019. Additionally, these data were joined with the regular season counting (hoopR) and advanced statistical data (kenpom) in order to analyze these regular season statistics as they relate to tournament outcomes. In our data set, each matchup is represented as a row which contains the outcome of the matchup, and the regular season averages of the winner's counting and advanced statistics alongside the loser's counting and advanced statistics, as well as an indication of whether the outcome of the matchup was an upset.

# Chapter 2: Exploratory Data Analysis

## 2.1 Can we identify over and under seeding?

One of the earliest questions we had about upsets in the NCAA tournament is "are upsets[1] a product of overseeding, and if so, can we spot if a team has potentially been overseeded?" To directly answer if an upset is solely a product of overseeding is a little tricky, but we are able to look at historic seeding data in the NCAA tournament and compare the offensive and defensive efficiency ratings (from kenpom) of teams of all seeds. Doing such allows us to identify "clusters" of seeds in the data and if a number seed 1, for example, is "hanging out" in a cluster of 3 seeds, let's say, then this is an automatic red flag that the 1 seed might be overseeded. It should be mentioned that offensive and defensive efficiency ratings in the below figure (**Figure 1**) are the amount of points that a team scores and allows, respectively, per 100 possessions. Once again, the plot covers data from the 2007 season, through the 2019 season.

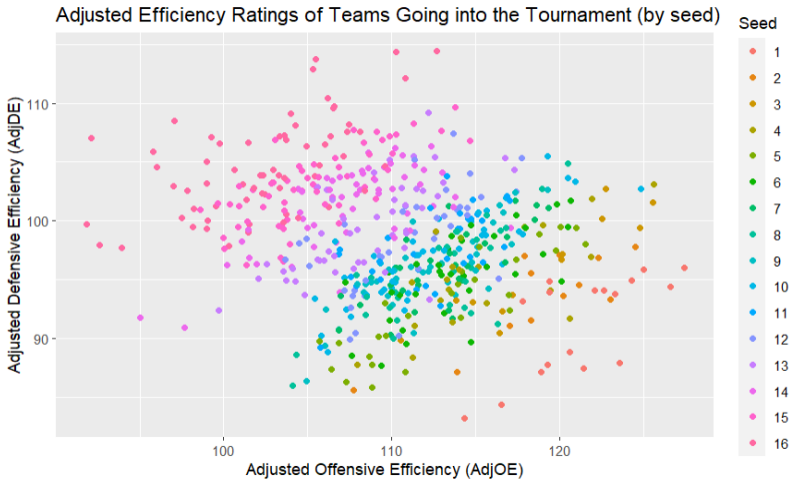


*Figure 1: Adjusted Efficiency Ratings of Teams Going into the Tournament (by seed0*

[1] We classify upsets as any March Madness game where a lower seed beats a higher seed (e.g., 16 seed beats a 1 seed, 12 seed beats a 5 seed, 9 seed beats and 8 seed, etc.)

In this figure we can notice that there is definitely a clear pattern in seeding relating to AdjOE and AdjDE (meaning that some clustering is present). A 1 seed is likely going to have an AdjOE above 110 and an AdjDE below 100. We can see that there is a 1 seed in our plot with an AdjOE below 110 which actually happens to be the 2013 Louisville basketball team that won the national championship. This leads us to believe that overseeding actually might not be as a large of an issue as we may have thought.

If you look at this question from the perspective of underseeding, however, we may begin to have some answers. For example, when looking at the 2011 Virginia Commonwealth team that made the Final Four, they had a pre-tournament AdjOE of 109.6 and an AdjDE of 101.4, metrics that are more characteristic of a 2 or 3 seed, NOT an 11 seed who had to win a play-in game to even make it to the Round of 64. Other examples of clearly underseeded teams include the 2018 Loyola-Chicago team that made the Final Four, having a 107.8 AdjOE and a whopping 96.2 AdjDE, and also the 2018 UMBC team that became the first 16 seed team in NCAA history to upset a number 1 one team (Virginia) in the tournament. This UMBC team had a whopping 113.8 AdjOE efficiency rating which certainly should have raised some more red flags for people than were originally brought up.

## 2.2 Team “power” versus seeding in the tournament

**Figure 2** is a box plot that shows the relationship between Team Power and their Seed ranking. The Power on the y-axis represents a composite score combining various performance indicators such as scoring efficiency, defensive strength, and other season-long statistics for each team. The Seed on the x-axis lists the tournament seeds from 1 to 16, with 1 being the highest (strongest teams based on regular season performance) and 16 the lowest. From this plot, it’s

pretty clear that higher-seeded teams (1-4) generally have higher median 'Team Power' scores compared to lower seeds. The presence of outliers, particularly in lower seeds, suggests that some lower-seeded teams have power scores comparable to much higher-seeded teams, potentially indicating they are underrated or have qualities that could lead to unexpected success.

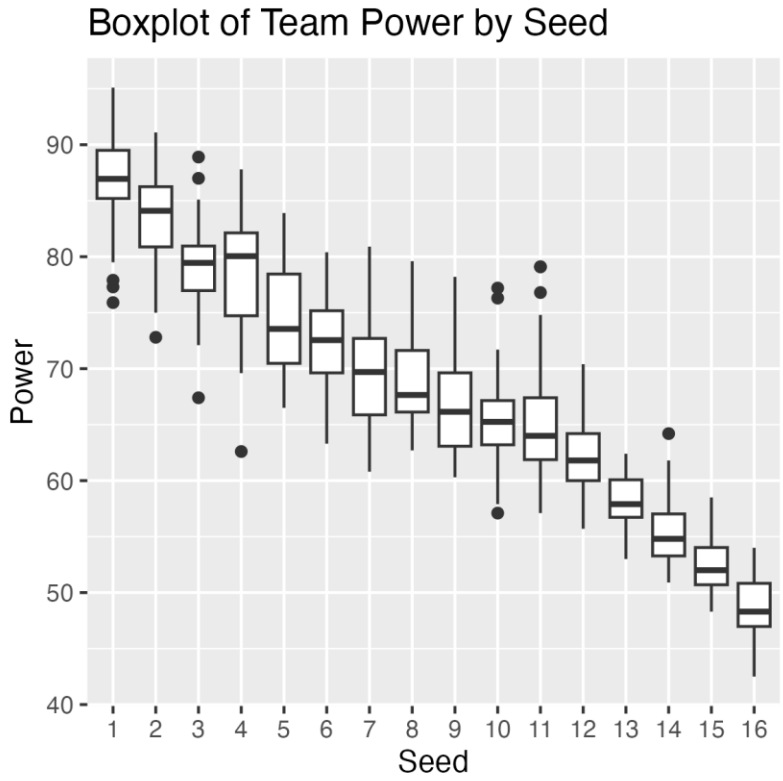


*Figure 2: Team Power by Seed*

## 2.3 Quantifying a team's tournament "path of difficulty"

Having examined the variability in team power across different seeds and noted the potential for lower-seeded teams to possess comparable power levels to their higher-seeded counterparts, we now shift our focus towards understanding the challenges these teams face throughout the tournament. A critical factor influencing game outcomes, particularly upsets, is

the 'Path Difficulty' that teams navigate. This measure reflects the series of challenges teams overcome, including the caliber of opponents and the margins by which games are won or lost.

To explore how the difficulty of a team's tournament path might correlate with their likelihood of causing an upset, we analyze the distribution of 'Path Difficulty' for games resulting in upsets compared to those that follow expected outcomes. This analysis not only enhances our understanding of what it takes for a lower-seeded team to succeed against odds but also probes whether the adversity faced en route to the tournament primes teams for unexpected victories.

The violin plot, **Figure 3**, shows a broader distribution and higher variability in path difficulty among games that ended in an upset, compared to a more concentrated distribution for games that did not result in an upset. This suggests that teams causing upsets often face more challenging paths, which could indicate a resilience or capability that isn't reflected in their seed alone. Such insights about path difficulty prompt us to further explore how specific in-game performance metrics might also influence the likelihood of upsets.

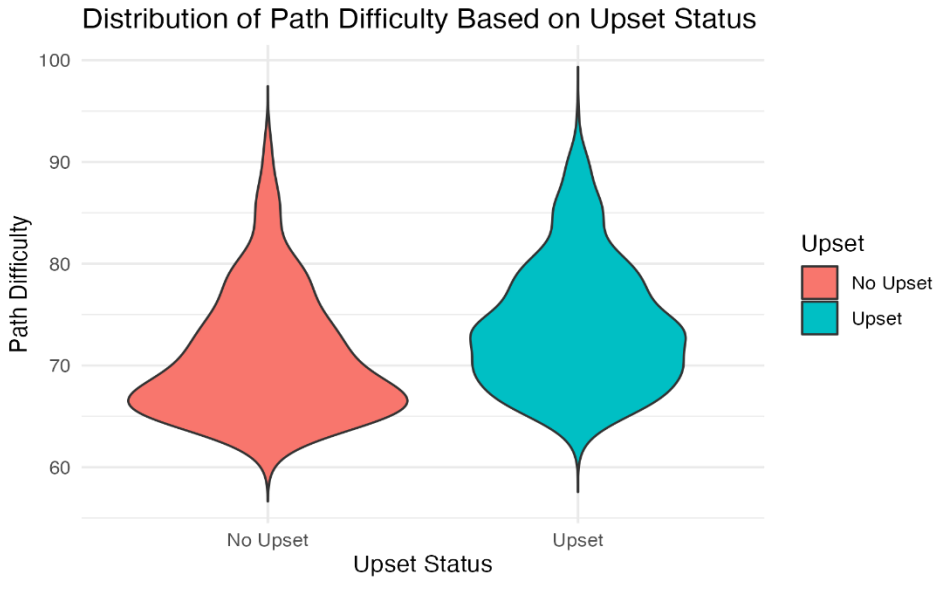


*Figure 3: Distribution of Path Difficulty Based on Upset Status*

## 2.4 Field Goal Percentage and Scoring

Moving from the broader context of tournament paths to the specifics of in-game action, we next turn our attention to field goal percentages, a direct measure of shooting efficiency. The histogram **Figure 4** provides a comparative look at the distribution of average field goal percentages between games that ended in an upset and those that followed expected outcomes. The vertical axis represents the frequency of games, while the horizontal axis shows the range of field goal percentages. Analyzing these percentages helps us understand if superior shooting is as influential in determining game outcomes as strategic execution or defensive prowess.

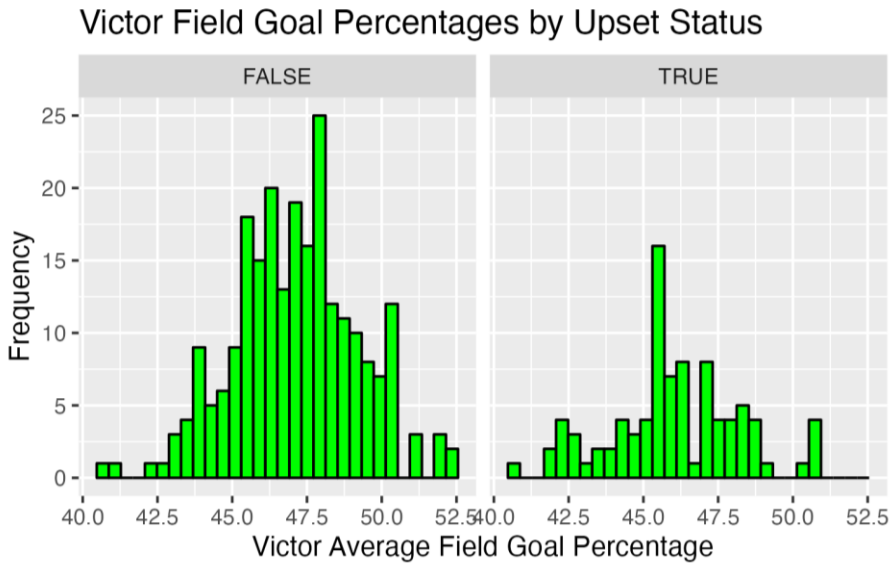


*Figure 4: Field Goal Percentages by Upset Status*

We observe from the range of these two distributions that most field goal percentages fall between 40% to 52.5%. Interestingly, the distribution for games with upsets does not show significantly higher field goal percentages compared to non-upset games. This indicates that simply shooting well may not be the determining factor in upsets, suggesting that other elements, such as defensive play or tactical decisions, could play more critical roles. This finding prompts

a deeper look into other factors that may contribute to upsets, such as turnover rates, defensive stops, or clutch performance under pressure, which we will be analyzing later in this paper.

Looking just at scoring in the below scatter plot (**Figure 5**), the plot depicts pre-conference average scores against tournament seeds offers an perspective on the alignment of regular-season performance with tournament expectations. The x-axis represents the pre-conference average score, which is indicative of a team's offensive output before entering the high-stakes environment of the tournament. The y-axis shows the tournament seed, with lower numbers representing higher seeded, presumably stronger, teams. The trend line in the plot reveals a general decrease in seed quality with higher pre-conference scores, suggesting that teams performing better offensively during the regular season tend to receive higher (better) seeds. However, the distribution of points indicates considerable overlap between teams that experienced upsets (marked in cyan) and those that did not (marked in pink). This overlap suggests that while offensive prowess contributes to better seeding, it is not a foolproof predictor of success in the tournament, emphasizing the complexity of factors leading to upsets.

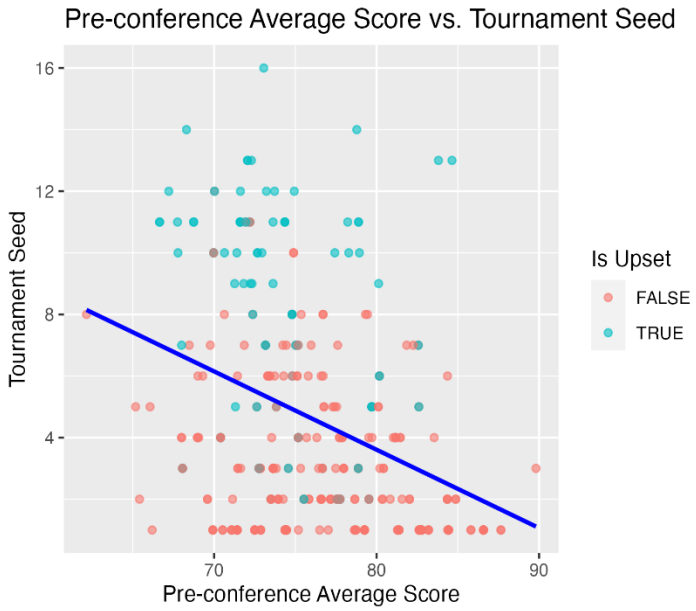


*Figure 5: Average Regular Season Score per game vs. Tournament Seed*

Given the limited predictive power of pre-conference scoring on its own, it becomes pertinent to consider more comprehensive measures of team strength and tournament difficulty, which leads us to the next analysis.

## 2.5 Combining Power, Path of Difficulty, and Scoring

In exploring the relationship between a team's overall strength ('POWER') and the challenges they face ('PATH') across different seeds, the scatter plot (**Figure 6**) provides a view of the tournament dynamics. The x-axis represents the composite 'POWER' score, encompassing various performance metrics that signify a team's overall capability. The y-axis displays the 'PATH' difficulty, quantifying the challenge level based on opponent strength and game margins encountered through the tournament. The colored data points, differentiated by seed, show a clear trend: higher 'POWER' often correlates with lower 'PATH' difficulty, particularly for top-seeded teams, indicated by the descending blue trend line. This trend suggests that stronger teams generally face easier paths, potentially due to seeding advantages. However, the variability in 'PATH' difficulty for teams with similar 'POWER' scores, especially across different seeds, points to the nuanced and unpredictable nature of tournament matchups. This analysis underscores the need to consider both the intrinsic strength of teams and the external challenges they face when predicting tournament outcomes. This nuanced understanding aids in appreciating the multifaceted nature of upsets, where even powerful teams can falter under challenging conditions.

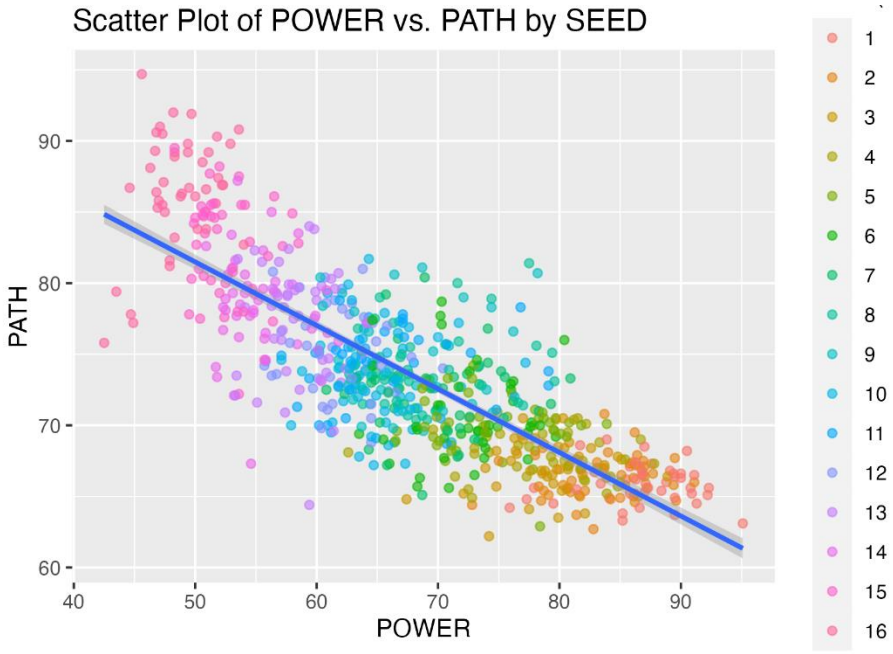


*Figure 6: Scatter Plot of Power vs. Path (by seed)*

Building upon our understanding of team power and path difficulty, we now focus on the regular season scoring performance of tournament game winners. This kernel density plot illustrates the distribution of average points scored by winners in games that ended in upsets versus those that followed the expected outcomes. The x-axis displays the regular season winner's average points scored, and the two differently colored curves represent whether the game was an upset or not. The density plot (**Figure 7**) reveals that both distributions have a significant overlap, with the peak for non-upset games slightly lower than for upsets, indicating a modestly better scoring performance on average. However, the broad overlap suggests that high scoring in the regular season does not distinctly differentiate between teams that will uphold or defy tournament expectations, which correspond to average field goal percentages between games **(Figure 4**). This observation reinforces the notion that while offensive efficiency is crucial, it alone does not guarantee success against the unique pressures and matchups of tournament play.

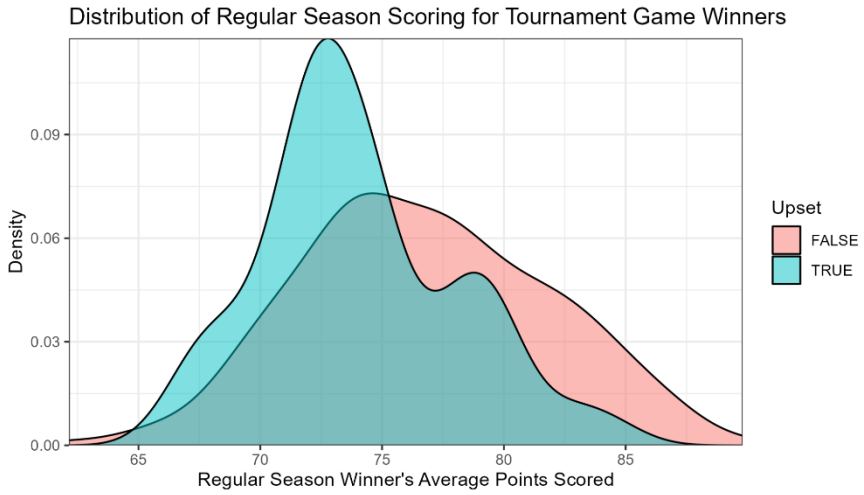


*Figure 7: Distribution of Regular Season Scoring for Tournament Game Winners*

To encapsulate the competitive dynamics of NCAA tournament upsets more comprehensively, we next examine the margins by which these surprising victories transpire. Analyzing the victory margins provides critical insight into not just the occurrence of upsets, but also the competitiveness of these encounters. This boxplot (**Figure 8**) categorizes games based on their upset status and illustrates the victory margins associated with each category. The x-axis differentiates games by whether an upset occurred, and the y-axis measures the margins of victory or defeat. The color scheme—red for expected outcomes and teal for upsets—visually segregates the typical games from the anomalies.

The median victory margin for upsets, hovering closer to zero, illustrates that these games are generally more tightly contested than non-upset matches. This finding aligns with the intuitive understanding that upsets are not merely flukes but often the result of underdog teams capitalizing on crucial moments within tightly contested scenarios. Moreover, the variability indicated by the spread and outliers in the upset category highlights that while many upsets occur

by narrow margins, there are notable instances where underdogs win by unexpectedly large margins.

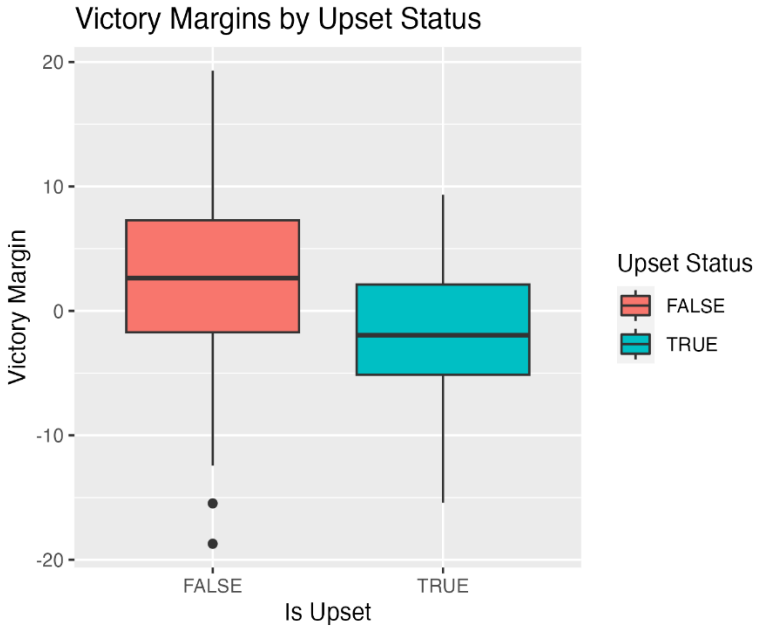


*Figure 8: Victory Margins by Upset Status*

These nuances are vital as they suggest that upsets may often be driven by factors beyond regular statistical measures like scoring averages or offensive efficiency—such as strategic ingenuity, superior game-day execution, or psychological resilience under tournament pressure. By closely examining these victory margins, we gain a richer understanding of the dramatic and unpredictable nature of upsets, which not only captivates audiences but also challenges predictors and analysts to look beyond conventional metrics to grasp the full complexity of tournament dynamics.

## 2.6 Next Steps

Our exploratory data analysis has provided valuable insights into the dynamics and complexities of NCAA tournament upsets, revealing that factors such as pre-conference

performance, power rankings, path difficulties, and even the tightness of victory margins contribute to the unpredictable nature of these games. While our visual and statistical explorations have highlighted key patterns and raised important questions, a more systematic approach is needed to quantitatively assess these findings and predict future outcomes.

# Chapter 3: Predicting Upsets in the Tournament

## 3.1 Methods

Predicting whether a March Madness upset will occur is a binary classification problem as either a given matchup results in an upset, or it does not. In order to derive how to most effectively "predict" whether an upset will occur, we applied several machine learning techniques, including logistic regression, linear discriminant analysis (LDA), support vector machines (SVM), and artificial neural networks (ANNs).

Binary logistic regression is a statistical method that, given a set of features (independent variables), assigns a probability of whether a given data point belongs to one of two outcomes (e.g., positive or negative). In the context of this project, a logistic regression model quantifies, given certain predictors (i.e., regular season statistics for each team), the probability a given March Madness matchup will result in an upset. The coefficients for each of the predictors are found through Maximum Likelihood Estimation (MLE), wherein the logistic model iterates through various coefficients until it finds a combination that maximizes the log-likelihood function. Logistic regression is a popular choice for dealing with binary classification tasks, and is known for its low computational complexity and greater interpretability compared to more complex methods. An important assumption for logistic regression is that there is no significant multicollinearity between the predictors of interest. None of the Variance Inflation Factor (VIF) values for the counting statistics of the winners and losers of each March Madness matchup exceeded 5, which is a common benchmark for VIF. As all of these VIF values are below the threshold of 5, we have evidence to suggest that this assumption is met, and that there is not

significant multicollinearity between the predictors. Thus, logistic regression proves to be an appropriate method to apply to understanding the nature of March Madness upsets.

Linear discriminant analysis (LDA) is another statistical method that traditionally specializes in binary classification. LDA aims to find a linear combination of the features in the data that best separate the two outcomes of interest by minimizing intra-class variation (separation of points in the same class) and maximizing inter-class variation (separation of classes from one another). One of the most important aspects of LDA is that it performs a dimensionality reduction on the data without sacrificing any delineations between classes which enhances interpretability, mitigates the possibility of overfitting, and has relatively low computational complexity. For these reasons, LDA was applied to attempt to delineate and predict whether a March Madness occurs.

Support vector machines (SVMs) are supervised machine learning models that attempt to find a hyperplane (decision boundary) that best separates the points within the classes of interest. SVMs, much like LDA aim to minimize intra-class variation (distance between points within a class) while maximizing inter-class variation (distance from the points of each class and the decision boundary/margin). SVMs on their own can only handle linearly separable data – when the points of each class can be separated by a single line – but can also handle non-linearly separable data using the “kernel trick”. The kernel trick employs the use of another mathematical function to transform the data into a higher dimensional space where they may be linearly separable. Through the tuning of the regularization hyperparameter “C”, one can adjust the permissibility of the decision boundary, which constitutes a tradeoff between model complexity (which can increase the likelihood of overfitting) and the classification error (how well the model fits the training data). Due to the flexibility of SVMs with regard to high dimensional data

combined with their high performance with binary classification, we employed SVMs in this project to delineate and predict March Madness upsets.

Artificial neural networks (ANNs) are very powerful machine learning tools that can be used for a variety of data science problems, including classification. ANNs consist of several layers of "nodes", each connected to one another consecutively. Throughout the model training process, weights are assigned to each connection which are optimized through a loss function and back propagation. Neural networks are notoriously prone to overfitting given their complexity and versatility, but this can be mitigated through the appropriate choice of hyperparameters and activation functions. Activation functions enable neural networks to more accurately model nonlinear relationships, and these functions transform the outputs of a given layer in the network. The choice of activation functions is dependent primarily on the context of the intended application of the neural network. Predicting upsets in March Madness involves predicting a binary output, and the sigmoid activation function is most appropriate for binary classification applications, so we employed the sigmoid activation function for both the hidden layers and output layer. Neural networks have been and continue to be at the forefront of machine learning applications and are extremely versatile and powerful, so we applied an artificial neural network to model March Madness upsets.

## 3.2 Results Overview

Overall, our LDA model gives us our strongest results with test accuracy of over 80% and a whopping 83.5% test precision, which is arguably more important as our goal is to develop a model that correctly predicts upsets, not just a model that tells us when an upset is NOT going to happen. Our LDA has an AUC of 65.4% which is slightly less than our logistic regression's

AUC of 68.3%, but once again, we are focusing on precision in this case. A test precision of 83.5% tells us that when our model predicts an upset, it is going to be correct 83.5% of the time which is EXTREMELY valuable. Unlike our ANN, our logistic and LDA models do 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.

## 3.3 Logistic Regression

For the logistic regression model, an 64-16-20 training-validation-test split was employed using the regular season counting statistics of the winners and losers of each tournament matchup as predictors. Additionally, 10 fold cross validation was implemented to enhance the robustness of the model using R's caret package. Initial results from a logistic regression with all features can be seen in the below figure:

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.25378    8.63737  -0.029  0.97656
victor_avg_score    -0.42801    0.10588  -4.043 5.29e-05 ***
victor_avg_assists   0.16941    0.13202   1.283  0.19943
victor_avg_blocks   -0.01855    0.14780  -0.126  0.90013
victor_avg_dreb      0.11223    0.13023   0.862  0.38882
victor_avg_fgp      -0.18007    0.13148  -1.370  0.17082
victor_avg_oreb     -0.27760    0.13390  -2.073  0.03816 *
victor_avg_steals    0.55598    0.17687   3.144  0.00167 **
victor_avg_fta      -0.07966    0.07989  -0.997  0.31873
victor_avg_3fgp      0.11998    0.09214   1.302  0.19284
victor_avg_to        0.10827    0.14994   0.722  0.47023
victor_avg_opp_score 0.31222    0.06890   4.531 5.87e-06 ***
loser_avg_score      0.08627    0.09599   0.899  0.36877
loser_avg_assists    0.06421    0.12628   0.508  0.61113
loser_avg_blocks     0.39382    0.15050   2.617  0.00888 **
loser_avg_dreb       0.15089    0.13015   1.159  0.24628
loser_avg_fgp        0.07713    0.12758   0.605  0.54548
loser_avg_oreb       0.12524    0.12914   0.970  0.33213
loser_avg_steals    -0.11844    0.16631  -0.712  0.47638
loser_avg_fta        0.11956    0.07573   1.579  0.11439
loser_avg_3fgp       0.09083    0.07932   1.145  0.25216
loser_avg_to        -0.16541    0.14020  -1.180  0.23808
loser_avg_opp_score -0.17120    0.06237  -2.745  0.00605 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 478.27  on 405  degrees of freedom
Residual deviance: 322.89  on 383  degrees of freedom
AIC: 368.89

Number of Fisher Scoring iterations: 6
```

*Figure 9: Logistic Regression Model Summary*

Overall, we can see that our statistically significant features are victor average opponent score, victor average score, victor average steals, loser average opponent score, loser average blocks, and victor average offensive rebounds. This is a model that weights defense slightly heavier than offense, and among non-scoring statistics, blocks, steals, and offensive rebounds are by far the most important features.

The best performing logistic regression model returned a validation accuracy of 0.8118 and testing accuracy of 0.7874 along with a test precision at 0.7272. Model performance metrics are summarized in Table X. Ultimately, predictions generated by this model on the testing data set returned an AUC of 0.6830, and an ROC shown in **Figure 10**. This curve is above the red dotted line, which indicates that the logistic classifier predicts whether matchups will result in upsets more accurately than a random guess on the test data set. There does not appear to be overfitting given that the training, validation, and test accuracies are all relatively close to one another as evidenced by the metrics reported in **Figure 11**Figure 11.

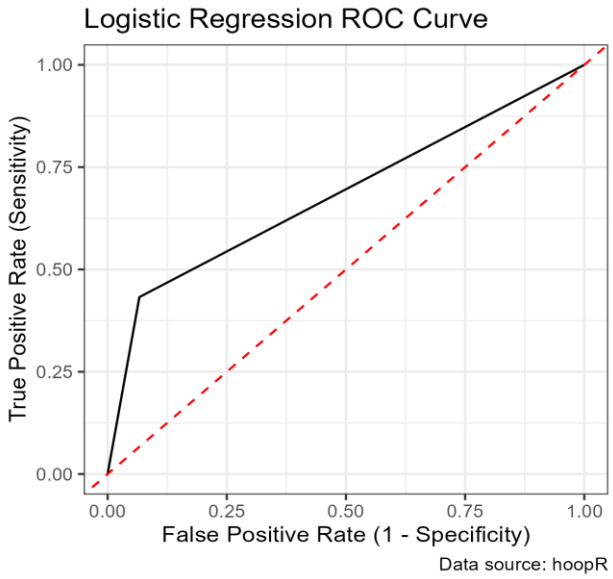


*Figure 10: Logistic Regression ROC Curve*

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

*Figure 11: Logistic Regression Performance Measures*

## 3.4 Linear Discriminant Analysis

Upon making the logistic regression model, it appears our data has a high degree of linear separability in a high dimensional space. We wanted to continue building on this finding by exploring other linearly separable classification methods. Using Linear Discriminant Analysis (LDA), we are further able to pin down which features explain the most variation in our target variable, upset. Such findings will help us pin down which, specifically, are the most important features that determine if an upset is going to happen in the NCAA Tournament. We used an 80-20 test-train split for our model.

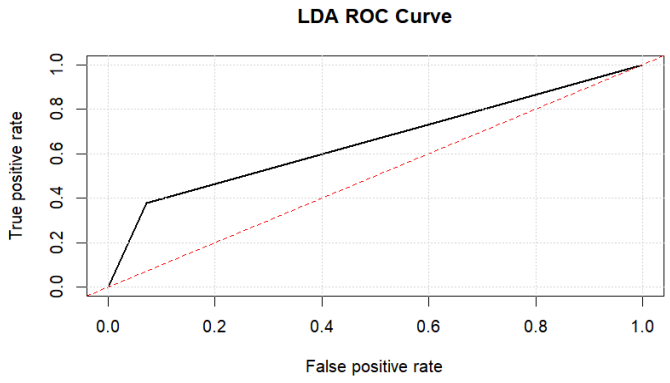


*Figure 12: LDA ROC Curve*

| Metric | Value |
|---|---|
| Training Accuracy | 0.8166 |
| Test Accuracy | 0.8031 |
| Test Precision | 0.8348 |
| AUC | 0.6539 |

*Figure 13: LDA Performance Measures*

With our LDA model, our output metrics are extremely promising. With a solid test accuracy of over 80% and 83.5% test precision, our LDA classifier shows great promise. It has an AUC of 65.4% which is slightly less than our logistic regression's AUC of 68.3%, but still really powerful. A test precision of 83.5% tells us that when our model predicts an upset, it is going to be correct 83.5% of the time which is EXTREMELY valuable. Once again, our data has a very high degree of linear separability over a high dimensional space.

One of the best parts about LDA is that we can directly quantify how each of our features in our data set explains the variation in our target variable, "upset" in this case. A few major features stick out initially. The losing team's (that team that was upset) average blocks throughout the season has the most "weight" on whether an upset is going to occur or not, followed then by the winning team's average offensive rebounds and average steals per game throughout the regular season. The winning team's average score per game in the regular season and average score allowed are then weighted the next heaviest, with the losing team's average steals per game showing to be significant.

| Variable | LD1 |
|---|---|
| victor avg score | -0.1877 |
| victor avg assists | 0.0976 |
| victor avg blocks | -0.0059 |
| victor avg dreb | -0.0337 |
| victor avg fgp | -0.0789 |
| victor avg oreb | -0.2280 |
| victor avg steals | 0.2149 |
| victor avg fta | 0.0552 |
| victor avg 3fgp | 0.0109 |
| victor avg turnovers | 0.0925 |
| victor avg opp score | 0.1482 |
| loser avg score | 0.0811 |
| loser avg assists | 0.0499 |
| loser avg blocks | 0.2704 |
| loser avg dreb | 0.0738 |
| loser avg fgp | 0.0379 |
| loser avg oreb | 0.0679 |
| loser avg steals | -0.1111 |
| loser avg fta | 0.0238 |
| loser avg 3fgp | 0.0333 |
| loser avg turnovers | -0.0813 |
| loser avg opp score | -0.0923 |

*Figure 14: LD1 Components*

What all of this information is able to tell us is that a higher seeded team that tends to be less defensively ept through blocking and stealing versus a lower seeded team that is a strong offensive rebounding team that also blocks and steals well, is the perfect recipe for an upset to take place. To see how heavily our linear discriminant coefficient values emphasize defensive play and rebounding over raw offensive power/talent, relatively speaking, is certainly interesting and great context to have when trying to gauge what teams in the tournament are primed to have a “Cinderella” run.

## 3.5 Support Vector Machines

Using the same predictors used in the logistic model, an SVM model (linear kernel) 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 and LDA model, as evidenced by the ROC curve plot and model performance metrics shown below.

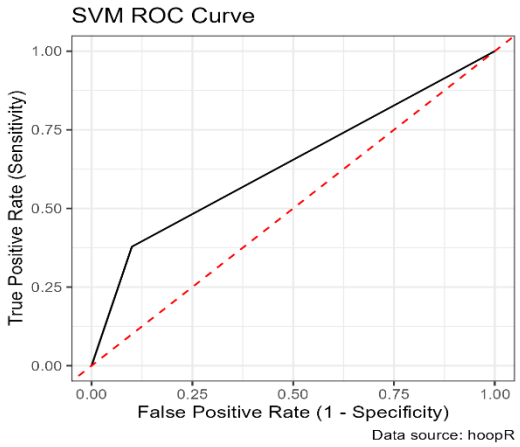


*Figure 15: SVM ROC Curve*

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

*Figure 16: SVM Performance Measures*

## 3.6 Artificial Neural Networks

Using the same training-validation-test split as performed in logistic regression and LDA with identical predictors, a one hidden layer neural network was fitted to these data using R's nnet package. The sigmoid activation function was employed for both the hidden layer (16

nodes) and output layer (1 node) as this model is intended to perform a binary classification task. Ultimately, the neural network performs exceptionally well on the training set, and more poorly on the validation and testing data sets. This, combined with the relatively low precision (compared to the other models discussed in this report), is an indication that this model is overfitting on the training data, and not performing well on unseen data. The ROC curve (and AUC value are the highest of any of the models we constructed, but due to the apparent overfitting these metrics are misleading.

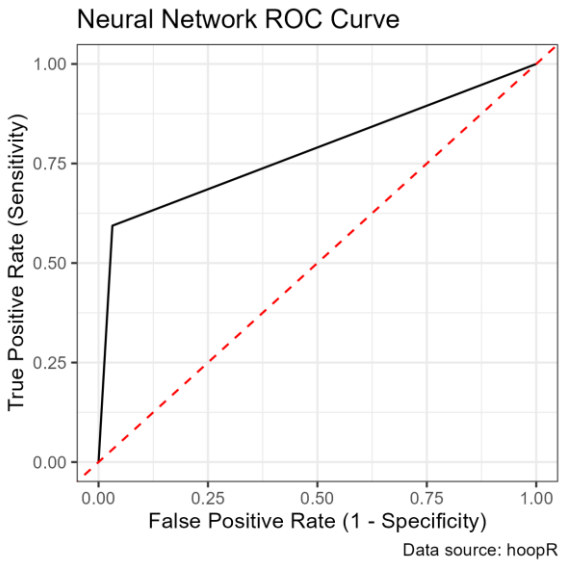


*Figure 17: ANN ROC Curve*

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

*Figure 18: ANN Performance Measures*

# Chapter 4: Conclusion

In conclusion, our investigation into the dynamics of upsets in the NCAA March Madness tournament unveils significant insights through the application of various statistical learning techniques. Among these methods, linear discriminant analysis (LDA) emerges as the standout performer, boasting a test accuracy surpassing 80% and an impressive precision of 83.5%. These findings highlight the crucial role of defensive metrics, particularly the average points allowed by lower-seeded teams, in effectively predicting tournament upsets. Furthermore, our study underscores the importance of precision in predicting upsets, emphasizing the significance of correctly identifying potential upsets rather than merely detecting non-upsets. This distinction underscores the practical utility of our predictive models in assisting analysts and coaches in making informed decisions during the tournament.

Looking ahead, our research paves the way for future endeavors aimed at enhancing the predictive power and robustness of our models. One avenue for improvement involves refining the selection and weighting of predictive features to better capture the nuanced dynamics underlying upsets. Additionally, exploring advanced machine learning techniques and incorporating additional data sources could further augment the accuracy and reliability of our models. Moreover, identifying threshold metrics and developing actionable insights based on model outputs can provide valuable guidance for analysts and coaches seeking to anticipate and strategize for potential upsets in future tournaments. By continuing to refine and iterate upon our predictive models, we can deepen our understanding of the complex factors driving March Madness upsets, ultimately empowering stakeholders with valuable insights for navigating the unpredictable landscape of collegiate basketball tournaments.

# References

CBS Sports Staff (2023, March 17). *Purdue vs. Fairleigh Dickinson prediction, odds: 2023 NCAA Tournament picks, March Madness bets by top experts*. https://www.cbssports.com/college-basketball/news/purdue-vs-fairleigh-dickinson-prediction-odds-2023-ncaa-tournament-picks-march-madness-bets-by-top-expert/

Gregory, S. (2023, March 17). *Fairleigh Dickinson Wasn't Even Supposed to Be in the NCAA Tournament. Here's How They Beat Purdue.* https://time.com/6264323/fairleigh-dickinson-beats-purdue-ncaa-tournament/

Gilani, Salem et al. (2024). hoopR: The SportsDataverse's R Package for Men's Basketball Data. https://hoopr.sportsdataverse.org/

Lancaster, R. (2021, November 22). *The Six Eras of College Basketball*. Sports Central. https://www.sports-central.org/sports/2021/11/22/the_six_eras_of_college_basketball.php

NCAA (2024, March 7). *Records for every seed in March Madness from 1985 to 2023.* \ https://www.ncaa.com/news/basketball-men/article/2024-03-07/records-every-seed-march-madness-1985-2023

Pomeroy, K. (2023). *2023 Pomeroy College Basketball Rankings*. https://kenpom.com/index.php?y=2023

Roy, M. (2023). *NCAA Tournament Results*. data.world. https://data.world/michaelaroy/ncaa-tournament-results