**Project Final Report**

**DSO 579**

**Cody Greene**

**Exploring the Causes of March Madness Upsets**

**Objective:** The objective of this study is to investigate the underlying factors that contribute to upsets in the March Madness tournament and to develop a predictive model for future tournaments

**Problem:** March Madness often experiences upsets where lower-seeded teams defeat higher-seeded teams. These upsets present challenges for individuals attempting to predict tournament outcomes.

**Research Questions:** What factors have contributed to the increase in upsets during March Madness, and can these factors be effectively predicted to anticipate upsets in the tournament?

**Null Hypothesis:**

H0: There is no significant data that can predict the cause of upsets in March Madness.

H1: There is significant data that can predict the cause of upsets in March Madness.

**Report Structure**

1. Datasets
2. Defining “Upset”
3. Upset by Round
4. Methods
   1. Feature Selection
   2. Predictive Models
      1. Logistic Regression
      2. Random Forest
      3. Gradient Boosting
5. Analysis of Barthag and Wins Above Bubble
6. Limitations
7. Conclusion
8. Relevant Literature
9. Feature Glossary

**Datasets**

Downloaded two datasets from Kaggle

1. Tournament\_Scores\_1981-2021.csv ([URL](https://www.kaggle.com/datasets/woodygilbertson/ncaam-march-madness-scores-19852021))

This dataset includes team names, seeds, and scores of each March Madness game.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **YEAR** | **ROUND** | **WSEED** | **WTEAM** | **WSCORE** | **LSEED** | **LTEAM** | **LSCORE** | **…** |
| 2021 | 6 | 1 | Baylor | 73 | 1 | Gonzaga | 68 | … |
| 2021 | 5 | 1 | Gonzaga | 71 | 11 | UCLA | 63 | … |

1. Tournament\_Team\_Data\_2008-2022.csv ([URL](https://www.kaggle.com/code/nishaanamin/predicting-upsets/input))  
   This dataset includes team data of the season participating in March Madness.  
   On top of basic stats like FG% and REB%, advanced stats from [KemPon Rankings](KemPon%20Rankings) and [BartTorvik Rankings](BartTorvik%20Rankings) are also featured.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **YEAR** | **SEED** | **TEAM** | **OF EFF** | **DF EFF** | **…** | **WIN %** |
| 2022 | 1 | Kansas | 1.17 | 0.85 | … | 0.760 |
| 2022 | 1 | Arizona | 1.16 | 0.86 | … | 0.729 |

For predictive analysis, merge these datasets and create Merged\_Data\_2008-2021.csv  
Prior to merging datasets, standardize school names, e.g. USC -> Southern California, so that the datasets can merge by using the school name as a key.  
Add suffix \_F for favorite team data and \_U for underdog team.  
Add a binary value column UPSET and drop unnecessary columns.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **YEAR** | **ROUND** | **TEAM\_F** | **SEED\_F** | **OF EFF\_F** | **DF EFF\_F** | **…** | **TEAM\_U** | **SEED\_U** | **OF EFF\_U** | **DF EFF\_U** | **…** | **UPSET** |
| 2021 | 6 | Baylor | 1 | 1.12 | 0.771 | **…** | UCLA | 11 | 1.05 | 0.910 | … | 0 |
| 2021 | 5 | Gonzaga | 1 | 1.19 | 0.698 | **…** | Houston | 6 | 1.09 | 0.830 | … | 1 |

**Defining “Upset**”  
Upset should not be a phenomenon that occurs frequently, and definitions in different studies are inconsistent.

The study [Predicting March Madness using Machine Learning](Predicting%20March%20Madness%20using%20Machine%20Learning) simply defines upset as lower seed beating a higher seed

The study [Use my upset prediction model to pick underdogs in your NCAA tournament bracket](https://towardsdatascience.com/use-my-upset-prediction-model-to-pick-underdogs-in-your-ncaa-tournament-bracket-87c4aa3935f5) define upset as a win by an underdog team seeded at least 4 slots lower than their opponent

Due to the inconsistency, this study is defining a new definition of upset based on historical probability.

A graph of a graph

Description automatically generated

In the above image, notice that most of the upset probability of each combination decreases significantly above the yellow line.  
That tiles above the line are the games where underdog team seeded at least 5 slots lower than their opponent.  
Therefore, upset definition in this study is defined as a win by underdog team against a team that’s seeded 5 slots or higher.

**Upset by Round**  
A graph with a green line

Description automatically generated

* Most upsets occur in 1st round. However, provability is not as high
* Upset probability increases in the 4th round. This indicates that once an underdog makes it to the 3rd round, they have the momentum to advance further
* There are no upsets in the championship within the data period. The last championship upset was in 1988, Kansas (6) vs Oklahoma (1)

**Feature Selection**Performed ANOVA and selected top 30 important features (high F-Score and low p-value) to before training predictive models.  
Started out with selecting top 50, and then top 40 features but the high number of features did not improve predictive model performance at all. Top 30 seems to be a good cut off point where it can provide a good performance for the models and save computing power.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Feature** | **F-Score** | **p-value** |
| **10** | BARTHAG\_F | 44.735574 | 6.899160e-11 |
| **7** | BARTTORVIK ADJUSTED EFFICIENCY\_F | 43.040659 | 1.515662e-10 |
| **3** | KENPOM ADJUSTED EFFICIENCY\_F | 37.841246 | 1.732121e-09 |
| **44** | BARTTORVIK ADJUSTED EFFICIENCY\_U | 37.252041 | 2.287699e-09 |
| **47** | BARTHAG\_U | 36.902496 | 2.698781e-09 |
| **40** | KENPOM ADJUSTED EFFICIENCY\_U | 35.701817 | 4.766732e-09 |
| **72** | WINS ABOVE BUBBLE\_U | 31.272375 | 3.953434e-08 |
| **48** | ELITE SOS\_U | 24.470318 | 1.077659e-06 |
| **35** | WINS ABOVE BUBBLE\_F | 23.860049 | 1.455085e-06 |
| **46** | BARTTORVIK ADJUSTED DEFENSE\_U | 22.929026 | 2.303543e-06 |
| **42** | KENPOM ADJUSTED DEFENSE\_U | 22.804567 | 2.449733e-06 |
| **2** | SEED\_F | 22.348260 | 3.070432e-06 |
| **39** | SEED\_U | 21.344945 | 5.052015e-06 |
| **36** | WIN %\_F | 20.246293 | 8.735209e-06 |
| **37** | POINTS PER POSSESSION OFFENSE\_F | 19.796110 | 1.094031e-05 |
| **45** | BARTTORVIK ADJUSTED OFFENSE\_U | 18.063276 | 2.612758e-05 |
| **4** | KENPOM ADJUSTED OFFENSE\_F | 17.800780 | 2.982834e-05 |
| **8** | BARTTORVIK ADJUSTED OFFENSE\_F | 17.317324 | 3.808664e-05 |
| **41** | KENPOM ADJUSTED OFFENSE\_U | 15.639963 | 8.933405e-05 |
| **38** | POINTS PER POSSESSION DEFENSE\_F | 10.979723 | 9.976003e-04 |
| **20** | OFFENSIVE REBOUND %\_F | 10.533068 | 1.262502e-03 |
| **32** | OP D REB %\_F | 10.533068 | 1.262502e-03 |
| **9** | BARTTORVIK ADJUSTED DEFENSE\_F | 8.339414 | 4.071594e-03 |
| **13** | 2PT %\_F | 7.057030 | 8.183798e-03 |
| **5** | KENPOM ADJUSTED DEFENSE\_F | 6.199591 | 1.314782e-02 |
| **16** | EFG %\_F | 5.294752 | 2.185859e-02 |
| **57** | OFFENSIVE REBOUND %\_U | 5.201907 | 2.304161e-02 |
| **69** | OP D REB %\_U | 5.201907 | 2.304161e-02 |
| **24** | 2PT % DEFENSE\_F | 4.818724 | 2.867620e-02 |
| **59** | BLOCK %\_U | 4.765842 | 2.956005e-02 |

**Logistic Regression Model**

Reason for choosing this model:Logistic regression model is a simple and widely-used linear classification algorithm that works well for binary classification like predicting the outcome of a sports game

Cross-validation results  
[0.81818182, 0.79545455, 0.72727273 0.79545455, 0.75, 0.79545455, 0.72727273, 0.68181818, 0.79545455, 0.77272727]

Average cv accuracy: 0.76590

A screenshot of a computer

Description automatically generated

Based on the confusion matrix, the model predicted non-upset games 81% (96 out of 119) right. And it predicted the upset games 38.5% (5 out of 13) right.

**Random Forest Model**

Reason for choosing this model:Random forest model is an ensemble model which can provide high accuracy and reduce overfitting

Cross-validation results  
[0.81818182, 0.79545455, 0.72727273 0.79545455, 0.75, 0.79545455, 0.72727273, 0.68181818, 0.79545455, 0.77272727]

Average cv accuracy: 0.74545

A blue and black graph

Description automatically generated

Top 3 important features in Random Forest Model are BARTHAG\_F, KENPOM ADJUSTED EFFICIENCY\_F and WINS ABOVE BUBBLE\_U

**Gradient Boosting Model (GBM)**

Reason for choosing this model:A model using GBM won the *NCAAW March Madness Kaggle Competition* in 2022. Therefore, this model can be applied to predict Men’s Tournament

Cross-validation results  
[[0.81818182, 0.77272727, 0.79545455, 0.72727273, 0.68181818, 0.77272727, 0.65909091, 0.70454545, 0.79545455, 0.70454545]

Average cv accuracy: 0.74318

A blue and white graph

Description automatically generated

Top 3 important features in GBM are BARTHAG\_F, and WINS ABOVE BUBBLE\_U and POINTS PER POSSESION DEFENSE\_F

Out of the three models, Logistic Regression has the highest accuracy at 0.76590. This means that this model can predict the outcome of potential upset games (where an underdog team is 5 slots or lower than the favorite team) 76.6% of the time. The result of all three models was close and the order could easily change with some additional data or analysis from different angles/perspectives. However, the reason why Logistic Regression won may be because it is specialized in predicting a binary outcome. On the other hand, Random Forest and GBM are well-rounded ensemble models that are not necessarily made exactly for binary outcomes.

**Analysis of Barthag and Wins Above Bubble**

Decided to investigate Barthag and Wins Above Bubble deeper because they were both in the top 3 of important features of Random Forest and GBM.

**Barthag definition:** Power Rating (Chance of beating the average Division I basketball team).

**Wins Above Bubble definition:** The expected winning percentage for an average \*bubble team in each game of a team's schedule subtracted from the total of the team's actual wins. The bubble teams for the NCAA tournament are on the cusp of making the field of 68, but an invitation isn’t guaranteed

Below is the relationship between Barthag and Upsets. Blue dots are non-upset games and red dots are upset games. We can see that a game of favorite team with barthag < 0.95 and underdog with barthag > 0.75 has a higher concentration of upsets.

A graph with red and blue dots

Description automatically generated

Below is the relationship between Wins Above Bubble and Upsets. We can see that a game of favorite team with WAB < 6 and underdog with WAB > 0 has a higher concentration of upsets

A diagram of a graph

Description automatically generated

**Limitations**

* + Data period
    - Used 2008-2021 data for this project. March Madness has a long history since 1939
  + Alternative data
    - Did not use data outside of team stats. It could produce a model with higher accuracy if able to collect alternative data e.g. coach experience, travel distance
  + Recency data
    - No data that takes recency into account. There is something to be said about peaking at the right time in the season
  + Resource and time
    - If there were more time and resources, it would be interesting to deep dive into the data more by testing additional prediction models or looking into the trends through different rounds

**Conclusion**

The objective of this study was to investigate the underlying factors that contribute to upsets in the March Madness tournament and to develop a predictive model for future tournaments. The problem is March Madness often experiences upsets where lower-seeded teams defeat higher-seeded teams. These upsets present challenges for individuals attempting to predict tournament outcomes. Through this study, the created model proved to be able to predict with a fairly high accuracy. For next year’s tournament, we can download data from KenPom and BartTorvic Rankings and apply those metrics to the model to predict upsets.

Below are key takeaways:

* Upset is defined as an underdog win in a game where **Higher Seed – Lower Seed >= 5**
* Round 4 (Elite 8) has the highest probability of upsets
* Out of the three tested models (logistic regression, random forest and gradient boosting) **logistic regression has the highest accuracy**
* **Barthag** and **Wins Above Bubble** features are important to the random forest and gradient boosting models

**References**

1. Worley, M. (2022, March 3). *Use my upset prediction model to pick underdogs in your NCAA tournament bracket*. Medium. <https://towardsdatascience.com/use-my-upset-prediction-model-to-pick-underdogs-in-your-ncaa-tournament-bracket-87c4aa3935f5>
2. *T-Rank - Customizable College Basketball Tempo Free Stats - T-Rank*. (n.d.). Www.barttorvik.com. Retrieved July 27, 2023, from <https://www.barttorvik.com>
3. *2021 Pomeroy College Basketball Ratings*. (n.d.). Kenpom.com. <https://kenpom.com/>
4. *Predicting Upsets*. (n.d.). Kaggle.com. Retrieved July 27, 2023, from <https://www.kaggle.com/code/nishaanamin/predicting-upsets#Upset-Count-Per-Tournament->
5. Worley, M. (2022, March 3). *Predicting Upsets in the NCAA Tournament with Machine Learning*. Medium. <https://towardsdatascience.com/predicting-upsets-in-the-ncaa-tournament-with-machine-learning-816fecf41f01>
6. *Advanced Metrics in College Basketball and How to use them | Hoops Amino*. (n.d.). Hoops | Aminoapps.com. Retrieved July 27, 2023, from <https://aminoapps.com/c/hoops/page/blog/advanced-metrics-in-college-basketball-and-how-to-use-them/RrNH_wuKmZrxjxL2VY8Ylkkwoo5p4z>
7. C.J. Moore. (2013, March 7). *The Ken Pomeroy Effect: How Advanced Metrics Have Revolutionized NCAA Basketball*. Bleacher Report; Bleacher Report. <https://bleacherreport.com/articles/1556819-the-ken-pomeroy-effect-how-advanced-metrics-have-revolutionized-ncaa-basketball>
8. *NCAAM March Madness scores 1985-2021*. (n.d.). Www.kaggle.com. Retrieved July 27, 2023, from <https://www.kaggle.com/datasets/woodygilbertson/ncaam-march-madness-scores-19852021>