Finding Math in the Madness: Predicting Upsets in the March Madness Basketball Tournament

Louisa Brown, Lily Feingold, Manu Skora, and Claire Woodrow
Furman University
Bracket structure
- 68 teams
- 4 regions
- 16 teams per region ranked by seed
- 6 rounds
2021 West Region

At least 5 seed difference for Favorite/Underdog matchup
Focus of our Research

• Historically, 23% of matchups with at least a 5-seed difference were upsets

• Goal: Determine which regular season characteristics of favorites and underdogs result in upsets more/less often than historical average

• Use these characteristics to predict future upsets
Understanding our Data Set

• Regular season and tournament data from 2007–2022
  • Regular season data – all D1 teams
  • Tournament data – only contains games with seed difference of at least five
• Use 2007–2021 for model training
• Reserve 2022 data for future testing
• No tournament in 2020
Simple Rating System (SRS)

- Used to form a rating, $r_i$, for each team $i$
- $r_i$ represents how much better team $i$ is than an average team on a neutral court
Simple Rating System (SRS)

For each team $i$:

$$r_i = \frac{\sum_{j \in O_i} pd_{i,j}}{\text{total games}_i} + \frac{\sum_{j \in O_i} r_j}{\text{total games}_i}$$

- Average margin of victory (MOV)
- Average opponent strength (SOS)
SRS example

3 Kansas Favorite $15.9267 = SRS_{KU}$

14 E Washington Underdog $3.7227 = SRS_{EWU}$

Predicted MOV: $SRS_{KU} - SRS_{EWU} = 12.2040$

Actual MOV: $12.0684$

Secret Sauce: Actual MOV - Predicted MOV = $-0.1356$
Histogram of Secret Sauce
Multi Variable Analysis
Analysis

- Use data to find historical percentage of upsets in each category
<table>
<thead>
<tr>
<th>Favorite Offense</th>
<th>Low</th>
<th>Medium</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Underdog Defense</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>22.81%</td>
<td>15.51%</td>
<td>2.99%</td>
</tr>
<tr>
<td>Medium</td>
<td>27.54%</td>
<td>26.67%</td>
<td>12.77%</td>
</tr>
<tr>
<td>High</td>
<td>45.10%</td>
<td>32.76%</td>
<td>25.53%</td>
</tr>
<tr>
<td></td>
<td>31.07%</td>
<td>25.00%</td>
<td>13.19%</td>
</tr>
</tbody>
</table>

Historical Upset Percentage: 23%
Fast Pace + Rebounds Decision Tree

Fast Favorite
- 18.87%
  - Fast Underdog
    - 11.02%
      - High Rebound Favorite
        - 6.90%
        - High Rebound Underdog
          - 10.00%
        - Low Rebound Favorite
          - 15.00%
          - Low Rebound Underdog
            - 3.57%
            - 24.14%
          - 6.45%
      - Slow Underdog
        - 25.34%
          - High Rebound Favorite
            - 21.92%
            - High Rebound Underdog
              - 29.73%
              - Low Rebound Underdog
                - 13.89%
                - 34.29%
          - Low Rebound Favorite
            - 29.17%
            - Low Rebound Underdog
              - 24.32%
Fast Pace + Rebounds Decision Tree

Fast Favorite: 18.87%

- Fast Underdog: 11.02%
  - High Rebound Favorite: 6.90%
  - Low Rebound Underdog: 10.00%
    - High Rebound Underdog: 3.57%
  - Low Rebound Favorite: 15.00%

- Slow Underdog: 25.34%
  - High Rebound Favorite: 21.92%
  - Low Rebound Underdog: 29.17%
    - High Rebound Underdog: 24.32%
  - Low Rebound Favorite: 13.89%

Slow Pace + Rebounds Decision Tree

Slow Favorite
- 26.67%

Fast Underdog
- 26.40%
  - High Rebound Favorite
    - 19.35%
    - High Rebound Underdog
      - 20.00%
  - Low Rebound Favorite
    - 33.33%
    - Low Rebound Underdog
      - 18.75%

Slow Underdog
- 26.67%
  - High Rebound Favorite
    - 26.67%
    - High Rebound Underdog
      - 25.81%
  - Low Rebound Favorite
    - 26.67%
    - Low Rebound Underdog
      - 27.50%
  - High Rebound Underdog
    - 25.81%
    - Low Rebound Underdog
      - 32.50%
  - Low Rebound Underdog
    - 26.67%
    - Low Rebound Underdog
      - 20.00%
Slow Pace + Rebounds Decision Tree

- **Slow Favorite**: 26.67%
  - **Fast Underdog**: 26.40%
    - **High Rebound Favorite**: 19.35%
    - **Low Rebound Favorite**: 33.33%
  - **Slow Underdog**: 26.67%
    - **High Rebound Favorite**: 26.67%
    - **Low Rebound Favorite**: 26.67%

**Pace**
- **High Rebound Underdog**: 20.00%
- **Low Rebound Underdog**: 18.75%
- **High Rebound Underdog**: 40.63%
- **Low Rebound Underdog**: 25.81%
Community Detection

Two different clustering algorithms

• *k*-means clustering
  o Used to cluster similar teams

• Louvain clustering algorithm
  o Used to cluster similar games
<table>
<thead>
<tr>
<th></th>
<th>Butterfingers</th>
<th>Offense Focused</th>
<th>Average Joe</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Power</strong></td>
<td>50%</td>
<td>23.81%</td>
<td>26.15%</td>
</tr>
<tr>
<td><strong>Defense Focused</strong></td>
<td>35.85%</td>
<td>20.0%</td>
<td>26.42%</td>
</tr>
<tr>
<td><strong>Lucky Team</strong></td>
<td>26.79%</td>
<td>9.09%</td>
<td>4.60%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Butterfinger-Power Matchups</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Alabama vs UCLA</strong> 2021</td>
</tr>
<tr>
<td><strong>Tennessee vs Oregon St.</strong> 2021</td>
</tr>
<tr>
<td><strong>Wisconsin vs Iowa St.</strong> 2022</td>
</tr>
</tbody>
</table>
Louvain Clustering Algorithm

Each node represents a game.

Cluster Number

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upset percentage</td>
<td>18.58%</td>
<td>23.53%</td>
<td>34.59%</td>
<td>16.44%</td>
</tr>
</tbody>
</table>
### Gonzaga vs Georgia State (2022)
- Michigan vs Texas Southern (2021)
- Kansas vs Detroit Mercy (2012)
- Villanova vs Radford (2018)
- Kansas vs Boston University (2011)
- Florida vs Jackson St (2007)
- Oklahoma vs CSU Bakersfield (2016)
- Kentucky vs Western Kentucky (2012)
- Oklahoma vs Morgan St (2009)
- Florida vs Northwestern St (2013)
- Kansas vs Western Kentucky (2013)

0 of the 10 most similar games were upsets

### Villanova vs Michigan (2022)
- Brigham Young vs Gonzaga (2011)
- SMU vs UCLA (2015)
- Memphis vs Saint Mary’s (2013)
- Duke vs California (2010)
- Villanova vs Saint Mary’s (2009)
- Villanova vs Saint Mary’s (2010)
- Texas A&M vs Utah St. (2010)
- Duke vs West Virginia University (2008)
- St. Johns vs Gonzaga (2011)
- Memphis vs Nevada (2007)

<table>
<thead>
<tr>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>34.59%</td>
<td>16.44%</td>
</tr>
</tbody>
</table>

4 of the 10 most similar games were upsets
Ensemble Model

Singular model made up of our 18 initial models, with preferential weighting given to models that are more predictive
Ensemble Model

• Purpose: pick models that offer new information other models lack

• After each iteration, each newly picked model has less voting power

• Newly picked model is the best model at correctly predicting the games that the previous models mispredicted

• Drawback: risk of overfitting our ensemble model to predict rare occurrences
Testing Our Model on 2022 Games

Correct predictions: 76.74%
## 2022 Games our Model Predicted as Upsets

<table>
<thead>
<tr>
<th>Game</th>
<th>Score</th>
<th>Is upset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saint Mary’s vs Indiana</td>
<td>82 - 53</td>
<td>0</td>
</tr>
<tr>
<td>Colorado St vs Michigan</td>
<td>63 - 75</td>
<td>1</td>
</tr>
<tr>
<td>Texas vs Virginia Tech</td>
<td>81 - 73</td>
<td>0</td>
</tr>
<tr>
<td>Alabama vs Notre Dame</td>
<td>64 - 78</td>
<td>1</td>
</tr>
<tr>
<td>LSU vs Iowa St</td>
<td>54 - 59</td>
<td>1</td>
</tr>
<tr>
<td>Baylor vs North Carolina</td>
<td>86 - 93</td>
<td>1</td>
</tr>
<tr>
<td>Tennessee vs Michigan</td>
<td>68 - 76</td>
<td>1</td>
</tr>
<tr>
<td>Texas Tech vs Notre Dame</td>
<td>59 - 53</td>
<td>0</td>
</tr>
<tr>
<td>Wisconsin vs Iowa St</td>
<td>49 - 54</td>
<td>1</td>
</tr>
</tbody>
</table>
Thank you!

We would like to acknowledge the following people who made this research possible.

Dr. Liz Bouzarth
Dr. John Harris
Dr. Kevin Hutson
### $k$-means Clusters of Favorite Teams

<table>
<thead>
<tr>
<th>Variables</th>
<th>B: Butterfingers Cluster</th>
<th>O: Offense-Focused Cluster</th>
<th>A: Average Joe Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rebounds</td>
<td>![Down Arrow]</td>
<td>![Up Arrow]</td>
<td>![Down Arrow]</td>
</tr>
<tr>
<td>Rebound prevention</td>
<td>![Up Arrow]</td>
<td>![Down Arrow]</td>
<td>![Up Arrow]</td>
</tr>
<tr>
<td>Turnover rate</td>
<td>![Down Arrow]</td>
<td>![Up Arrow]</td>
<td>![Up Arrow]</td>
</tr>
<tr>
<td>Offense</td>
<td>![Down Arrow]</td>
<td>![Up Arrow]</td>
<td>![Up Arrow]</td>
</tr>
</tbody>
</table>
$k$-means Clusters of Underdog Teams

<table>
<thead>
<tr>
<th>Clusters</th>
<th>SOS</th>
<th>Rebounds</th>
<th>Defense</th>
<th>Field Goals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power Cluster</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>Defense-Focused Cluster</td>
<td>↓</td>
<td>↑</td>
<td>↑</td>
<td>↓</td>
</tr>
<tr>
<td>Lucky Team Cluster</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
</tr>
</tbody>
</table>
Current Models

Single and Multi-Variable Analyses
- Pace
- Rebounds
- Turnovers
- Three pointers
- Strength of Schedule
- Offense
- Defense

Clustering
- Similar teams: $k$-means and Louvain
- Similar games: Louvain

SRS Scores
- Simple SRS
<table>
<thead>
<tr>
<th>Underdog SOS</th>
<th>Low</th>
<th>Medium</th>
<th>High</th>
<th>Historical Upset Percentage: 23%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>10.53%</td>
<td>13.85%</td>
<td>10.81%</td>
<td>11.86%</td>
</tr>
<tr>
<td>Medium</td>
<td>24.0%</td>
<td>14.81%</td>
<td>31.91%</td>
<td>23.30%</td>
</tr>
<tr>
<td>High</td>
<td>28.13%</td>
<td>31.58%</td>
<td>36.67%</td>
<td>32.04%</td>
</tr>
</tbody>
</table>
Lift chart