Shouvik Dutta, Sheldon H. Jacobson* and Jason J. Sauppe

Identifying NCAA tournament upsets using Balance Optimization Subset Selection

DOI 10.1515/jqas-2016-0062

Abstract: The NCAA basketball tournament attracts over 60 million people who fill out a bracket to try to predict the outcome of every tournament game correctly. Predictions are often made on the basis of instinct, statistics, or a combination of the two. This paper proposes a technique to select round-of-64 upsets in the tournament using a Balance Optimization Subset Selection model. The model determines which games feature match-ups that are statistically most similar to the match-ups in historical upsets. The technique is then applied to the tournament in each of the 13 years from 2003 to 2015 in order to select two games as potential upsets each year. Of the 26 selected games, 10 (38.4%) were actual upsets, which is more than twice as many as the expected number of correct selections when using a weighted random selection method.

Keywords: basketball; optimization; predictive modeling.

1 Introduction

The men's college basketball championship tournament, henceforth referred to as *the tournament*, is held annually by the National Collegiate Athletic Association (NCAA). The tournament attracts a tremendous amount of attention nationally from the public and the media, which has resulted in the event being commonly known as *March Madness*. People all over the country engage in the tournament both by supporting their favorite team and by attempting to predict the outcome. Sites such as ESPN (2015) and Yahoo (2015) host bracket competitions, where people submit their predictions for the outcome of each game in the tournament. In 2015 alone, people submitted approximately 70 million brackets to various sites (Wartenberg 2015). Accurately predicting the outcome of games can also be financially lucrative, with Americans wagering approximately $9 billion in 2015 (Marino 2015). In 2014, Quicken Loans partnered with Yahoo to offer $1 billion to anybody who could create a bracket with every game predicted correctly (Yahoo 2014). Both the pride from being correct and financial opportunities have incentivized many individuals and companies to develop models to predict the outcomes of the tournament games.

The world of sports forecasting can be a daunting one for those people not familiar with the sport and current teams. Newcomers attempting to learn about the teams in the tournament are faced with a copious amount of statistics, team rankings, and expert opinions. To help users create a bracket, news and sports sites such as ESPN and Fivethirtyeight.com (FiveThirtyEight 2015) make their own predictions publicly on how the tournament will proceed. However, while they do disclose some components and relationships that go into their predictive models, a large portion of the models are proprietary. Even the revealed portions of the models involve a multitude of factors that are prohibitive for a newcomer to obtain and use. The use of a few key statistics that are easy to acquire and understand would allow both experts and novices to make forecasts based on the same data.

This paper proposes a technique to select potential upsets using only a small number of publicly available statistics by identifying match-ups in the current year that exhibit characteristics similar to those exhibited by historical round-of-64 upsets. The differences in season statistics between the two teams in each historical upset are used to build a profile of past upsets, which is then compared to first round games in the current year to find match-ups that are most similar to historical upsets. By limiting the potential characteristics to game statistics, the technique can be back-tested to ascertain its accuracy. Testing was done by generating predictions for each year from 2003 to 2015 using only data that would have been available at the time for each year. This technique is shown to outperform both the random selection of upsets, as well as the Kaggle competition-winning technique provided by Lopez and Matthews (2015), and the results obtained are reproducible using freely available information from TeamRankings.com (TeamRankings 2015).

By examining historical upsets, our technique is able to identify characteristics that allow a weaker team to beat a stronger team, and then find games in a given year's tournament that exhibit those characteristics. Taking the specific match-up in each game into account allows us

***Corresponding author: Sheldon H. Jacobson,** University of Illinois – Computer Science, Urbana, IL, USA, e-mail: shj@illinois.edu
Shouvik Dutta: University of Illinois – Industrial Engineering, Urbana, IL, USA
Jason J. Sauppe: University of Wisconsin–La Crosse, Computer Science, La Crosse, WI, USA

to identify upsets with greater accuracy than weighted random selection would allow.

This paper is organized as follows. Section 2 describes current techniques for predicting the outcome of games. Section 3 describes the method by which Balance Optimization Subset Selection is used to choose potential upsets. Section 4 summarizes the results of the proposed technique by providing the predictions made by the technique. Section 5 provides concluding remarks.

2 Background

2.1 Rating systems

Several team rating systems that quantify the skill of teams have been introduced and popularized, including ESPN's *basketball power index (BPI)* (Oliver 2013), Ken Pomeroy's *pythagorean ratings* (Pomeroy 2012), and Jeff Sagarin's *predictor ratings* (Sagarin 2015). These rating systems focus on assigning a numeric rating to each team that estimates how successful that team will be in games. The premise behind these systems is that a team with a higher rating is stronger than a team with a lower rating, where the difference between the rating of the two teams indicates the difference between the strength of the teams.

The basketball power index (BPI) was introduced in 2013 by ESPN and touted as "a little more refined than any other existing power ranking."(Oliver 2013) While the exact formula for calculating a team's BPI is not reported in the literature, ESPN does reveal some of the components of the ranking. The BPI includes such information as a team missing an important player during a game, whether the game is home or away, whether the game was a blowout or a close game (Oliver 2013), the pace of the game, and the strength of a team's schedule (how strong a team's opponents were) (Volner 2013).

Ken Pomeroy, owner and operator of kenpom.com, scores teams based on a *pythagorean winning percentage* (Pomeroy 2012), which is the expected fraction of games a team should win against an average team. To calculate this percentage, he uses the *adjusted offensive efficiency (AdjO)* and *adjusted defensive efficiency (AdjD)* of a team. The adjusted offensive efficiency is an estimate of the number of points a particular team would score per 100 possessions against an average team (as assessed by Pomeroy). The adjusted defensive efficiency is an estimate of the points allowed per 100 possessions by a team against an average team. The method of computing these adjusted values is not reported in the literature. The adjusted offensive efficiency and adjusted defensive efficiency are then combined into the pythagorean winning percentage using the formula

$$\text{pyth} = \frac{AdjO^{10.25}}{AdjO^{10.25} + AdjD^{10.25}}. \tag{1}$$

These ratings systems are used to determine the outcome of games. In a match-up between two teams, the team with the higher rating is predicted to have a higher likelihood of winning a game. The magnitude of the difference in the ratings is also used to determine how likely each team is to win. For example, Fivethirtyeight.com combines seven different ratings to predict the likelihood of one team beating another (Silver 2015).

2.2 Match-up analysis

An alternative to predicting game outcomes by comparing the rating of two teams is to compare the two teams in a match-up directly.

The tournament is divided into four regions, and each team in the tournament is given a *seed*, which is an estimate of the rank of a team in their respective region of the bracket determined by the selection committee prior to the tournament. The team deemed by the committee to be the strongest in each region is given a seed of 1 and the team deemed to be the weakest in each region is given a seed of 16. We define a team with a small numeric seed as having a *high seed* and teams with a large numeric seed as having a *low seed*. Therefore, a team with seed one is the highest seeded team in its region and a team with seed 16 is the lowest seeded team in its region. The games of interest to this paper are upsets, which are games in which a low seeded team beats a high seeded team. Upsets are defined by ESPN as games where the difference between the seed of the winning team and the seed of the losing team is at least five (Keating 2013). ESPN looks for potential upsets by looking for teams that are stronger than their seed would suggest or by finding match-ups where the weaker seeded team has a strength that could exploit the weakness of a stronger team (Brenner and Keating 2015). ESPN defines four categories of high seed teams that are capable of losing (Brenner and Keating 2015):

- Power giants: Strong offensive rebounding, average defensive rebounding, do not force many turnovers
- Gambling giants: Strong offensive rebounding, weak defensive rebounding, force many turnovers
- Pack-line giants: Average offensive rebounding, do not force many turnovers, good defensive rebounding

– Generic giants: Generally skilled, not specifically strong in offensive rebounding or generating turnovers

ESPN also defines four categories of low seed teams that have the potential to upset (Brenner and Keating 2015):
– Generic killers: Decent teams with no especially strong rebounding, turnovers, shooting, or defense
– Slow killers: Strong offensive rebounding, limit opponent shooting, neither generate steals nor shoot 3-point shots
– Perimeter killers: Strong 3-point shooters, generate lots of steals, weak offensive rebounding, weak at limiting opponent shooting
– High-possession killers: Limit opponent shot accuracy, strong offensive rebounding, do not shoot many 3-point shots

These categories are then analyzed to see which *Giants* (high seed teams) could fall to which *Giant killers* (low seed teams). ESPN's conclusions are shown in Figure 1, where an arrow from a Giant to a Giant killer means that the Giant is weak against that Giant killer and an arrow from a Giant killer to a Giant means that the Giant killer is strong against that Giant.

ESPN does not elaborate on exactly how each team is placed into any of the above categories.

There has also been research into quantitatively predicting the probability of one team winning against another. Kaggle (2015), a website that hosts data science competitions, ran a college basketball tournament prediction contest in 2014 and 2015. In Kaggle's competition, participants were asked to compute the probability of each team beating each other team in the tournament, but were only scored on those matches that actually occurred. Because there are 68 teams in the tournament (including the play-in matches), participants made probabilistic predictions for each of the 2278 potential team pairings that could occur. Each match was weighted equally, unlike traditional bracket scoring where predicting the winner of the tournament is worth several times as many points as predicting the outcome of a round-of-64 match. The advantage of Kaggle's system is that it made predicting round-of-64 upsets correctly more advantageous than a traditional bracket where the later rounds are much more important due to later rounds typically being worth more points. The Kaggle competition winners (Lopez and Matthews 2015) used a logistic regression model incorporating the Las Vegas point spread (the expected margin of each game) given by Covers (Covers 2015) and Ken Pomeroy's efficiency ratings.

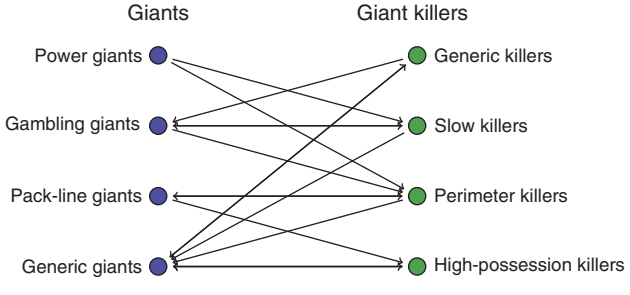

Figure 1: Which Giants might lose to which Giant killers and which Giant killers might win against which Giants.

The downside to the existing methodologies for predicting the outcome of a game is that they are difficult to replicate due to their opaque nature. The method for calculating the Las Vegas point spread is not publicly available, and neither is the exact formula for ESPN's BPI. Also, while using factors such as strength of schedule seems useful, it introduces its own biases because it is not an objective statistic, but rather relies on people to determine the relative strengths of teams who have never played against each other. ESPN outlines qualities of Giants that may fall and the Giant Killers that may upset them, but does not elaborate on the methods used to identify these teams. This paper aims to provide an approach to predicting upsets that uses only statistics that can be obtained from watching the games, and also self-identifies the important factors in predicting upsets without user intervention.

3 Methodology

In this paper, we define an upset as a team seeded 13, 14, or 15 winning a game in the round-of-64 (compared to ESPN's 11 or higher). A *k-seed game* refers to a game in which one of the teams has seed k. We exclude 16-seeded teams because no 16-seed has ever won a round-of-64 game. This allows the focus to be placed on identifying characteristics of a much smaller number of upsets than if 11 and 12 seeded teams were included. Between 1985 and 2015, 52 of the 372 games played involving 13–15 seeds have been upsets. This is an average of 1.67 upsets per year. Therefore, we aimed to identify two match-ups each year that are similar to past historical upsets. All raw data was obtained from TeamRankings.com (TeamRankings 2015) and are obtained as of Selection Sunday for each year (so no tournament game statistics are included in the data). The proposed technique can be described as a series of four steps.

1. Computing match-up statistics: Compare the two teams playing in each game using team statistics.
2. Identifying useful match-up statistics: Use an extra-trees classifier to select a set of match-up statistics $\mathcal{S}$ that are strong indicators of historical performance.
3. Finding similar match-ups: For each subset of match-up statistics $S \subseteq \mathcal{S}$ with $|S| = 4$, use Balance Optimization Subset Selection to identify three match-ups that are similar to historical upsets on the match-up statistics in S.
4. Combining models: Identify combinations of match-up statistics $S \in \binom{\mathcal{S}}{4}$ that performed well historically and use them to select two match-ups as potential upsets.

The following subsections will elaborate on each of the four listed steps.

3.1 Computing match-up statistics

Because both players and play styles in college basketball can change from year to year, we use ordinal rankings instead of the raw statistic value when looking at team statistics. For example, instead of using 58.4% as Notre Dame's Two Point Percentage in 2015, we use the fact that it was the highest Two Point Percentage of any team in 2015, and assign it a value of 1. Gonzaga, having the 2nd highest Two Point Percentage in 2015, would be assigned a value of 2 for that statistic. Relative ranking allows teams from different years to be compared while accounting for the way that the league as a whole may change. Relative rankings also have the advantage of forcing the range of values for each statistic to be the same (1 to the number of teams barring ties). In the event that multiple teams have the same value for a statistic, they are assigned a rank that is the average of the ranks of that value. For example, if three teams shared the third highest value for a statistic, the three teams would all be assigned $(3+4+5)/3 = 4$ for that statistic. The team with the next highest value for that statistic would be assigned a rank of 6.

Rather than looking at a team's statistics, it is more useful to look at how a team's statistics compare to those of its opponent. One team averaging a very high number of steals may signal a high number of scoring opportunities, but if the opponent team averages a similar number of steals, neither team is likely to gain an advantage over each other by relying solely on steals. To find the differences between the teams, we subtract the ordinal ranking of each statistic for the higher-seeded team from those of the lower-seeded team for each game. Comparing the teams in the match-up reveals gaps in the statistics of the teams, such as if one team shoots many more three point shots or one team forces many more turnovers than the other. These match-up statistics, rather than team statistics, are then used to find potential upsets. By observing which statistics have gaps that lead to upsets in historical games, games in the future can be identified as having the potential for being an upset.

3.2 Identifying useful match-up statistics

TeamRankings.com (TeamRankings 2015) tracks 115 different statistics for each team going back to the 1997-1998 season. The first step is to find a small set of statistics that are correlated with upsets. Trying every combination of statistics is infeasible, as there are $2^{115}-1$ possible combinations. To find a subset of statistics that are useful for selecting potential upsets, an extra-trees classifier (Geurts, Ernst, and Wehenkel 2006) was trained on round-of-64 tournament matches from 1998 to 2015, where the classifier was built using all 115 match-up statistics. An extra-trees classifier builds a large number of decision trees that, in our case, use the various match-up statistics to differentiate upsets from non-upsets. The extra-trees classifier was chosen for its resistance to overfitting (Geurts et al. 2006) and because it allows us to measure the importance of each feature for the purposes of classification. The resistance to overfitting is due to the nature of the classifier, which trains each tree on a random subset of the data using a random subset of input features. Because each training example and feature is excluded in many of the trees, the classifier avoids learning to overfit the training set provided a sufficiently large number of trees.

The classifier was built using 100,000 decision trees with $\sqrt{115}$ features sampled for each split and two samples required to split each node (as suggested in Geurts et al. (2006)). The importance of each feature can then be extracted from the classifier using *Gini Importance* (Breiman et al. 1984). The implementation used for both the classifier and feature importance was the extra-trees classifier in Python's scikit-learn library. Based on preliminary experiments testing different numbers of match-up statistics with subsequent steps of the technique, we opted to use the resulting 15 most important features for identifying upsets. These features were (in descending order of Gini Importance):

1. Effective Possession Ratio
2. Games Played
3. Extra Scoring Chances per Game
4. Opponent Floor Percentage

5. Personal Fouls per Possession
6. Opponent Steals per Defensive Play
7. Assist / Turnover Ratio
8. Personal Fouls per Defensive Play
9. Opponent Steals per Possession
10. Opponent Average Scoring Margin
11. Average Scoring Margin
12. Opponent Three Point Percentage
13. Steals per Defensive Play
14. Steals per Possession
15. Average 2nd Half Margin

These statistics are defined in Appendix A. Let $\mathcal{S}$ be the set of these match-up statistics. It should be noted that the extra-trees classifier was trained using games from 1998 to 2015 rather than training the classifier to identify features for each year separately. The decision to train the extra-trees classifier on games from 1998 to 2015 instead of only on games from 1998 to the year for which upsets were being identified was made due to the limited quantity of available historical data.

Finding a suitable set of statistics to use could have also been done by enumeration and validation, but the classifier was used to avoid enumerating $2^{115}-1$ different potential combinations of statistics. None of the decision trees generated by the extra-trees classifier are being used in any way to select potential upsets.

3.3 Finding similar games

The next step of the technique is to use the set of 15 match-up statistics $\mathcal{S}$ to select potential upsets. The Balance Optimization Subset Selection (BOSS) framework (Nikolaev et al. 2013) is used to do the selection. BOSS is a framework that allows the selection of a small *control group* from a larger *control pool* that is similar to a *treatment group*, where similarity is determined by a defined balance measure.

The BOSS framework was originally designed as a framework for conducting observational studies. In an observational study, one has access to a set of units that were exposed to treatment along with a set of units that were not (the controls) and the goal is to estimate the effect of the treatment. A significant difficulty is that exposure to treatment is almost always non-random, which makes it difficult to determine if estimated effects are due to the treatment itself or other confounding factors called *covariates*. The traditional approach to resolve this difficulty is to use matching methods (Rubin 1973, 1980), which pair each treatment unit with a control unit that possesses similar covariate values. As the number of covariates increases, finding exact matches becomes more difficult. Propensity score methods (Rosenbaum and Rubin 1983) are one potential solution to this problem, but they still require the propensity score (a scalar value representing likelihood of treatment) to be estimated in order to be used. When matching methods fail to achieve exact matches, it becomes difficult to determine which set of matched pairs is best. A commonly accepted practice is to select the set of matched pairs that features the best level of *covariate balance*, which is typically a measure of aggregate similarity (e.g., difference of covariate means, two-sample Kolmogorov-Smirnov tests applied to the covariates' marginal distributions) between the treatment units and the matched controls (Diamond and Sekhon 2013; Rosenbaum and Rubin 1985).

Whereas matching methods typically match first and then assess covariate balance afterwards, the BOSS framework drops the matching component in favor of directly optimizing covariate balance. That is, BOSS is designed to identify a control group that possesses optimal balance with respect to the treatment group, without requiring the construction of matched pairs. While BOSS can be used to optimize any measure of covariate balance, typical measures focus on the marginal distributions of the covariates. This is because balance on the marginal distributions of the covariates is a relaxation of the requirement for exactly matched pairs, and as such is more likely to be achievable, particularly when faced with limited data (Sauppe, Jacobson, and Sewell 2014).

For selecting potential upsets, the control pool consisted of the 13, 14, and 15-seed round-of-64 games and the treatment group consisted of historical upsets. The control group selected by BOSS would be the set of match-ups most similar to the historical upsets according to the defined balance measure. For each combination of four match-up statistics $S \subset \mathcal{S}$, BOSS is used to select three match-ups as potential upsets. These will then be narrowed to two final selections in the next step of the technique.

BOSS requires a balance measure by which games in the current year can be compared to upsets in the past to measure the similarity between the control group and the treatment group. The balance measure used was a combination of (1) the difference between the empirical distribution of the treatment and control group for each statistic $i \in S$ and (2) the relative difference between the sum of each statistic for each game in the control and treatment group. The difference between the empirical distributions was measured using the Kolmogorov-Smirnov (KS) test statistic, which defines the distance K between

two sets f_1 and f_2 with empirical distributions F_1 and F_2 respectively over a set of values V as

$$K(f_1, f_2, V) \equiv \max_{v \in V} |F_1(v) - F_2(v)|. \quad (2)$$

The Kolmogorov-Smirnov statistic measures the maximum vertical distance between two empirical distributions.

Because the KS statistic only measures the difference in the height of the empirical distributions, it is possible for the first and last values in the empirical distribution to be significantly further apart on one distribution than the other while retaining the same KS statistic value. For example, the KS statistic between {1, 2, 3, 4} and {1, 2, 3, 5} would be the same as the KS statistic between {1, 2, 3, 4} and {1, 2, 3, 1000}. In order to prevent that last value from being very far away from the rest of the distribution, we also use the relative difference between the distributions to include the horizontal difference between the two sets. The relative difference constraints force the horizontal spread to be similar in both distributions.

The relative difference R between the sum of the sets f_1 and f_2 is defined as:

$$R(f_1, f_2) \equiv \frac{|\frac{1}{|f_1|}\sum_{g \in f_1} g - \frac{1}{|f_2|}\sum_{g \in f_2} g|}{\frac{1}{|f_2|}\sum_{g \in f_2} g}. \quad (3)$$

The *covariates* in this problem are the different match-up statistics being used. Let the following terms be defined as:

- T: Treatment Group
- C: Control Pool
- G: Control Group
- $\mathbf{X}$: Set of covariates
- S_i: Values of set S for covariate i
- V_i: Set of unique values in $T \cup C$ for covariate $i \in \mathbf{X}$

We then set our balance measure $M(G)$ for control group G to be

$$M(G) = \sum_{x \in \mathbf{X}} \max\{K(T_i, G_i), R(T_i, G_i)\}. \quad (4)$$

We then find group G with size $|G| = 3$ such that $M(G)$ is minimized. The three teams in G will be the three selections for the set of covariates S. Due to the small size of both T and C, we solved BOSS by enumerating all possible sets of G and choosing the one with the smallest $M(G)$. BOSS can also be solved via a Mixed Integer Program (MIP), the formulation for which is presented in the appendix.

BOSS was run once for each combinations of four statistics out of the 15 chosen by the classifier (1365 combinations total) on years from 2001–2015 using each combination of statistics as covariates. The earliest year used was 2001 because detailed data for years prior to the 1997–1998 season were unavailable from TeamRankings.com and forming a treatment group requires historical upsets, so some years would have to be used to build a small treatment group. The upsets in years from 1998 to 2000 were used to form the treatment group for 2001.

3.4 Combining models

Solving BOSS with each $S \in \mathcal{S}$ produced 1365 sets of three match-ups each (one set of three match-ups from each combination of four match-up statistics).In order to finally select two match-ups as potential upsets, the results of those 1365 BOSS solutions must be combined to yield two match-ups. In order to do this, a reasonable action is to take the two match-ups that appear most frequently across the set of BOSS solutions. However, not all combinations of match-up statistics are equally informative or valuable. Therefore, only those combinations of match-up statistics that proved to provide accurate solutions historically were included. The subset of combinations to use was chosen by evaluating the performance of each combination of match-up statistics using historical data and selecting the ones with the best past performance. The performance of a given combination of match-up statistics was measured as the number of upsets that BOSS selected correctly over the entire range of years used when the given combination was used as covariates. To formalize this, let $n_{S,y}$ be the number of upsets in year y that were included in the BOSS solution when optimizing over the match-up statistics S (so $n_{S,y} \in \{0, 1, 2, 3\}$). Then let $N_{S,y} = \sum_{y' < y} n_{S,y'}$ be the historical performance of S in predicting upsets up through but not including year y. For any given year y, let $N^\star_y = \max_{S \subseteq \mathcal{S}} N_{S,y}$ be the largest number of selected upsets across all match-up statistics. Then let $P_y = \left\{S \in \binom{\mathcal{S}}{4} : N_{S,y} \geq N^\star_y - \tau\right\}$ be the set of high-performing match-up statistics in year y, where τ is a tolerance parameter. The statistics in P_y are then used to select upsets for year y. The tolerance value τ was determined by testing values between one and 20 and choosing the value that yielded the most correct selections, where a correct selection is the selection of a game that was an actual upset. The value of τ was determined separately for each year.

The steps to select two teams for year Y from the results of BOSS are as follows:

1. Find the number of correct selections made by BOSS for each combination of four match-up statistics $S \in \binom{\mathcal{S}}{4}$ from year 2001 to year $Y-2$. Let the number of correct selections for S be $N_{S,Y-2}$.

2. Find the single combination of four match-up statistics for which BOSS made the most correct selections when run from year 2001 to year $Y-2$. Let this be $N^{\star}_{Y-2}$.
3. Set τ to one. Find all combinations of match-up statistics for which BOSS selected at least as many upsets correctly as the number correctly selected by the best combination found in the previous step minus the tolerance value ($N^{\star}_{Y-2} - \tau$) for years 2001 to $Y-2$. Let P_{Y-1} be the set of the selected combinations.
4. Use the two teams most frequently selected by BOSS for year $Y-1$ when run on each combination of match-up statistics in P_{Y-1} as the two selections for year $Y-1$.
5. Repeat steps (1)–(3) for each τ between one and 20. Select the value of τ that resulted in the most correct selections for year $Y-1$.
6. Use the value of τ found in step (4) to make selections for year Y using steps (1)–(3) but iterating to year $Y-1$ instead of $Y-2$ in step 2.

Algorithm 1 provides the pseudo-code for selecting the subset of combinations of match-up statistics to use and combining the teams chosen using each combination in that subset into the two selected games for a target year Y where Y is a year after 2002. Lines 3–12 iterate through each year prior to Y and determine the number of correct selections that are made for each potential tolerance value. Lines 5–8 find the number of correct selections made using each combination of match-up statistics by comparing the games selected by BOSS using that combination of match-up statistics to the actual historical upsets that occurred. The combinations that performed within the tolerance of the best single combination are used to generate the final two selections. The final selections are the two match-ups that are most selected by the chosen combinations. The tolerance value that results in the most accurate selections for years between years 2002 and $Y-1$ is then used to select two games as potential upsets for year Y in lines 14–20. If multiple tolerance values generated the same number of correct selections, the maximum of those tolerance values was used. In the event that there was a tie for the most frequently occurring team or the second most frequently occurring team, more than two teams would have been selected. However, such a tie never occurred, so this contingency was never used.

Combining the results of multiple models or instances of a model, known as *ensembling*, has been shown to frequently reduce errors due to a specific failing in individual models (Opitz and Maclin 1999). Due to the small size of the dataset, one specific set of covariates may be high performing, but a high performance by a single covariate combination may be due to coincidence because the outcome of each game is essentially a random variable. Ensembling multiple models or instances of a model should make the resultant ensemble more resistant to overfitting the dataset, but determining the amount of overfitting is difficult due to the limited number of years for which there is historical team data.

Algorithm 1 Generating selections for year Y using BOSS

1: $\mathbb{S} \leftarrow \{\text{all } S \subset \mathcal{S} : |S| = 4\}$
2: **for** $\tau \in \{1, 2, \ldots, 20\}$ **do**
3: $\quad g_\tau \leftarrow \{\}$
4: $\quad$ **for** year $y \in \{2002, 2003, \ldots, Y-1\}$ **do**
5: $\quad\quad$ **for** each $s \in \mathbb{S}$ **do**
6: $\quad\quad\quad N_{s,y} \leftarrow$ number correct selections by s using BOSS for years $\{2001, .., y-1\}$
7: $\quad\quad$ **end for**
8: $\quad\quad N^{\star}_{y} \leftarrow \max_{s \in \mathbb{S}} \{N_{s,y}\}$
9: $\quad\quad P_y \leftarrow \{s \in \mathbb{S} : N_{s,y} > N^{\star}_{y} - \tau\}$
10: $\quad\quad g_\tau = g_\tau \cup \{$the two games most frequently selected by all $s \in P_y$ using BOSS for year $y\}$
11: $\quad$ **end for**
12: **end for**
13: $\tau_{opt} \leftarrow \tau$ such that g_τ contains the most correct upsets for years $\{2002, \ldots, Y-1\}$
14: **for** each $s \in \mathbb{S}$ **do**
15: $\quad N_{s,Y} \leftarrow$ number correct selections by s using BOSS for years $\{2001, \ldots, Y-1\}$
16: **end for**
17: $N^{\star}_{Y} \leftarrow \max_{s \in \mathbb{S}} \{N_{s,Y}\}$
18: $P_Y \leftarrow \{s \in \mathbb{S} : N_{s,Y} > N^{\star}_{Y} - \tau_o pt\}$
19: return the two games most frequently selected by all $s \in P_Y$ using BOSS for year Y

4 Results

The technique presented in Section 3 was used to select two potential upsets per year for years between 2003 and 2015. The games selected as upsets are listed in Table 1. The number of upsets that occurred each year is shown in Table 2. Table 3 lists the selection frequency and accuracy for each seed separately.

In total, the presented technique selected 10 upsets correctly out of 26 picks (38.4%) over 13 years.

Analysis of the results lead to several observations about the tendencies of the technique. We selected two

Table 1: Games selected.

Year	Game seed	Winning team	Losing team	Upset selected correctly
2015	14	Georgia St	Baylor	True
2015	14	UAB	Iowa State	True
2014	13	San Diego St	N Mex State	False
2014	13	Michigan St	Delaware	False
2013	14	Marquette	Davidson	False
2013	13	La Salle	Kansas St	True
2012	14	Georgetown	Belmont	False
2012	15	Lehigh	Duke	True
2011	15	N Carolina	LIU-Brooklyn	False
2011	15	Notre Dame	Akron	False
2010	14	Ohio	Georgetown	True
2010	15	W Virginia	Morgan St	False
2009	13	Xavier	Portland St	False
2009	13	Cleveland St	Wake Forest	True
2008	13	Siena	Vanderbilt	True
2008	13	Pittsburgh	Oral Roberts	False
2007	13	S Illinois	Holy Cross	False
2007	13	Virginia	Albany	False
2006	14	Gonzaga	Xavier	False
2006	13	Bradley	Kansas	True
2005	14	Oklahoma	Niagara	False
2005	13	Vermont	Syracuse	True
2004	13	Maryland	TX El Paso	False
2004	13	Kansas	IL-Chicago	False
2003	13	Tulsa	Dayton	True
2003	14	Xavier	Troy	False

Shaded rows indicate games selected correctly.

Table 2: Number of upsets per year.

Year	Number of actual upsets	Number selected correctly
2015	2	2
2014	1	0
2013	3	1
2012	3	1
2011	1	0
2010	2	1
2009	1	1
2008	2	1
2007	0	0
2006	2	1
2005	2	1
2004	0	0
2003	1	1

upsets correctly one time, one upset correctly eight times, and zero upsets correctly four times. However, out of the 4 years where we got zero correct, 2 years had zero upsets actually occur. Therefore, we selected at least one upset correctly in nine out of the 11 years that had at least one upset occur. Moreover, given that we choose exactly two potential upsets per year, we can observe from the historical record that the maximum number of upsets that we could have chosen correctly is 18. Therefore, we selected 10 out of the 18 possible upsets that we could have selected correctly. The games chosen also tend to favor stronger seeds, as we pick a 13-seed fourteen times, a 14-seed eight times, and a 15-seed four times. Selecting higher-seed teams with higher frequency is reasonable because the 13-seeds are more likely to win than the 15-seeds.

Table 3: Selection accuracy by seed.

Seed	Number of actual upsets	Number selected	Number selected correctly
13	10	14	6
14	7	8	3
15	3	4	1
Total	20	26	10

Another interesting observation is that the tolerance value determined when using Algorithm 1 for each year remained constant for all years from 2009 onward. A constant tolerance could be evidence of some level of stability, because the fact that it stayed constant for seven consecutive years suggests that is likely to be the correct value to use in future years. However, due to the limited number of years of data available, the value for further years should be determined using the algorithm until this theory can be further tested.

To assess the variability of our technique, let X_i denote the number of correctly predicted upsets in the ith year of analysis (i.e., year 1 is 2003, year 2 is 2004, and so on). Then we have

$$\bar{X} = \frac{1}{13}\sum_{i=1}^{13} X_i = \frac{10}{13} \approx 0.769 \tag{5}$$

and

$$S^2 = \frac{1}{13-1}\left(\sum_{i=1}^{13} X_i^2 - 13\cdot\bar{X}^2\right) \approx 0.359, \tag{6}$$

so the sample standard deviation for number of correct selections per year is approximately 0.599. Assuming that the X_i are independent, the variance in the number of correct selections across all 13 years is approximately 4.667, with a standard deviation of 2.16.

To compare the performance of our technique to predictions made by randomly choosing games, we determined the expected number of correct selections when two teams were randomly selected as predicted upsets. We can either randomly select two teams each year with equal probability or, because we know the historical frequency of each seed winning a round-of-64 game, we can

randomly select two possible upsets each year using the historical frequency of an upset occurring for each seed as weights. The weights used for weighted random selection each year were calculated using upsets that occurred prior to that year. For example, when randomly selecting teams as predictions for 2010, the frequency of upsets from 1985 to 2009 was used for weighting. Each game was then determined to be an upset by modeling it as a Bernoulli variable with the probability being the fraction of historical games of that seed match-up that resulted in upsets. The following proposition establishes the number of upsets that would be correctly predicted through (1) random selection where each team has the same probability of being selected and (2) random selection where the probability of each team being selected is weighted based on the historical frequency of that team's seed resulting in an upset.

In order to compute the expected value and variance, let the following terms be defined:

- U_y: Random variable representing the number of upsets selected correctly in year y
- x_{sy}: Number of games that seed s has won before year y
- $G_{s,y}$: Number of round-of-64 games played by s-seed teams before year y
- n_y: Total number of upsets that occurred prior to year y.

The variables are dependent on the year y because the weighted probabilities used each year are computed by using the frequency of upsets occurring before that year. Therefore, these probabilities change each year as new upsets occur each year of the tournament.

Proposition 1. *When choosing two 13, 14, or 15-seeded match-ups as upsets randomly for each year between y_1 and y_2 where the probability of choosing each team is the historical frequency with which upsets occur for that seed, the expected number of upsets selected correctly is*

$$E[U_{[y_1,y_2]}] = \sum_{y\in[y_1,y_2]} \left(4\left(\frac{x_{sy}}{4n_y} + \frac{3x_{sy}}{4n_y}\frac{x_{sy}}{4n_y - x_{sy}}\right) + 4\sum_{i\in\{13,14,15\}:i\neq s} \frac{x_{iy}}{4n_y}\frac{x_{sy}}{4n_y - x_{iy}}\right)\left(\frac{x_{sy}}{G_y}\right), \tag{7}$$

and the variance is

$$Var[U_{[y_1,y_2]}] = \sum_{y\in[y_1,y_2]} \left(4\left(\frac{x_{sy}}{4n_y} + \frac{3x_{sy}}{4n_y}\frac{x_{sy}}{4n_y - x_{sy}}\right) + 4\sum_{i\in\{13,14,15\}:i\neq s} \frac{x_{iy}}{4n_y}\frac{x_{sy}}{4n_y - x_{iy}}\right)\left(\frac{x_{sy}}{G_y}\right) + 8\left(\sum_{s_i=s_j} \frac{x_{s_iy}}{4n_y}\frac{3x_{s_iy}}{4n_y - x_{s_iy}}\left(\frac{x_i}{G_y}\right)^2 + \sum_{s_i\neq s_j} \frac{x_{s_iy}}{4n_y}\frac{4x_{s_jy}}{4n_y - x_{s_iy}}\frac{x_i x_j}{G_y^2}\right). \tag{8}$$

Proof. The following equations will compute the expected value and variance when using weighted random selection. A modification to use uniform random selection is provided at the end of the proof.

The probability of randomly selecting a specific team with seed s is

$$P(\text{selecting team with seed s}) = P(\text{choose team first}) + P(\text{not choose first and choose second}) \tag{9}$$

Because there are four teams with each seed, this probability is multiplied by four. Also, the probability depends on year y. Therefore,

$$P(\text{selecting team with seed } s \text{ in year } y) = 4\left(\frac{x_{sy}}{4n_y} + \frac{3x_{sy}}{4n_y}\frac{x_{sy}}{4n_y - x_{sy}}\right) + 4\sum_{i\in\{13,14,15\}:i\neq s} \frac{x_{iy}}{4n_y}\frac{x_{sy}}{4n_y - x_{iy}}. \tag{10}$$

The probability of the selected team being an actual upset is

$$P(\text{seed } s \text{ correct in year } y) = \frac{x_{sy}}{G_y}. \tag{11}$$

Therefore for each year, the expected number of correctly predicted upsets is

$$E[U_y] = \sum_{s\in\{13,14,15\}} P(\text{selecting team with seed } s \text{ in year } y) * P(\text{seed } s \text{ correct in year } y) \tag{12}$$

Because the results of each year are independent, as upsets occurring one year do not depend on upsets occurring in the previous years, we can add the expected number of upsets each year to arrive at the expected number of upsets over a range of years. When x_s, n, and G are determined by historical data prior to each year, we find the expected number of upsets from y_1 to y_2 (including y_2) to be

$$E[U_{[y_1,y_2]}] = \sum_{y\in[y_1,y_2]} E[U_y]. \tag{13}$$

To compute the variance, we can rewrite the expected number of upsets as

$$E[U_y] = P(U_y = 1) + 2P(U_y = 2). \tag{14}$$

To compute the variance, $E[U_y]^2$ is required.

$$\begin{aligned} E[U_y]^2 &= 1^2 * P(U_y = 1) + 2^2 * P(U_y = 2) \\ &= E[U_y] + 2P(U_y = 2) \end{aligned} \tag{15}$$

To find $P(U_y = 2)$, we let (s_i, s_j) be all possible seed pairs where i and j are each drawn from {13, 14, 15} with replacement . Then

$$\begin{aligned} P(U_y = 2) =& 4 \left(\sum_{s_i = s_j} \frac{x_{s_i y}}{4n_y} \frac{3x_{s_i y}}{4n_y - x_{s_i y}} \left(\frac{x_i}{G_y} \right)^2 \right. \\ &\left. + \sum_{s_i \neq s_j} \frac{x_{s_i y}}{4n_y} \frac{4x_{s_j y}}{4n_y - x_{s_i y}} \frac{x_i x_j}{G_y^2} \right). \end{aligned} \tag{16}$$

This allows us to express the variance as

$$\begin{aligned} Var[U_y] = E[U_y] + 8 & \left(\sum_{s_i = s_j} \frac{x_{s_i y}}{4n_y} \frac{3x_{s_i y}}{4n_y - x_{s_i y}} \left(\frac{x_i}{G_y} \right)^2 \right. \\ &\left. + \sum_{s_i \neq s_j} \frac{x_{s_i y}}{4n_y} \frac{4x_{s_j y}}{4n_y - x_{s_i y}} \frac{x_i x_j}{G_y^2} \right). \end{aligned} \tag{17}$$

Because the result of each year is independent, we can say that

$$Var[U_{[y_1, y_2]}] = \sum_{y \in [y_1, y_2]} Var[U_y]. \tag{18}$$

The above equations express the expected number and variance of correct selections when using weighted random selection. In order to compute the expected number and variance of correct selections when each seed is equally likely to be chosen, modify (10) to be

$$\begin{aligned} P(\text{select team with seed } s \text{ in year } y) &= 4 \left(\frac{1}{12} + \frac{11}{12} \frac{1}{11} \right) \\ &= \frac{8}{12}, \end{aligned} \tag{19}$$

and (16) to be

$$P(U_y = 2) = \sum_{s_i = s_j} \frac{4}{12} \frac{3}{11} \left(\frac{x_i}{G_y} \right)^2 + \sum_{s_i \neq s_j} \frac{4}{12} \frac{4}{11} \frac{x_i x_j}{G_y^2}. \tag{20}$$

with the other equations suitably modified because we want each seed to be chosen with probability 1/3 instead of having them depend on the historical frequency of upsets by each seed. □

Table 4: Number of randomly selected upsets correctly chosen.

Year	Expected number upsets selected correctly	Variance
2015	0.331	0.276
2014	0.337	0.280
2013	0.335	0.278
2012	0.343	0.284
2011	0.344	0.285
2010	0.339	0.281
2009	0.341	0.282
2008	0.330	0.276
2007	0.345	0.286
2006	0.340	0.281
2005	0.333	0.277
2004	0.35	0.289
2003	0.354	0.291

By using available historical data and uniform random selection and the result from the proposition, the expected number of upsets to be chosen correctly when two upsets are selected per year between 2003 and 2015 is 3.26 with a variance of 2.93. These values change to 4.42 and 3.66, respectively, when using weighted random selection. The year-by-year expected value and variance for weighted random selection can be found in Table 4.

Our technique selected 10 upsets over the 13 year period between 2003 and 2015. Therefore, our technique produced a number of correct selections that is 2.92 standard deviations above the expected number of correct predictions if weighted random selection were used. If uniform random selection is used, our technique produces a number of correct selections that is 3.94 standard deviations above the expected value of 3.26 correct selections. This is a key comparison to establish the performance of our technique, because what we observe is that our technique performs significantly better than if we used a form of random selection. This means that using our technique to select potential upsets is a more reliable way of identifying potential upsets than choosing match-ups randomly. However, the variance of our technique, 4.667, is significantly higher than that of random selection.

We also applied our technique to the 11 and 12 seeded games instead of the 13, 14, and 15 seeded games. For the 11 and 12 seeded games, we modified Algorithm 1 to select three teams instead of two. Table 5 provides the games selected. When run on the 11 and 12 seeds, there were cases where we had ties; the last step of our technique found multiple games that were selected with the same frequency. In order to resolve this, we count the number of correct selections by weighting each correct selection by the frequency with which it would be chosen if the ties were resolved by random selection. For example, if two

Table 5: Games selected.

Year	Game seed	Winning team	Losing team	Upset selected correctly
2015	11	UCLA	S Methodist	True
2015	11	Butler	Texas	False
2015	12	Utah	Ste F Austin	False
2014	12	Saint Louis	NC State	False
2014	11	N Carolina	Providence	False
2014	12	N Dakota St	Oklahoma	True
2014	11	Tennessee	U Mass	True
2013	11	Arizona	Belmont	False
2013	11	Memphis	St Marys	False
2013	12	Oregon	Oklahoma St	True
2012	12	New Mexico	Lg Beach St	False
2012	11	Cincinnati	Texas	False
2012	11	NC State	San Diego St	True
2011	12	Richmond	Vanderbilt	True
2011	12	Arizona	Memphis	False
2011	11	Cincinnati	Missouri	False
2010	11	Tennessee	San Diego St	False
2010	11	Old Dominion	Notre Dame	True
2010	12	Cornell	Temple	True
2009	12	W Kentucky	Illinois	True
2009	11	Marquette	Utah State	False
2009	11	UCLA	VCU	False
2008	12	Notre Dame	Geo Mason	False
2008	11	Oklahoma	St Josephs	False
2008	12	W Kentucky	Drake	True
2007	12	USC	Arkansas	False
2007	12	Butler	Old Dominion	False
2007	11	Louisville	Stanford	False
2007	11	Vanderbilt	Geo Wshgtn	False
2007	11	Winthrop	Notre Dame	True
2006	11	WI-Milwkee	Oklahoma	True
2006	11	Indiana	San Diego St	False
2006	12	Washington	Utah State	False
2005	12	GA Tech	Geo Wshgtn	False
2005	12	Villanova	New Mexico	False
2005	11	Texas Tech	UCLA	False
2004	12	Illinois	Murray St	False
2004	12	Syracuse	BYU	False
2004	11	Vanderbilt	W Michigan	False
2003	11	Maryland	NC-Wilmgton	False
2003	11	Oklahoma St	U Penn	False
2003	11	Missouri	S Illinois	False
2003	12	Butler	Miss State	True

Shaded rows indicate games selected correctly.

teams were tied for third-most-selected and one of them was a correct upset selection, those two teams would be combined into a score of 0.5. In the event of a three-way tie for second with one correct selection, the resulting score would be 1/3. However, the accuracy of the model for the 11 and 12 seeded games was comparable to that expected by weighted random selection. If weighted random selection were used, the expected number of correct selections would be 12.71 with a variance of 8.56, while our technique selected 10.67 upsets correctly. One hypothesis as to why this might be the case is that there is enough information that can be drawn from the covariates of upsets in the past that makes it possible to predict upsets in the future better than randomly selecting teams for 13–15 seeded games, while the 11 and 12 seeded games do not contain enough distinguishing information in the statistics available to us. Because the gap between the seeds in the 11 and 12 seeded games is not as large, the inherent randomness of the games might be overwhelming the information that the covariates provide about what causes an upset to occur for those seeds.

As another way to assess performance, we attempted to estimate the drop-off of our technique by selecting four teams each year instead of two, thereby providing a "next-best" scenario. Selecting four teams each year resulted in 13 of 52 correct selections (compared with 10 of 26 when selecting two each year). This is a significant drop in accuracy, reducing our success rate from 38% to 25% with a marginal drop of three correct selections of 26 additional selected teams, equating to an 11% marginal success rate.

In order to determine the computational effectiveness of our technique, each step of the technique was run 10 times on a computer with an Intel Xeon E3-1246 quad-core processor at 3.5 GHz with 16 GB of memory. The runtimes for each step are listed in Table 6.

As noted above, we made a compromise in the training procedure where we trained the extra-trees classifier on data from all years, rather than just the years prior to the target year. This could potentially cause data leakage, but was done due to the limited amount of available data. To address this, we also performed some experiments where we trained the classifier on a subset of the available data. First, we tried training the classifier using all the games except those from the target year. For example, if we were attempting to select games for 2013, the classifier was trained on games from 1998–2012 and 2014–2015. This led to us selecting 8 of 26 games correctly (compared to 10 using all the games as described above). We also tried training on all the games before the target year, which for a target year of 2013 means we trained the classifier

Table 6: Runtimes for each step of technique.

Step	Min (s)	Mean (s)	Max (s)
Compute match-up statistics	1	1	1
Identify useful match-up statistics	138	154	170
Find similar match-ups	2331	2425	2607
Combine models	25	27	30

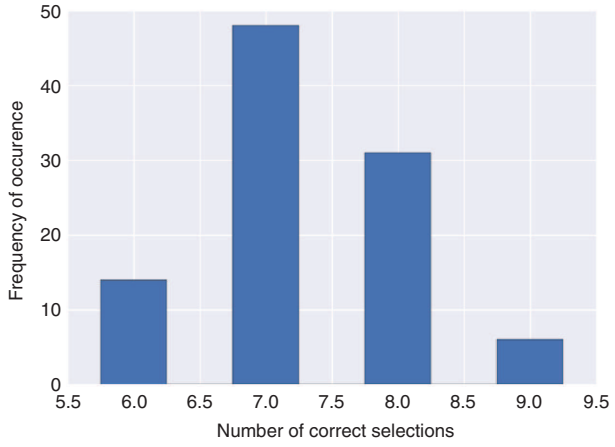

Figure 2: Frequency of number of correct selections by Lopez and Matthews (2015) model.

on games from 1998 to 2012. However, due to the small number of upsets, we only tested this method for years 2011–2015, because years prior to 2011 had very few upsets in the training set. In those five years, we selected 2 of 10 games correctly, compared to 4 using the full dataset for training.

In order to further evaluate the efficacy of our technique, we compare it to other methods found in the literature. We compared our results to the technique described in Lopez and Matthews (2015), which predicts the probability of each team winning against each other team. Therefore, to make a fair comparison, we used the Lopez and Matthews (2015) model to generate the probabilities of each low seed winning their first-round game and selected the two games where the lower seeded team had the highest probability of winning. The model was trained separately for each year using all games that occurred prior to that year's tournament. One note is that their model uses the home team stats and away team stats as inputs to their logistic regression. In the event of a neutral game, we randomly assigned one team as home and one team as away in addition to marking the game as neutral using the neutral indicator in their model. Due to this randomization, we ran their model 100 times and for each run counted how many upsets would have been selected correctly. Their model gives an average of 7.46 upsets correctly out of 26 selections, with a maximum of nine and a minimum of six. Figure 2 shows a histogram of the number of upsets correctly predicted by the Lopez and Matthews (2015) technique. Because our technique selected 10 upsets correctly out of 26 selections, our technique was better over the given time period. However, their approach has a variance of 0.669 in the number of correctly predicted upsets, which is significantly better than ours.

Bryan, Steinke, and Wilkins (2006) did an analysis on predicting round-of-64 upsets using a regression model where they define an *upset* as a game where the lower-seeded team wins and a *nonupset* as a game where the higher-seeded team wins. Their results were that 41.8% of the games they selected as upsets were actually upsets and 80.99% of the games they selected as nonupsets were actually nonupsets between 1994 and 2005 and 36.36% of the games they selected as upsets were actually upsets and 80.26% of the games they selected as nonupsets were actually nonupsets between 2000 and 2005. However, they declared their model as successfully predicting an upset if "it predicts a probability of upset greater than the historic proportion of games at the given seed difference that resulted in an upset". They also considered upsets as a 10, 11, 12, or 13-seed winning a game, whereas we consider an upset as a 13, 14, or 15-seed winning. Because we choose games specifically as upsets rather than those where the weaker team is more likely to win than the historical average and have a different definition for what constitutes an upset, the results are not directly comparable.

5 Conclusions

This paper presents a technique to select round-of-64 NCAA tournament upsets using game statistics. The technique identifies important statistics and uses those statistics to find match-ups similar to historical upsets. The performance of the technique was tested by using the technique to select potential upsets for the years 2003-2015. The technique was shown to significantly outperform random selection, both when the random selection was done with a uniform random distribution and when the distribution was weighted by the historical frequency of each seed winning a round-of-64 game.

There are several limitations to the technique used in this paper. First, the identification of important match-up statistics was done using all the years of data rather than identifying the important statistics using only the data that were available in each year. This was done due to the limited amount of data available; however, in the future there will be enough data for the choice of statistics to not vary from year to year. Furthermore, the technique as presented is limited to choosing a fixed number of potential upsets each year. This is a limitation because historical data shows that the number of upsets that occurs in a year can vary from zero to three, and this technique does not account for that. BOSS could be adapted to identify the most likely upset or the three most likely upsets by changing the

desired size of the control group; additional modifications would be needed for BOSS to decide whether no upsets should occur. Another limitation to the technique is its high variability; reducing this variability is left as a direction for future research. Finally, due to the limited number of years for which we have data, it is difficult to optimize the parameters of the technique. Because the number of upsets being selected is so small, even a small variation in the number of upsets selected correctly presents itself as a large change in the accuracy percentage. This means that some of the parameters chosen might not be optimal, but rather happened to perform slightly better on the small amount of data we had. As more years of data become available, the technique will be able to be tested more thoroughly.

Some potential areas for future work are experimentation with different match-up statistics, different methods for matchup-statistic combination selection (namely modifications to Algorithm 1), and the adaptation of the technique to be able to select a varying number of games as potential upsets. Additional match-up statistics to consider could include statistics such as the distance teams have to travel for each game or the injury status of each team at the time of the game. The primary improvement for this technique, however, will come with the availability of more data that will allow for further experimentation and testing. More data will allow the determination of accuracy to be more robust and less sensitive to each individual upset. Conference tournaments provide one potential source of additional data; including this is left as a direction for future work. Given the small size of the control group, another direction for future work is to develop efficient algorithms for solving BOSS directly without the use of MIP models. This paper enumerated through all combinations of possible control groups as a substitute to solving the MIP due to the small size of the data, but on larger problems alternative methods could be useful.

A Appendix

A.1 Definitions of statistics used

Statistics and their definitions from TeamRankings.com (TeamRankings 2015).

- Possession: One instance of a team controlling the ball until it scores, loses the ball, or commits a violation
- Steal: One instance of a defensive player forcing a turnover by acquiring or deflecting the ball from an offensive player
- Assist: One instance of a player passing the ball to a teammate in a way that directly leads to a field goal
- Effective possession ratio: (Possessions + offensive rebounds − turnovers)/Possessions
- Games played: Number of games a team has played in the current season before the tournament begins
- Extra scoring chances per game: Offensive rebounds + opponent turnovers − opponent offensive rebounds − turnovers
- Opponent floor percentage: Fraction of the opponent team's possessions that result in at least one point.
- Personal fouls per possession: Fouls/Possessions
- Opponent steals per defensive play: Opponent steals/Opponent defensive plays
- Assist/turnover ratio: Assists/Turnovers
- Personal fouls per defensive play: Personal fouls/Defensive plays
- Opponent steals per possession: Opponent steals/Opponent possessions
- Opponent average scoring margin: Average number of points between the opponent team and other teams they have played against (where positive is a victory and negative is a loss)
- Average scoring margin: Average number of points between the team and their opponents (where positive is a victory and negative is a loss)
- Opponent three point percentage: Three pointers made/Three pointers attempted
- Steals per defensive play: Steals/Defensive plays
- Steals per possession: Steals/Possessions
- Average 2nd half margin: Average difference between the number of points the team scores in the 2nd half and the number of points their opponents score in the 2nd half.

A.2 MIP formulation

Let the following terms be defined:

- T: Treatment group
- C: Control pool
- G: Control group
- $\mathbf{X}$: Set of covariates
- T_i: Set of unique values in T for covariate $i \in \mathbf{X}$
- C_i: Set of unique values in C for covariate $i \in \mathbf{X}$
- V_i: Set of unique values in $T \cup C$ for covariate $i \in \mathbf{X}$
- $V_{i,j}$: jth smallest value in V_i
- K_i: $K_i = K(T_i, G_i, V_i)$ for $i \in \mathbf{X}$: Kolmogorov-Smirnov statistic of the Treatment and Control groups for covariate $i \in \mathbf{X}$:

- R_i: $R_i = R(T_i, G)$ for $i \in \mathbf{X}$: Relative difference in sum between the Treatment and Control group for covariate $i \in \mathbf{X}$
- y_i: $y_i = \max(K_i, R_i)$: Larger of the Kolmogorov–Smirnov statistic and the relative difference between the treatment and control group for covariate $i \in \mathbf{X}$
- α_c: Binary variable which is 1 (0) if game $c \in C$ is (not) included in the control group, otherwise 0. The games with value 1 make up G.
- $x_{t,i}$: Value of covariate $i \in \mathbf{X}$ for $t \in T$
- $x_{c,i}$: Value of covariate $i \in \mathbf{X}$ for $c \in C$.
- $z_{i,j}$: $z_{i,j} = |\{x_{c,i} : x_{c,i} < V_{i,j},\ c \in C\}|$: Number of units in control pool with value less than jth smallest value of V_i for each covariate $i \in \mathbf{X}$
- $T_{i,j}$: $T_{i,j} = |\{x_{t,i} : x_{t,i} < V_{i,j},\ t \in T\}|$: Number of units in the treatment group with value than the jth smallest value of V_i for each covariate $i \in \mathbf{X}$
- β: $\beta = \sum_{c \in C} \alpha_c / |T|$: Constant that relates the size of the treatment group to the control group

Then, BOSS can be formulated as a mixed integer linear program (MILP):

$$\min \sum_{i \in \mathbf{X}} y_i \tag{21a}$$

such that

$$z_{i,1} = \sum_{\substack{c \in C \\ \text{such that } x_{c,i} = V_{i,1}}} \alpha_c \quad \forall i \in \mathbf{X} \tag{21b}$$

$$z_{i,j-1} + \sum_{\substack{c \in C \\ \text{such that } x_{c,i} = V_{i,j}}} \alpha_c = z_{i,j}\ \forall i \in \mathbf{X}, j \in \{2, 3, \ldots, |V_i|\} \tag{21c}$$

$$\frac{z_{i,j}}{\beta|T|} - \frac{T_{i,j}}{|T|} \le y_i \text{ for all } i \in X, j \in \{1, 2, \ldots, |V_i|\} \tag{21d}$$

$$\frac{T_{i,j}}{|T|} - \frac{z_{i,j}}{\beta|T|} \le y_i \text{ for all } i \in X, j \in \{1, 2, \ldots, |V_i|\} \tag{21e}$$

$$\frac{\sum_{c \in C} x_{c,i}\alpha_c - \beta \sum_{t \in T} x_{t,i}}{\beta \sum_{t \in T} x_{t,i}} \le y_i \quad \text{for all } i \in \mathbf{X} \tag{21f}$$

$$\frac{\beta \sum_{t \in T} x_{t,i} - \sum_{c \in C} x_{c,i}\alpha_c}{\beta \sum_{t \in T} x_{t,i}} \le y_i \quad \text{for all } i \in \mathbf{X} \tag{21g}$$

$$\sum_{c \in C} \alpha_c = \beta|T| \tag{21h}$$

$$\alpha_c \in \{0, 1\} \quad \text{for all } c \in C \tag{21i}$$

Equation (21a) minimizes the sum of the maximum of the Kolmogorov-Smirnov (KS) statistic and the relative difference for each covariate. Constraint (21c) sets the value of the empirical distribution at each point, and constraints (21d) and (21e) set the difference in empirical distributions at each point to be less than or equal to the KS statistic for that covariate. Constraints (21f) and (21g) set the relative difference constraints. Because the goal was to have BOSS select three teams, β was chosen depending on the size of the treatment group such that the control group would consist of three teams.

References

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. *Classification and Regression Trees*. Belmont, CA: Wadsworth.

Brenner, J. and P. Keating. 2015. "Upsets! We Got Upsets!" http://espn.go.com/espn/feature/story/_/id/12348280/upsets-got-upsets.

Bryan, K., M. Steinke, and N. Wilkins. 2006. "Upset Special: Are March Madness Upsets Predictable?" SSRN Scholarly Paper ID 899702, Social Science Research Network, Rochester, NY. http://papers.ssrn.com/abstract=899702.

Covers. 2015. "NCAA College Basktball Odds & Betting Lines – Spreads & Totals." http://www.covers.com/odds/basketball/college-basketball-odds.aspx.

Diamond, A. and J. S. Sekhon. 2013. "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies." *Review of Economics and Statistics* 95:932–945.

ESPN. 2015. "ESPN: The Worldwide Leader in Sports." http://espn.go.com/.

FiveThirtyEight. 2015. "2015 March Madness Predictions | FiveThirtyEight." http://fivethirtyeight.com/interactives/march-madness-predictions-2015/#mens.

Geurts, P., D. Ernst, and L. Wehenkel. 2006. "Extremely Randomized Trees." *Machine Learning* 63:3–42. URL http://link.springer.com/article/10.1007/s10994-006-6226-1.

Kaggle. 2015. "March Machine Learning Mania 2015 | Kaggle." https://www.kaggle.com/c/march-machine-learning-mania-2015.

Keating, P. 2013. "Explaining the New Giant Killers Formula - College Basketball". http://espn.go.com/mens-college-basketball/story/_/id/9022008.

Lopez, M. and G. Matthews. 2015. "Building an NCAA Men's Basketball Predictive Model and Quantifying Its Success." *Journal of Quantitative Analysis in Sports* 11:5–12.

Marino, J. 2015. "NCAA tournament gambling projection - Business Insider." http://www.businessinsider.com/ncaa-tournament-gambling-projection-2015-3.

Nikolaev, A. G., S. H. Jacobson, W. K. T. Cho, J. J. Sauppe, and E. C. Sewell. 2013. "Balance Optimization Subset Selection (BOSS): An Alternative Approach for Causal Inference with Observational Data." *Operations Research* 61:398–412. http://pubsonline.informs.org/doi/abs/10.1287/opre.1120.1118.

Oliver, D. 2013. "BPI - The College Basketball Power Index explained." http://espn.go.com/mens-college-basketball/story/_/id/7561413/bpi-college-basketball-power-index-explained.

Opitz, D. and R. Maclin. 1999. "Popular Ensemble Methods: An Empirical Study." *Journal of Artificial Intelligence Research* 11:169–198.

Pomeroy, K. 2012. "The kenpom.com Blog." http://kenpom.com/blog/index.php/weblog/entry/ratings_glossary.

Rosenbaum, P. R. and D. B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70:41–55.

Rosenbaum, P. R. and D. B. Rubin. 1985. "Constructing a Control Group Using Multivariate Matched Sampling Models that Incorporate the Propensity Score." *The American Statistician* 39:33–38.

Rubin, D. B. 1973. "Matching to Remove Bias in Observational Studies." *Biometrics* 29:159–183.

Rubin, D. B. 1980. "Bias Reduction Using Mahalanobis-Metric Matching." *Biometrics* 36:293–298.

Sagarin, J. 2015. "College Basketball Ratings Page." http://sagarin.com/sports/cbsend.htm.

Sauppe, J. J., S. H. Jacobson, and E. C. Sewell. 2014. "Complexity and Approximation Results for the Balance Optimization Subset Selection Model for Causal Inference in Observational Studies." *INFORMS Journal on Computing* 26:547–566. http://pubsonline.informs.org/doi/abs/10.1287/ijoc.2013.0583.

Silver, N. 2015. "How FiveThirtyEight's March Madness Bracket Works." http://fivethirtyeight.com/features/march-madness-predictions-2015-methodology/.

TeamRankings. 2015. "Sports Predictions, Rankings & Stats - TeamRankings." https://www.teamrankings.com/.

Volner, D. 2013. "Ride ESPN's BPI to Bracket-Making Success - ESPN Front Row." http://www.espnfrontrow.com/2014/03/bpi/.

Wartenberg, S. 2015. "How Many People will Fill Out March Madness Brackets? | The Columbus Dispatch." http://www.dispatch.com/content/blogs/the-bottom-line/2015/03/how-many-people-will-fill-out-march-madness-brackets.html.

Yahoo. 2014. "$1 Billion Offered for Perfect Tournament Bracket - Yahoo Sports." http://sports.yahoo.com/news/1-billion-offered-perfect-tournament-200547143--ncaab.html.

Yahoo. 2015. "Yahoo Sports – Sports News, Scores, Rumors, Fantasy Games, and More." http://sports.yahoo.com/.