Knapsacks and Basketball: A Cost-Driven Analysis of NCAA Men's Tournament Teams

Aaron Ludkowski

Advisor: Dr. Stephen M. Stigler

Approved _____

Date _____

Month-day, 2024

Abstract

The NCAA Division I men's basketball tournament, informally known as March Madness, is a yearly tournament that pits 68 of the top American college basketball teams against each other from late March to mid-April. As the tournament has risen in popularity, so, too, has interest in predicting how teams will fare. While the most well-known type of prediction contest has participants fill out a full "bracket", predicting the outcome of each of the 63 tournament matches, other sorts of contests have also been devised. Professor James Stapleton of Michigan State University devised a contest where participants would choose a short list of teams, scoring points equal to the sum of the number of games that teams on the short list won. This paper analyzes what teams tend to be good choices for this sort of contest, and methods for optimizing a team selection process. Discussion includes analysis on how Professor Stapleton's contest can be extended to the more common bracket contest.

Contents

1. Introduction

While the NCAA men's basketball tournament has taken place every year since 1939 (excluding 2020), the modern tournament as we know it – with 64 teams, all in a single-elimination bracket – first began in 1985. Given the premise that any one of the participating teams has a chance to win the championship, the tournament attracts considerable attention and excitement, especially from students and faculty at participating universities. As part of this excitement, it is common for fans to fill out full "brackets", predicting the outcome of the 63 matches that comprise the tournament. According to the American Gaming Association, 56.3 million Americans – almost 17% of the entire American population – were expected to participate in a bracket contest in 2023 [1].

However, the bracket contest is not the only predictive contest associated with March Madness. Professor James Stapleton of Michigan State University devised a contest for statisticians which is more like fantasy sports than the traditional bracket-predicting format [2]. In Dr. Stapleton's contest, the goal is to select a short list of teams, with participants scoring points equal to the sum of the number of games won by teams on the list, i.e.

$$\text{Score} = \text{SUM}(\#\text{ of games won}) \qquad (1)$$

Participants would have 100 points to allocate when choosing teams, with higher-seeded[1] teams (e.g. 1 seeds, 2 seeds) costing more points to select. For example, a player picking a 1 seed, two 2 seeds, and a 15 seed would cost that player 25 + (2)(19) + 1 = 64 points, out of the 100 point maximum. The table of costs [3] associated with each seed is presented below, with s representing seed number and C(s) representing the cost to select a team with that seed.

Table 1. The cost associated with each seed.

s	1	2	3	4	5	6	7	8
C(s)	25	19	13	12	11	10	8	5
s	9	10	11	12	13	14	15	16
C(s)	5	4	4	3	2	2	1	1

Notably, as there are four teams for each seed, picking every 1 seed would require the full allocation of 100 points. This dissuades participants from only picking the theoretical best teams, and pushes them to consider teams that have the potential to perform better than expected. Given this, optimizing the selection of teams is a twofold task:

[1] "Higher" and "lower" seeds are difficult terms to use, as a seed is *numerically lower* (e.g. 1, 2) for teams that are more *highly ranked*. For consistency, in this paper, "higher seeds" refers to *better* seeds, while "lower seeds" refers to *worse* seeds (e.g. 14, 15, 16).

1. Optimizing the selection of better-seeded teams, specifically picking teams that are unlikely to be upset early on
2. Optimizing the selection of worse-seeded teams, specifically picking teams that have the greatest potential to pull off an “upset” (an unexpected victory)

The goal of this paper is to analyze different strategies in optimizing this team selection process, and also examine if this analysis can lend any insight into the more common bracket prediction contest.

2. Methodology

2.1 Data Collection

The data used for this paper was primarily collected from barttorvik.com [4], one of the largest analytics websites for college basketball, run by basketball analyst Bart Torvik. While other sources like kenpom.com [5] and ESPN [6] were considered, barttorvik was chosen due to the following:

1. Bart Torvik provides all relevant data for free. For most serious data analytics purposes, getting advanced analytics from kenpom or ESPN requires a subscription or a fee.
2. Bart Torvik provides information on what stage of the tournament each team reached (e.g. Champion, Final Four, Elite Eight, etc.). This is highly relevant for this project, as this is the response variable we want to work with.
3. Finally, and *most crucially*, barttorvik allows us to pull data as it appeared on a certain day in the past (e.g. the day before the tournament starts). This is critical, as, otherwise, for past years, our data would include games played during the very tournament that we are trying to make predictions about!

Using barttorvik does come with one notable disadvantage: its datasets only go as far back as 2008 (compared to, for example, kenpom, which goes back to 2002). While more data is often desirable, older data may be less useful, given how the tournament has changed over the years. In addition, 2008 is still quite far back; when including all data from 2008 to 2024, there are over 1000 teams included in our dataset – a very respectable sample size.

All data was compiled into an Excel spreadsheet. A list of all variables included in the spreadsheet can be found in Appendix B, but a general overview is as follows:

- Year
- Tournament Seed
- Team Name
- **Number of Wins** (response)
- Conference
- Games Played / Game Record
- Offense Rating, Defense Rating, Overall Rating, "Wins Above Bubble" (from barttorvik)
- Raw Team Statistics (from barttorvik), with some example variables being:
    - 3-point percentage
    - Turnover percentage
    - Free throw percentage
- Opponent Statistics, including:
    - First-round opponent's Overall Rating
    - First-round opponent's "Wins Above Bubble"
    - Second-round opponent's Overall Rating
- "Blue Blood" status

While most of the variables above are likely self-explanatory, some require some further explanation:

- As a reminder, "Number of Wins" refers to the number of wins in the NCAA tournament (a team that loses in the first round would have 0 wins, while the champion would have 6 wins).
- Offense Rating, Defense Rating, and Overall Rating are all calculated on barttorvik.com using a proprietary formula that incorporates relevant raw statistics, as well as relevant context for each game a team plays (e.g. doing well against a highly-ranked team is more impressive than doing well against a lower-ranked team).
- "Wins Above Bubble" is a relatively new statistic [7] that evaluates how many more games a team won during their season compared to how a "bubble" team would theoretically do with the same schedule.
    - A "bubble" team is a team that is only barely good enough to make the tournament as an "at-large" team[2]. Bubble teams that make the tournament are typically teams seeded 10 or 11.
- "Blue Blood" status may be the most subjective variable included in this list. A "Blue Blood" is informally defined as a team with a long, significant history of success in college basketball, with ample resources to consistently attract some of the best players. In 2018, sportswriter Jon Bois noted that the top 4% of NCAA teams achieved 58% of all Final Four berths [8]; to try to be objective about this variable, I labeled the twelve members of this 4% as "Blue Bloods" for the purposes of this analysis.

[2] There are two ways to qualify for the NCAA tournament: automatically qualify as the champion of one of the ~32 Division I conferences, or qualify as one of the remaining "at-large" teams, chosen by a committee of basketball experts.

2.2 Data Filtering

Beyond the variables listed above, some other potential explanatory variables were considered, but were ultimately discounted:

- "Power conference" status, similar to "Blue Blood" status, is related to the concept of some schools having more resources and basketball prestige than others. Power conferences are conferences that tend to have many teams qualify as "at-large" teams, and are traditionally defined as one of the following conferences:
    - Atlantic Coast Conference
    - Big East Conference
    - Big Ten Conference
    - Big Twelve Conference
    - Pacific 12 Conference
    - Southeastern Conference
- A binary variable indicating whether a team has won the tournament within the last ten years was considered. For example, if North Carolina won the tournament in 2009, they would be labelled a "Recent Champ" for all years up to, and including, 2019.
    - This variable and the "Power conference" variable were removed for being similar not only to each other, but, more significantly, to the "Blue Blood" variable. With the "Blue Blood" predictor being the strongest, the above two variables were removed.
- Some raw statistical attributes obtained from barttorvik.com were deemed insignificant and eliminated during the variable selection process. These include:
    - Turnover percentage
    - Defensive[3] turnover percentage
    - Defensive rebound percentage
    - Defensive free throw percentage
    - Defensive three point percentage
    - Three point rate[4]
    - Defensive three point rate
- The 64-team bracket is divided up into 4, smaller sub-brackets of 16 teams (with each of these four smaller brackets having one team of each seed). Each sub-bracket is held in a different region of the United States – the South, the West, the East, or the Midwest.[5]
    - A categorical variable related to whether a certain region had an effect on teams did not end up being significant. Similarly, whether a team was playing in their "home" region (as defined by the US Census Bureau [9]) or not did not seem to affect their predicted wins.

[3] In this list, "defensive" refers to the average of the relevant statistic for teams *this team was defending against*. For example, if a team conceded many three point shots, their defensive three point percentage would be a higher (worse) value.

[4] Three point *percentage* was included as a predictor. Three point *rate* is a slightly different statistic that is typically highly correlated with three point percentage, and was removed as a result.

[5] Technically, in 2011, the "South" region was labelled as "Southeast", and the "Midwest" region was labelled as "Southwest", due to games in those regions being played in New Orleans and San Antonio, respectively.

- Time did not seem to be an important predictor. Overall, from year-to-year, there did not seem to be wide shifts in how similar teams performed.
- A two-step modeling approach was attempted, where an initial model would be trained, before a second model would be constructed using all of the same predictors as before, as well as a new predictor consisting of the *predicted number of wins* for a team's first round opponent.
    - This predictor overwhelmed the second model so much that, almost always, only high seeds (teams seeded 1-4) were included in the final team lists, which was not the case when this predictor was not included. This predictor was scrapped due to its overwhelming effect.

2.3 Model Building

Three main predictive models were considered for this project, with the idea of establishing a range of complexity:

Table 2. Three different models were considered, each with advantages and disadvantages.

Model	Relative Complexity	Variable Importance	Relative Running Time
Ordinal GLM	Low	Easy to obtain	Medium
Random Forest	Medium	Possible to obtain	Low
XGBoost	High	Difficult/impossible to obtain	Long

Ordinal GLM was included primarily as a comparison tool to the other models – in theory, given the complexity of the dataset, more advanced machine learning methods like random forest and XGBoost should have an advantage over the relatively humbler GLM. One advantage, however, of the ordinal GLM model was that it was absolutely guaranteed to predict values in the range of the response (i.e. between 0 and 6), as predictions were calculated by taking the expected number of wins for each team (i.e. taking the sum of 0 times the predicted probability a team would get 0 wins, 1 times the predicted probability a team would get 1 win, and so on).

This was possible due to treating the response as *ordinal* (categorical with a specified ordering) – this was reasonable, due to the response being an integer between 0 and 6. Teams are already oftentimes described by the number of games they win in a tournament – for example, great importance is placed on teams making the "Sweet Sixteen", "Elite Eight", and so on. Finally, this model is especially helpful in identifying important variables – simply performing a t-test for each variable in the model would be a quick and easy way to identify significant variables.

Random forest was considered as a somewhat more complex "intermediate" model, which would be powerful enough to do well, but still be understandable and relatively usable by a non-statistician (with proper guidance). The random forest algorithm works by aggregating the results of several decision trees, each of which predicts how many wins a team will get based off of the variables examined in the tree (for this project, the default setting of aggregating 500 trees was used). This is good for our purposes, since a random forest can work well when working with many explanatory variables, as it can account for more complex relationships (e.g. while low-seeded teams are rarely predicted to do well, a random forest might be able to identify specific niche situations where they could win).

Finally, XGBoost happens to be the most complex model of the three, but also the most powerful. The pertinent aspects of the model for this paper are that it is known to be one of the best algorithms for tabular data (i.e. data with variables in rows and columns, as opposed to, e.g., text or visual data), and that it runs *iteratively* – an initial model is modified a set number of times, each time reducing its *training error* (its error on a pre-defined training set of data). It can also be somewhat awkward to implement, requiring data to be in a specific format before use.

However, its power is well-known, and was included as an "ideal" model to compare to the two simpler models.

When making predictions for a given year, the GLM and random forest models were trained on all data from years prior (e.g. if we were predicting results for 2016, we would train the GLM and random forest models on all data from 2008 to 2015). Meanwhile, the XGBoost model was trained on all data from over 3 tournaments prior, then validated on a *test set* consisting of data from the past 3 tournaments (e.g. if we were predicting results for 2016, we would train the XGBoost model on data from 2008 to 2012, and use the iteration of the model that had the lowest *test error* for years 2013 to 2015).

2.4 Team Selection

Using one of the three models above, a predicted number of wins was generated for each team in the dataset. As half the teams in the dataset achieved 0 wins (given the single-elimination nature of the tournament), most predictions tended be to low; it was extraordinarily rare for teams to be predicted to win 4 or more games. This was expected, and even hoped for – even the best teams can lose early in the tournament, so no team is a certain bet to win the tournament, or even make the Final Four.

With this list of predictions, we can use the **knapsack optimization** scheme to select an ideal list of teams. The knapsack scheme is as follows [11]:

1. Assign a **weight** and a **cost** to each team
    a. For our purposes, the **cost** is C(*s*), as seen in Table 1, while the **weight** is the predicted number of wins, as calculated by our model
2. Find the greatest sum of weights possible for each possible cost level
    a. E.g., for cost level 1, we find the largest predicted value for teams seeded 15-16; for cost level 2, we compare the largest predicted value for teams seeded 13-14, and compare with the largest sum of predicted values for two teams seeded 15-16
3. Iterate through step 2 until we find the greatest sum of weights possible for the maximum cost level – in our case, 100
4. Return this list of teams whose sum of weights is greatest for cost level 100

The adagio package in R was used to implement steps 2 through 4. A strong advantage of using the adagio package is its quick running time; depending on the implementation of the above approach, the running time of the knapsack algorithm can be highly variable. The adagio package implements the algorithm in a quick enough way to be usable for our purposes [10].

While the main focus of this project was using the knapsack algorithm to select well-performing team lists given a *predicted* number of wins for each team, we can use the *actual* number of wins to produce a theoretical best team list for each year using the knapsack algorithm. In Table 3 below, the theoretical best team lists for 2022, 2023, and 2024 are listed. The full list of theoretical best team lists for all years from 2009 to 2024 can be found in Appendix A.

Table 3. The theoretical best team lists for 2022, 2023, and 2024.

Seed	2022: 35 pts	Wins	Seed	2023: 38 pts	Wins	Seed	2024: 30 pts	Wins
1	Kansas	6	3	Kansas State	3	1	UConn	6
2	Duke	4	3	Gonzaga	3	3	Illinois	3
4	Arkansas	3	4	UConn	6	4	Duke	3
5	Houston	3	5	San Diego State	5	4	Alabama	4
8	North Carolina	5	5	Miami (FL)	4	6	Clemson	3
9	Memphis	1	6	Creighton	3	8	Utah State	1
10	Miami (FL)	3	7	Michigan State	2	10	Colorado	1
11	Notre Dame	1	8	Arkansas	2	11	Duquesne	1
11	Michigan	2	9	Florida Atlantic	4	11	NC State	4
11	Iowa State	2	10	Penn State	1	11	Oregon	1
12	New Mexico State	1	11	Pittsburgh	1	12	James Madison	1
12	Richmond	1	13	Furman	1	13	Yale	1
15	Saint Peter’s	3	15	Princeton	2	14	Oakland	1
			16	Fairleigh Dickinson	1			

3. Results

3.1 Initial Results

After running all three models, a predicted "best" team list was generated for each model for each year. In addition, for comparison purposes, two simpler models were created for each year: a model that selected only the four 1 seeds, and a model that selected every team seeded 9 through 16 (which, imperfectly, only uses 88 of the 100 points available). Below is a table of the scores of each model for each year from 2013 to 2024, listed alongside the theoretical best score for that year.

Table 4. The listed scores for each model, with each year's best in boldface.

Year	GLM Score	Rand. For. Score	XGBoost Score	All 1s	All Underdogs	Theor. Best Score
2013	**26**	18	16	11	16	35
2014	**25**	16	10	10	12	35
2015	**21**	13	14	16	6	28
2016	**17**	14	**17**	14	**17**	31
2017	16	**19**	14	15	8	30
2018	13	15	9	11	**17**	35
2019	13	12	**15**	14	13	29
2021	18	**21**	16	15	17	34
2022	24	20	**25**	11	16	35
2023	18	21	**23**	5	11	38
2024	15	11	**16**	15	14	30

Clearly, no one model does the best for all, or even a majority of, the years. XGBoost seems to do well in more recent years (for which more data is available to be used for training), while GLM seems to do well further back in time (for which less data is available). Interestingly, the simple "All 1s" model never is the best choice – we can always do better. (In fact, for all years except 2019, the GLM model at worst ties it). Relatedly, when looking at the five models, the GLM *never* has the worst score for a year. It is a fairly consistent, reliable model, compared to the other models' more variable performances.

However, we fortunately find ourselves in somewhat of a unique situation. While it might normally be difficult to combine the results of different models, here we have a (relatively) easy way to do that: re-utilize the knapsack algorithm, but allot a subsection of the total 100 points to each of the three main models we've used thus far. For example, we could allot 50 points to our GLM model, 25 points to our random forest model, and the remaining 25 points to our XGBoost model, and combine the team lists chosen by each model into one full team list. This allows us to rectify some of the issues identified in this section with a combined model.

3.2 Combined Model

The combined model was designed to be a combination of the three **component models** - ordinal GLM, random forest, and XGBoost. Constructing the ideal combined model was a multi-stage process. The process was very similar to traditional parameter tuning – a list of parameters that could vary was selected, and then a list of different values of those parameters was used. In particular, these were the parameters that were allowed to vary:

4. **GLM Cap** – The maximum sum of costs the GLM model could use when selecting teams
    a. The values tested for this variable were every other integer from 1 to 97, inclusive. (I.e. 1, 3, 5, 7, … 93, 95, 97).
5. **Random Forest Cap** – The maximum sum of costs the random forest model could use when selecting teams
    a. The values tested for this variable were, similarly to the GLM Cap, every other integer from 1 to 97, inclusive.
6. **XGBoost Cap** – The maximum sum of costs the XGBoost model could use when selecting teams
    a. This was always set to be equal to 100 minus the sum of the GLM Cap and the Random Forest Cap, so that our combined model would be able to use the full 100 points.
7. **Order** – Once a team has been selected by a model, it could not be selected by a later model. The order variable determined which order the models chose teams (e.g. GLM first, random forest second, XGBoost third).
    a. All potential orderings were tested.
8. **Focus –** Some models might be better than other models at, for example, selecting teams that can pull off an upset. Using this idea, the three models were set to "focus" on a certain subset of seeds:
    a. A "high" focus meant that the model would only select teams seeded 1-4.
    b. A "mid" focus meant that the model would only select teams seeded 5-9.
    c. A "low" focus meant that the model would only select teams seeded 10-16.
    d. No two models were assigned the same focus; thus, each focus was assigned to exactly one of the models. All focus combinations meeting this criterion were tested.

Testing every combination of the above parameters would result in 86,436 possible models. As running a model and obtaining its score takes non-negligible time (on average 2.5 seconds), getting the scores for a given year of all possible models would take approximately 60 hours. Given that it would be desirable to have results from multiple years, computation time quickly balloons. So, the most logical next step is to cut down on the number of parameter combinations we test. The following restrictions were implemented, with numbers having been slowly adjusted until the list of parameter combinations was reasonably small:

- Whenever focus was set to "high" for a model, the cap for that model must be greater than 40.
- Whenever focus was set to "mid" for a model, the cap for that model must be greater than 18.
- Whenever focus was set to "low" for a model, the cap for that model must be less than 23.
- The cap for the model that performed the worst in general – random forest - must be below 38.
- The cap for the model that consistently performed the best in general – GLM – must be above 18.

These constraints were chosen to narrow down our list of 86,436 parameter combinations to a more manageable number. After the above restrictions were implemented, we were left with 3,072 valid parameter combinations left. Getting all scores for a certain year would now take just over 2 hours – in the end, scores were obtained for all years between 2018 and 2024 (inclusive), so model tuning took around 12.5 hours in total.

The process for constructing a combination model, given a set of parameters, was as follows:

1. First, create a temporary dataset consisting of all training data and all teams in the test data that the first component model can actually pick.
    a. A team is selectable by a model if it both:
        i. Has a seed corresponding to the model's focus, and
        ii. Does not cost more points than the model's points cap
2. Run the first component model and save its picks. Its picks are stored and are immediately used for two purposes:
    a. If the component model did not spend a number of points equal to its points cap, the leftover points are transferred to the points cap of the *next* component model. (e.g. if a component model had a cap of 40 points, and spent 39, the next component model has an extra point to spend).
    b. We remove teams selected by the first model from the test sets the remaining models will use. This ensures that a team cannot be picked twice. (This is why the orders in which the component models run matters).
3. Repeat Steps 1 and 2 for the second component model.
4. Repeat Steps 1 and 2 for the third component model.
5. Combine the three lists of team picks we have (one from each component model) into one full list of teams. Return this full list of teams.

This algorithm was performed for each of the 3,072 parameter combinations, for each year from 2018 to 2024. The results were stored in a spreadsheet, where each row matched one of the 3,072 parameter combinations.

From here, this dataset was filtered down to only include models that performed better every year than the worst-performing model for each year. As an example, referring back to Figure 3.1, any model that scored 11 points or less in 2024 was removed from the dataset, as that was the lowest score of the three main models in 2024. This whittled down the dataset from 3,072 models to 186 remaining models. At this point, it may be helpful to analyze some common trends among these 186 reasonably successful models.

- In all but 7 models, XGBFocus was "high". In all but 26 models, GLMFocus was "mid". In all but 26 models, RFFocus was "low". This seems to indicate that, among these three component models:
    - XGBoost is best at picking 1-4 seeds that will avoid early upsets and last long in the tournament.
    - Ordinal GLM is best at picking 5-9 seeds.
    - Random Forest is best at picking 10-16 seeds that can pull off an upset.
- The highest score of any of the 186 models for 2024 was 18 points. It was the only year where none of the 186 models managed at least 20 points; even 2019, which had a lower theoretical highest score than 2024, saw models get up to 22 points.
- None of the 186 models were *always* better than the median score of the three component models; however, four models did *at least* as well as the median in 5 out of 6 years, with the one below-median year being 2022 (with each model scoring 21 points, compared to the median score of the component models, 24).
    - These four models were incredibly similar: each had a CapRF of 19 points, a CapGLM of either 27 or 29 points, a CapXGB of either 54 or 52 points, and an order of XGBoost -> GLM -> Random Forest or Random Forest -> XGBoost -> GLM. Each model also used its most common focus, as per above.

Out of the above 186 models, the following two "ideal" combined models will be compared with the component models from before:

1. Combined Model 1: The model that has the highest combined score of the 4 models discussed above (that did at least as good as the median basic model score in 5 out of 6 years), and
2. Combined Model 2: The model that has the highest **total proportional score**; the total proportional score is calculated by summing the model's score from each year over the theoretical best score for each year (i.e. the model's score for 2018 over 35, the model's score for 2019 over 29, etc.)

Table 5. A comparison of the above two models with the three component models.

Year	GLM Score	Rand. For. Score	XGBoost Score	Comb. M1	Comb. M2
2013	**26**	18	16	17	16
2014	**25**	16	10	15	22
2015	**21**	13	14	17	20
2016	**17**	14	**17**	13	12
2017	16	**19**	14	16	11
2018	13	15	9	15	**17**
2019	13	12	15	19	**22**
2021	18	**21**	16	**21**	18
2022	24	20	**25**	21	23
2023	18	21	**23**	22	**23**
2024	15	11	**16**	**16**	15

Some closing comments on the above:

- Besides 2016 and 2017, Combined Model 2 does impressively well, achieving the highest score in three years and being within 3 points of the top score in 5 further years.
- Both combined models tend to be more consistent than the simpler models; although both models are mostly based on XGBoost (with CapXGB being above 50 for both models), thanks to the incorporation of results from GLM and random forest, both models perform reasonably well for XGBoost's worst years (like 2014 and 2018).

3.3 Variable Importance

Earlier on, in Figure 2.1, the ease of obtaining variable importance information was identified as a strength of GLM and random forest models. As such, for this section, we will be focusing on what variables are important for these two models.

For ordinal GLM, t-tests are very easily performed for each variable. Below are the coefficient values and p-values for the ordinal GLM models predicting on years 2013 (the model's best year), 2022 (the model's best post-pandemic year), and 2024 (the most recent year).

Table 6. The coefficients and p-values for each predictor in selected years' GLM models. Significant p-values and coefficients are highlighted in yellow.

Variable	2013 Coef.	2013 p-val.	2022 Coef.	2022 p-val.	2024 Coef.	2024 p-val.
Seed	0.0104	0.8989	0.0846	0.1059	0.0265	0.5809
Adj. Offense	0.0848	0.2261	0.1063	0.0176	0.1286	0.0011
Adj. Defense	-0.1699	0.0167	-0.1436	0.0013	-0.1555	0.0002
Avg. Rating	-1.2096	0.7264	0.2922	0.8947	-1.3741	0.4971
Field Goal%	0.2722	0.5858	0.3149	0.3056	0.3018	0.2823
Df. Fd. Goal%	0.0962	0.4660	0.0360	0.6358	0.0612	0.8182
ORB	0.0553	0.1511	0.0258	0.2139	0.0204	0.2858
Free Throws	-0.0087	0.7438	0.0279	0.0635	-0.0230	0.0940
Two Point %	-0.3383	0.3189	-0.2242	0.2617	-0.2085	0.2536
Df. 2P%	-0.0258	0.7918	0.0256	0.6443	0.0416	0.4132
3P%	-0.1140	0.6427	-0.1981	0.2147	-0.1980	0.1802
WAB	0.1320	0.1827	0.1042	0.0583	0.0793	0.1220
Opp1 Avg.	-3.1017	0.1202	-3.2354	0.0071	-2.1113	0.0601
Opp2 Avg.	-2.0047	0.5397	-3.7698	0.0776	-3.2438	0.0999
Opp1 WAB	-0.0189	0.8260	-0.0353	0.4825	-0.0303	0.5184
"Blue Blood"	0.8905	0.0137	0.7229	0.0005	0.7527	0.0001

While some results above may look strange (most prominently: a team's defense seeming to be a significant *hindrance* to their chances at winning games in the tournament, and a team's Seed not seeming to matter much), it's important to remember that, in these models, coefficients assume that *all other variables are held constant*. For two teams with near-identical statistics, it's somewhat encouraging that their seed does not matter too much; a good enough team, it seems, can overcome a bad seed (which makes sense, considering the number of upsets in the tournament each year). Similarly, a lower defensive rating may correspond to faster play, which may give an edge to teams with otherwise similar statistics.

Finally, one of the most interesting observations is the effect of being a "Blue Blood". Being one of the few teams to have storied, consistent success in the tournament seems to translate to further success, which makes some degree of success – these programs, in theory, would be able to attract the best players and coaches, have the best facilities, and have the most resources to spend on a

basketball program. When comparing two similar teams, it seems that those resources may make a significant difference.

To close out this section, it may also be helpful to look at importance plots from the random forest models. The plots below will focus on 2017 and 2021, the random forest model's best years.

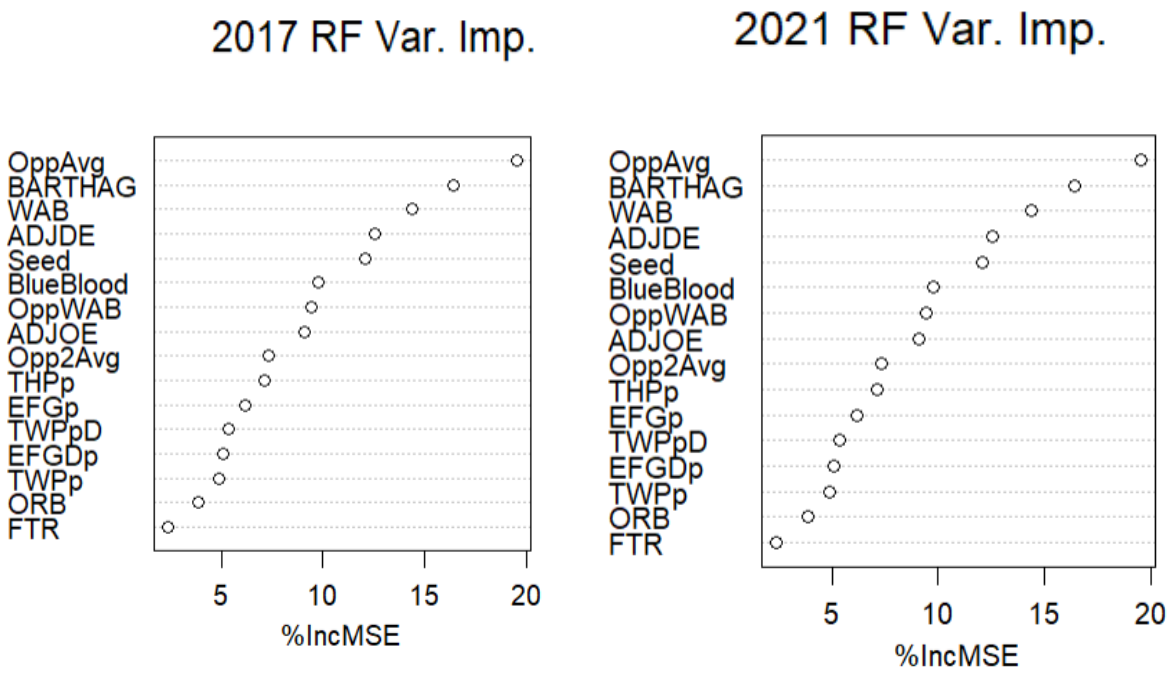

Figure 1: Above are two plots of variable importance for the random forest models predicting results in 2017 and 2021, respectively. The variable importance statistic used here is the percentage increase in MSE caused by removing a specific variable.

The plots above indicate a somewhat different understanding of variable importance, but the interpretations are not contradictory. In our analysis of GLM coefficients, each variable is evaluated on how important it is when holding *all other variables equal*. In this sense, we see that "Blue Blood" status and defensive rating are important *when comparing two otherwise very similar teams*. In the above plots, due to the heavily interconnected nature of random forest (as random forests are comprised of decision trees, where each variable is evaluated with respect to other variables' values), it makes sense that we seem to be looking at plots of *overall* variable importance. The above plots would most likely be more helpful to someone broadly evaluating what teams among the field would be good choices.

In this sense, one of the most interesting aspects of the plots above is how a team's *first round opponent's* average rating is more important than the rating of the team itself. This is a somewhat surprising result, but it does make some sense: with how many upsets happen every year, even to very strong, highly-rated teams, it makes sense that avoiding potential upsets would be important for highly-ranked teams. We also see the value in the index statistics (BARTHAG, Wins above Bubble, Adjusted Defense, and Adjusted Offense) as compared to the menial raw statistics like free throw rate and offensive rebound rate, which don't seem to affect the models too much.

4. Discussion and Conclusion

One of the most surprising aspects of this study was the success of the relatively simple ordinal GLM model, compared to the other two, more complex models. The idea of simpler models being fairly successful in predicting March Madness results is not entirely new; Chartier et al. found that simple linear models could produce brackets ranking near the top percentile on ESPN [12]. However, the combined model (consisting of ordinal GLM, random forest, and XGBoost) was able to introduce complexity without sacrificing consistency, indicating that more complex methods may be able to work when tempered by simpler methods.

In terms of general advice for picking teams in a bracket contest, results from the random forest models indicate that it is most important to focus on the rating of a team's *first round opponent*. For example, if a certain team faces a relatively strong low seed in the first round, it may be unwise to pick that team to go too far into the tournament. For teams that are somewhat evenly matched, based off the GLM results, it may be worth considering if one of the teams is a "Blue Blood"; this status may give a team an edge (e.g. in an 8 vs. 9 seed first-round matchup).

However, one of the strongest take-home messages from these results is that a team's fate in the tournament is never certain. Models that perform well in some years can perform poorly in others. Teams that seem certain to lose just need thirty minutes of game time to make history. As of 2024, nobody has yet perfectly predicted the March Madness bracket. So, perhaps the best advice of all would be to enjoy the tournament, whether you make a bracket yourself, participate in a specialized contest like Professor Stapleton's, or just watch from home. The NCAA tournament is, after all, called "March Madness" for a reason.

Appendix A: Theoretical Best Team Lists

Below are full tables of the theoretical best team lists for all years from 2009 to 2024. Note that some years may have more than one theoretical best team list, as teams of the same seed (or cost) that perform equally well can be interchanged. The team list returned when using the adagio package in R is what is displayed below for each year.

Seed	2009: 27 pts	Wins	Seed	2010: 33 pts	Wins	Seed	2011: 36 pts	Wins
1	North Carolina	6	1	Duke	6	3	BYU	2
2	Michigan State	5	3	Baylor	3	3	UConn	6
3	Villanova	4	5	Michigan State	4	4	Kentucky	4
3	Missouri	3	5	Butler	5	4	Wisconsin	2
8	Oklahoma State	1	6	Tennessee	3	5	Arizona	3
10	USC	1	9	Northern Iowa	2	8	George Mason	1
10	Michigan	1	10	Missouri	1	8	Butler	5
10	Maryland	1	10	Georgia Tech	1	8	Michigan	1
11	Dayton	1	10	Saint Mary's	2	9	Illinois	1
12	Wisconsin	1	11	Washington	2	10	Florida State	2
12	Arizona	2	11	Old Dominion	1	11	Marquette	2
12	Western Kentucky	1	12	Cornell	2	11	Gonzaga	1
			13	Murray State	1	11	VCU	4
						12	Richmond	2

Seed	2012: 32 pts	Wins	Seed	2013: 35 pts	Wins	Seed	2014: 35 pts	Wins
1	Kentucky	6	1	Louisville	6	2	Wisconsin	4
2	Kansas	5	3	Florida	3	3	Iowa State	2
3	Baylor	3	3	Marquette	3	4	Michigan State	3
4	Louisville	4	4	Michigan	5	6	Baylor	2
7	Florida	3	4	Syracuse	4	7	UConn	6
10	Purdue	1	9	Wichita State	4	8	Memphis	1
10	Xavier	2	10	Iowa State	1	8	Gonzaga	1
11	NC State	2	11	Minnesota	1	8	Kentucky	5
11	Colorado	1	12	Ole Miss	1	10	Stanford	2
12	South Florida	1	12	California	1	11	Dayton	3
13	Ohio	2	12	Oregon	2	11	Tennessee	2
15	Lehigh	1	13	La Salle	2	12	Stephen F. Austin	1
15	Norfolk State	1	15	Florida Gulf Coast	2	12	Harvard	1
						12	North Dakota St.	1
						14	Mercer	1

Seed	2015: 28 pts	Wins	Seed	2016: 31 pts	Wins	Seed	2017: 30 pts	Wins
1	Duke	6	2	Villanova	6	1	North Carolina	6
3	Notre Dame	3	2	Oklahoma	4	3	Oregon	4
4	Louisville	3	6	Notre Dame	3	4	Florida	3
6	Xavier	2	7	Wisconsin	2	7	South Carolina	4
7	Wichita State	2	8	Saint Joseph's	1	7	Michigan	2
7	Michigan State	4	9	UConn	1	8	Wisconsin	2
8	Cincinnati	1	9	Providence	1	8	Northwestern	1
8	NC State	2	10	VCU	1	8	Arkansas	1
10	Ohio State	1	10	Syracuse	4	10	Wichita State	1
11	Dayton	1	11	Wichita State	1	11	USC	1
11	UCLA	2	11	Northern Iowa	1	11	Xavier	3
14	Georgia State	1	11	Gonzaga	2	11	Rhode Island	1
			12	Yale	1	12	Middle Tennessee	1
			12	Little Rock	1			
			13	Hawaii	1			
			15	Middle Tennessee	1			

Seed	2018: 35 pts	Wins	Seed	2019: 29 pts	Wins	Seed	2021: 34 pts	Wins
1	Villanova	6	1	Virginia	6	1	Baylor	6
3	Michigan	5	2	Michigan State	4	2	Houston	4
3	Texas Tech	3	3	Texas Tech	5	3	Arkansas	3
7	Nevada	2	3	Purdue	3	6	USC	3
7	Texas A&M	2	5	Auburn	4	7	Oregon	2
8	Seton Hall	1	10	Minnesota	1	8	Loyola-Chicago	2
9	Kansas State	1	10	Florida	1	10	Maryland	1
9	Florida State	3	12	Liberty	1	11	UCLA	4
9	Alabama	1	12	Murray State	1	11	Syracuse	2
11	Loyola-Chicago	4	12	Oregon	2	12	Oregon State	3
11	Syracuse	2	13	UC Irvine	1	13	Ohio	1
13	Buffalo	1				13	North Texas	1
13	Marshall	1				15	Oral Roberts	2
16	UMBC	1						

Seed	2022: 35 pts	Wins	Seed	2023: 38 pts	Wins	Seed	2024: 30 pts	Wins
1	Kansas	6	3	Kansas State	3	1	UConn	6
2	Duke	4	3	Gonzaga	3	3	Illinois	3
4	Arkansas	3	4	UConn	6	4	Duke	3
5	Houston	3	5	San Diego State	5	4	Alabama	4
8	North Carolina	5	5	Miami (FL)	4	6	Clemson	3
9	Memphis	1	6	Creighton	3	8	Utah State	1
10	Miami (FL)	3	7	Michigan State	2	10	Colorado	1
11	Notre Dame	1	8	Arkansas	2	11	Duquesne	1
11	Michigan	2	9	Florida Atlantic	4	11	NC State	4
11	Iowa State	2	10	Penn State	1	11	Oregon	1
12	New Mexico State	1	11	Pittsburgh	1	12	James Madison	1
12	Richmond	1	13	Furman	1	13	Yale	1
15	Saint Peter's	3	15	Princeton	2	14	Oakland	1
			16	Fairleigh Dickinson	1			

Some thoughts:

- The average length of a list (measured by the number of teams it contains) is 13.2667. The shortest list was 2019's, with 11 teams. The longest list was 2016's, with 16 teams.
- 1 seeds only appear on this list if they are champions for that year. While 1 seeds have finished as runner-ups (e.g. Purdue in 2024, Gonzaga in 2021, North Carolina in 2016, etc.), a 1 seed must finish as champion to be considered an "optimal" pick. 2 seeds need 4 wins or more to appear on the above list, 3-5 seeds need 3 wins or more, and 6-7 seeds need 2. 8 seeds or worse need just 1 win to be considered an "optimal" pick, although not all such teams appear on the above lists (due to the 100-point cap restricting how many teams can be picked).
    - These ideas are formalized more in Appendix C.
- 11 seeds appear on all but one list (2019). The 6 vs. 11 first-round matchup is the most common upset, with the 11 seeds winning more than the 6 seeds over the past 10 tournaments: 22 teams seeded 11 have won their first-round matchup since 2014 (inclusive), while only 18 teams seeded 6 have won.

Appendix B: Full Variable List

A full, detailed list of variables included in the dataset is as follows:

- **Wins** – The response variable. This was treated as an ordinal variable for the ordinal GLM model, whereas it was treated as a numerical variable in the random forest and XGBoost models. It takes values from 0 to 6: 0 for teams that lost their first-round game, and 6 for the tournament's champion.
    - Thus, half the teams have 0 wins, a quarter have 1 win, an eighth have 2 wins, and so on.
    - First Four games are not counted for this. A team who plays a First Four game, and goes on to win the entire tournament, would have 6 wins, not 7. (First Four losers are also not included in the dataset – only 64 teams are included per year).
- **Year** – Ranges from 2008 to 2024 (including all years between except for 2020, which had no tournament)
    - This variable was originally considered as a predictor, but instead was used simply for separating the test set (the year of interest we are trying to predict for) from the training set (data from all previous years)
- **Seed** – Ranges from 1 to 16, with 1 being the best and 16 being the worst. For each year, there are four 1 seeds, four 2 seeds, four 3 seeds, etc.
- **Team** – The name of the team, e.g. "Iowa State", "Northwestern", "USC", etc. This is not a predictor; this variable is only used to construct the "Blue Blood" variable.
    - Some teams are named differently from year to year (e.g. "UConn" vs "Connecticut", "NC State" vs "North Carolina State"), but as this variable is only used to construct the "Blue Blood" variable, and all name variations are accounted for, this is not a large issue for our purposes.
- **Conf** – Describes what conference a team is from. This variable is not used anywhere in our analysis.
- **G**– Stands for "Games played", this variable measures how many games a team has played before the tournament. Not every team plays the same number of games in a season (this usually varies from conference to conference). This variable was not used in our analysis.
- **REC** – The raw record of a team, in the format Wins-Losses. A team can obviously only win or lose as many games as they play in a season (undefeated or winless seasons are rare, but not unheard of). If you're curious, the team with the worst record to make the first round of the tournament since 2008 was 16 seed Cal Poly in 2014, with a record of 12 wins to 19 losses.
- **ADJOE -** The adjusted offense of a team, as calculated by Bart Torvik. This statistic uses a proprietary formula to take in multiple offensive attributes (most likely things like points scored, free throw rating, three point shot percentages, etc.) and combine them into one overall rating for a team.
- **ADJDE –** The adjusted defense of a team, as calculated by Bart Torvik. Similarly to ADJOE, this statistic uses a proprietary formula to take in different defensive

attributes (most likely things like opponents' shot rates, blocks, etc.) and combine them into one overall rating for a team.

- **BARTHAG –** The average rating of a team, as calculated by Bart Torvik using a proprietary formula. This most likely combines most of a team's raw statistics into one number.
- For the following variables in this section, there is an *offensive* rating and a *defensive* rating. The names of the offensive ratings will be listed, with the defensive rating having the same name but including a capital D in the spreadsheet. The offensive rating refers to *the team's* raw statistics, while the defensive rating refers to *the team's opponents'* raw statistics:
- **EFGp** – A positive number, indicating a team's shooting success, calculated as follows:

$$\frac{2FG + (1.5 * 3FG)}{FGA}$$

- Where:
  - 2FG = the number of successful 2-point shots
  - 3FG = the number of successful 3-point shots
  - FGA = the number of attempted 2 and 3-point shots (i.e. field goal shots) [13]
- **TOR** – Turnover percentage. It is approximately the number of turnovers per 100 plays. This takes a value between 0 and 100.
- **ORB** – Offensive rebound percentage. It is approximately the percentage of unsuccessful shot attempts recovered by the offense of a team.
- **DRB** – Defensive rebound percentage. It is approximately the percentage of unsuccessful shot attempts from the *opposing* team recovered by a team.
- **FTR** – Free throw rate. It is calculated by dividing the number of free throw attempts of a team by the number of field goal attempts.
- **TWPp** – Two point shot percentage. This takes a value from 0 to 100, measuring the number of 2-point shots that are successful.
- **THRp** – Three point shot percentage. This takes a value from 0 to 100, measuring the number of 3-point shots that are successful.
- **THRPR** – Three point shot rate. This is the percentage of all field goal shot attempts that are 3 points (as opposed to 2 point attempts).

The rest of these variables are standalone variables that do not have the *offensive* and *defensive* split.

- **ADJT** – This is the adjusted tempo rating, which is very similar to ADJOE and ADJDE, in that it is calculated using a proprietary formula accounting for raw variables (potentially the speed in which a team scores).
- **WAB** – As explained above, WAB, known as "Wins above Bubble" measures a team's performance relative to a "bubble" team. It can take positive or negative values, measuring how many more games a team won during their season compared to how a "bubble" team would theoretically do with the same schedule.
- **OppAvg** – This is simply a team's first round opponent's BARTHAG value.
- **OppWAB** – This is a team's first round opponent's WAB value.
- **Opp2Avg** – This is a team's projected second round opponent's BARTHAG value (e.g., a 1 seed would have an Opp2Avg value of the BARTHAG of the 8 seed it could face in the second round, and the 16 seed that 1 seed faced would have the same Opp2Avg value)
- **BlueBlood** – This is a binary variable. It is "1" if the team is one of the following twelve teams, which accounted for 58% of all Final Four teams as of 2018:
    - Arizona
    - Duke
    - Florida
    - Kansas
    - Kentucky
    - Louisville
    - Michigan
    - Michigan State
    - North Carolina
    - Syracuse
    - UCLA
    - UConn
- **Cost** – This is a variable from 1 to 25, reflecting C(s) for each team as described in Table 1. This was not used for prediction purposes – its only use was in implementing the knapsack algorithm.

Appendix C: Analyzing the Cost Values

In Appendix A, we noticed that some seeds required a certain number of wins to show up in the theoretical best team lists – in other words, to be an "optimal" pick. The table below summarizes these findings:

Table A. The minimum number of wins for a selected team to be "optimal".

SEEDS	NUMBER OF WINS REQUIRED
1	6
2	4
3-5	3
6-7	2
8-16	1

It may be worth discussing further the specific costs for each seed displayed in Table 1, and trying to understand how Dr. Stapleton came to these cost values. Starting with the baseline of having 100 points to spend, 1 seeds costing 25 points make good intuitive sense: 1 seeds are the most likely to do well, being the best teams, so it makes sense that picking all 1 seeds would use up the full 100 points. Likewise, no seed should cost 0 points (as then every list would contain them), so making 16 seeds cost 1 point also makes good sense. The other 14 seeds' values require further exploration.

The obvious idea is that a seed's cost is relative to its expected number of wins, with an expectation calculated by taking the average number of wins for teams of a given seed for all previous tournament years. This method of calculating costs, however, is somewhat difficult for us to use, as we don't know Dr. Stapleton devised the contest. Expectations will change from year to year; for example, the expected number of wins for a 16 seed was 0 before 2018. Now it is above 0, as two 16 seeds have won their first-round game (in 2018 and 2023).

However, we know that the contest was devised after 1985 (as that is when teams began to be seeded), and before 2014 (as the contest has been run at the University of Illinois since at least 2014) [2]. Given that the exact year is uncertain, we could use current expectations, as the numbers – while certainly different – should at least be fairly close.

The table below displays the expected number of wins for each seed, alongside the expected number of wins divided by the seed's cost (i.e. expected number of wins per point). Expectations are pulled from BracketOdds [14].

Table B. The expected number of wins in relation to seed and cost.

SEED	EXPECTED # OF WINS	COST	EXPECTATION / COST
1	3.30	25	0.132
2	2.33	19	0.123
3	1.84	13	0.142
4	1.56	12	0.130
5	1.15	11	0.105
6	1.04	10	0.104
7	0.90	8	0.1125
8	0.71	5	0.142
9	0.62	5	0.124
10	0.60	4	0.150
11	0.67	4	0.1675
12	0.51	3	0.170
13	0.25	2	0.125
14	0.16	2	0.080
15	0.10	1	0.100
16	0.013	1	0.013

In the above table, seeds (besides 16 seeds) have an expectation / cost between 0.1 and 0.17, which is fairly similar. 12 seeds appear to be the "most bang for your buck", having the highest expectation / cost (i.e. having the most expected wins relative to cost). Meanwhile, even when costing only 1 point, 15 and 16 seeds do not especially seem worth it; 5, 6, and 14 seeds have similarly low expectation / costs.

We can recalculate the costs to account for the number of expected wins above. Keeping 1 seeds at 25 points and 16 seeds at 1 point, we calculate costs for every other seed, keeping their expectation / costs as close to 0.132 as possible (while keeping costs as integers):

Table C. The expected number of wins alongside the new, updated costs.

SEED	EXPECTED # OF WINS	OLD COST	NEW COST	EXPECTATION / NEW COST
1	3.30	25	25	0.132
2	2.33	19	18	0.129
3	1.84	13	14	0.131
4	1.56	12	12	0.130
5	1.15	11	9	0.128
6	1.04	10	8	0.130
7	0.90	8	7	0.129
8	0.71	5	5	0.142
9	0.62	5	5	0.124
10	0.60	4	5	0.120
11	0.67	4	5	0.134
12	0.51	3	4	0.125
13	0.25	2	2	0.125
14	0.16	2	1	0.160
15	0.10	1	1	0.100
16	0.013	1	1	0.013

The above "new" costs are the integers that get expectation / cost closest to 0.132 for all seeds. In general, low seeds' costs tend to go up (which makes sense, as upsets have most likely become more common since Dr. Stapleton created his contest), and mid-seeds costs tend to go down (which is most likely related to this increase in upsets). As it stands, only 5 and 6 seeds have costs changed by a magnitude of 2 (the greatest magnitude of change). Every other cost is either unchanged or changed by only one point; this lends credence to the idea that Dr. Stapleton used a similar method in creating his original list of costs for each seed.

References

[1] *68 million Americans to Wager on March madness*. American Gaming Association. (2023, April 4). https://www.americangaming.org/new/68-million-americans-to-wager-on-march-madness/

[2] *Department of Statistics - NCAA Contest*. David Unger. (n.d.). https://publish.illinois.edu/dunger/ncaa-contest/

[3] STAT DEPARTMENT NCAA BASKETBALL CONTEST (n.d.). https://github.com/aaludkow/STATDeptNCAAContest/blob/main/NCAA%20Contest%20Rules.pdf

[4] Rank - customizable college basketball tempo free stats - T-rank. T. (n.d.). https://barttorvik.com/

[5] 2024 Pomeroy College Basketball Ratings. (n.d.). https://kenpom.com/

[6] ESPN Internet Ventures. (n.d.). *Serving sports fans. anytime. anywhere.* ESPN. https://www.espn.com/analytics/

[7] Groel, C. (2020, March 5). *Wins above bubble and the at-large case for Stephen F. Austin*. Medium. https://medium.com/top-level-sports/wins-above-bubble-and-the-at-large-case-for-stephen-f-austin-d6ebea0cd904

[8] Bois, J. (2018, March 11). *The NCAA Tournament is a Loser Machine | Chart Party*. YouTube. https://www.youtube.com/watch?v=4a1TUszkMfI &t=389s

[9] Census regions and divisions of the United States. (n.d.). https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf

[10] Andonov, R., Poirriez, V., & Rajopadhye, S. (2000). Unbounded knapsack problem: Dynamic Programming Revisited. *European Journal of Operational Research*, *123*(2), 394–407. https://doi.org/10.1016/s0377-2217(99)00265-9

[11] Borchers, H. W. (2023, October 26). *Knapsack: 0-1 KNAPSACK PROBLEM IN ADAGIO: Discrete and global optimization routines*. knapsack: 0-1 Knapsack Problem in adagio: Discrete and Global Optimization Routines. https://rdrr.io/rforge/adagio/man/knapsack.html

[12] Chartier, T., Kreutzer, E., Langville, A., & Pedings, K. (2010). Bracketology: How can math help? *Mathematics and Sports*, 55–70. https://doi.org/10.5948/upo9781614442004.006

[13] *Glossary*. Basketball Reference. (n.d.). https://www.basketball-reference.com/about/glossary.html

[14] *Seed distributions for March madness: A tool for Bracketologists (2025)*. BracketOdds. (n.d.). https://bracketodds.cs.illinois.edu/seedadv.html