Analyzing March Madness Statistics and Bracket Success

Brian Janger

```
library(tidyverse)
```

```
basketball <- read_csv("data/ncaa_data.csv")
basketball2 <- read_csv("data/ncaa_data_2.csv")
marchmadness <- inner_join(basketball, basketball2)
marchmadness <- subset(marchmadness, select = -c(X9, X12, X15, X18, MP, SRS,
                                                  X21, W_1, L_1, W_2, L_2, W_3, L_3, Rk))
marchmadness <- marchmadness %>%
  filter(grepl("NCAA", School, fixed = TRUE) == T)
marchmadness$School <- gsub(' NCAA', '', marchmadness$School)
scores <- read_csv("data/startingvalues.csv")
marchmadness <- inner_join(marchmadness, scores)
marchmadness <- marchmadness %>%
  mutate(seed = 17 - powerScore) %>%
  mutate(ORB_gm = ORB/G) %>%
  mutate(TRB_gm = TRB/G) %>%
  mutate(AST_gm = AST/G) %>%
  mutate(STL_gm = STL/G) %>%
  mutate(BLK_gm = BLK/G) %>%
  mutate(TOV_gm = TOV/G) %>%
  mutate(Fouls_gm = PF/G) %>%
  mutate(PPG = Tm./G) %>%
  mutate(PAPG = Opp./G) %>%
  subset(select = -c(ORB,TRB,AST,STL,BLK,TOV,PF,Tm.,Opp.))
marchmadness <- marchmadness %>%
  mutate(pointdiff = PPG - PAPG)

# View(marchmadness)
write.csv(marchmadness, 'marchmadness.csv')
```

Developing a Base Formula for March Madness Strength

To develop a base model for all of the statistics we want to use for creating a team strength metric (which will be used to determine the winners of the various games in the March Madness bracket), we need to find a way to evenly weigh all of the various statistics we want to use in the model.

The main thing that we need to keep track of is that there are three main types of statistics that we want to incorpoate in our model, listed below:

- Per-game statistics, such as points per game, points allowed per game, rebounds per game, etc.
- Percentage based statistics, such as field goal percentage, three point percentage, free throw percentage, win percentage, etc.

- Miscellaneous statistics that could have an effect on team success, such as pace, **strength of schedule**, free throws per field goal attempted, etc.

It is important for the base model to not overweigh any of the statistics (as the model would then be biased toward a certain statistic), so a method for this must be devleoped.

The Importance of Strength of Schedule

Strength of schedule (SOS) is a vital statistic for all of the teams in the NCAA March Madness tournament since the strength of a team's opponents throughout the season has a drastic effect on all of the other statistics of a team. For example, a team with a below average SOS (i.e. < 0) could potentially have better statistics all around than a team with a more difficult SOS, even if the second team is a better team. The opponents that a team plays has a large effect on its statistics, so using SOS as a coefficient in the model or as a "balancing factor" is very important.

Below is a preliminary model idea for the power, P, of a team in the March Madness tournament based on per-game statistics which does not account for strength of schedule:

$$P_1 = (PPG - PAPG) + (STL\_gm - TOV\_gm) + TRB\_gm - Fouls\_gm$$

Applying this basic model to the data, we can sort the teams by their associated power rankings:

```
marchmadness <- marchmadness %>%
  mutate(power_pergame = PPG - PAPG + STL_gm - TOV_gm + TRB_gm - Fouls_gm)

marchmadness %>%
  arrange(desc(power_pergame)) %>%
  subset(select = c(seed, School, power_pergame)) %>%
  head(10)
```

```
## # A tibble: 10 x 3
##     seed School            power_pergame
##    <dbl> <chr>                     <dbl>
##  1     1 Gonzaga                    40.5
##  2     2 Houston                    39.5
##  3    14 Colgate                    38.9
##  4     1 Baylor                     33.8
##  5     2 Iowa                       33.1
##  6    14 Abilene Christian          30.1
##  7    11 Utah State                 29.9
##  8    11 Drake                      29.9
##  9     3 Arkansas                   29.7
## 10    12 UC-Santa Barbara           29.6
```

We can see that some high-seeded teams top the rankings of this metric without even accounting for SOS. If we add each team's SOS to this model, we can get the following rankings:

```
marchmadness <- marchmadness %>%
  mutate(power_pergame_SOS = power_pergame + SOS)

marchmadness %>%
  arrange(desc(power_pergame_SOS)) %>%
  subset(select = c(seed, School, power_pergame_SOS)) %>%
  head(10)
```

```
## # A tibble: 10 x 3
##     seed School         power_pergame_SOS
##    <dbl> <chr>                      <dbl>
##  1     2 Iowa                        44.5
##  2     1 Gonzaga                     43.8
##  3     2 Houston                     43.6
##  4    14 Colgate                     40.4
##  5     1 Illinois                    39.9
##  6     1 Michigan                    39.4
##  7     1 Baylor                      39.2
##  8     2 Alabama                     36.8
##  9     3 Arkansas                    36.0
## 10     8 North Carolina              35.0
```

We can see that the top-ranked teams are closer to the top, but can see a lot of surprises here, the main one being Colgate, a 14 seed, which is rated as the 4th best team in the tournament based on these statistics. This is probably an indicator that there is a lot of inaccuracy in this model (throughout the season, Colgate routinely demolished bad teams and got great statistics, and their low SOS isn't low enough to counter these statistics). However, a lot of the lower-seeded teams were pushed down in the rankings due to their low strength of schedule, making this model a bit more accurate than the last.

Introducing Percentages and Efficiency

A lot of the variables we want to look at in this model are not accounted for in the per-game model alone. We also want to see how percentage statistics, such as shooting percentage, steal, block, and turnover percentage, and win percentage, can affect a bracket model and a bracket's accuracy.

An introductory model (which does not account for SOS) is shown below:

$$P_2 = FG\% * 3P\% * FT\% * (AST\% + STL\% + BLK\% + ORB\% - TOV\%)$$

```
marchmadness <- marchmadness %>%
  mutate(power_percentage = FGprct*threeptprct*FTprct*W_Lprct*
           (ASTprct+STLprct+BLKprct+ORBprct-TOprct))

marchmadness %>%
  arrange(desc(power_percentage)) %>%
  subset(select = c(seed, School, power_percentage)) %>%
  head(10)
```

```
## # A tibble: 10 x 3
##     seed School            power_percentage
##    <dbl> <chr>                        <dbl>
##  1     1 Baylor                       13.2
##  2     1 Gonzaga                      12.8
##  3    14 Colgate                      12.3
##  4     2 Houston                      11.1
##  5     1 Michigan                     10.6
##  6    13 Liberty                      10.1
##  7    12 UC-Santa Barbara             10.0
##  8     8 Loyola (IL)                   9.97
##  9    14 Abilene Christian             9.66
## 10     2 Iowa                         9.62
```

```
marchmadness <- marchmadness %>%
  mutate(power_percentage_SOS = FGprct*threeptprct*FTprct*W_Lprct*
           (ASTprct+STLprct+BLKprct+ORBprct-TOprct)+SOS)

marchmadness %>%
  arrange(desc(power_percentage_SOS)) %>%
  subset(select = c(seed, School, power_percentage_SOS)) %>%
  head(10)
```

```
## # A tibble: 10 x 3
##     seed School        power_percentage_SOS
##    <dbl> <chr>                        <dbl>
##  1     1 Michigan                      23.0
##  2     1 Illinois                      21.7
##  3     2 Iowa                          21.0
##  4     2 Ohio State                    20.9
##  5     4 Purdue                        19.0
##  6     4 Florida State                 18.8
##  7     1 Baylor                        18.6
##  8     9 Wisconsin                     18.3
##  9     2 Alabama                       18.1
## 10     4 Virginia                      18.0
```

Introducing Miscellaneous Statistics

There are also a few statistics that don't necessarily fit into the previous categories, but they would be a good thing to potentially add to the model. Statistics such as the pace a team plays at, the percentage of field goal attempts that are three pointers, and the ratio of free throws to field goal attempts are all statistics we could add to the model to see if it has any effect on bracket success.

A Different Percentage Model

$$P_3 = 10 * (FG\% + 3P\% + FT\% + \frac{(AST\% + STL\% + BLK\% + ORB\% - TOV\%)}{100})$$

```
marchmadness <- marchmadness %>%
  mutate(power_percentage_2 = 10*( FGprct+threeptprct+FTprct+W_Lprct+
           (ASTprct+STLprct+BLKprct+ORBprct-TOprct)/100))

marchmadness %>%
  arrange(desc(power_percentage_2)) %>%
  subset(select = c(seed, School, power_percentage_2)) %>%
  head(10)
```

```
## # A tibble: 10 x 3
##     seed School           power_percentage_2
##    <dbl> <chr>                        <dbl>
##  1     1 Baylor                        35.3
##  2     1 Gonzaga                       35.2
##  3     2 Houston                       34.8
##  4    14 Colgate                       34.7
```

```
##  5    14 Abilene Christian              33.8
##  6    12 UC-Santa Barbara               33.6
##  7    12 Winthrop                       33.6
##  8     1 Michigan                       33.6
##  9     2 Iowa                           33.2
## 10     8 Loyola (IL)                    33.2
```

```
marchmadness <- marchmadness %>%
  mutate(power = powerScore + power_pergame +
           power_percentage_2 + SOS)

marchmadness %>%
  arrange(desc(power)) %>%
  subset(select = c(seed, School, power))
```

```
## # A tibble: 68 x 3
##     seed School   power
##    <dbl> <chr>    <dbl>
##  1     1 Gonzaga   95.1
##  2     2 Houston   93.5
##  3     2 Iowa      92.7
##  4     1 Baylor    90.4
##  5     1 Michigan  89.0
##  6     1 Illinois  88.2
##  7     2 Alabama   83.7
##  8     3 Arkansas  82.4
##  9     3 Kansas    79.9
## 10    14 Colgate   78.0
## # ... with 58 more rows
```

When we alter the percentage-based statistics, we will use this base model instead (with `power_percentage_2`) to prevent any other statistics from affecting the others, since the first base model does have that limitation.

Models and Brackets

The Base Model

Using the two non-SOS-weighted models above, we can create a base model where all statistics are weighted evenly (i.e. all coefficeints on the variables are 1). We will then account for SOS in this model to avoid double counting SOS, causing a heavier SOS weighting.

We will also introduce one last variable here: `powerScore`. Since it is a given that the better seeded teams have a higher chance of winning, we accounted for this in our model, where `powerScore = 17 - seed`. Therefore, 1-seeded teams will start with a `powerScore` of 16 while 16-seeded teams will have a score of just 1, giving higher seeded teams a "head start" against the lower seeded teams.

$$P = powerScore + P_1 + P_2 + SOS$$

```
marchmadness <- marchmadness %>%
  mutate(power = powerScore + power_pergame +
           power_percentage + SOS)
```

```
marchmadness %>%
  arrange(desc(power)) %>%
  subset(select = c(seed, School, power))
```

```
## # A tibble: 68 x 3
##     seed School   power
##    <dbl> <chr>    <dbl>
##  1     1 Gonzaga   72.7
##  2     2 Houston   69.7
##  3     2 Iowa      69.1
##  4     1 Baylor    68.4
##  5     1 Michigan  66.0
##  6     1 Illinois  64.9
##  7     2 Alabama   59.6
##  8     3 Arkansas  58.2
##  9     3 Kansas    55.7
## 10    14 Colgate   55.7
## # ... with 58 more rows
```

Using the base model, which only accounts for per-game statistics and percentage based statistics, we can see that Gonzaga is the favorite to win the tournament as they are the strongest team. Colgate, a 14 seed, still appears as the 10th best team in the tournament, so that means there could be upset potential in the bracket with them as well (however, their first round matchup is Arkansas, ranked 8th).

Other notable low-seeded teams with higher-than-expected ranking are 8 seed North Carolina (ranked 20th) and 9 seed St. Bonaventure (ranked 23rd).

Some high-seeded teams appear very weak according to the model, though. 4 seeed Purdue is ranked 22nd, 4 seed Oklahoma State is ranked 30th, and 3 seed Texas is ranked 25th, showing that these teams have upset potential in the second or third rounds of the tournament.

Using the base model, the following bracket is created:

insert base model bracket pdf here

We notice that there are not very many upsets in this bracket, which is odd considering March Madness is known for the crazy upsets that happen every year. The point of this project is to create as many brackets as possible, so we can see how our bracket performs if we remove seed advantage (by down-weighting the `powerScore` variable).

Base Model with Downweighted Seed Advantage

Here is the base model with `powerScore` halved, so 1 seeds have a score of 8 while 16 seeds have a score of 0.5.

```
marchmadness <- marchmadness %>%
  mutate(power = powerScore/2 + power_pergame +
           power_percentage + SOS)

marchmadness %>%
  arrange(desc(power)) %>%
  subset(select = c(seed, School, power))
```

```
## # A tibble: 68 x 3
##     seed School   power
##    <dbl> <chr>    <dbl>
##  1     1 Gonzaga   64.7
##  2     2 Houston   62.2
##  3     2 Iowa      61.6
##  4     1 Baylor    60.4
##  5     1 Michigan  58.0
##  6     1 Illinois  56.9
##  7    14 Colgate   54.2
##  8     2 Alabama   52.1
##  9     3 Arkansas  51.2
## 10     3 Kansas    48.7
## # ... with 58 more rows
```

Bracket:

Base Model with High SOS Weighting

Some argue that statistics are unfounded when SOS isn't accounted for, so we will increase the SOS weighting by 3x in this model.

```
marchmadness <- marchmadness %>%
  mutate(power = powerScore + power_pergame +
           power_percentage + 3*SOS)

marchmadness %>%
  arrange(desc(power)) %>%
  subset(select = c(seed, School, power))
```

```
## # A tibble: 68 x 3
##     seed School        power
##    <dbl> <chr>         <dbl>
##  1     2 Iowa           91.8
##  2     1 Michigan       90.9
##  3     1 Illinois       90.2
##  4     2 Ohio State     81.3
##  5     2 Alabama        80.3
##  6     1 Gonzaga        79.4
##  7     1 Baylor         79.1
##  8     2 Houston        78.0
##  9     3 Kansas         76.3
## 10     3 West Virginia  75.1
## # ... with 58 more rows
```

Bracket:

Base Model with Downweighted Seed Advantage, High SOS Weighting

Now we will combine the two adjustments on the previous model to make a model with both weight adjustments.

```
marchmadness <- marchmadness %>%
  mutate(power = powerScore/2 + power_pergame +
           power_percentage + 3*SOS)

marchmadness %>%
  arrange(desc(power)) %>%
  subset(select = c(seed, School, power))
```

```
## # A tibble: 68 x 3
##     seed School        power
##    <dbl> <chr>         <dbl>
##  1     2 Iowa           84.3
##  2     1 Michigan       82.9
##  3     1 Illinois       82.2
##  4     2 Ohio State     73.8
##  5     2 Alabama        72.8
##  6     1 Gonzaga        71.4
##  7     1 Baylor         71.1
##  8     2 Houston        70.5
##  9     3 Kansas         69.3
## 10     3 West Virginia  68.1
## # ... with 58 more rows
```

Bracket:

Experiment: Base Model with Extreme SOS Upweighting

As an experiment, we can see what the bracket looks like when we give SOS a 6 times weighting:

```
marchmadness <- marchmadness %>%
  mutate(power = powerScore + power_pergame +
           power_percentage + 6*SOS)

marchmadness %>%
  arrange(desc(power)) %>%
  subset(select = c(seed, School, power))
```

```
## # A tibble: 68 x 3
##     seed School        power
##    <dbl> <chr>         <dbl>
##  1     1 Michigan       128.
##  2     1 Illinois       128.
##  3     2 Iowa           126.
##  4     2 Ohio State     122.
##  5     4 Purdue         112.
##  6     2 Alabama        111.
##  7     9 Wisconsin      108.
##  8    10 Rutgers        108.
##  9     3 West Virginia  107.
## 10     3 Kansas         107.
## # ... with 58 more rows
```

The following bracket has a lot of upsets!!

Bracket:

Base Model with Emphasis on Steals

The following model gives a four times weighting to the steals per game and steal% categories.

```
marchmadness <- marchmadness %>%
  mutate(power = powerScore + power_pergame +
           power_percentage_2 + SOS +
           3*STL_gm + 3*STLprct/10)

marchmadness %>%
  arrange(desc(power)) %>%
  subset(select = c(seed, School, power))
```

```
## # A tibble: 68 x 3
##     seed School        power
##    <dbl> <chr>         <dbl>
##  1     1 Gonzaga        123.
##  2     2 Houston        122.
##  3     1 Baylor         121.
##  4     2 Alabama        113.
##  5     2 Iowa           112.
##  6     3 Arkansas       110.
##  7     1 Illinois       108.
##  8     1 Michigan       104.
##  9     3 West Virginia  103.
## 10     3 Kansas         103.
## # ... with 58 more rows
```

Bracket:

Base Model with Emphasis on Three Point Percentage and Attempt Rate

The following model gives an eight times weighting to the three point shooting percentage category and adds in three point attempt rate as a factor to consider.

```
marchmadness <- marchmadness %>%
  mutate(power = powerScore + power_pergame +
           power_percentage_2 + SOS +
           70*threeptprct + 100*threePAr*threeptprct)

marchmadness %>%
  arrange(desc(power)) %>%
  subset(select = c(seed, School, power))
```

```
## # A tibble: 68 x 3
##     seed School    power
##    <dbl> <chr>     <dbl>
```

```
##  1     1 Baylor     136.
##  2     2 Iowa       135.
##  3     2 Houston    134.
##  4     1 Gonzaga    132.
##  5     1 Michigan   129.
##  6     1 Illinois   126.
##  7     2 Alabama    125.
##  8    14 Colgate    121.
##  9     4 Virginia   120.
## 10     5 Creighton  119.
## # ... with 58 more rows
```

Bracket:

Base Model with Emphasis on Free Throw Percentage and Attempt Rate

The following model gives an eight times weighting to the free throw shooting percentage category and adds in free throw attempt rate as a factor to consider.

```
marchmadness <- marchmadness %>%
  mutate(power = powerScore + power_pergame +
           power_percentage_2 + SOS +
           70*FTprct + 100*FTr*FTprct)

marchmadness %>%
  arrange(desc(power)) %>%
  subset(select = c(seed, School, power))
```

```
## # A tibble: 68 x 3
##     seed School        power
##    <dbl> <chr>         <dbl>
##  1     1 Gonzaga        173.
##  2     2 Houston        167.
##  3     1 Michigan       166.
##  4     2 Iowa           165.
##  5     1 Illinois       164.
##  6     2 Ohio State     160.
##  7     3 Arkansas       158.
##  8     1 Baylor         158.
##  9     3 West Virginia  156.
## 10     5 Colorado       156.
## # ... with 58 more rows
```

Base Model with Emphasis on All Shooting Categories

The following model gives an eight times weighting to all shooting categories and adds in three point rate and free throw rate as categories to consider in the model. The model also interacts FG and 3P categories with SOS, but not FT categories since the opponent doesn't affect free throws.

```
marchmadness <- marchmadness %>%
  mutate(power = powerScore + power_pergame +
```

```
          power_percentage_2 + SOS +
          70*FTprct + 100*FTr*FTprct +
          SOS*(70*threeptprct + 100*threeptprct*threePAr +
          70*FGprct))

marchmadness %>%
  arrange(desc(power)) %>%
  subset(select = c(seed, School, power))
```

```
## # A tibble: 68 x 3
##     seed School         power
##    <dbl> <chr>          <dbl>
##  1     2 Ohio State     1138.
##  2     1 Michigan       1087.
##  3     1 Illinois       1084.
##  4    10 Maryland       1061.
##  5     2 Iowa           1020.
##  6     9 Wisconsin      1011.
##  7     4 Purdue          988.
##  8    10 Rutgers         980.
##  9    11 Michigan State  917.
## 10    12 Georgetown      904.
## # ... with 58 more rows
```

Bracket:

Base Model with Emphasis on Rebounding

The following model gives an eight times weighting to team rebounds per game as well as offensive rebound percentage to see what teams are stronger in the rebounding category.

```
marchmadness <- marchmadness %>%
  mutate(power = powerScore + power_pergame +
          power_percentage_2 + SOS +
          7*TRB_gm + 7*ORBprct/10)

marchmadness %>%
  arrange(desc(power)) %>%
  subset(select = c(seed, School, power))
```

```
## # A tibble: 68 x 3
##     seed School              power
##    <dbl> <chr>               <dbl>
##  1     2 Houston              409.
##  2     8 North Carolina       406.
##  3     2 Iowa                 395.
##  4     1 Illinois             394.
##  5     2 Alabama              389.
##  6     3 Arkansas             384.
##  7     1 Gonzaga              382.
##  8    11 Utah State           380.
```

```
##  9     6 Southern California  377.
## 10    14 Colgate              376.
## # ... with 58 more rows
```

Base Model with Emphasis on Defensive Statistics

The following model gives an eight times weighting to all defensive statistics included in the base model.

```
marchmadness <- marchmadness %>%
  mutate(power = powerScore + power_pergame +
           power_percentage_2 + SOS +
           7*STL_gm + 7*STLprct/10 +
           7*BLKprct/10)

marchmadness %>%
  arrange(desc(power)) %>%
  subset(select = c(seed, School, power))
```

```
## # A tibble: 68 x 3
##     seed School                power
##    <dbl> <chr>                 <dbl>
##  1     2 Houston                170.
##  2     1 Baylor                 169.
##  3     1 Gonzaga                166.
##  4     2 Alabama                160.
##  5     3 Arkansas               156.
##  6    11 Syracuse               150.
##  7     6 San Diego State        148.
##  8     9 Georgia Tech           147.
##  9     2 Iowa                   145.
## 10    10 Virginia Commonwealth  145.
## # ... with 58 more rows
```

Base Model with Pace and SOS

6 times SOS weighting, Pace included.

```
marchmadness <- marchmadness %>%
  mutate(power = powerScore + power_pergame +
           power_percentage_2 + 5*SOS +
           Pace/10)

marchmadness %>%
  arrange(desc(power)) %>%
  subset(select = c(seed, School, power))
```

```
## # A tibble: 68 x 3
##     seed School        power
##    <dbl> <chr>         <dbl>
##  1     1 Illinois       146.
##  2     1 Michigan       146.
```

```
##  3     2 Iowa           145.
##  4     2 Ohio State     139.
##  5     2 Alabama        133.
##  6     4 Purdue         130.
##  7     3 Kansas         128.
##  8     3 West Virginia  128.
##  9     9 Wisconsin      126.
## 10    10 Rutgers        126.
## # ... with 58 more rows
```

New Model 1

This is a new model based on offensive efficiency, pace, turnover percentage, and strength of schedule. Essentially, this model keeps track of how efficient a team is and how fast a team plays, all while being able to limit the number of turnover they have on offense.

```
marchmadness <- marchmadness %>%
  mutate(power = ORtg/10 + Pace/10 - TOprct + SOS)

marchmadness %>%
  arrange(desc(power)) %>%
  subset(select = c(seed, School, power))
```

```
## # A tibble: 68 x 3
##     seed School        power
##    <dbl> <chr>         <dbl>
##  1     2 Iowa           18.7
##  2     2 Ohio State     18.3
##  3     9 Wisconsin      17.6
##  4    10 Rutgers        16.0
##  5     1 Michigan       15.9
##  6     5 Villanova      15.7
##  7     1 Illinois       15.1
##  8    10 Maryland       14.5
##  9     3 West Virginia  14.2
## 10     4 Purdue         13.6
## # ... with 58 more rows
```