5-2022

ATTEMPTING TO PREDICT THE UNPREDICTABLE: MARCH MADNESS

Coleton Kanzmeier
ckanzmeier@unomaha.edu

ATTEMPTING TO PREDICT THE UNPREDICTABLE: MARCH MADNESS

Can past tournaments offer guidance to the 2022 tournament?

Coleton Kanzmeier

May 2022

Bachelor of Arts in Mathematics

Dr. Andrew Swift and Dr. Steven From

KEYWORDS: Machine Learning, March Madness, Prediction, Logistic Regression, Data Manipulation, Data Cleaning

Table of Contents

Abstract

Each year, millions upon millions of individuals fill out at least one if not hundreds of March Madness brackets. People test their luck every year, whether for fun, with friends or family, or to even win some money. Some people rely on their basketball knowledge whereas others know it is called March Madness for a reason and take a shot in the dark. Others have even tried using statistics to give them an edge. I intend to follow a similar approach, using statistics to my advantage. The end goal is to predict this year's, 2022, March Madness bracket. To achieve the best possible results, I will use team and individual statistics to help form logistic regression models and formulate new statistics that have not been used or thought of before. Rather than jumping right into the 2021-2022 season, I look into past years' statistics and tournaments to see how well my logistic regression models perform and see what differences if any, there are in variables used year to year. The 2019-2020 and 2020-2021 seasons will not be present (no tournament and no fans, respectively). After evaluating past years' models, I make rules to provide the best possibility of upsets to occur, based on what was seen in the 2011-2019 tournaments.

Introduction

The NCAA each year hosts a Division I college basketball tournament, also known as March Madness. The tournament currently consists of 68 teams, 8 of which have a play-in game to participate in the round of 64. These 68 teams are capable of making the tournament in one of two ways: an automatic bid or an at-large bid. An automatic bid is reserved for all teams in Division I that win their conference tournament. The at-large bids are then distributed by a committee that determines what teams are most deserving based on records, quality of wins, and other relevant factors. Once the teams are locked in, the committee then splits up the 68 teams (with 4 play-in games) into 4 regions with the teams seeded 1-16 in each region (the play-in games make a team play in for a given seed – typically an 11, 12, or 16). The tournament then consists of 6 rounds, excluding the play-in games, that span over 3 weekends. The first weekend includes the round of 64 and the round of 32. The second weekend hosts the sweet 16 and elite 8, while the third weekend crowns a champion with the final 4 and championship game. Having a little over half a week off between weekends adds an interesting component, allowing teams to rest and better prepare for their next opponent. In addition to extra time off between games, teams also travel to new locations which may aid in their success or hurt it. To the naked eye, there seems to be no rhyme or reason as to why certain teams make it to the second weekend, the third weekend, or even why they win it all. Some years like 2011 and 2014 have lower seeds dominating their way to the final 4, championship game, or even winning it all like 7 seeded Connecticut did in 2014. Some years give one underdog while the rest of the field is dominated by the best teams, such as 2012, 2015, 2016, 2017, and 2018. Or, a year may truly be run by the best teams like 2019 where a 5 seed in the final 4 was the lowest seed that advanced to that point. There seems to be a multitude of outcomes, based on past tournaments, that seem random at first glance. The goal of this paper is to use the results of the past decade of tournaments (2011-2019) to help form predictions for the 2022 NCAA tournament. The 2020 tournament is not included because it was canceled due to the Covid-19 pandemic and the 2021 tournament is not included because of the absence of fans, which is not a comparable season to use data for this project.

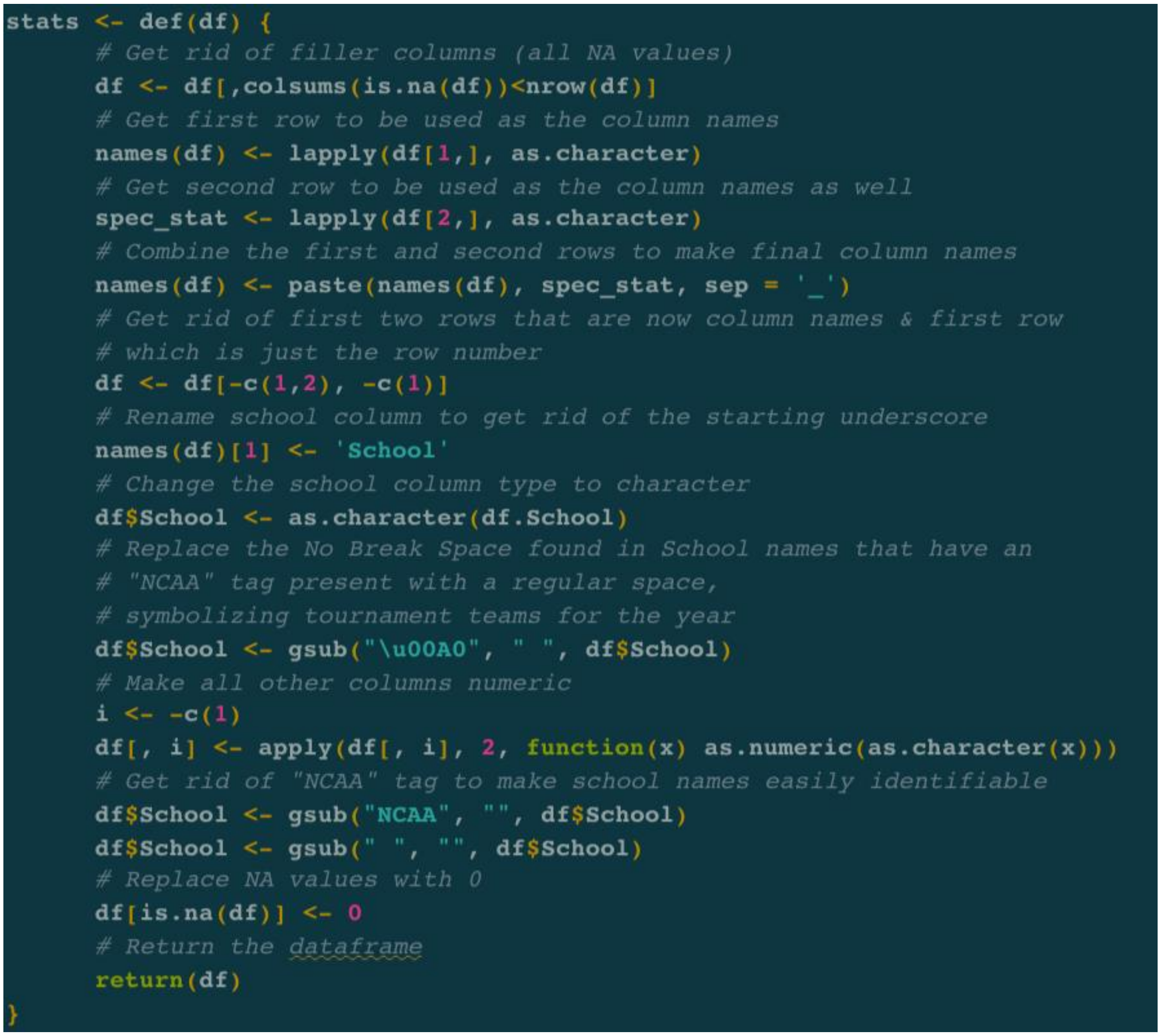
To show the reproducibility of my approach to predicting the 2022 NCAA tournament, I provide the steps taken to collect the data necessary, the cleaning that was performed, and then the approach to modeling that was used. After that, I will go through further research that I would like to undergo given more time and/or more access to other statistics.

Data Pulling & Cleaning

The data needed to develop appropriate logistic regression models are individual statistics data, team statistics data, and game results data. The individual and team statistics are added to the game results data before the regular season and conference tournament game data is used to form a training set within the modeling phase.

Team Statistics

All team statistics were pulled from sports-reference.com/cbb. The four datasets included on this website are basic school stats, basic opponent stats, advanced school stats, and advanced opponent stats. There are a total of 74 variables once combined, ranging from general statistics such as wins/losses, the strength of schedule, and a simple rating system to basic stats (for both the team and their opponents) like field goals (attempted), turnovers, fouls, and (offensive) rebounds, all the way to advanced stats such as the pace of the team, effective field goal percentage, and total rebound percentage. To combine the four datasets, they need to be appropriately cleaned because the CSV files not only have multiple row headers but also some missing values. The missing values are filled to the numerical value '0' because they appear in the conference wins/losses columns where the correlating team does not appear in a conference (which creates a missing value once called into RStudio). The rest of the cleaning of the team statistics includes renaming the columns to appropriately deal with the multiple row headers, getting rid of unnecessary rows and columns, and giving each column its correct variable type. All of the changes can be seen in Figure 1, which includes comments describing what the following line of code accomplishes. Once the function in Figure 1 is run on all four datasets for each given year, that specific year can be merged to provide a data frame with 74 variables (excluding the school's name). Note that the 2010-11 data set still needs to have school names changed to match the 2012-19 seasons for consistency (i.e., Alabama-Birmingham to UAB). All of these instances were found after intersecting school names from each year and finding what schools were not included).

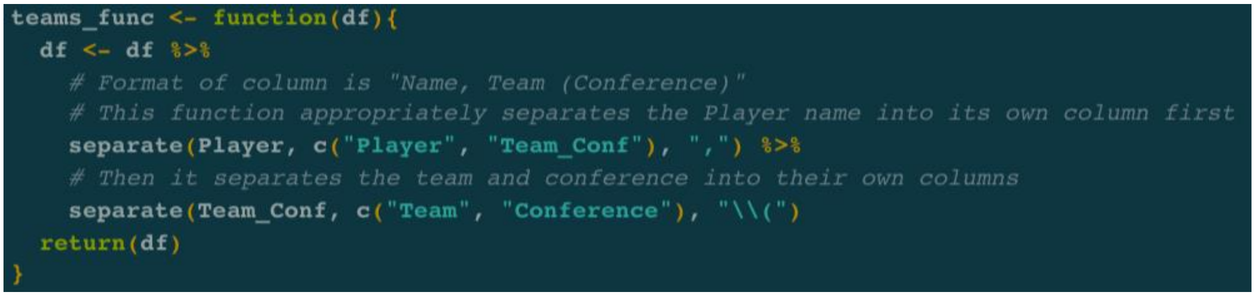


```
stats <- def(df) {
    # Get rid of filler columns (all NA values)
    df <- df[,colsums(is.na(df))<nrow(df)]
    # Get first row to be used as the column names
    names(df) <- lapply(df[1,], as.character)
    # Get second row to be used as the column names as well
    spec_stat <- lapply(df[2,], as.character)
    # Combine the first and second rows to make final column names
    names(df) <- paste(names(df), spec_stat, sep = '_')
    # Get rid of first two rows that are now column names & first row
    # which is just the row number
    df <- df[-c(1,2), -c(1)]
    # Rename school column to get rid of the starting underscore
    names(df)[1] <- 'School'
    # Change the school column type to character
    df$School <- as.character(df.School)
    # Replace the No Break Space found in School names that have an
    # "NCAA" tag present with a regular space,
    # symbolizing tournament teams for the year
    df$School <- gsub("\u00A0", " ", df$School)
    # Make all other columns numeric
    i <- -c(1)
    df[, i] <- apply(df[, i], 2, function(x) as.numeric(as.character(x)))
    # Get rid of "NCAA" tag to make school names easily identifiable
    df$School <- gsub("NCAA", "", df$School)
    df$School <- gsub(" ", "", df$School)
    # Replace NA values with 0
    df[is.na(df)] <- 0
    # Return the dataframe
    return(df)
}
```

Figure 1

Individual Statistics

All individual statistics were pulled from stats.ncaa.org. The five individual statistics that were pulled include points per game, rebounds per game, assists per game, assist to turnover ratio, and steals per game. These data sets were exported as excel files which are then saved as CSV files to be pulled into RStudio. The goal of these individual statistics was to create a factor variable where a team is given a “1” if they have a top 100 scorer, assister, rebounder, stealer, or assist to turnover ratio player. Since the format of the dataset has the player's name, school name, and conference of the team all in the same column, I utilized the function seen in Figure 2 to appropriately split off the school's name. This was done to be able to create a variable within the team statistic dataset from the newly retrieved individual statistic dataset.

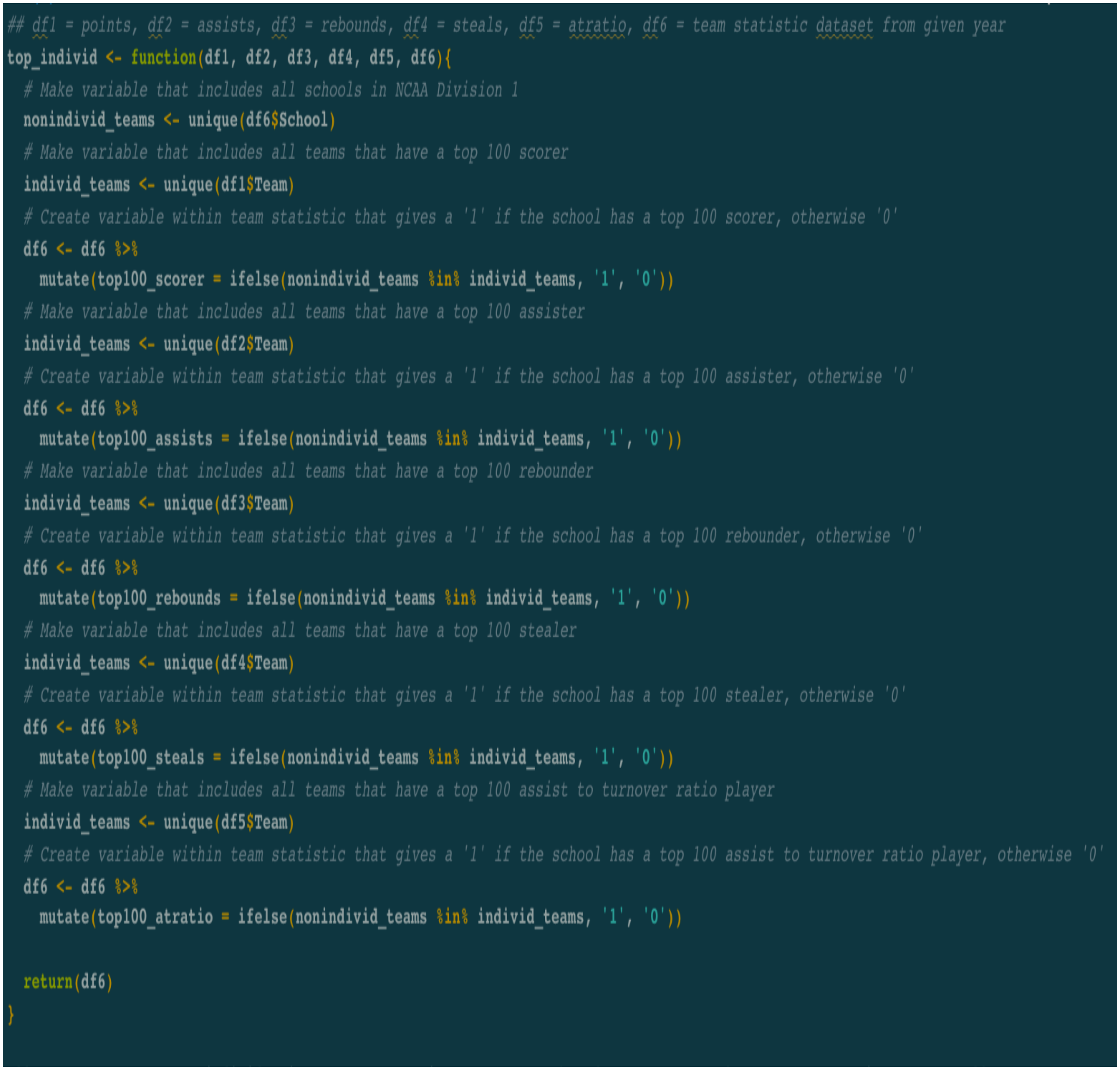


```
teams_func <- function(df){
  df <- df %>%
    # Format of column is "Name, Team (Conference)"
    # This function appropriately separates the Player name into its own column first
    separate(Player, c("Player", "Team_Conf"), ",") %>%
    # Then it separates the team and conference into their own columns
    separate(Team_Conf, c("Team", "Conference"), "\\(")
  return(df)
}
```

Figure 2

After the school's name is in its column, I make a unique list of the team names and compare them to the school's names within the team statistics because there are discrepancies. Using the dplyr library, I use the gsub() function within the mutate() function to make sure all school names are the same within both datasets (I.e., NIU to Northern Illinois, Pennsylvania State to Penn State, etc.). Once this is finished, the new variables can be created within the team statistic dataset. This can be accomplished by using the function seen in Figure 3.

The last individual statistic used to create a new variable on the team statistic dataset is the ESPN Top 100 player list. This list is comprised of the top 100 high school seniors and their commitments. Since there is not a dataset for this but rather just lists online, I took notes by hand marking the number of incoming ESPN Top 100 players a school had. This data was then used to create a new, numerical variable that listed the number of ESPN Top 100 players a school had incoming.



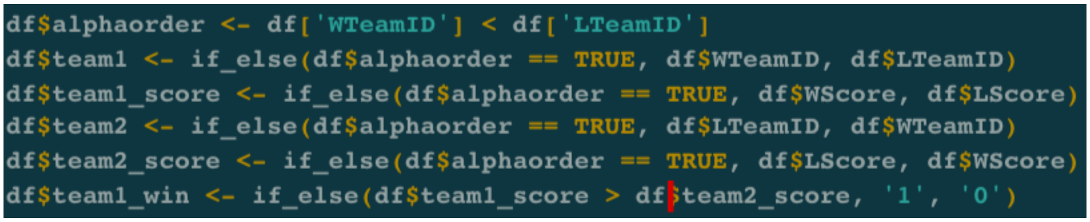
```
## df1 = points, df2 = assists, df3 = rebounds, df4 = steals, df5 = atratio, df6 = team statistic dataset from given year
top_individ <- function(df1, df2, df3, df4, df5, df6){
  # Make variable that includes all schools in NCAA Division 1
  nonindivid_teams <- unique(df6$School)
  # Make variable that includes all teams that have a top 100 scorer
  individ_teams <- unique(df1$Team)
  # Create variable within team statistic that gives a '1' if the school has a top 100 scorer, otherwise '0'
  df6 <- df6 %>%
    mutate(top100_scorer = ifelse(nonindivid_teams %in% individ_teams, '1', '0'))
  # Make variable that includes all teams that have a top 100 assister
  individ_teams <- unique(df2$Team)
  # Create variable within team statistic that gives a '1' if the school has a top 100 assister, otherwise '0'
  df6 <- df6 %>%
    mutate(top100_assists = ifelse(nonindivid_teams %in% individ_teams, '1', '0'))
  # Make variable that includes all teams that have a top 100 rebounder
  individ_teams <- unique(df3$Team)
  # Create variable within team statistic that gives a '1' if the school has a top 100 rebounder, otherwise '0'
  df6 <- df6 %>%
    mutate(top100_rebounds = ifelse(nonindivid_teams %in% individ_teams, '1', '0'))
  # Make variable that includes all teams that have a top 100 stealer
  individ_teams <- unique(df4$Team)
  # Create variable within team statistic that gives a '1' if the school has a top 100 stealer, otherwise '0'
  df6 <- df6 %>%
    mutate(top100_steals = ifelse(nonindivid_teams %in% individ_teams, '1', '0'))
  # Make variable that includes all teams that have a top 100 assist to turnover ratio player
  individ_teams <- unique(df5$Team)
  # Create variable within team statistic that gives a '1' if the school has a top 100 assist to turnover ratio player, otherwise '0'
  df6 <- df6 %>%
    mutate(top100_atratio = ifelse(nonindivid_teams %in% individ_teams, '1', '0'))

  return(df6)
}
```

Figure 3

Game Results Data

The game results data was pulled from two different sources. For the 2010-19 seasons, the data was pulled from Kaggle.com, the Google Cloud & NCAA® ML Competition 2019-Men's. The 2021-22 season game results data was pulled from masseyratings.com. The reason masseyratings.com was utilized is that Kaggle got their game results data from masseyratings.com as well. This allowed for the format to be identical without having to clean the data from 2010-19 (since Kaggle already cleaned the data here). The first thing that needed to be done to prepare the game results data for modeling was to substitute the arbitrary school numbers with the school names. This allows for the merging of the team statistic dataset, along with the newly created variables, via the school's name column in each dataset. After changing the school number to the school's name, I made new columns titled "Team1" and "Team2", where Team1's school name comes first alphabetically in the matchup. The reason for this is that if I left WTeamID and LTeamID, it would create perfect separation in the modeling process, not allowing for a logistic model to be created (it would always predict the WTeamID column to be the winning team based on the training set). This approach can be seen in Figure 4, pictured below. Upon completion, the team statistics dataset is merged with the game results data and is ready for the modeling stage.



```
df$alphaorder <- df['WTeamID'] < df['LTeamID']
df$team1 <- if_else(df$alphaorder == TRUE, df$WTeamID, df$LTeamID)
df$team1_score <- if_else(df$alphaorder == TRUE, df$WScore, df$LScore)
df$team2 <- if_else(df$alphaorder == TRUE, df$LTeamID, df$WTeamID)
df$team2_score <- if_else(df$alphaorder == TRUE, df$LScore, df$WScore)
df$team1_win <- if_else(df$team1_score > df$team2_score, '1', '0')
```

Figure 4

Modeling

The goal of the modeling process is to attempt a unique approach to choosing winners in each round using predication intervals and certain seeding matchups. My models will be fit using logistic regression, which is a method that predicts a binary outcome, in my case "winning team" or "losing team". By overfitting the rules from the 2010-19 seasons, I am creating a binary outcome for game matchups from the prediction intervals that are given. Albeit overfitting is not encouraged because of its high error rate, this will allow me to see if there are any patterns within the prediction intervals to explain when and why upsets happen in March Madness, an approach that I have not seen used before.

2010-19 Tournament & Rules

When it came to creating models to find rules for upsets and whom to choose in certain seeding matchups, I created four different models. The first model was a forward selection model that was scoped on the entire variable dataset (referred to as the full-forward model). The second model was a self-selected variable model to test the human element to pick the best variables at times (referred to as the self-selected model). The last two models utilized forward selection and backward elimination on the self-selected dataset (referred to as self-forward and self-backward models). These models were trained on the regular season and conference tournament games for the given year. After that, they were tested in the NCAA tournament. This is where I provided a 95% prediction interval on each game. The point of this was to find cutoffs within the prediction intervals where upsets would occur. Because of the uniquity of this approach, it was done by eyeballing the 95% prediction intervals to help determine rules to best identify upsets. Here are the rules that were formed:

- First Round
    - Take all 1 and 2 seeds
    - Take 14 seed over 3 seed if 14 goes over 40% in the self-backward model
    - Take 9 seed over 8 seed if 9 goes over 57% in the self-backward model
    - Take 12 seed over 5 seed if 12 goes over 34% in the self-backward model
    - Take 13 seed over 4 seed if 13 goes over 28% in the self-backward model
    - Take 11 seed over 6 seed if 11 goes over 39% in the self-forward model
    - Take 10 seed over 7 seed if 10 goes over 48% in the self-backward model
- Second Round
    - 1v16 winner vs 8v9 winner
        - Take 8 seed if it goes over 35% in the self-forward model
        - Take 9 seed if it goes over 28% in the self-forward model
        - Otherwise, take the 1 seed
    - 5v12 winner vs 4v13 winner
        - In 5v4, take 5 if it goes over 51% in the self-backward model
        - In 12v4, take 12 if it goes over 46% in the self-backward model
        - In 13v5 or 13v12, take 13 if it goes over 49% in the self-backward model
    - 6v11 winner vs 3v14 winner
        - In 11v3, take 11 if it goes over 29% in the self-forward model
        - In 6v3, take 6 if it goes over 54% in the self-forward model

        - In 14v11, take whatever team has better odds
        - In 14v6, take the 6 seed
    - 7v10 winner vs 2v15 winner
        - In 7v2, take 7 if it goes over 33% in the self-forward model
        - In 10v2, take 10 if it goes over 35% in the self-forward model
        - Otherwise, take the 2 seed
- Sweet 16 (Round 3)
    - Upper Half of Regions (1, 4, 5, 8, 9, 12, 13, and 16 seeds)
        - If any team playing a 1 seed goes over 36% in the self-backward model, take that team
        - Never take a 13 seed
        - Take 8 or 9 seed to win if they go over 39% in the self-backward model
    - Lower Half of Region (2, 3, 6, 7, 10, 11, 14, and 15 seeds)
        - Take the lower seed if they go over 50% in the self-forward model
        - Take 10 seeds over the 11 seeds
- Elite 8, Final 4, and Championship Game (Rounds 4-6)
    - Take better odds
        - Exception 1: If a seed is 2-4 seeds lower than the team they are facing, take the lower seed if they go over 50% odds at all in the self-backward model

- Exception 2: If a seed is 5+ seeds lower than the team they are facing, take the lower seed if they have over 40% odds at all in the self-backward model

The sole purpose of these rules is to determine how well the 2010-19 tournaments could help predict the 2022 tournament.

2022 Tournament Results

The approach to the 2022 tournament was to utilize the rules formed above to see the potential outcome of the tournament. Scoring for all brackets is done by first-round correct is 1 point, second round correct is 2 points, third round correct is 4 points, quarter final games correct is 8 points, semi-finals game correct is 16 points, and correct champion is worth 32 points. This allows for a maximum of 192 points. To provide a comparison to how the other four models performed, full-forward, self-selected, self-forward, and self-backward models, they were all predicted using a 0.5 cutoff. Figure 5 is the bracket made using the rules that were formed. Figure 6 shows the full-forward model. Figure 7 shows the self-selected model. Figure 8 shows the self-forward model. Figure 9 shows the self-backward model. Finally, Figure 10 shows the actual results from the NCAA tournament. In all of the Figures, highlighted school names are correct picks.

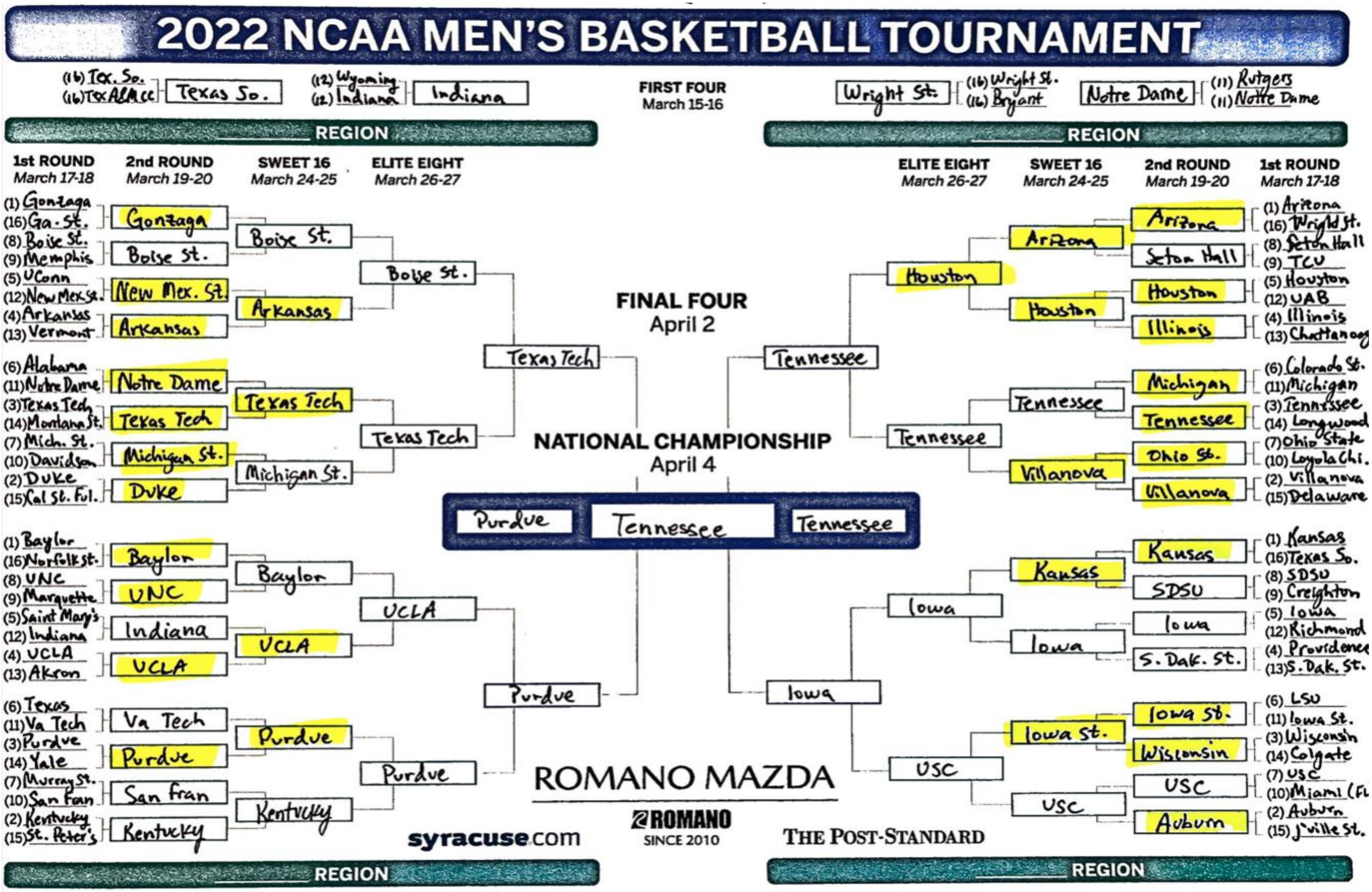


Figure 5: Rules Model, 44/192 points

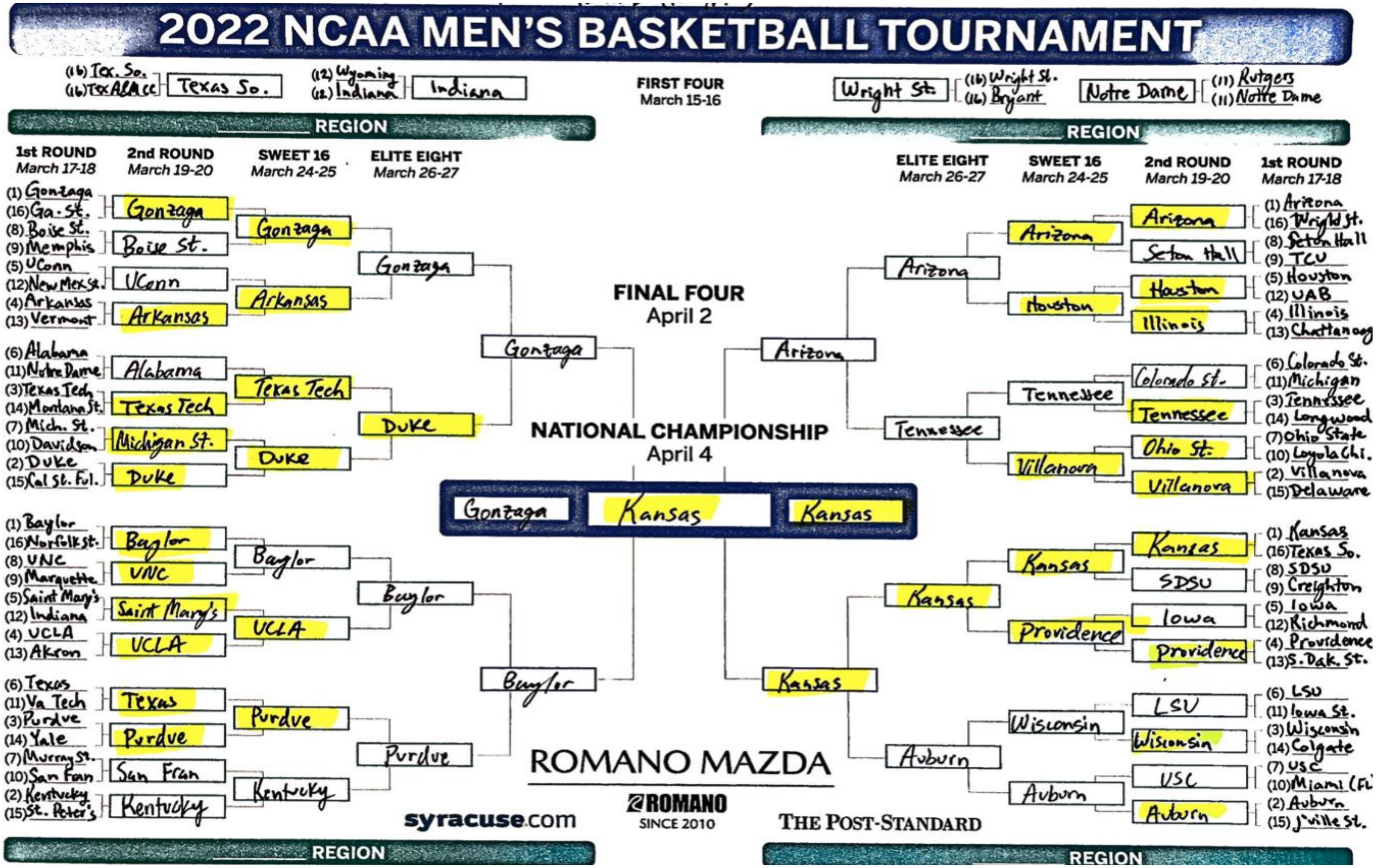
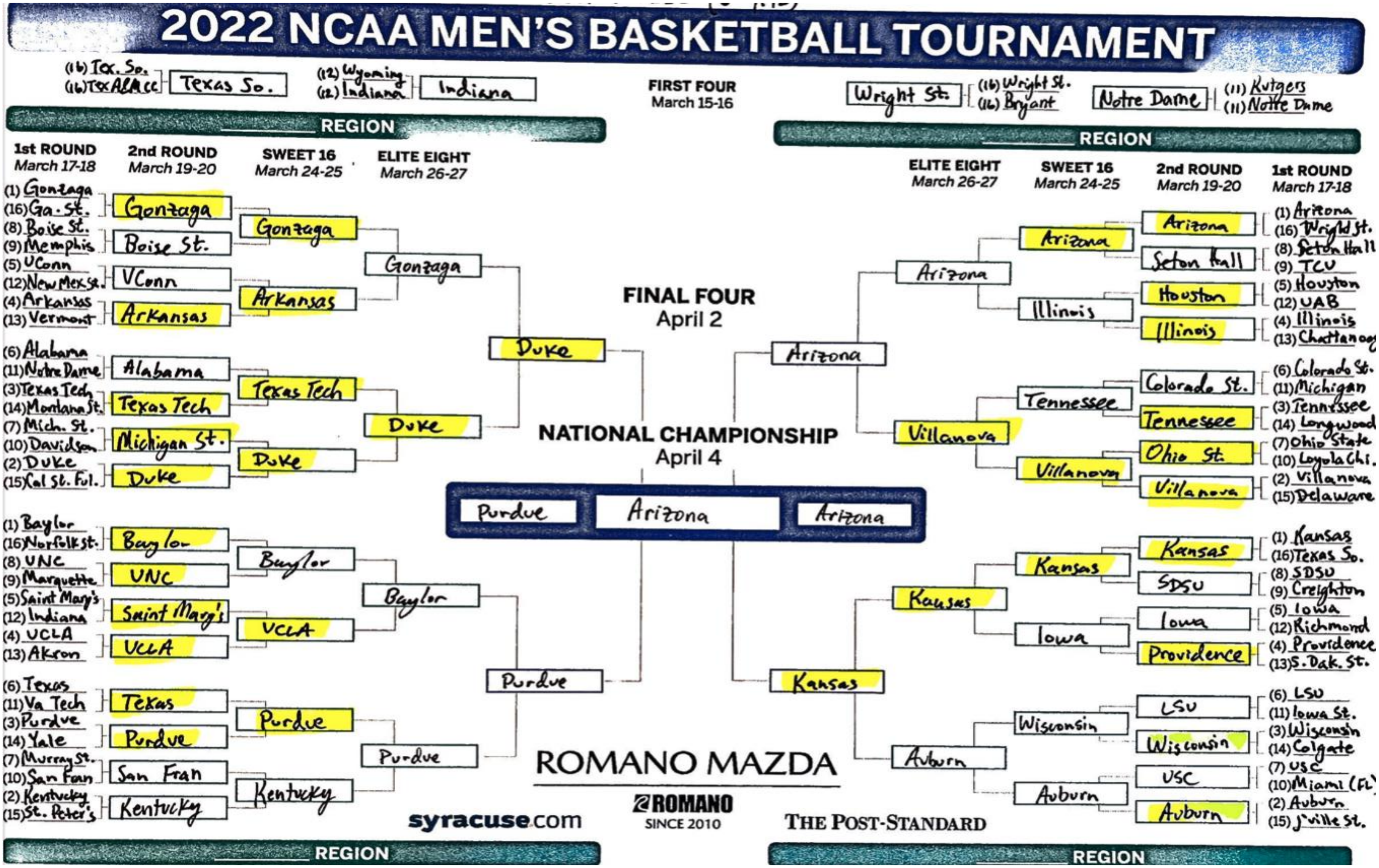


Figure 6: Full-Forward Model, 106/192 points



2022 NCAA MEN'S BASKETBALL TOURNAMENT

FIRST FOUR
March 15-16

(16) Tex. So.
(16) Tex A&M CC
Texas So.

(12) Wyoming
(12) Indiana
Indiana

(16) Wright St.
(16) Bryant
Wright St.

(11) Rutgers
(11) Notre Dame
Notre Dame

REGION

1st ROUND March 17-18
2nd ROUND March 19-20
SWEET 16 March 24-25
ELITE EIGHT March 26-27

(1) Gonzaga
(16) Ga. St.
(8) Boise St.
(9) Memphis
(5) UConn
(12) New Mex St.
(4) Arkansas
(13) Vermont
(6) Alabama
(11) Notre Dame
(3) Texas Tech
(14) Montana St.
(7) Mich. St.
(10) Davidson
(2) Duke
(15) Cal St. Ful.

Gonzaga
Boise St.
UConn
Arkansas
Alabama
Texas Tech
Michigan St.
Duke

Gonzaga
Arkansas
Texas Tech
Duke

Gonzaga
Duke

Duke

REGION

1st ROUND March 17-18
2nd ROUND March 19-20
SWEET 16 March 24-25
ELITE EIGHT March 26-27

(1) Arizona
(16) Wright St.
(8) Seton Hall
(9) TCU
(5) Houston
(12) UAB
(4) Illinois
(13) Chattanoog
(6) Colorado St.
(11) Michigan
(3) Tennessee
(14) Longwood
(7) Ohio State
(10) Loyola Chi.
(2) Villanova
(15) Delaware

Arizona
Seton Hall
Houston
Illinois
Colorado St.
Tennessee
Ohio St.
Villanova

Arizona
Illinois
Tennessee
Villanova

Arizona
Villanova

Arizona

FINAL FOUR
April 2

NATIONAL CHAMPIONSHIP
April 4

Purdue
Arizona
Arizona

(1) Baylor
(16) Norfolk St.
(8) UNC
(9) Marquette
(5) Saint Mary's
(12) Indiana
(4) UCLA
(13) Akron
(6) Texas
(11) Va Tech
(3) Purdue
(14) Yale
(7) Murray St.
(10) San Fran
(2) Kentucky
(15) St. Peter's

Baylor
UNC
Saint Mary's
UCLA
Texas
Purdue
San Fran
Kentucky

Baylor
UCLA
Purdue
Kentucky

Baylor
Purdue

Purdue

REGION

(1) Kansas
(16) Texas So.
(8) SDSU
(9) Creighton
(5) Iowa
(12) Richmond
(4) Providence
(13) S. Dak. St.
(6) LSU
(11) Iowa St.
(3) Wisconsin
(14) Colgate
(7) USC
(10) Miami (FL)
(2) Auburn
(15) J'ville St.

Kansas
SDSU
Iowa
Providence
LSU
Wisconsin
USC
Auburn

Kansas
Iowa
Wisconsin
Auburn

Kansas
Auburn

Kansas

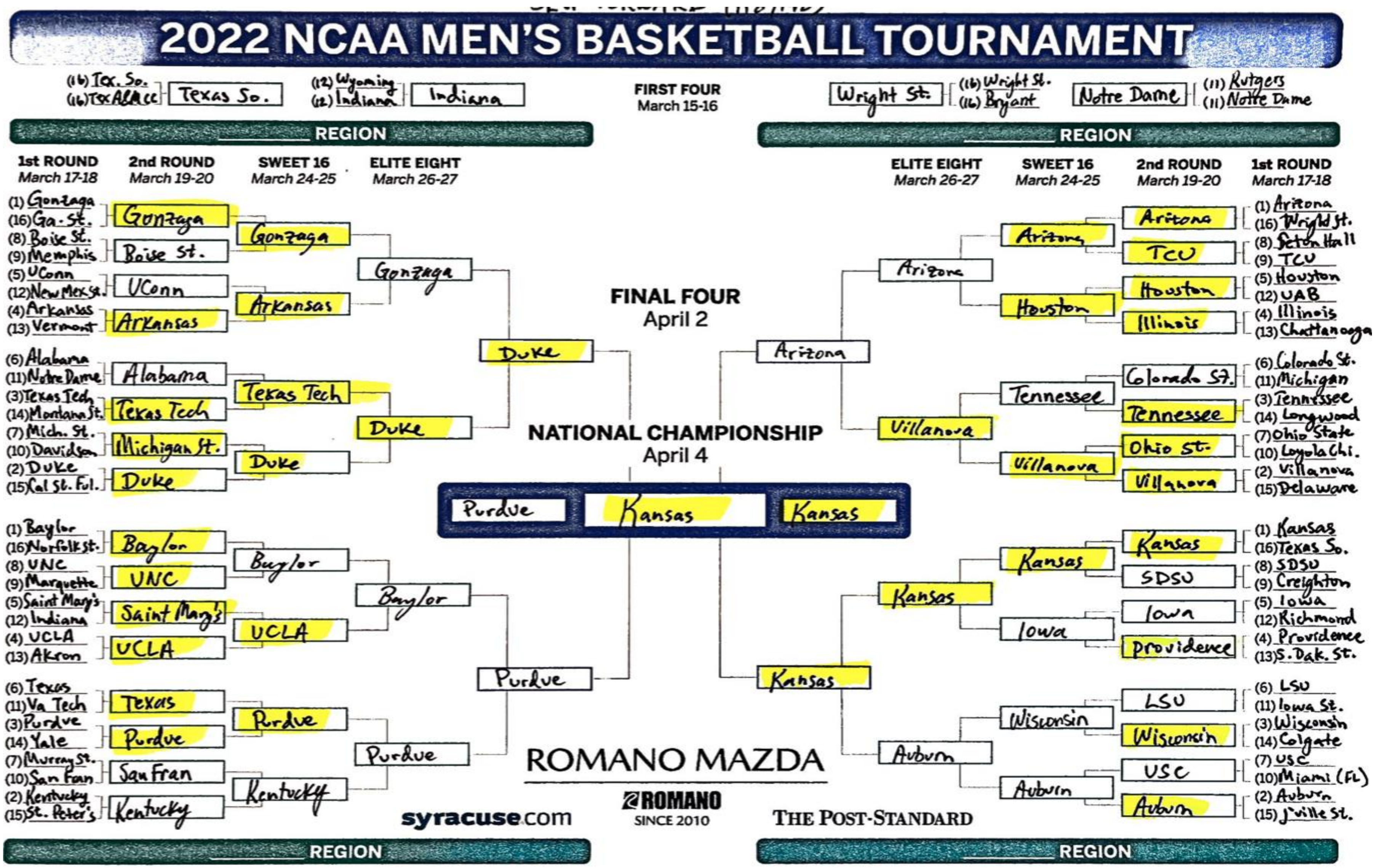
REGION

ROMANO MAZDA
ROMANO
SINCE 2010

syracuse.com

THE POST-STANDARD

Figure 7: Self-Model, 67/192 points



2022 NCAA MEN'S BASKETBALL TOURNAMENT

FIRST FOUR
March 15-16

(16) Tex. So.
(16) Tex A&M cc
Texas So.

(12) Wyoming
(12) Indiana
Indiana

Wright St.
(16) Wright St.
(16) Bryant

Notre Dame
(11) Rutgers
(11) Notre Dame

REGION

1st ROUND March 17-18
2nd ROUND March 19-20
SWEET 16 March 24-25
ELITE EIGHT March 26-27

(1) Gonzaga
(16) Ga. St.
(8) Boise St.
(9) Memphis
(5) UConn
(12) New Mex St.
(4) Arkansas
(13) Vermont
(6) Alabama
(11) Notre Dame
(3) Texas Tech
(14) Montana St.
(7) Mich. St.
(10) Davidson
(2) Duke
(15) Cal St. Ful.

Gonzaga
Boise St.
UConn
Arkansas
Alabama
Texas Tech
Michigan St.
Duke

Gonzaga
Arkansas
Texas Tech
Duke

Gonzaga
Duke

Duke

REGION

(1) Baylor
(16) Norfolk St.
(8) UNC
(9) Marquette
(5) Saint Mary's
(12) Indiana
(4) UCLA
(13) Akron
(6) Texas
(11) Va Tech
(3) Purdue
(14) Yale
(7) Murray St.
(10) San Fran
(2) Kentucky
(15) St. Peter's

Baylor
UNC
Saint Mary's
UCLA
Texas
Purdue
San Fran
Kentucky

Baylor
UCLA
Purdue
Kentucky

Baylor
Purdue

Purdue

REGION

ELITE EIGHT March 26-27
SWEET 16 March 24-25
2nd ROUND March 19-20
1st ROUND March 17-18

(1) Arizona
(16) Wright St.
(8) Seton Hall
(9) TCU
(5) Houston
(12) UAB
(4) Illinois
(13) Chattanooga
(6) Colorado St.
(11) Michigan
(3) Tennessee
(14) Longwood
(7) Ohio State
(10) Loyola Chi.
(2) Villanova
(15) Delaware

Arizona
TCU
Houston
Illinois
Colorado St.
Tennessee
Ohio St.
Villanova

Arizona
Houston
Tennessee
Villanova

Arizona
Villanova

Arizona

REGION

(1) Kansas
(16) Texas So.
(8) SDSU
(9) Creighton
(5) Iowa
(12) Richmond
(4) Providence
(13) S. Dak. St.
(6) LSU
(11) Iowa St.
(3) Wisconsin
(14) Colgate
(7) USC
(10) Miami (FL)
(2) Auburn
(15) J'ville St.

Kansas
SDSU
Iowa
Providence
LSU
Wisconsin
USC
Auburn

Kansas
Iowa
Wisconsin
Auburn

Kansas
Auburn

Kansas

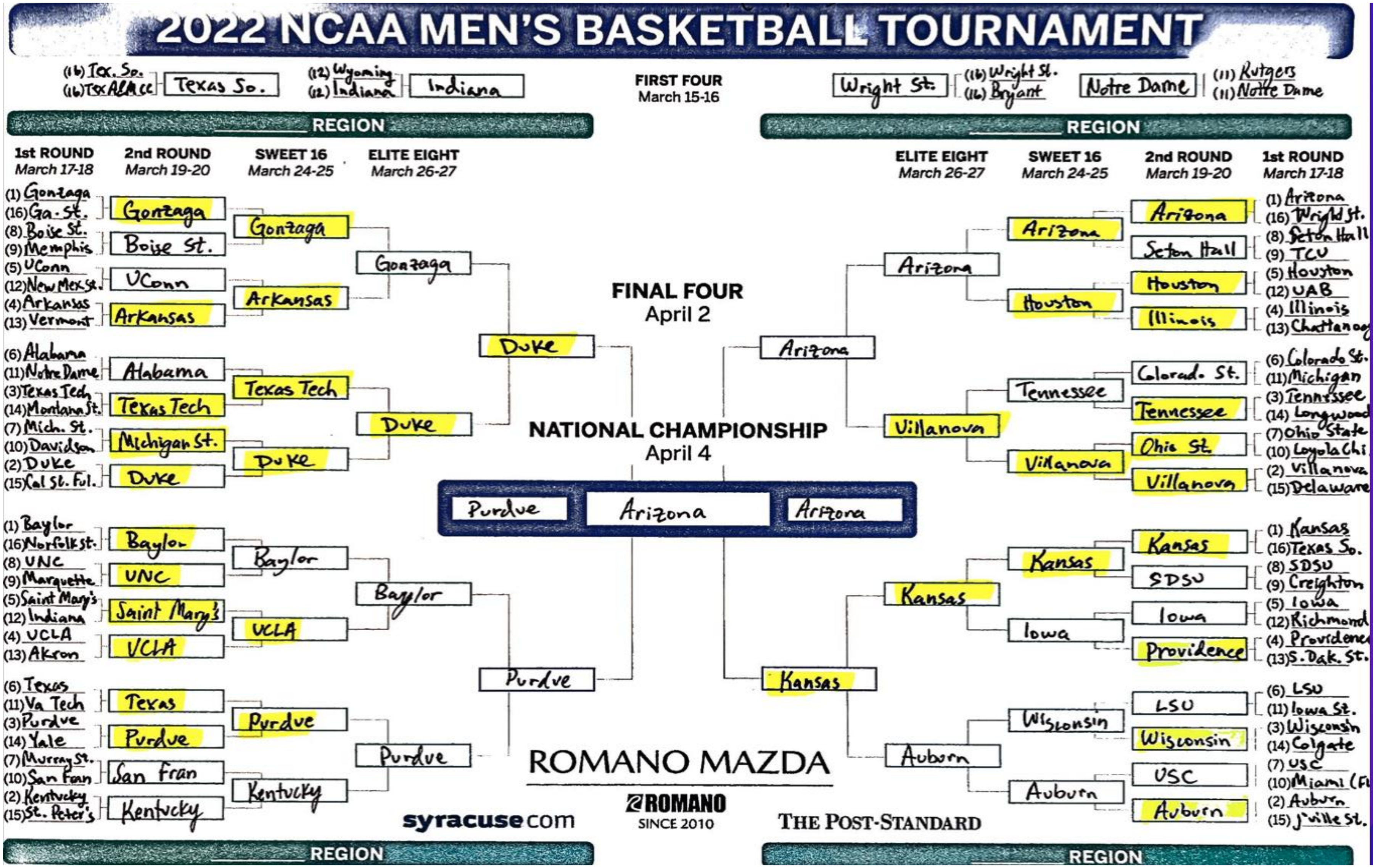
FINAL FOUR
April 2

NATIONAL CHAMPIONSHIP
April 4

Purdue
Kansas
Kansas

ROMANO MAZDA
ROMANO SINCE 2010
syracuse.com
THE POST-STANDARD

Figure 8: Self-Forward Model, 118/192 points



2022 NCAA MEN'S BASKETBALL TOURNAMENT

FIRST FOUR
March 15-16

(16) Tex. So.
(16) Tex A&M CC
Texas So.

(12) Wyoming
(12) Indiana
Indiana

(16) Wright St.
(16) Bryant
Wright St.

(11) Rutgers
(11) Notre Dame
Notre Dame

REGION

1st ROUND
March 17-18

(1) Gonzaga
(16) Ga. St.
(8) Boise St.
(9) Memphis
(5) UConn
(12) New Mex St.
(4) Arkansas
(13) Vermont
(6) Alabama
(11) Notre Dame
(3) Texas Tech
(14) Montana St.
(7) Mich. St.
(10) Davidson
(2) Duke
(15) Cal St. Ful.

2nd ROUND
March 19-20

Gonzaga
Boise St.
UConn
Arkansas
Alabama
Texas Tech
Michigan St.
Duke

SWEET 16
March 24-25

Gonzaga
Arkansas
Texas Tech
Duke

ELITE EIGHT
March 26-27

Gonzaga
Duke

Duke

REGION

ELITE EIGHT
March 26-27

Arizona
Villanova

SWEET 16
March 24-25

Arizona
Houston
Tennessee
Villanova

2nd ROUND
March 19-20

Arizona
Seton Hall
Houston
Illinois
Colorad. St.
Tennessee
Ohio St.
Villanova

1st ROUND
March 17-18

(1) Arizona
(16) Wright St.
(8) Seton Hall
(9) TCU
(5) Houston
(12) UAB
(4) Illinois
(13) Chattanoog
(6) Colorado St.
(11) Michigan
(3) Tennessee
(14) Longwood
(7) Ohio State
(10) Loyola Chi
(2) Villanova
(15) Delaware

Arizona

FINAL FOUR
April 2

NATIONAL CHAMPIONSHIP
April 4

Purdue
Arizona
Arizona

1st ROUND

(1) Baylor
(16) Norfolk St.
(8) UNC
(9) Marquette
(5) Saint Mary's
(12) Indiana
(4) UCLA
(13) Akron
(6) Texas
(11) Va Tech
(3) Purdue
(14) Yale
(7) Murray St.
(10) San Fran
(2) Kentucky
(15) St. Peter's

2nd ROUND

Baylor
UNC
Saint Mary's
VCLA
Texas
Purdue
San Fran
Kentucky

SWEET 16

Baylor
UCLA
Purdue
Kentucky

ELITE EIGHT

Baylor
Purdue

Purdue

REGION

Kansas

ELITE EIGHT

Kansas
Auburn

SWEET 16

Kansas
Iowa
Wisconsin
Auburn

2nd ROUND

Kansas
SDSU
Iowa
Providence
LSU
Wisconsin
USC
Auburn

1st ROUND

(1) Kansas
(16) Texas So.
(8) SDSU
(9) Creighton
(5) Iowa
(12) Richmond
(4) Providence
(13) S. Dak. St.
(6) LSU
(11) Iowa St.
(3) Wisconsin
(14) Colgate
(7) USC
(10) Miami (Fl
(2) Auburn
(15) J'ville St.

ROMANO MAZDA
ROMANO SINCE 2010
syracuse.com
THE POST-STANDARD

REGION

REGION

Figure 9: Self-Backward Model, 69/192 points

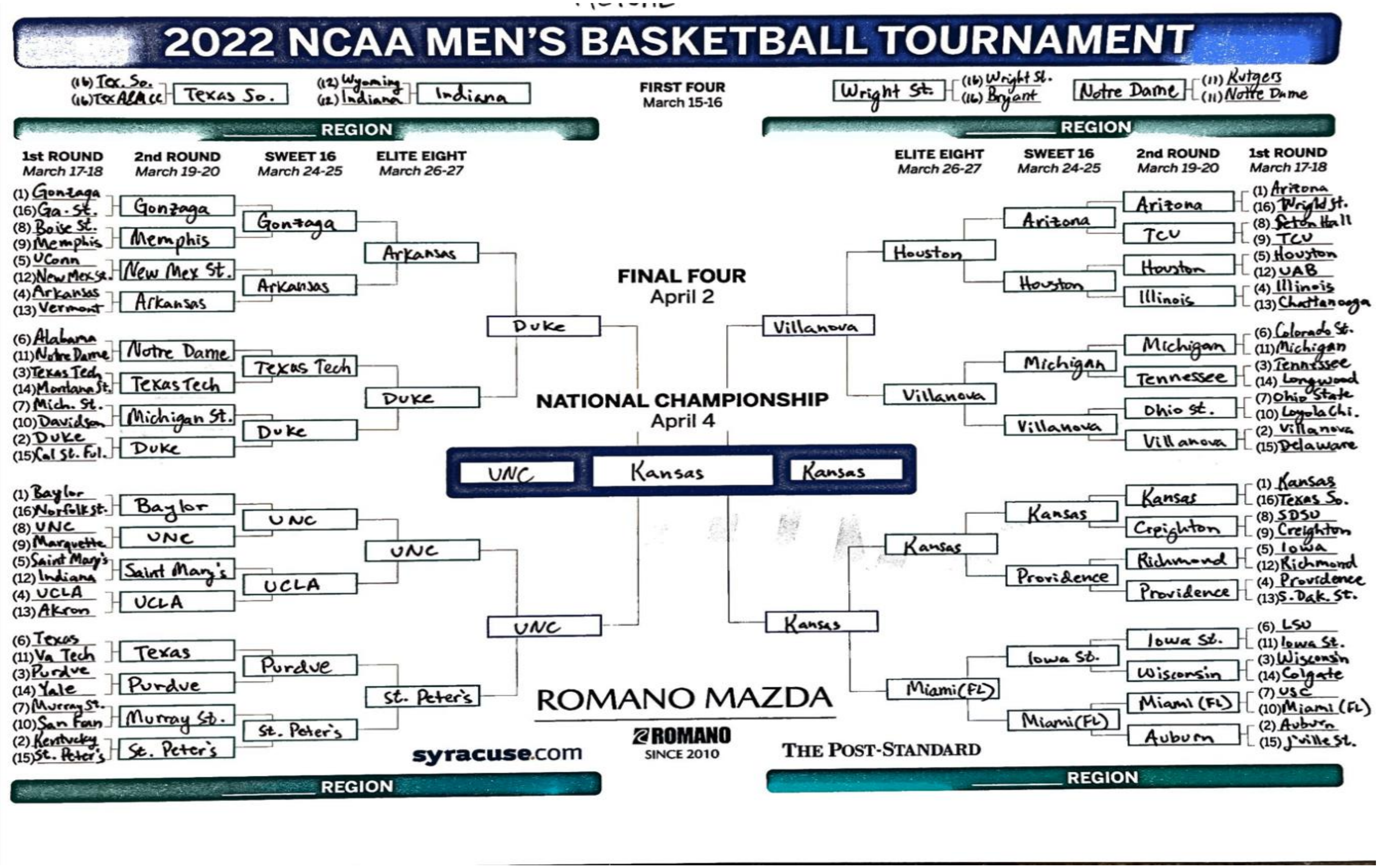


Figure 10: Actual Bracket

Results/Conclusions

The Rules bracket did the worst out of the five brackets shown above. This was interesting considering the goal of being able to predict upsets better, but it was not surprising. Since the rules were extremely overfitted for the 2010-19 tournaments, the risk of poor results was present. It was interesting that it predicted some upsets that none of the other brackets did, such as New Mexico State beating Connecticut, Notre Dame beating Alabama, Houston beating Arizona (and going to the elite 8), and Iowa State going to the Sweet 16. However, because of the greediness of the Rules, it also predicted some upsets that never happened, such as USC and Boise State going to the elite 8 when they both lost in the first round. It seems as if using rules for the first round (and perhaps the second round) may work out well. However, after the first weekend, the best teams left are the ones who win. There were a few outliers this year, including the first-ever 15 seed to the elite 8 in Saint Peter's and 8 seeded North Carolina to the championship game. Overall, upsets do seem to be predictable in a sense, but the later in the tournament, the harder it is to predict an upset.

If I were to make predictions on next year's tournament, I would use the rules on the first and second rounds before choosing the better teams from the sweet 16 through the championship. I believe that this would give the best result outside of pure luck. It would allow accounting for upsets that inevitably happen within the tournament. On top of that, it would account for the fact that highly ranked seeds are ranked where they are for a reason. I still believe that this approach may not be successful in any given year because of what seems to be an immeasurable upset factor.

Future Work

There is some future work that I did not get to within this project that could potentially help make predictions easier in the future. I believe that adding other variables that could account for experience on a team (multiply a player's minutes per game by their year in school and average it out for the team), win streaks, conference tournament finishes, injuries throughout the year, or in the tournament, and even how the transfer portal affects teams. On top of that, I would like to pursue other models besides logistic regression, such as decision trees and random forests.

Sources

“College Basketball Statistics and History: College Basketball at Sports.” *Reference.com,* https://www.sports-reference.com/cbb/.

“Google Cloud & NCAA® ML Competition 2019-Men's.” *Kaggle,* https://www.kaggle.com/c/mens-machine-learning-competition-2019/data.

Massey Ratings - Sports Computer Ratings, Scores, and Analysis, Ken Massey, https://masseyratings.com/scores.php?s=379387&sub=11590&all=1&mode=3&format=1.

Massey Ratings - Sports Computer Ratings, Scores, and Analysis, Ken Massey, https://masseyratings.com/scores.php?s=379387&sub=11590&all=1&mode=3&format=2.

NCAA Statistics, https://stats.ncaa.org/rankings/change_sport_year_div.