Final Project Step 2

Ramsey King

2021-05-15

Introduction

The problem that I have is that I am overly competitive. Every year, my family competes together in a bracket for March Madness, with the winner getting to be treated by the family to a lunch of their choosing. For many years, I have been the far and away winner of this pool (competition). This year, however, I was unsuccessful in winning the family pool. I would like to have my title back. In wanting to return to glory, I have decided that I will use exploratory data analysis, statistics, and R programming to see if patterns exists that will help me choose the Final Four teams of the men's basketball tournament. By consistently and correctly predicting the Final Four teams of the tournament, it will all but ensure that I continue my reign as the family basketball bracket picking champion.

Why would anyone besides me think that this is important? The ability to pick teams for the college basketball tournament has turned into a lucrative exercise. Even ESPN will award a winner $1 million if they are able to choose correctly every game during the tournament. By the way, there is a 1 in 9,223,372,036,854,775,808 chance of doing this, but it's greater than 0. By being able to choose the teams consistently that will reach the Final Four, there is real money to be had. If you do a Google search on "March Madness Bracket Contests," you will see grand prizes of cash money and trips to Hawaii and Las Vegas.

This is a data science problem because it includes a lot of probability, and there are many metrics that are kept with college basketball teams throughout the course of a season upon which statistical analysis can be performed.

Research questions

1. Is there a pattern that exists that will predict the Final Four Teams?
2. What are the most important metrics that will help determine the success of a college basketball team in the tournament?
3. What are some metrics that may not seem important initially but then prove to have statistical significance?
4. If a pattern exists, will it truly be causation, or is it just a random happening?
5. What is the smallest number of predictors that can produce the best model?

Approach

I will gather data from 1985-2016 or 2017 to see if a predictive pattern exists to select the Final Four teams.

How your approach addresses (fully or partially) the problem.

My approach will address the problem because I will have historical data from the last 20 years on basketball teams that have made the tournament and how they have fared. Based on the teams that have made the Final Four, there may be patterns that exists (number of wins, strength of schedule, ranking, etc.) that well help predict how future teams will perform in the tournament.

Data (Minimum of 3 Datasets - but no requirement on number of fields or rows)

Data sets that will be used are:

- Big_Dance_CSV
- ncaa-team-data
- NCAAInstitutions
- RPIStats

Big_Dance_CSV data set:

The Big Dance CSV data set was obtained from **Big_Dance_CSV.**

The original purpose of the Big Dance CSV data set was to provide "every NCAA tournament game result since 1985 (when the tournament was expanded to the 64 team bracket). The dataset contains the year, round (1-6), seed of the teams (1-16), region (1-4) and the scores."

ncaa-team-data

The ncaa team data data set was obtained from **ncaa-team-data**

The description of the data set is "I scraped data from sports-reference.com and made it tidy."

NCAA Institutions

The NCAA Institutions data set was obtained from Wikipedia. The data was collected and then put into a CSV file to be used for joining the other data sets together. The link to the Wikipedia file is **List of NCAA Division I institutions**

RPIStats

The RPIStats dataset was obtained from multiple sources and combined into one. For the years 1994 through 2017, the data was obtained from **College RPI**. For the years 1985 through 1993, the data was obtained from **RPI Archive**.

Because Big Dance CSV includes data going back to 1985 (which is the first season that the tournament went to 64 teams), I wanted to have RPI information going back that far as well, so both data sets will need to be joined together into one dataset.

Required Packages

The packages that will be needed are ggplot2, dplyr, QuantPsyc, readExcel. I am sure that there will be other packages needed, but I am not aware that I will need them at this time.

Plots and Table Needs

I will need to use histograms, scatterplots, and basic tables for to describe tabulated results.

Questions for future steps

I need to learn how to join tables together in R. I am not quite comfortable with that at the moment. I will also need to figure out how to create other plot types outside of histogram and scatter plots to help visualize the data and findings. As I get more involved in the project, I am sure that I will discover other things that I do not know yet that I will need to figure out.

Step 2

How to import and clean my data

There has been (what I consider) a LOT of cleanup and manipulating of the data to work my way to the final data set. First of all, I have changed all the csv files (Big_Dance_CSV, ncaa-team-data, NCAAInstitutions,

and RPIStats) to xlsx (Excel) files. The reason I have done this is because there is functionality in xlsx files that are easier to work with than csv files, and xlsx files "remember" formulas better (at least that is what I have experienced so far). A lot of this cleanup and manipulation has been done within Excel, but there will be some manipulation that will take place in R. I will describe the data prep for the specific csv/xlsx files.

Big_Dance_CSV data set: For the Big_Dancd_CSV data set, the following changes were made:

- changed the file extension from CSV to XLSX (as mentioned above)
- added the following columns:
  - TeamW,
  - TeamWSeed,
  - TeamWCommonName,
  - TeamWID,
  - TeamL,
  - TeamLSeed,
  - TeamLCommonName,
  - TeamWLID

The TeamW/TeamL columns were added to help identify which teams won the game played on the row and which team lost.

The TeamCommonName columns were added to be able to connect to the NCAAInstitutions data set.

The NCAAInstitutions data set will be the data set to identify Teams between all the other data sets.

ncaa-team-data

- added common name and TeamID columns to be able to tie to the NCAAInstitutions dataset.

RPIStats and RPI_Dataset The data from rpi95_04 and RPIStats were combined together into RPI_Dataset. The RPI_Dataset contains RPI (Ratings Percentage Index) information for all teams from 1985 through 2016. The Ratings Percentage Index is a metric used in college sports to help determine a team's relative team strength. It is a metric that is used in ranking and seeding the basektball teams for the single elimination tournament.

What does the final data set look like?

The final data set will be a collection of all the basketball teams that played in the NCAA single elimination championship tournament from 1985 to 2016 and will contain the following columns:

- CommonNameSchool (school name that will link to NCAAInstitutions dataset)
- W (Wins)
- L (Losses)
- WL (Winning %)
- srs (Sports Reference Statistic - give definition of this)
- sos (Strength of schedule)
- ncaa_result (How the team fared in the NCAA single elimination tournament)
- ncaa_numeric (a numerical representation of how the team did in the NCAA tournament)
- TeamID (A team ID nummber to link to the NCAAInstitutions dataset)
- Year
- Seed
- TeamW, TeamL (specific game info in big_dance_csv data set)
- TeamWSeed, TeamLSeed
- RPI
- OWP, OOWP (from 2003 through 2016)

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## # A tibble: 6 x 25
##   RPI_Rank   OWP  OOWP   RPI  Year TeamCommonName TeamID conf     rk     w     l
##      <dbl> <dbl> <dbl> <dbl> <dbl> <chr>           <dbl> <chr> <dbl> <dbl> <dbl>
## 1        1 0.620 0.567 0.652  1985 Michigan          173 Big ~    32    28     5
## 2        2 0.600 0.551 0.646  1985 St. John's        257 Big ~    32    31     5
## 3        3 0.594 0.594 0.629  1985 Georgetown        103 Big ~    32    24     8
## 4        4 0.592 0.556 0.615  1985 Duke               80 ACC      32    37     3
## 5        5 0.635 0.556 0.612  1985 Maryland          164 ACC      32    19    14
## 6        6 0.606 0.545 0.610  1985 Georgia Tech      105 ACC      32    27     7
## # ... with 14 more variables: wl <dbl>, srs <chr>, sos <chr>, pts_for <chr>,
## #   pts_vs <chr>, pts_total <chr>, ap_pre <dbl>, ap_high <dbl>, ap_final <dbl>,
## #   pts_diff <chr>, ncaa_result <chr>, ncaa_numeric <dbl>, season <chr>,
## #   coaches <chr>
```

Questions for future steps.

What information is not self-evident?

The information that is not self evident is if there is a pattern that exists that will predict the final four teams. I also may need to think about the seeding to make sure that I only choose teams that can make the final four (don't want two teams from the same region making the final four, that's impossible)

What are different ways you could look at this data?

I plan on doing scatterplots for all metrics that I am interested in against the teams that made the final four. For example, Wins vs. RPI, Wins vs. SOS, Wins vs. SRS, etc.

How do you plan to slice and dice the data?

One way that I can slice the data is that I can create a subset of the data set from 2003 forward. This will allow me to include the OWP & OOWP metrics in the predictions.

How could you summarize your data to answer key questions?

Whatever pattern exists, if any, that will be used to answer the five questions will be presented in this section.

What types of plots and tables will help you to illustrate the findings to your questions?

scatterplots, histograms, summary tables

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

I do not plan on incorporation any machine learning techniques at this time to answer my research questions. This may change as I get into the later portions of the course.