Final Project Step 1

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.