Research Class 1

December 9, 2017

The materials covered in this class are:

1. Describe the problem: predicting the winning probability of NCAA basketball playoff games based on the regular season performance of participating teams.
2. Introduce the codes we are going to use and set up a GitHub repository.
3. Propose an organization of the data we will collect.

1 Problem Descriptions

National Collegiate Athletic Association (NCAA) Division I (D-I) men's basketball tournament is held annually through March until early April. It has been one of the most famous annual sporting events, and generating annual revenues in hundreds of million dollars just from broadcasting. In this tournament, 68 teams play single elimination games in 7 rounds. Since 2011, in the first round, eight teams play in pairs to compete for four at-large seats called "first four". After that, 64 teams are divided into four different regions and six rounds of games will be played to decide the championship. A bracket of tracking game results in 2017 is shown in Figure 1.

Before the tournament begins, every team will play 30 to 35 regular season games. About 60% of the regular season games are played against an opponent within the same conference. After the regular season and before the NCAA tournament, conference championship tournaments will be played. The performance of regular season games and conference championship games will determine whether the team receive a bid to the NCAA tournament.

There is a massive interest in the sports gambling industry to create pools for people to submit their prediction of the entire bracket. It is generally impossible to predict the winners right for all 63 games, but it is possible to use data analytics tools to enhance the probability to get a bracket with higher scores. The scoring system of the bracket is described as follows. Getting the result of each first round game right is worth 10 points and that point will double for every next round. For the championship game alone it is worth 320 points. Within a pool the player with the highest total score will win some prize.

In our project, we collect the game data from all regular season and conference championship games from 2008 to 2017. We aim to train an analytics model using the data of 2008 to 2016 and verify its performance using the data of 2017. When it comes to prediction, we use the whole year's regular season data as independent variables. Given the name of 68 teams of a specific year, we aim to predict the winning probability of all possible ($\frac{68\times 67}{2} = 2278$) match-ups.

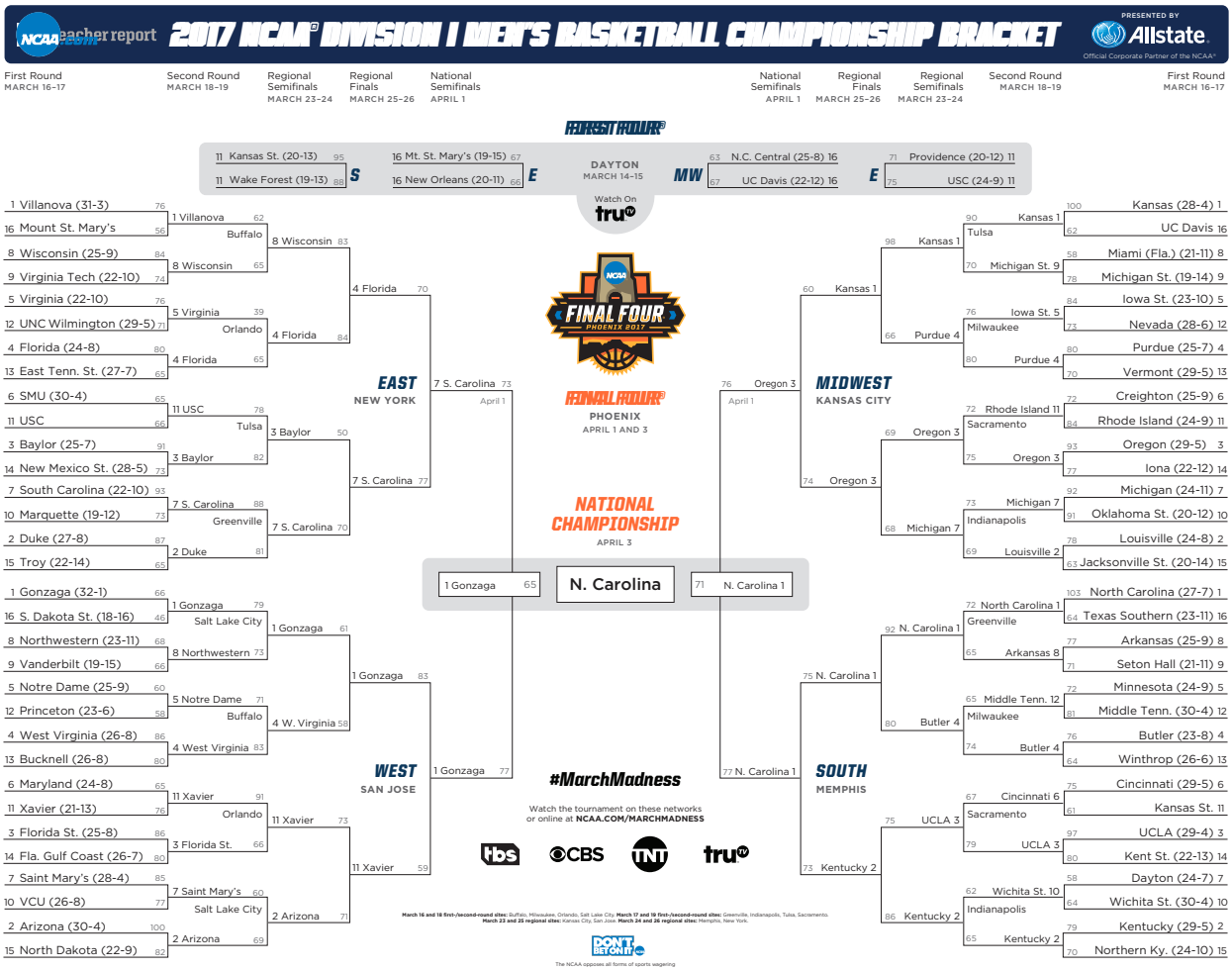

Figure 1: 2017 Bracket of NCAA D-I Men's Basketball Tournament

2 GitHub and Codes

We will use GitHub to share the codes and collected data. First you need to register a GitHub account on `www.github.com`. Git is a free and open source version control system for developers to collaborate in big coding projects. The tutorial of how to use Git can be found here: `https://www.atlassian.com/git`. A good IDE is very helpful for beginners to use Git. GitKraken is recommended for this purpose: `https://www.gitkraken.com/`.

The first action is to clone the Git repository I have created:`https://github.com/haoxiangyang89/NCAA-Analytics.git`. Once you make some changes to the files in the repository, follow the rules to check whether there is a conflict because sometimes I might be editing the same file as well. Pull and push will be the most used action on Git.

3 Data Organization

We will build spreadsheets using python codes. Some useful resources are listed here to understand the pattern of the html file of the website we are scraping from:

- Regular expression:
  - Comprehensive introduction: `https://www.regular-expressions.info/`;

  - Python RegEx syntax: `https://docs.python.org/3/library/re.html`;
  - Codes in the archive: Everyday_Run.py line 114 - 146.

- Beautiful soup: for structured xml files: `https://www.crummy.com/software/BeautifulSoup/bs4/doc/`.

The format of the spreadsheet should be as follows:

Team Name	Home	FG Made	FG Attempt	3PT Made	3PT Attempt	FT Made	FT Attempt	OREB
	DREB	REB	AST	STL	BLK	TO	PF	PTS

The data should be stored in the "Data" folder under the Git repository, in the format of .csv file. .csv file is a type of data format in which the text is separated by comma.