kaggle Host Competitions Datasets Scripts Jobs Community ▾ Rupak Chakraborty Logout

Completed • $15,000 • 248 teams

March Machine Learning Mania

Tue 7 Jan 2014 – Tue 8 Apr 2014 (21 months ago)

Competition Details » Get the Data » Make a submission

Data Files

File Name	Available Formats
regular_season_results	.csv (2.15 mb)
sample_submission	.csv (26.70 kb)
tourney_results	.csv (26.60 kb)
tourney_slots	.csv (20.32 kb)
seasons	.csv (1022 b)
teams	.csv (5.34 kb)
tourney_seeds	.csv (13.41 kb)

If you are unfamiliar with the format and intricacies of the NCAA tournament, we encourage reading its wikipedia page before diving into the data. The data description and schema may seem daunting at first, but it's not as complicated as it looks.

As a reminder, you are encouraged to incorporate your own sources of data. We have provided team-level historical data to jump-start the modeling process, but there is also a world of player-level and game-level data that may be useful.

We extend our gratitude to Kenneth Massey for his work gathering and providing the historical data.

What to predict

Stage 1 - You should submit predicted probabilities for every possible matchup in the past 5 NCAA tournaments.

Stage 2 - You should submit predicted probabilities for every possible matchup before the 2014 tournament begins.

Refer to the Timeline page for specific dates. In both stages, the sample submission will tell you which games to predict.

File descriptions

Below we describe the format and fields of the "essential" data files. Optional files may

Dashboard

Home
Data
Make a submission

Information
Description
Evaluation
Rules
Prizes
FAQs
Organizers
Timeline
Winners

Forum

Leaderboard
Public
Private

My Team
GitHub

My Submissions

Leaderboard

1. One shining MGF
2. Jason_ATX
3. zachtrexler
4. Frederocks
5. Nathan Weir
6. SJBeard
7. Adam Agata
8. KazAnova
9. EDDIEDUNKS
10. Yale Bulldogs

Forum (94 topics)

Please Explain How Submissions Are Scored
8 months ago

Printable bracket in R
9 months ago

Best Meta-Prediction?
10 months ago

End-of-competition thread
11 months ago

Removal of cheaters
21 months ago

Would you compete again?
21 months ago

teams

players

entries

be added to the data while the competition is running. You can assume that we will provide the essential files for the current season. You should not assume that we will provide optional files for the current season. To avoid confusion, we will keep the current season data (for stage 2) separate from the historical data (stage 1).

teams.csv

This file identifies the 356 different college teams that are present in at least one of the seasons from 1995-1996 through 2013-2014. The other data files that identify teams, such as when game-by-game results are listed or tournament seeds are listed, will always reference the teams by their id number rather than by their name. This makes the files more compact and also eliminates any possible spelling issues. This is the only file that actually contains the text names of the college teams.

- "id" - this number uniquely identifies the college team. All id values are three digit numbers, starting with 501, and the id's are assigned in alphabetical order, so for instance 501 is "Abilene Chr", and 502 is "Air Force", ... and 855 is "Youngstown State", the last team name alphabetically. The one exception to the alphabetical sequencing is Incarnate Word (#856), which was added recently. Having numbers so high avoids any possible confusion with game scores and ensures that all team id's will be exactly three digits long. There are four teams that are not present in the historical data, but nevertheless exist in the teams file: 501 (Abilene Chr), 609 (Grand Canyon), 656 (MA Lowell), and 856 (Incarnate Word). These three teams are present for the first time in the 2013-2014 data (the current season), and so they have already been assigned id numbers.
- "name" - this is the text name of the team, as determined in Kenneth Massey's historical game data. If a team name changed from one year to the next, then the new name is applied historically as well. For instance, the name of one college team used to be "MD Baltimore Co", and is now "UMBC", but since we know those are the same team, that is just represented as team 813, with a current team name of UMBC. Therefore, if you want to know what a team's results were in previous years, you don't have to look for different spellings of the team name, but instead you can just look for games played by that same team id in the previous years.

seasons.csv

This file identifies the 18 different seasons included in the historical data, along with certain season-level properties.

- "season" - indicates the letter used to uniquely identify each season. For instance, season Q represents the 2011-2012 season, meaning the college basketball season that started in late 2011 and ended with the final tournament during March/April 2012.
- "years" - indicates the years spanned by each season. For instance, you can see that season Q was played during the years 2011-2012.
- "dayzero" - tells you the date corresponding to daynum=0 during that season. All game dates have been aligned upon a common scale so that the championship game of the final tournament is on daynum=154. Working backward, the national semifinals are always on daynum=152, the "play-in" games are on days 134/135, Selection Sunday is on day 132, and so on. All game data includes the day number in order to make it easier to perform

date calculations. If you really want to know the exact date a game was played on, you can combine the game's "daynum" with the season's "dayzero". For instance, since day zero during the 2011-2012 season was 10/31/2011, if we know that the earliest regular season games that year were played on daynum=7, they were therefore played on 11/07/2011.

- "regionW/X/Y/Z" - by convention, the four regions in the final tournament are always named W, X, Y, and Z. Whichever region's name comes first alphabetically, that region will be Region W. And whichever Region plays against Region W in the national semifinals, that will be Region X. For the other two regions, whichever region's name comes first alphabetically, that region will be Region Y, and the other will be Region Z. This allows us to identify the regions and brackets in a standardized way in other files. For instance, during the 2012 tournament, the four regions were East, Midwest, South, and West. Being the first alphabetically, East becomes W. Since the East regional champion (Ohio State) played against the Midwest regional champion (Kansas) in the national semifinals, that makes Midwest be region X. For the other two (South and West), since South comes first alphabetically, that makes South Y and therefore West is Z. So for this season, the W/X/Y/Z are East,Midwest,South,West.

regular_season_results.csv

This file identifies the game-by-game results for all 18 seasons of historical data, from season A (1995-6) through season R (2012-3). Each year, it includes all games played from daynum 0 through 132 (which by definition is "Selection Sunday", the day that tournament pairings are announced). Each row in the file represents a single game played.

- "season" - this is the one-letter identifier of the season, corresponding to the "season" column in the "seasons.csv" file.
- "daynum" - this integer always ranges from 0 to 132, and tells you what day the game was played on. It represents an offset from the "dayzero" date in the "seasons.csv" file. For example, the first game in the file was daynum=16. Combined with the fact from the "season.csv" file that day zero was 10/30/1995, that means the first game was played 16 days later, or 11/15/1995. There are no teams that ever played more than one game on a given date, so you can use this fact if you need a unique key. In order to accomplish this uniqueness, we had to adjust one game's date. In March 2008, the SEC postseason tournament had to reschedule one game (Georgia-Kentucky) to a subsequent day, so Georgia had to actually play two games on the same day. In order to enforce this uniqueness, we moved the game date for the Georgia-Kentucky game back to its original date.
- "wteam" - this identifies the id number of the team that won the game, as listed in the "teams.csv" file. No matter whether the game was won by the home team or visiting team, "wteam" always identifies the winning team.
- "wscore" - this identifies the number of points scored by the winning team.
- "lteam" - this identifies the id number of the team that lost the game.
- "lscore" - this identifies the number of points scored by the losing team.
- "numot" - this indicates the number of overtime periods in the game, an integer 0 or higher. Note that this information is only available for season J (2004-5) or later. For seasons A thru I, all games will have numot=NA.
- "wloc" - this identifies the "location" of the winning team. If the winning

team was the home team, this value will be "H". If the winning team was the visiting team, this value will be "A". If it was played on a neutral court, then this value will be "N". Sometimes it is unclear whether the site should be considered neutral, since it is near one team's home court, or even on their court during a tournament, but for this determination we have simply used the Kenneth Massey data in its current state, where the "@" sign is either listed with the winning team, the losing team, or neither team.

tourney_results.csv

This file identifies the game-by-game NCAA tournament results for all 18 seasons of historical data, from season A (1995-6) through season R (2012-3). The data is formatted exactly like the "regular_season_results" data, except that all games are assumed to be on a neutral court, and therefore the "wloc" column is not included. Note that these games also include the play-in games (which always occurred on day 134/135) for those years that had play-in games.

tourney_seeds.csv

This file identifies the seeds for the final 64 teams in each NCAA tournament, for all 18 seasons of historical data. The losers of play-in games are not listed, though their games are included in the "tourney_results.csv" file. Therefore there are exactly 64 rows for each year, and a total of 64x18=1152 rows.

- "season" - this is the one-letter identifier of the season, corresponding to the "season" column in the "seasons.csv" file
- "seed" - this is a three-character identifier of the seed, where the first character is either W, X, Y, or Z (identifying the region the team was in) and the next two digits (either 01, 02, ..., 15, or 16) tells you the seed within the region. For example, the first record in the file is seed W01, which means we are looking at the #1 seed in the W region (which we can see from the "seasons.csv" file was the East region). This seed is also referenced in the "tourney_slots.csv" file that tells us which bracket slots face which other bracket slots in which rounds.
- "team" - this identifies the id number of the team, as specified in the "teams.csv" file.

tourney_slots.csv

This file identifies the mechanism by which teams are paired against each other, depending upon their seeds. Because of the existence of play-in games for particular seed numbers, the pairings have small differences from year to year. If there were N teams in the tournament during a particular year, there were N-1 teams eliminated (leaving one champion) and therefore N-1 games played, as well as N-1 slots in the tournament bracket, and thus there will be N-1 records in this file for that season. There were 64 tournament teams in seasons A-E, 65 tournament teams in seasons F-O, and 68 tournament teams in seasons P-R.

- "season" - this is the one-letter identifier of the season, corresponding to the "season" column in the "seasons.csv" file
- "slot" - this uniquely identifies one of the tournament games. For play-in games, it is a three-character string identifying the seed fulfilled by the winning team, such as W16 or Z13. For regular tournament games, it is a four-character string, where the first two characters tell you which round the game is (R1, R2, R3, R4, R5, or R6) and the second two characters tell

you the expected seed of the favored team. Thus the first row is R1W1, identifying the Round 1 game played in the W bracket, where the favored team is the 1 seed. As a further example, the R2W1 slot indicates the Round 2 game that would have the 1 seed from the W bracket, assuming that all favored teams have won up to that point. The slot names are different for the final two rounds, where R5WX identifies the national semifinal game between the winners of regions W and X, and R5YZ identifies the national semifinal game between the winners of regions Y and Z, and R6CH identifies the championship game. The "slot" value is used in other columns in order to represent the advancement and pairings of winners of previous games.

- "strongseed" - this indicates the expected stronger-seeded team that plays in this game. For Round 1 games, a team seed is identified in this column (as listed in the "seed" column in the "tourney_seeds.csv" file), whereas for subsequent games, a slot is identified in this column. In the first record of this file (slot R1W1), we see that seed W01 is the "strongseed", which during the 1996 tournament would have been Massachusetts. Whereas for games from Round 2 or later, rather than a team seed, we will see a "slot" referenced in this column. So in the 33rd record of this file (slot R2W1), it tells us that the winners of slots R1W1 and R1W8 will face each other in Round 2. Of course, in the last few games of the tournament - the national semifinals and finals - it's not really meaningful to talk about a "strong seed" or "weak seed", but those games are represented in the same format for the sake of uniformity.
- "weakseed" - this indicates the expected weaker-seeded team that plays in this game, assuming all favored teams have won so far. For Round 1 games, a team seed is identified in this column (as listed in the "seed" column in the "tourney_seeds.csv" file), whereas for subsequent games, a slot is identified in this column. So in the first record of this file (slot R1W1), we see that seed W16 is the "weakseed". So we can look this up in the "tourney_seeds.csv" file to see that in season A, the W16 seed was team 809, which we can then find (from the "teams.csv" file) to be UCF. Thus we know that this game was Massachusetts against UCF during the 1996 tournament. And just like the "strongseed" column, for games from Round 2 or later, rather than a team seed, we will see a "slot" referenced in this column.