**Objective**

To predict the outcomes of this year's US men's college basketball tournament using a combination of rich historical data and computing power. This competition consists of two stages. For this project, we have chosen to work on the first stage.

In the first stage of the competition, Kagglers will rely on results of past tournaments to build and test models. Using the constructed models the objective is to predict probabilities for every possible matchup in the past 4 NCAA tournaments (2013-2016).

**Expected Output**

List of every possible matchup between the tournament teams. team1 vs. team2 is the same as team2 vs. team1. For example, in a tournament of 68 teams (64 + 4 play-in teams), you will predict (68\*67)/2 = 2278 matchups.

Each game has a unique id created by concatenating the season in which the game was played, the team1 id, and the team2 id. For example, "2013\_1104\_1129" indicates team 1104 played team 1129 in the year 2013. Goal is to predict the probability that the team with the lower id beats the team with the higher id.

The resulting submission format looks like the following, where "pred" represents the predicted probability that the first team will win:

**id,pred**

2013\_1103\_1107,0.5

2013\_1103\_1112,0.5

2013\_1103\_1125,0.5

**3. Experimental Analysis**

One of the important phases of the project is to come up the ideal set of predictors that give us the best result. In layman’s terms, from the data provided by Kaggle, we need to be able to identify what are the indicators for any given team winning a game in March Madness.

Based on our initial discussions, we have identified the following as the key indicators that capture if a team will do well in each year’s NCAA tournament:

(Please note that all indicators are specific to each season)

* Win percentage of the team in Regular Season.
* Seeding in the NCAA tournament.
* Average point difference with which they won their games in the Regular Season.
* Average point difference with which they lose their games in the Regular Season.
* Home games win percentage in Regular Season.
* Away games win percentage in Regular Season.
* Their record in the last 10 games of the Regular season (Shows the teams momentum coming into the NCAA Tournament).

These are just an initial set of indicators, and more could be added or removed. Our objective is the use experimental analysis to identify the ideal combination of indicators that give us the best model.

Also, please understand that these indicators only need to be calculated for the teams that made it to the NCAA tournament that season. There is no point calculating the indicators for teams that didn’t make the tournament that year as they would not be helpful in training our model. In addition, the model will be trained only on the matchups that happened in that year’s NCAA tournament.

After arriving at the best model, we can start to predict probabilities for every possible matchup in the past 4 NCAA tournaments (2013-2016).

**Example:** Predicting probabilities for every possible matchup in 2013

This will be the testing or validation phase where we use the trained model to predict outcomes or tune parameters for our model. We must create the above-mentioned indicators for all the teams based on their 2013 Regular Season performance. We will then pass these indicators for every possible matchup through our trained model. The model will predict the winning team when a team with certain type of indicators plays a team with a certain type of indicators. The model can do this because it has been trained using historical data, where its might have seen 2 teams with similar Regular season indicators and it knows who won when they met in the NCAA tournament.

**Calculating the Indicators per Season**

|  |  |
| --- | --- |
| **Indicator** | **How to calculate?** |
| Win percentage of the team in Regular Season | RegularSeasonCompactResults.csv can be used to calculate the win percentage of a team during each regular season. This csv captures all the matchups between teams in across the 32 divisions in the Regular season. |
| Seeding in the NCAA tournament. | TourneySeeds.csv captures the seeding of each team in their respective region. Each region consists of 16 teams and this is the starting point of the NCAA tournament. There are 4 regions. |
| Average point difference with which they won their games in the Regular Season / Average point difference with which they lose their games in the Regular Season. | RegularSeasonCompactResults.csv has the fields Wscore (winning teams score) and Lscore (Losing teams score). This can be used to calculate the average point difference they with which they win or lose by for each Regular Season. |
| Home games win percentage in Regular Season / Away games win percentage in Regular Season. | RegularSeasonCompactResults.csv has a field Wloc (Win Location). H is home, A is Away and N is Neutral. This field will help us in calculating this indicator. Sometimes it is unclear whether the site should be considered neutral, since it is near one team's home court. |
| Their record in the last 10 games of the Regular season | This also can be calculated from RegularSeasonCompactResults.csv, by just taking the last 10 records for each season per team. |

**4. Coding Language / Technique to be used**

We will be using Spark to process our data, build our indicators and train our model. Scala will be language of choice. We will be heavily reliant on Spark’s MLlib to perform our predictions and derive conclusions.

As for the machine learning techniques to be used, we are planning to train our data using the following classifiers:

* Random Forests
* Decision Trees
* SVM
* GLM

We may have to limit our choice of classifiers based on the time constraint. But we aim to be comprehensive and see which classifiers and indicators perform best.