Data-Driven March Madness

1 Background

March Madness is one of the most exciting times in sports. There are thousands of tournaments and millions of people trying to predict the impossible perfect bracket, most of which are built using qualitative basketball knowledge. However, there are large amounts of historical data out there that may provide greater insights on the probability that a team will win in the tournament. In addition, it is vital that the bracket is optimized according to the scoring mechanism of one's tournament. Can one build a purely data-driven model to build a bracket that not only predicts winners, but also optimizes to the scoring mechanism?

2 Milestones

2.1 Project Selection

Form teams of 2 or 3 and select a project from the provided list.

2.2 Literature Study

Go through the following resources and write down a half-page to full page of notes about the different approaches each model uses:

1. Kaggle page for yearly march madness competition with rules/descriptions/forums discussing strategies:
   `https://www.kaggle.com/c/march-machine-learning-mania-2016`
2. Harvard Sports Analysis Club released a potential strategy they developed:
   `https://harvardsportsanalysis.wordpress.com/2011/05/18/quantifying-intangibles-a-network-analysis-prediction-model-for-the-ncaa-tournament/`
3. Nate Silver, creator of the website 538, builds a model every year and releases notes about how it is built:
   `http://fivethirtyeight.com/features/how-fivethirtyeight-is-forecasting-the-2016-ncaa-tournament/`
4. Brady T. West, A Simple and Flexible Rating Method for Predicting Success in the NCAA Basketball Tournament:
   `http://www-personal.umich.edu/~bwest/final_version_inpress.pdf`

2.3 Data Exploration and Cleaning

1. The main dataset for this project is the history of games played between 1985-2015 found on the Kaggle website:
   `https://www.kaggle.com/c/march-machine-learning-mania-2016/data`
2. Additionally here is a Github project by a fellow Harvard undergraduate that scrapes additional information from NCAA results for SportsReference:
   `https://github.com/mdgoldberg/sportsref`
3. Here is another Github project that scrapes boxscores from ESPN:
   `https://github.com/pathow/NCAA_MBB`
4. Here is an additional Github project that has capabilities for building individual player profiles:
   `https://github.com/rodzam/ncaab-stats-scraper`

Perform the following exploration steps:

1. Decide on a suitable database to store the data, and on a computing resource to process the data.
2. Look through the various data resources and determine the pros/cons of each.
3. Think about and plan various ways you could model the data and the tradeoffs that each would provide. i.e. on the team level, player level, general model for a team, general model for a basketball program (meaning a model for Kansas, Duke, Uconn, etc...)

2.4 I5 Proposal

Propose methodologies and ideas to be implemented, tested and interpreted for your final project.

Implement Baselines:

1. Decide on a *performance metric* to evaluate prediction. Consider the scoring structure of a march madness bracket, and the discussions on Kaggle.
2. Decide on how you will train/test your data by splitting it up to avoid overlapping training/testing data.
3. Do some *feature extraction*. Sift through all of the data that is available and extract features that you believe will impact a teams chances in the tournament. Start with basic statistics such as score differential, strength of schedule, etc... Then try to think about some other clever ways you could gather more features thinking about things like momentum, depth of a team, injuries, etc...
4. Implement the following baseline techniques:
   (a) Start with a model that chooses the winner based off of the pre-defined NCAA ranking seed.

(b) Train a linear regression model with basic features from the datasets (mainly a score-based ranking model).

(c) Train a linear regression model with basic and additional features, discussed in item number 3.