Duke University
March 7th, 2017
Team 14: Mary Ziemba, Melanie Krassel, Alex Deckey, David Duquette, Camila Vargas
CS 216 - Proposal write up

MMM - March Madness Mania

It's that time of year where everyone turns their attention towards college basketball. The NCAA Men's Basketball Tournament takes place every year in March, pitting the top 68 teams in the country against each other in a whirlwind two-weeks of upsets, buzzer-beaters, and Cinderella stories. The excitement around the tournament and its seemingly unpredictable nature has led many around the country to search for the perfect bracket. Warren Buffett has even offered $1 Million to the first person to perfectly predict the tournament. Unfortunately, there are over 9.2 quintillion possible brackets to be made. Looking at historical data we want to attempt to predict the outcomes of this year's tournament.

We will explore several questions as we work with historical data on regular season results and on tournament results. Some of these questions include: Which are the factors that best predict individual games? What are the key factors that best predict upsets specifically? What factors have been for and against Duke in some key games? Can we determine if a team is "hot" and will that contribute to our predictive model? These questions, and others that will come up as we explore the data, will guide our project.

To explore all these questions, we will use a group of data sets that have been published on Kaggle by Kenneth Massey. The files include regular season results, tournament results and tournament slots. In addition to that we will scrape ancillary data.

- **Regular Season Detailed Results:** We will be working with data from regular-season games, which determine a team's seed in the tournament. The data includes includes the day a game was played, the winning and losing teams, the location of the game, and game statistics--including attempted and successful field goal attempts, rebounds, assists, and other game-related statistics.
- **Tourney Detailed Results:** This dataset includes the same data described above, except for games in the NCAA tournament.
- **Tourney Seeds**: This file identifies the seeds for all teams in each NCAA tournament, for all seasons of historical data.
- **Tourney Slots:** This file identifies the mechanism by which teams are paired against each other, depending upon their seeds. It codifies each game in the tournament; for example, the second-round West region game played between the winner of the first round games between the #1 and #16 and #8 and #9 teams is codified as R2W1.
- **Various ancillary sources:** We would like to explore sources of data that are ostensibly less-related to the outcome of a basketball game--the size of the student population, the distance of the school to the "neutral" site where the game is played, the percentage of

seniors on the team, the distance the team traveled to play the game, and other sources that may become relevant as we explore the data.

There are multiple mechanisms we will be using to evaluate the project's success. Because there will be an actual outcome for March Madness, we can compare our predictions to the real outcomes. Additionally, we can use historical data and compare that to past outcomes (as we will be using hold-out data, so our training sets for prediction will leave out some years so that we can test it against those years).

Another mechanism we can use is comparing our project to similar projects that other Kagglers are working on simultaneously. There is a competition for predicting what the bracket itself will be, so there will be a vast number of related projects to compare against. You can learn more about the larger "March Machine Learning Mania 2017" competition being held by Kaggle here.

The steps to solve our problem

1. Collect ancillary data (such as number of freshmen in the team).
2. Set up sql databases based on the csv files that we have.
3. Explore the historical data from 2003 to 2016 by looking into the variables we have, finding meaningful ranges, visualizing and really getting to know our numbers.
4. Start exploring prediction models.
5. As the season plays out, compare real results from this year to the past and to our expectations, based on our predictive model.