Vincent Sheu, Winnie Wong, Rosie Abe, J.P. Saunders, and Li Lin
Statistics 133, Spring 2012 (Ibser/Kuang)
04.16.2012
Final Project: Prospectus

There are 345 schools distributed among 32 conferences in Division I of the governance structure established by the National Collegiate Athletics Assocation (NCAA). Every school year during a season spanning the fall, winter, and spring quarters and/or semesters, these schools play biweekly games against other schools. In conjunction with the rise of college basketball in prominence and stature, an industry for sports gambling has also arisen. In the modern age, sports books based in Las Vegas (and elsewhere) have evolved to publish lines on virtually every game occurring on American soil. Legality issues aside, office and other private "pools" predicting the outcome of the NCAA Division I Basketball Tournament, also known as "March Madness", are thriving. The tournament itself, and all the advertising and other associated revenue streams, combine to form the NCAA's single largest source of income, generating over $1 billion in annual profits.

As Vegas has risen, so have computer models ranking teams and predicting the outcomes of games. The prior is relatively easy; in fact, the NCAA has itself published a crude formula known as the Rating Percentage Index (RPI), which takes into account a team's winning percentage (25%), opponents' winning percentages (50%), and those opponents' opponents' winning percentage (25%). Minor adjustments are made based on location of games and other factors. Independently, statisticians such as Ken Pomeroy and Jeff Sagarin have developed sophisticated computer models to analyze basketball. These and other models claim to be more accurate at ranking than the RPI, as well as predict the future outcomes of games – something the RPI does not profess to do.

Generally, two forms of betting occur: the moneyline and the spread. In bets involving a moneyline, prospective clients select who they believe to be the winning team. The odds are quoted such that underdog triumphs result in a higher payout; favorite triumphs result in a lower payout. In bets involving the spread, the quoted information for a game is the "line". Bets are made on either side of the line; if a client's bet wins, s/he effectively doubles his/her money; otherwise s/he loses the money placed on the bet. This second form is the prediction many statistical models strive to beat.

Because there is less data to analyze at the beginning of the season, models tend to be less accurate; therefore, we will look at games only occurring after January 1 of each season we have data for. In addition, the new year tends to mark the beginning of conference play, in which teams from conferences exclusively play other teams of the same conference, giving models a plethora of connected data from which to analyze and predict results for.

Our project seeks to analyze the predictive power of different statistical models. Specifically, we will be looking to answer the following questions:

- How accurate is the Vegas line? What % of games does it correctly predict the winner for?
- How often do any of the prediction models for which we have significant amounts of data for beat the spread? That is, how often do any of these engines correctly predict the outcome of games in relation to the quoted line?
- (Time permitting) Are any schools perennially overrated or underrated by Vegas? That is, are there any schools who consistently fall on one side of the quoted line?

To aid our analysis, we will scrape data off of ThePredictionTracker.com, an online web resource run by Todd Beck. The Prediction Tracker has data for all NCAA Division I Men's Basketball Games from the 2007-2008 season to the present (http://www.thepredictiontracker.com/basketball.php), tracking home/away teams, final scores, lines, and predictions from several major predictive models such as Pomeroy, Sagarin ELO, Sonny Moore, and Sagarin Predictive.

Data has been archived as comma separated value-type files; therefore, we can use R's read.csv function to import data for analysis. In order to answer our main questions, we will need to convert the

Vincent Sheu, Winnie Wong, Rosie Abe, J.P. Saunders, and Li Lin
Statistics 133, Spring 2012 (Ibser/Kuang)
04.16.2012
Final Project: Prospectus

score results to single integers, which we can then compare to each of the numbers given by the predictive models. To answer the first two clusters of questions, we will need to create a data frame ordered by game index number (numeric), providing the result (numeric), whether Vegas correctly predicted the winner (logical), what the Vegas line was (numeric), what each Prediction model's prediction was in relation to the Vegas line (numeric), whether each Prediction model's prediction successfully predicted the outcome of the game in relation to the Vegas line (logical). To answer the (time permitting) last cluster of questions, we will need a set of variables representing the rating of each school. For every result where a school beats the Vegas line, we can increase that school's variable; for every loss, we can decrease that school's variable.

Currently, we plan to generate the following plots:

- For each year, a scatterplot, where the x-axis is the game index number, the y-axis is score differential, and colored points indicate different predictions from different models, as well as the final result. This graph can be plotted using the data of the main data frame.
- For each year, a graph plotting accuracy of predictions when compared to the Vegas line as a function of time (in our world, game index number). This graph can also be generated off the main data frame, by a function that calculates accuracy and plots it immediately.
- (Time permitting) If we do the last question, we are thinking of generating a Google Earth representation, with placemarks coded to individual school locations. Overratings and underratings could be represented by different placemark types. We are very close to throwing this idea out, because the only sources of data which may have latitude/longitude locations for Universities of the United States are either behind paywalls, or part of massive enough data dumps to be consistently crashing computers (650 Mb+ sized files).

Tasks are preliminarily split up as such:

- Inception: Vincent, Winnie, Rosie
- Initial Research : Vincent, Winnie, Rosie, JP
- Prospectus: Vincent
- Data verification, initial importing of data and organization: Li, JP, Winnie (Sunday, April 22)
- Generation of final data tables: Vincent, Rosie, Winnie (Sunday, April 29)
- Graph Visualizations:, Vincent, Winnie (Tuesday, May 1)
- Google Earth Representation: JP, Li, Rosie (Tuesday, May 1)
- Final Report Writing: Vincent, Rosie, Winnie (Wednesday, May 2)
- Presentation: All

The following challenges are immediately present: the sheer amount of data present is breathtaking. 345 schools are in Division I basketball. Each one plays around 35 games in a season, over half of which take place after January 1st. This translates to thousands of games a season, multiplied by five seasons worth of data. We will try to verify as much data as reasonably possible, but most of our project will probably be based on numbers that we assume to be correct. Our data comes from five different sources, and we will need to organize and manage everything down to a reasonable format we can work with.

The "time permitting" section will also prove extremely challenging; if we choose to do this section, we will have to organize data by school, producing a 345-length list of school records across multiple seasons. Schools may be incorrectly spelled, or represented in other ways entirely; for example,

Vincent Sheu, Winnie Wong, Rosie Abe, J.P. Saunders, and Li Lin
Statistics 133, Spring 2012 (Ibser/Kuang)
04.16.2012
Final Project: Prospectus

our institution legally has assumed the names University of California, Berkeley; University of California at Berkeley; Berkeley; California; University of California; UC Berkeley; Cal; and Cal Berkeley alone. In athletic competitions, we play as the University of California Golden Bears, sometimes shortened to California Golden Bears, Cal Golden Bears, or Cal Bears. Reputable outside sources have also referred to our school as UCB; Berkeley University; University of Berkeley; Cal-Berkeley; University of California-Berkeley; Berkeley-University of California; Berkeley Bears, Berkeley Golden Bears, California Bears, University of California Bears, and virtually every other variation I can and cannot think of involving some combination of the words “University”, “California, “Berkeley”, and “Bears”. Regular expressions may help with some schools, but not with others. Abbreviations could refer to multiple schools; for example, USC correspond to University of Southern California or University of South Carolina; in real world terms, your geographical location determines which school you are actually referring to. The latitude and longitude data specifying each school does not appear to be readily available from any reputable source in a format useful for web-scrapping. The only sources found so far are either behind a search engine, paywall, or come in a large file so big that Notepad and MS Word refuse to even attempt to open the file at all.

We would appreciate any advise, assistance, or even encouragement with regards to our project. Thank you for your consideration. (Go Bears!)