An Iterative Strength Rating Based Model for the Prediction NCAA Basketball Games

Philip Tan & Jeff Harrison

CSE 140: Homework 9

Summary of Research Questions & Results:

1. What method of predicting college basketball games should we use to obtain the best results?
    a. After careful research we determined that the Iterative Strength Ranking was by far the most successful, along with the added bonus of being easy to implement.
2. Can we alter the basic algorithm to produce more accurate predictions of the NCAA tournament?
    a. The standard method of using an Iterative Strength Rating to rank teams allowed us to successfully predict the outcome of 47 out of the 63 games played; a success rate of 74.6%. We produced 7 additional models that utilized several changes to the algorithm. The most successful model predicted the outcome of 52 out of the 63 games played; a success rate of 82.5%.

Motivation & Background:

With March Madness coming up, the world of college basketball eagerly anticipates the top teams that will be competing for the NCAA title. Countless fans are preparing their brackets with their favorite teams. Of course, many are betting on outcomes as well, so we thought we'd take a crack at predicting these outcomes. We hope to use win-loss ratios and score data from the NCAA and ESPN websites to create code that will predict winners of basketball games. While our focus is small to begin with, extrapolations of our code may be used in the near future to predict more significant basketball events like the NCAA title (or even the NBA title!).

While on the whole, athletics is unimportant to "the progress of mankind", the collection of data and statistics is very important. Accurate data collection and its analysis and

extrapolation to future events is key to many things. The most obvious of course is the economy. In general markets fluctuate depending on consumer sentiment and whatever political or environmental events happen. However, some things do remain the same every year. For example, in many countries, Christmas causes an increase in consumer spending on gifts and in the US, Thanksgiving marks an increase in turkey sales. These are annual patterns and analysis of sales data would tell you that these patterns will continue as long as people continue to observe such holidays. That was a simple example, but of course, we can also apply the idea of data collection and analysis for predictions of future results to scientific research. For instance, biologists studying bird migration may want to look at a particular species of bird and estimate how many birds will migrate with each season. By recording data of bird populations from previous seasons, scientists would be able predict future bird populations. In studying environmental effects on bird populations, scientists could look at pollution, resources, environmental destruction, or a multitude of other potential sources of increase or decrease in population.

Coming back to basketball however, there is monetary motivation for predicting game winners. The sports betting business is a multi-billion dollar industry. People such as Ken Pomeroy(1) have developed algorithms for predicting the results of NCAA basketball games quite successfully. With his algorithm, Pomeroy was able to successfully predict the outcome of over 77% of NCAA basketball games. When you compare this to the Las Vegas Oddsmakers(2) at 75% you come up with a 2% advantage over the Casino. This 2% edge is all it takes to rake in millions of dollars by placing wagers on sports games. If we could somehow improve upon this edge there would no doubt be a high demand for our results. This is a potential moneymaker!

Dataset:

While the NCAA keeps detailed statistics, these do not include the results of single game statistics or results. In order to perform our analysis we will need to calculate things such as Strength of Schedule, for which we will require a summary of the results of every game the team played. There were no databases immediately available so it became necessary to write a small program in python to parse the ESPN website for team schedules. Using a Python package called BeautifulSoup(3) we were able to mine game data directly from the ESPN website. In

general, ESPN game schedules and results can be found at the following address: http://espn.go.com/mens-college-basketball/schedule.

The dataset used for this analysis is a schedule of all the games played during the 2012 NCAA Basketball season. The information contained in this schedule includes the opponent, the score, the result of the game (a win or a loss), and whether the game was at home or away. In addition to the information included in the schedule, we calculated the point margin between the two teams to get some sense as to the magnitude of the victory(or defeat).

In the initial search for data it became obvious that there were no existing databases that suited our needs. In order to get a list of all the games played in the season we wrote a program to grab data from ESPN's website. The program utilized various modules including BeautifulSoup and Requests to access and clean the data. In order to make this program usable to a wider audience(such as yourself) we used the built in module "pickle" to create a representation of the final data structure that stored the season information. Pickle allows the user to store large data structures as text files and to access them later without the need for length computation time.

In addition to the schedule, the program reads the data from an outside file containing a list representing the 63 games played in the 2012 NCAA Tournament as well as the actual results of those games. Using this data we were able to test how well our predictions matched up to the final results of the Big Dance.

Methodology:

In order to predict the outcome of the 2012 NCAA Tournament we would first need a method of ranking all the teams in the NCAA. There are numerous ways of doing this but the method we finally settled on was to use an Iterative Strength Rating algorithm to determine where the teams stood in relationship to each other. The Iterative Strength Ranking (ISR) is a procedure for determining how we can use the result of a game to determine whether or not a team will be successful. The first step in this procedure is to assign to each team in the entire NCAA, a base rating of some sort. For our analysis every team started with a base rating of 1000 points. Next, we go through every team in the NCAA and look at their entire schedule. For every game in this schedule we determine whether it was a win or a loss, and adjust their rank

accordingly. If the result of a game is a win we assign a new ranking to the team that is equal to the rank of their opponent added to a bonus for winning the game. For our analysis this constant bonus was 50 points. If the result of the game is a loss, instead of adding this constant bonus, we subtract it. After we go through every team and every game in the NCAA schedule we are left with a different ranking than when we started. Now, using this mapping from team to ranking we just created as the new "base ranking", we repeat all the steps mentioned before, except we do not reset the base rating to 1000. We continue this process until two consecutive iterations produce the exact same ranking. When this occurs we have reached the end of our process and will use this ranking to determine the winner of any given match-up. In theory, the higher the ranking of the team, the better the team is and the more likely they are to win against a lower ranked opponent.

The method just described is a very well known procedure and produces a very accurate prediction of the NCAA tournament. However, there are no prizes to be had by repeating a well known procedure. In order to try and improve upon the predictions of this method we first needed to identify specific areas for improvement. The two areas that we singled out were the value of the points assigned for a win or loss, and the method by which a winner is determined. In an attempt to improve upon the Standard ISR, we introduced several variations by adjusting the way points were assigned. These algorithms are entirely of our own design and include the "Close-Game ISR", the "Strength of Victory ISR", and the "March is More ISR". A summary of the changes made to the standard ISR is as follows:

1. Close-Game ISR- The theory that drove the creation of this ranking system is the idea that it is not as good to win by a small margin than it is to win by a large one, and that it is somewhat better to lose by a small margin than by a large one. Using this driving theory we adjusted the weight of close wins and close losses.
2. Strength of Victory ISR- The theory behind this ranking system is the idea that a superior team will win its games by large margins. For this system we increased the weight of winning a game by 14 or more points.
3. March is More ISR - The theory that drove the creation of this ranking system is the idea that good teams win game later in the year. For this ranking system we increased the weight of winning a game towards the end of the season.

In addition to developing new ways of allocating points, we also developed a new way of determining the winner of a match-up. In the Standard ISR, the winner of a game is determined by which team has the higher ranking. We thought that this method could be improved and introduced a new system for choosing the winner that we call “Smart Winner”.

The primary motivation behind this algorithm is research done by Georgia Tech into the science of Quantitative Analysis in college basketball(4). In order to determine the result of a College Basketball game we used a process known as Markov Regression. The basic principle behind this is that if two teams have vastly different rankings it is fairly easy to say that the team with the better ranking should win. However, when teams are very close in ranking we noticed that the Standard ISR did not provide a very good prediction of the result. We came up with a mathematical relationship between the difference in two teams rankings and the probability that the team will win the game. For a large separation in points there should be no difference between the Standard ISR and our new method. For close games, we used this probability to produce elements of a transition matrix and applied the mathematics of Markov Regression. Using this method we were able to vastly improve on the original ISR in addition to several of our custom made ISR models.

Results:

The results can be split up into two categories pertaining to the two methods we used to predict winners. The first one is “WINNER TAKES ALL”. The winner of a match-up is selected based solely on which team has the higher rank. Results are as follows:

ISR:	**Right:**	**Wrong:**	**Correct:**
Standard	47	16	74.6%
Close Game	50	13	79.37%
Strength of Victory	48	15	76.19%
March is More	49	14	77.78%

The second method is called “SMART WINNER”. The winner of a match-up is selected using a Markov Chain. Results are as follows:

ISR:	Right:	Wrong:	Correct:
Standard	52	11	82.54%
Close Game	46	17	73.02%
Strength of Victory	47	16	74.6%
March is More	50	13	79.37%

On the whole, we saw that the Standard Iterative Strength Ranking using the "Smart Winner" function was the most effective prediction model. As such, we used this to analyze all the data for the 2012 NCAA basketball season and predicted the outcome. This yielded 8066 correct results and 2574 incorrect results corresponding 75.81% correctness.

As answered briefly in the 'Summary of Research Questions and Results' section, we determined that the Standard ISR method was the most effective. For the first method where the winner of a match-up is selected based solely on which team has the higher rank, we saw that we were able to produce new methods that gave us slightly more accurate predictions. However for the second method, the standard ISR gave us an incredibly high correctness percentage.

While the result of this analysis seems to indicate that we have found a very effective method for predicting the 2012 tournament, it is unclear as to how successful this model would be when applied to any other year. Perhaps the remarkable figure of nearly 83% correct predictions is an outlier in an otherwise ineffective system. It would certainly be worth investigating this result in more depth as it represents a significant improvement over the methods used by professionals within this field. An additional source of bias in this result is the fact that all changes to the algorithm were produced in order to improve the prediction of this single tournament. A successful model was one that most accurately predicted this particular tournament.

Reproducing Our Results:

To reproduce our results, follow the procedures below:

1. Download all the included documents and ensure that they all exist within the same folder. The folder should contain NCAApredictions.py, 2012tourney.txt, actual-2012-tournament.txt, and full-schedule.txt
2. The program requires NumPy in order to carry out some of its calculations. Make sure NumPy is installed. NumPy can be installed from the following website: www.numpy.org
3. Run NCAApredictions.py using either your command line or IDLE and be amazed as the program displays a comparsion of the different predictions.

Collaboration:

None.

Reflection:

This assignment helped us see the power in being able to use code to manipulate data and extract information from it. While we should have seen this during the entire course, it is not till one considers the possibilities of having such powers that we truly become afraid of ourselves.

In Plato's Allegory of the Cave(6), Socrates describes a cave inhabited by prisoners who are chained and unable to turn their heads. The prisoners can only see the wall of the cave. Behind them a fire produces light from which puppeteers cast shadows on the wall that the prisoners face. The prisoners cannot see the puppets but can only perceive the shadows and echoes cast by these objects. To these prisoners, the shadows are reality because they would not know the true thing that causes the shadows. If unchained however, the prisoners would be terrified to see that they have only been experiencing the shadows this entire time.

If a man freed from his chains looked at the fire, would he turn back to the shadows? If this man were taken out of the cave, would he want to returned to the cave? He would be unable to see the truth because he is now blinded by the natural light outside the cave! One might suggest that once the man's eyes adjust to the light, he would see that the sun is the true provider of light. One may extrapolate this allegory to our class. We have faced the fire with each assignment and at last we have been dragged out into the light. In a sense, we have

unleashed upon ourselves something of true beauty and value. Some may even say we now have enhanced ability to seek additional truths of the universe. With great power comes great responsibility(7).

On a more serious note, we wish we had known how to use external Python packages and modules to aid in mining and extraction from publicly available data. Knowing a few of these packages would have been immensely helpful.

Additional Notes:

While our initial project proposal contained a long list of research questions to answer, we realized after TA feedback that the scope of our project was too large and ambitious given the time constraints. For that reason we narrowed our focus down simply to the prediction of NCAA games.

While predictions are potentially highly accurate, the authors of this study do not recommend the practice of gambling or betting on basketball games. Gamble at your discretion. The authors are not liable for social, physical, monetary losses, or legal penalties following predictions made in this study.

References:

1. Pomeroy, K.R., Advanced Analysis of College Basketball. Multiyear Pomeroy College Basketball rankings. http://kenpom.com/ (March 6, 2013)

2. Orendorf D., Johnson T., First-Order Probabilistic Models for Prediction the Winners of Professional Basketball Games (2007). http://www.ics.uci.edu/~dorendor/basket.pdf (March 7, 2013)

3. Richardson, L., Beautiful Soup. Python coding language library for parsing HTML and XML from websites. http://www.crummy.com/software/BeautifulSoup/ (March 6, 2013)

4. Kvam, P., Sokol, J.S., A Logistic Regression/Markov Chain Model for College Basketball Rankings (2006),Naval Research Logistics 53, pp. 788-803 https://wiki.engr.illinois.edu/download/attachments/185991190/ncaa.pdf?version=1&modificationDate=1274977989000 (March 7, 2013)

5. ESPN. Men's College Basketball. http://espn.go.com/mens-college-basketball/schedule (March 12, 2013)

6. Cohen, S.M. The Allegory of the Cave (2006). http://faculty.washington.edu/smcohen/320/cave.htm (March 14, 2013)

7. IMDB. Spider-Man (2002). http://www.imdb.com/title/tt0145487/quotes (March 14, 2013)