Vincent Sheu, Winnie Wong, Rosie Abe, J.P. Saunders, and Lindy (Li) Lin Chang
Statistics 133, Spring 2012 (Ibser/Kuang)
05.09.2012
Final Project Write-Up
"Beating Vegas"

Introduction

There are 345 schools distributed among 32 conferences in Division I of the governance structure established by the National Collegiate Athletics Assocation (NCAA) for men's basketball. 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 90% of the NCAA's revenue, generating over $500 million annually [1].

As Vegas has risen, so have computer models ranking teams and predicting the outcomes of games. The prior is relatively easy to approximate; in fact, the NCAA itself has 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%)[2]. 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 (minus any fees an institution may charge); otherwise s/he loses the money placed on the bet. This second form is the prediction many statistical models strive to beat.

In our project, we analyzed the predictive power of several different statistical models and prediction systems: Jeff Sagarin, Sagarin ELO, Sagarin Predictive, Sonny Moore, Jon Dokter, and StatFox. Because there is less data to analyze at the beginning of each season, models tend to be less accurate; therefore, in our analysis, we only looked at games occurring after January 1 of each season. The new year also 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.

Along the way, we looked at the following sets of questions:

Vincent Sheu, Winnie Wong, Rosie Abe, J.P. Saunders, and Lindy (Li) Lin Chang
Statistics 133, Spring 2012 (Ibser/Kuang)
05.09.2012
Final Project Write-Up
"Beating Vegas"

- How accurate is both the Vegas opening line and final line? What % of games does it correctly predict the winner for? How accurate are each of the prediction systems at predicting the correct winner?

- 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 and opening line?

Materials and Methods

To aid our analysis, we scraped data off of The Prediction Tracker [3], an online web resource run by Todd Beck, a sports statistician. The Prediction Tracker has data for all NCAA Division I Men's Basketball Games from the 2007-2008 season to the present tracking home/away teams, final scores, lines, and predictions from several major predictive models [4].

Data was archived as comma separated value-type files [5] [6] [7] [8] [9] [11]. To verify the data and check its credibility, we manually checked all the game scores with other two websites: ESPN, and Sports Reference [12][13]. We then used R's read.csv function to import data for analysis. In order to answer our main questions, we converted the score results to single integers, which we then compared to each of the numbers given by the predictive models. Each prediction for each game was then assigned a logical value of "TRUE" or "FALSE", with "TRUE" representing a correct prediction and "FALSE" representing an incorrect prediction.

The accuracy of each prediction system against the spread was determined as a function of time. This examined how the system was predicting scores up to each subsequent game. The results of this analysis were visualized with a scatter plot. To determine the accuracy of each prediction system as a whole, the sum of the prediction logical values was taken and divided by the total number of games to obtain an average value. This "accuracy" value was calculated for predictions against the line and opening line. These two values were then plotted in a bar graph for each prediction system per season. We wrote a function that combined these actions into one function that generated a bar graph for each season.

The "straight up" accuracy of each prediction system was also calculated. This was a measure of whether each system could predict game winners, without margin of victory taken into account. For example, if Duke played North Carolina and Sagarin predicted North Carolina -10, Sagarin picked North Carolina to win by 10, and by extension, picked North Carolina to win. Wilcoxon–Mann–Whitney rank-sum tests were run to compare the accuracy of each system to a series of coin flips.

We then built a Google Earth representation with Placemarks on each school. Placemark data indicated a frequency for how often a school beat the spread for its own games. In order to initiate this portion of the project, we acquired data from GeoNames [10], a free online database of the latitudes

and longitudes of many locations, to obtain the latitude and longitude of every Division I university. This web scraping required the use of a function that filled out a 'get' form over the names of all of the universities. The function to obtain the latitude and longitude data was a multi-part one that parsed HTML, obtained the table with relevant latitudes and longitudes, and then extracted the latitude and longitude of the first location classified as a "School" in the table. Our final function for this portion of web scraping returned a data frame that contained all school names and their respective latitudes and longitudes.

In order to run the function with proper input, we compiled the names of all Division I schools from the list of final rankings in the 2011-2012 season (obtained from Sports Reference) [11]. We then formatted the names of the Division I universities using regular expressions to search for their locations in GeoNames. A large number of schools did not initially turn up, and we were forced to make increasingly interesting modifications find our results. This included modifications such as the adding of "University" to the end school names where was not already present and changing "St." to "State University" for other schools. However, because of the very specific input that the GeoNames website requires to accurately search its database, manual editing had to be done to the Schools data set for a number of special cases. Ultimately, there were a small number of schools whose names did not return any data from the GeoNames form; these names did not appear on the Google Earth map.

Once all the latitudes and longitudes were accumulated into a data frame, they were edited into decimal form--the form of latitude and longitude that Google Earth requires. This required the use of more regular expressions to take out the unnecessary characters that were extracted with the geonames data (degrees symbol, minutes and seconds symbols), and an arithmetic function that converted degrees, minutes, and seconds into a decimal form of latitude and longitude. Following the standard formatting of decimal latitudes and longitudes, latitudes were all positive values because all NCAA schools lay north of the Equator. Likewise, longitude values were all negative because all NCAA schools lay west of the Prime Meridian.

For the prediction section, in addition to using gsub on school names to match the longitude and latitude data frame, we used the grep function in R to pick each school's prediction from each season and wrote a function that calculated the probability of the correct predictions, such as (number of times the home team beat the spread (home.beat.line = 1), given that it was the home team) + (number of times the away team beat the spread (home.beat.line = 0), given that it was the away team) and divided by the total number of games that school competed in. Then, we used merge functions to link both coordinates and predictions for each school. In order to visualize in Google Earth, we created a XML file and several XML nodes. We then wrote a function that added our data into XML nodes and use a for loop to generate the schools. In the end, we visualized each school in Google Earth successfully. [13]

Vincent Sheu, Winnie Wong, Rosie Abe, J.P. Saunders, and Lindy (Li) Lin Chang
Statistics 133, Spring 2012 (Ibser/Kuang)
05.09.2012
Final Project Write-Up
"Beating Vegas"

Results and Discussion

	row.names	Line.Accuracy	Opening.Line.Accuracy	Sagarin.Accuracy	Sagarin.ELO.Accuracy	Sagarin.Predictive.Accuracy	Sonny.Moore.Accuracy	Jon.Dokter.Accuracy	StatFox.Accuracy
1	2011-2012	0.7428838	0.7461503	0.7456836	0.7293514	0.7438171	0.7442837	0.7456836	0.7442837
2	2010-2011	0.7162041	0.7063563	0.7170994	0.7117278	0.7153089	0.7068039	0.7224709	0.7094897
3	2009-2010	0.7293869	0.7234672	0.7353066	0.7171247	0.7399577	0.7365751	0.7424947	0.7319239
4	2008-2009	0.7048220	0.6994142	0.7070753	0.6958089	0.7097792	0.7129338	0.7120324	0.7084272
5	2007-2008	0.7255937	0.7218997	0.7271768	0.7139842	0.7255937	0.7266491	0.7319261	0.7203166
6	Overall	0.7236551	0.7192336	0.7264186	0.7135225	0.7269713	0.7254053	0.7309322	0.7229182

Figure 1: Straight Up Prediction Results (PredictionResults.pdf)

When looking at straight up results of games over five seasons, all of the models we looked at - Sagarin, Sagarin ELO, Sagarin Predictive, Sonny Moore, Jon Dokter, and StatFox -were able to correctly predict the winners for a majority of the games. Given that all six systems were independently developed, it is remarkable that they were as close to each other as they were: predictive accuracy ranged from 71.3% (Sagarin ELO) to 73.09% (Jon Dokter). However, none of the systems were able to correctly beat the spread at a reliable level over the five seasons of data we looked at. Sagarin, Sagarin ELO, Sagarin Predictive, Sonny Moore, and Jon Dokter all averaged between 48.7% and 49.9% accuracy over the course of five seasons, and no system managed to do better than 52.73% (Jon Dokter against the Opening Line in 2009-2010) over the course of an entire season. Results from the Wilcoxon-Mann–Whitney rank-sum tests indicated that all but one of the models were not much different than a series of coin flips. StatFox clocked a p-value of 1.437E-19, indicating that the predictions were significantly different from random coin-flipping.

	row.names	Sagarin	Sagarin.ELO	Sagarin.Predictive	Sonny.Moore	Jon.Dokter	StatFox
1	11-12 Against Opening Line	0.4923005	0.4871675	0.4853010	0.4969669	0.4983668	0.4526365
2	11-12 Against Line	0.4820345	0.4955670	0.4778348	0.4932338	0.4815679	0.4512366
3	10-11 Against Opening Line	0.4910474	0.4803044	0.4820949	0.4986571	0.5026858	0.4592659
4	10-11 Against Line	0.4798568	0.4937332	0.4825425	0.4897046	0.4852283	0.4623993
5	09-10 Against Opening Line	0.5090909	0.5040169	0.5264271	0.5090909	0.5272727	0.4799154
6	09-10 Against Line	0.4993658	0.4942918	0.5145877	0.4913319	0.5014799	0.4659619
7	08-09 Against Opening Line	0.5033799	0.5024786	0.5114917	0.5011266	0.5105904	0.4470482
8	08-09 Against Line	0.4772420	0.4966201	0.4826498	0.4984227	0.4826498	0.4420910
9	07-08 Against Opening Line	0.4691293	0.4707124	0.4744063	0.4981530	0.4833773	0.4490765
10	07-08 Against Line	0.4622691	0.4622691	0.4701847	0.5071240	0.4701847	0.4522427
11	Average over Five Seasons	0.4874263	0.4894989	0.4917557	0.4982959	0.4951640	0.4566139

Figure 2: Accuracy of Each Prediction System Against the Spread (PredictionResults.pdf)

One interesting piece of data came up with regards to StatFox. Although StatFox was a middling predictor of straight up results, ranking 6th out of 8 systems (including the Opening Line and Line), StatFox lost to the spread so consistently (averaging 45.66% accuracy over five seasons) that picking against StatFox made more sense than picking with StatFox. In fact, StatFox was so inaccurate at predicting against the spread over the last five seasons (Low: 44.21% in 2008-2009 against the Line; 47.99% in 2009-2010 against the Opening Line) that one would have beaten the spread predicting

against StatFox for every single season we had data for. So far, the evidence does not support a scenario of one being able to apply the law of large numbers and generate consistent income based on the accurate predictive power of any one system.

	row.names	p.sag	p.sage	p.sagp	p.moore	p.dok	p.fox
1	P Values for Test Against Coin Flipping	0.1222382	0.1368137	0.1582726	0.2192635	0.07269495	1.437475e-19

Figure 3: P-Values for Wilcoxon-Mann_Whitney Tests of each System's predictions against a randomly generated sequence of 1's and 0's (P.Values.For.Predictions.Against.Spread.pdf)

Models with high "straight up" accuracies but low accuracies (against the spread) are good at predicting the winner but not within the spread. Jon Dokter had the highest "straight up" accuracy of 73.1% and 2nd highest accuracy of 49.4%. It is therefore a pretty good model on both counts. Sonny Moore had a moderate 72.5% "straight up" accuracy but the highest accuracy against the spread (49.8%). On the other hand, even though StatFox also had a moderate 72.2% "straight up" accuracy, it had the lowest accuracy against the spread (45.6%). Lastly, Sagarin ELO had the lowest "straight up" accuracy of 71.4% but a relatively high accuracy against the spread (48.9%). To test whether one was a good predictor of the other, we attempted to linearly model the two sets of accuracies.

A: Call:
lm(formula = sacc ~ acc)

Residuals:
```
      1       2       3       4       5       6
1.995e-03 -1.108e-02  2.200e-03  3.772e-07  5.863e-03  1.020e-03
```

Coefficients:
```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.68398    0.09294   7.360  0.00182 **
acc          0.08311    0.19122   0.435  0.68624 (a)
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 0.006461 on 4 degrees of freedom
Multiple R-squared: 0.0451 (b), Adjusted R-squared: -0.1936
F-statistic: 0.1889 on 1 and 4 DF, p-value: 0.6862 ©

Vincent Sheu, Winnie Wong, Rosie Abe, J.P. Saunders, and Lindy (Li) Lin Chang
Statistics 133, Spring 2012 (Ibser/Kuang)
05.09.2012
Final Project Write-Up
"Beating Vegas"

B:

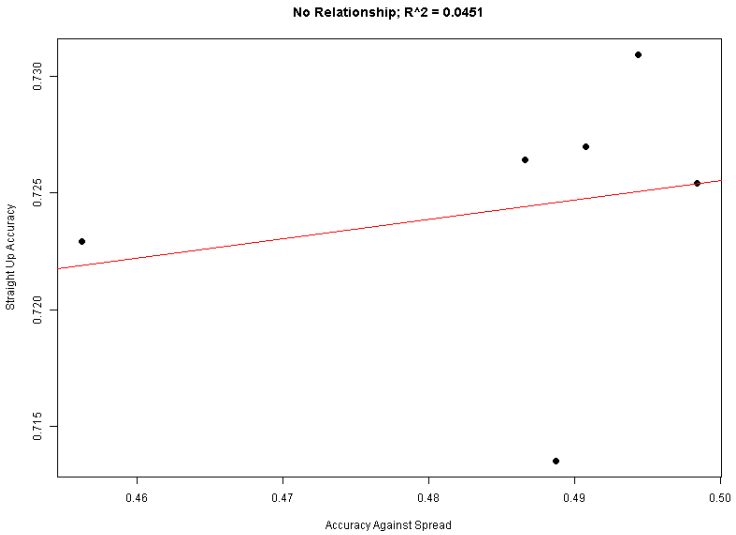
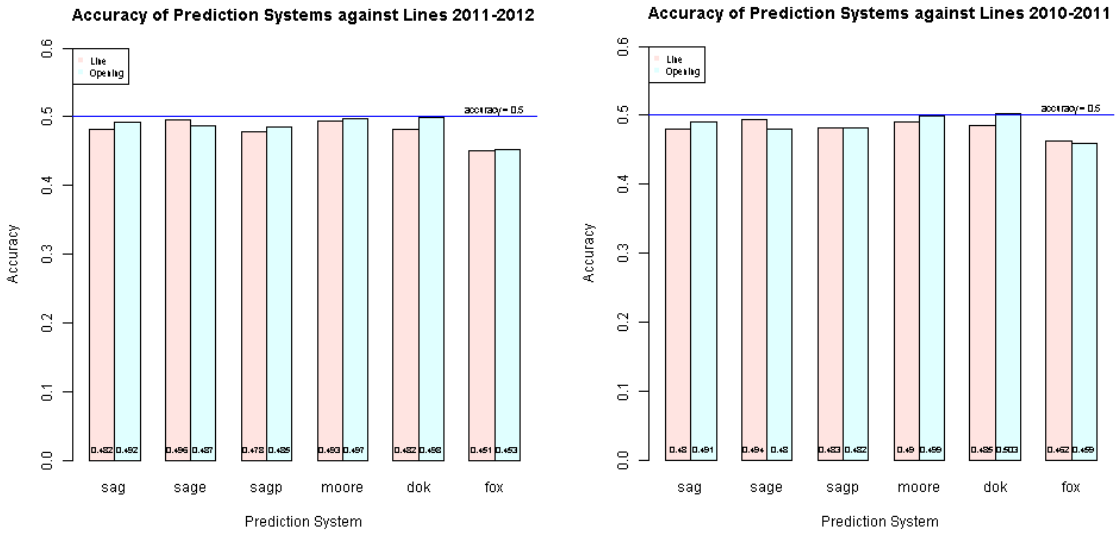


Figure 4: Linear Regression of Straight Up Accuracy against Accuracy (NoRelationship.png)

Some important numbers to look at: (a) is extremely high. This implies that the coefficient determining slope is zero. The intercept is nonzero (low p-value), but this is inconsequential. (b), the R-squared value, is miniscule, implying that accuracy is not well correlated with straight up accuracy. These results together confirm our initial suspicions: the data supports the observation that the accuracy and "straight up" accuracy are quite independent of each other. Both parameters need to be taken into account to form a conclusion about the overall accuracy of each model.



Vincent Sheu, Winnie Wong, Rosie Abe, J.P. Saunders, and Lindy (Li) Lin Chang
Statistics 133, Spring 2012 (Ibser/Kuang)
05.09.2012
Final Project Write-Up
"Beating Vegas"

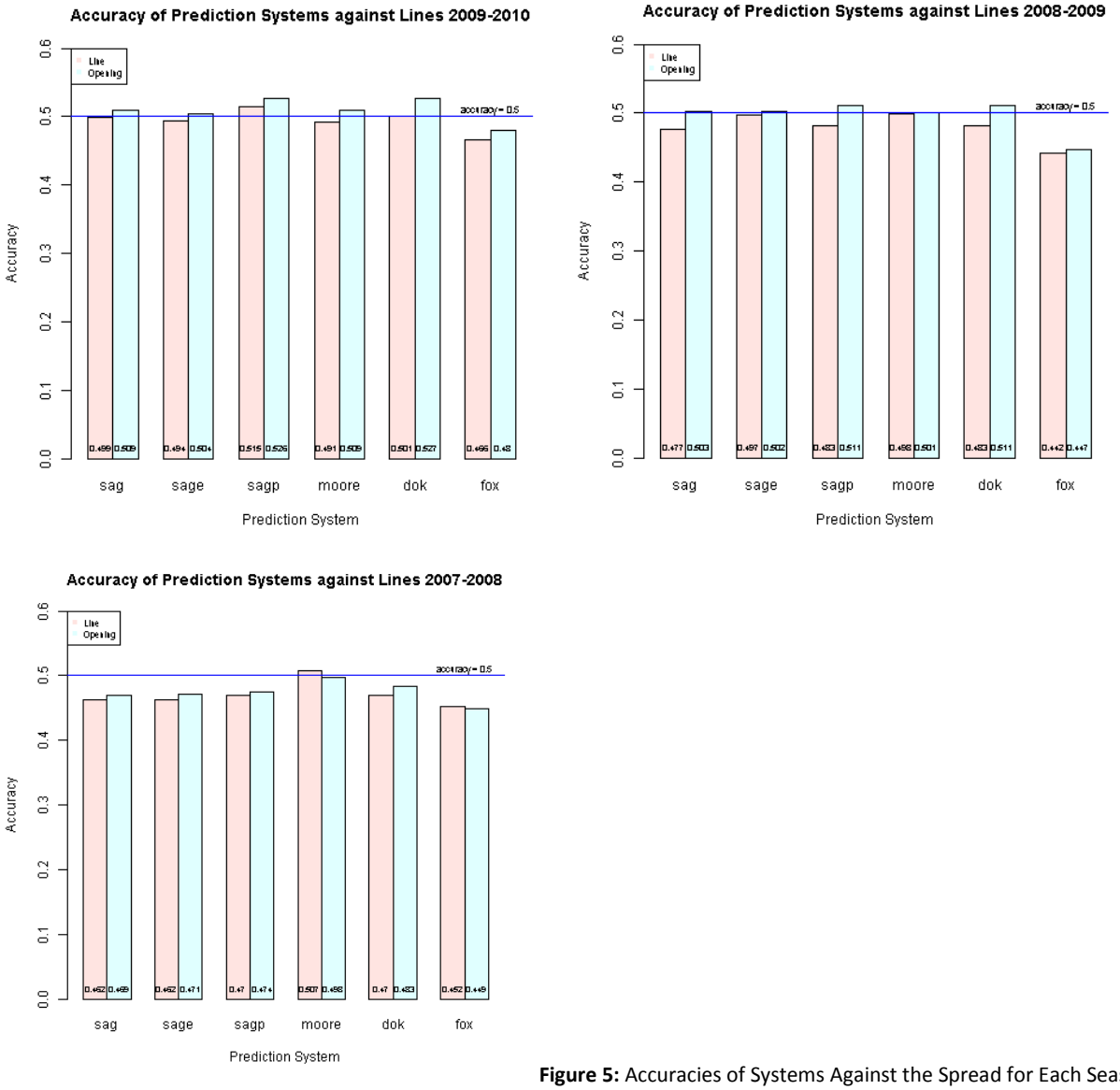
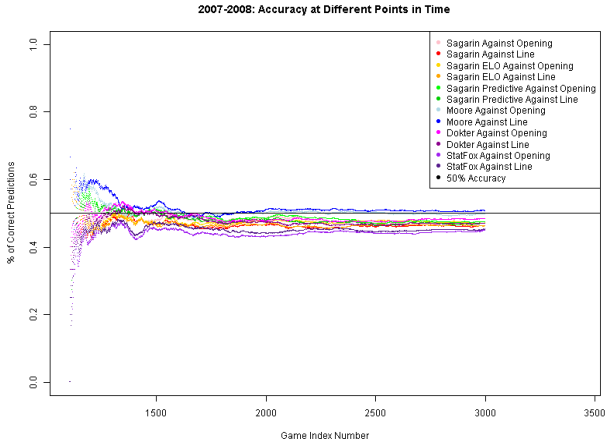
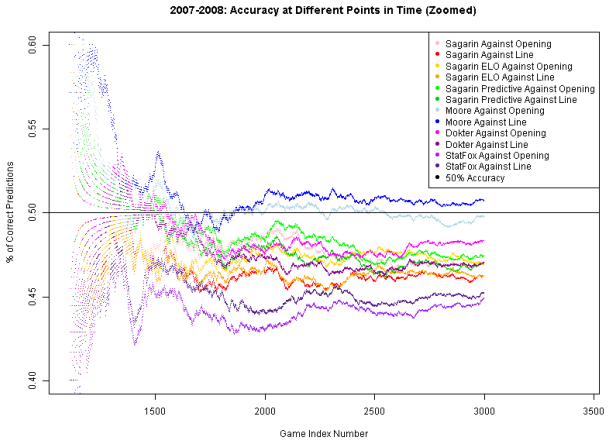


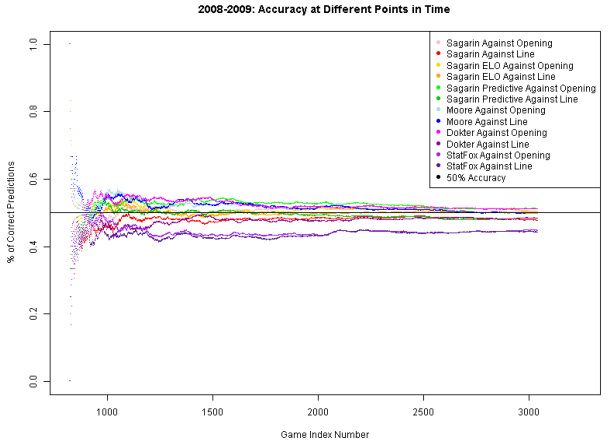
Figure 5: Accuracies of Systems Against the Spread for Each Season (accuracy1112.png, accuracy1011.png, accuracy0910.png, accuracy0809.png, accuracy0708.png)

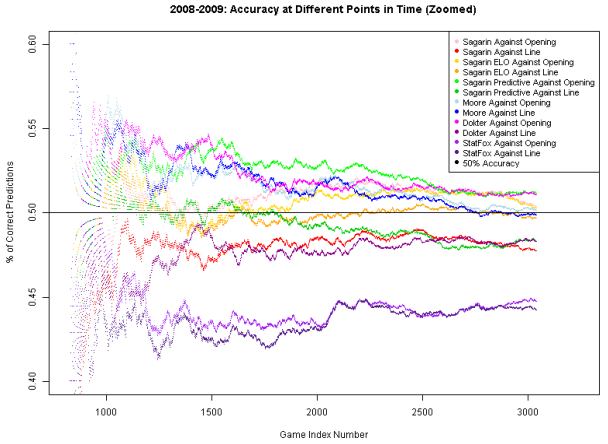
We also tracked the accuracy of each system at each point in time. The "scatter plot/line plot" of accuracy of each system against either a line or opening line looks at the following question: If you had trusted any one system on every single prediction against the spread for every game up to game X, how many games would we have beaten the spread on? By the results of the scatter plots, it appears that while initial accuracies can easily destroy the 50% threshold, eventually, the law of large numbers applies and accuracies drop to their usual 48-49%'s. StatFox's, again it appears, never ever recover from their inaccuracy.

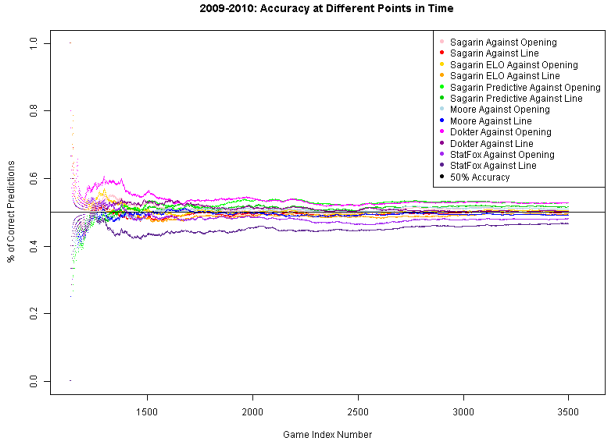
Vincent Sheu, Winnie Wong, Rosie Abe, J.P. Saunders, and Lindy (Li) Lin Chang
Statistics 133, Spring 2012 (Ibser/Kuang)
05.09.2012
Final Project Write-Up
“Beating Vegas”

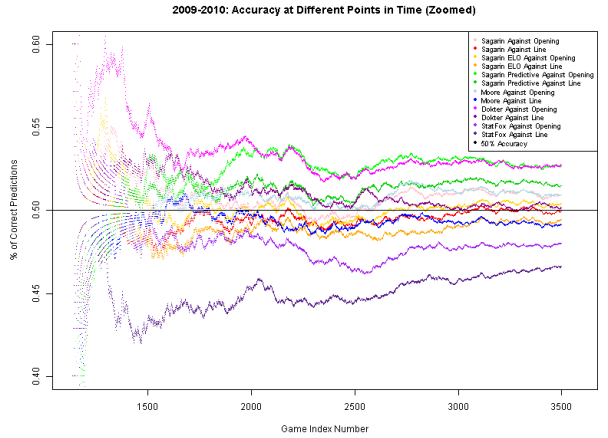












Vincent Sheu, Winnie Wong, Rosie Abe, J.P. Saunders, and Lindy (Li) Lin Chang
Statistics 133, Spring 2012 (Ibser/Kuang)
05.09.2012
Final Project Write-Up
"Beating Vegas"

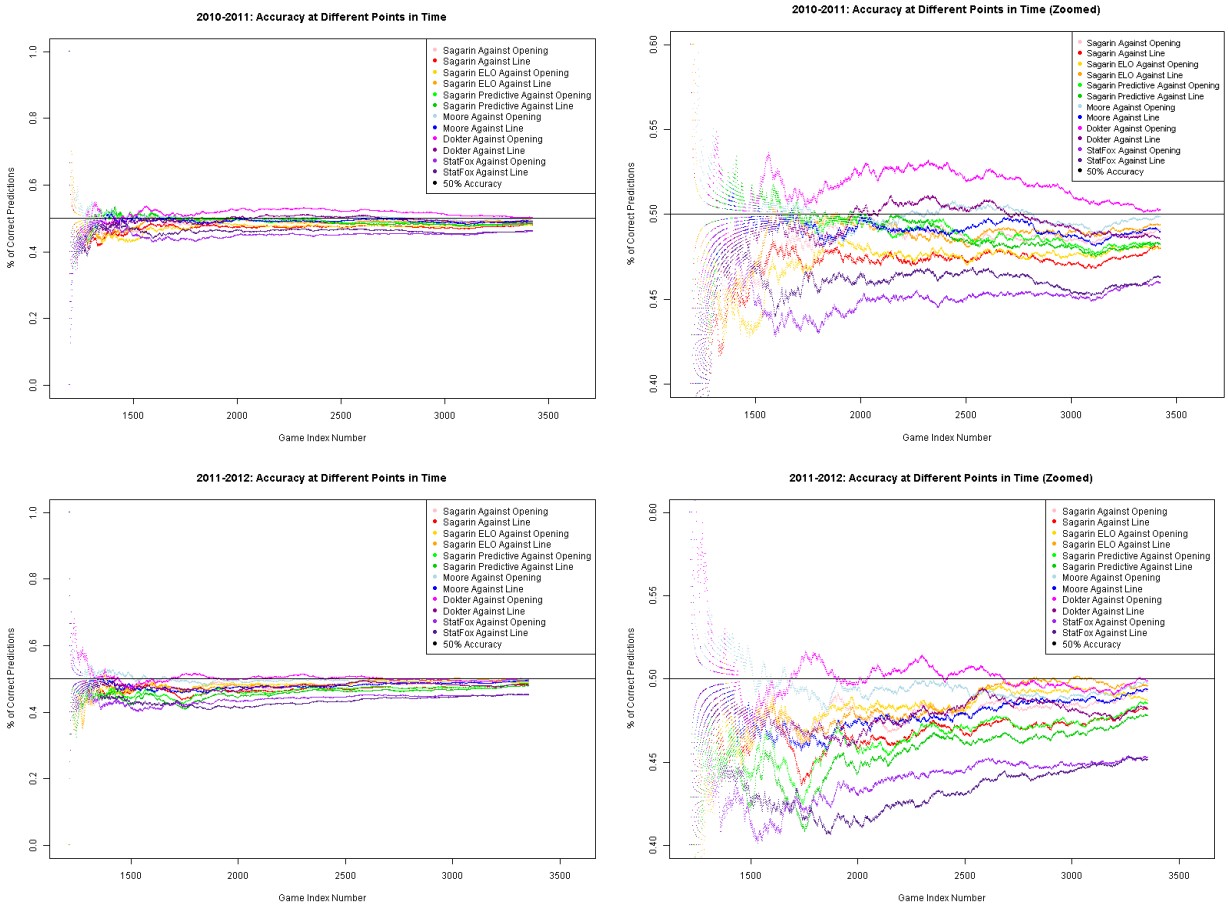


Figure 6: Accuracy of Each Prediction System for each season as a function of time (Game Index Number) (0708Accuracy.png, 0708AccuracyZoomed.png, 0809Accuracy.png, 0809AccuracyZoomed.png, 0910Accuracy.png, 0910AccuracyZoomed.png, 1011Accuracy.png, 1011AccuracyZoomed.png, 1112Accuracy.png, 1112AccuracyZoomed.png)

Limitations

There were numerous limitations with this project, from the data we used to the results we produced. The data we used does not take into account the fact that teams change throughout the years and within a season. Players can get injured or switch teams and predictive models usually do not immediately take these changes into account. We also limited our data to games played after the new year of each season. Generally, teams open up the season by playing "cupcakes": teams from local colleges that are extremely weak and easy to defeat. This results in a large number of games that are relatively easy to predict: Kentucky will probably beat East Western North Polytechnic State College for Agriculture, Mining, and Farming by at least 20 points every time.

Image fidelity is a major limitation with regards to the bar and scatter plots. Because there is so much data of similar values, it is sometimes difficult to discern differences between each model. We

tried to address this by generating a zoomed in version for the scatter plots, but this wasn't a practical solution for the bar plots.

There are also limitations in our Google Earth representation of how often schools beat the spread. For example, schools with larger fan bases tend to have more fans who bet on their schools, which can significantly skew the line from the opening line, as the line moves in response to the betting patterns of the public. Examples of such schools are University of Kansas and University of Kentucky, two universities with large, national, fervent fan bases. These deviations may or may not be accounted for in the Google Earth representation. In addition, schools are abbreviated in many ways.

The GeoNames database we used was limited in capturing all the schools properly. Many modifications had to be made with regular expressions to obtain an appropriate input name for GeoNames. Because of this, some of the school names in our Google Earth may not match their conventional names because we had to add different key words to the search. Also, the functions that cleaned data encountered problems in older versions of R. We believe our method of retrieving schools' latitudes and longitudes is justified because proper input of a school name would generally ensure that the campus's location appears first among all other schools. After checking the obtained latitudes and longitudes, there were few cases that yielded erroneous latitudes and longitudes. Some of the errors that arose included "California" being changed to "California University" which then gave output of the latitude and longitude of the University of Southern California. Additional manual editing was done to the Schools vector to minimize these sorts of errors (we changed "California University" to "University of California Berkeley"). Due to the size of the data set, we could not manual check every single school name; however, our final Google Earth representation does not have alarming errors that we feel make our project inaccurate.

Future Directions

In the future, a prediction system that would use functions to weigh the results of all of the predictors to generate a final prediction about each game could be created. After running this predictor over all the games, we could compare the results of our predictor against all of the original ones. Weights could be adjusted to generate optimal predictions. Ideally, we could create a system that is more accurate than each of the individual prediction systems.

Other future directions would include updating the Google Earth representations. We would like to be able to take into account whether or not a school is consistently over or under rated and visualize this as part of our Placemark (e.g. overrated teams in red; underrated teams in blue).

Additionally, some of our errors were due to the structure of the GeoNames website. Due to the fact that other location databases were often behind paywalls, we were forced to choose a database that we believed had several flaws (discussed earlier). To work around these flaws, we were forced to

make undesirable modifications to some of our data to get any results. To overcome these in the future, we could use data from a pay-for-use database that has all of the schools' names and a better search results form.

Conclusions

How accurate is both the Vegas opening line and final line? What % of games does it correctly predict the winner for? How accurate are each of the prediction systems at predicting the correct winner? See **AccuracySummary.pdf**.

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 and opening line? See **PredictionResults.pdf.**

References and Citations

1. Ourand, J and Smith, M. "NCAA, TV talk about bigger men's tourney". *Sports Business Journal.* Accessed 25 April 2012. <http://www.sportsbusinessdaily.com/Journal/Issues/2009/12/20091207/This-Weeks-News/NCAA-TV-Talk-About-Bigger-Mens-Tourney.aspx>
2. Johnson, Greg. "RPI formula altering for 2013 season." NCAA. Accessed 25 April 2012. <http://www.ncaa.com/news/baseball/2011-08-03/rpi-formula-altering-2013-season>
3. Beck, Todd. The Prediction Tracker, Accessed April 25 2012 <http://www.thepredictiontracker.com/basketball.php>
4. http://home.comcast.net/~tlbeck/ncaabb11.csv
5. http://home.comcast.net/~tlbeck/ncaabb10.csv
6. http://home.comcast.net/~tlbeck/ncaabb09.csv
7. http://home.comcast.net/~tlbeck/ncaabb08.csv
8. http://home.comcast.net/~tlbeck/ncaabb07.csv
9. GeoNames, Accessed 28 April 2012 <http://geonames.org/>
10. SR College Basketball 2012 Standings, Accessed 28 April 2012 <http://www.sports-reference.com/cbb/seasons/2012-standings.html#standings::none>
11. ESPN, Accessed 26 April 2012 <http://espn.go.com/frontpage/prestitial>
12. SR College Basketball, Accessed 26 April 2012 <http://www.sports-reference.com/cbb/>
13. R Code.R. Saved as a file in the ZIP archive