# CS229 Final Report - Machine Learning Madness

Elliot Chanen, John Gold

December 2014

## 1 Introduction

March Madness is the NCAA Men's Division I Basketball Championship tournament that happens every March. The tournament, organized by the National Collegiate Athletic Association (NCAA), was founded in 1939 by the National Association of Basketball Coaches. It started as an 8 team single elimination tournament and has since been expanded, most notably to 64 teams in 1985. As the tournament has grown, so has its national popularity. March Madness has become one of the most famous annual sporting events in the United States, partially because of its enormous television contract with CBS, but mainly because of March Madness pools. For years fans have been entering gambling-related contests to see who can predict the tournament most correctly. Some people have even gone as far as saying that filling out a tournament bracket has become a "national pastime." Even President Obama, famously, fills out a bracket every year.

Ignoring the four "play in" games (which is done in most pools), there is a 64 team pool which means there are $2^{63}$, or 9.2 quintillion possible brackets. Needless to say it is very difficult to predict every game correctly, but people continue to research and try their best. Last year, Warren Buffett offered a one billion dollar prize to any person(s) who could correctly predict the outcome of the tournament. No one was able to claim the prize. We hope to create a model which best allows us to fight for that prize.

## 2 Data

We used data from two main sources, sports-reference.com and kenpom.com, both of which track college basketball statistics. The data is organized by division 1 team, and has seasonal statistics in many categories. Sports-Reference has basic information such as wins, losses, rebounds, points scored, points allowed, and so on. Kenpom was created by a statistician, and uses propietrary stats built from other factors, such as offensive and defensive efficiency, that try to represent teams more wholistically. There was a march madness kaggle competition and they provided a nicely formatted list of regular season and tournament games from previous years which we used as well.

# 3 Features

There are many statistics compiled for every basketball game, so it was difficult to decide which ones to use. One of the simplest but most useful factors we could think of was average margin of victory. Generally, the better teams will win more games, and specifically win those games by more points. Another feature we used was shooting percentage, and this statistic can get at two important factors of the game. Primarily, teams with a higher shooting percentage have better players, and secondly, it can be explained by taking more open shots which is a proxy for ball movement, a feature we wanted to include but has limited statistical data. We ended up compling 14 different features of each team, so our initial input vector was in $\mathbb{R}^{28}$ because their are two teams for each game.

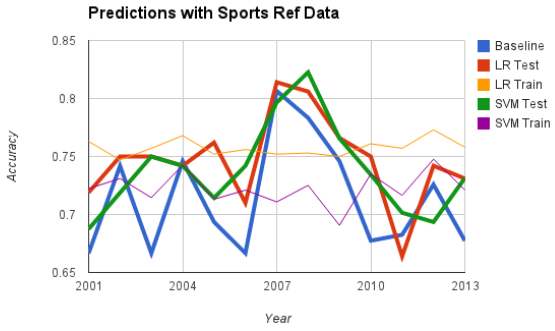
However, we wrote a script for feature selection and it turned out some of the statistics were largely irrelevant and did not change our classifiers enough to make a difference. Additionally, we found that the same subset of features was not optimal for both the support vector machine and logistic regression, so there was tuning there as well. In the end we used seven basic statistics from the sports reference data for the logistic regression, and just three for the support vector machine. For the kenpom data, we used a five dimensional feature vector for each team, and kept those features consistent for both models.

# 4 Models

We were faced with a classification problem, a binary decision of win or loseleveraged the python library SciKit Learn to run our regression and support vector machine models. We defined our training data as the regular season games, of which there are roughly 5,000 each season, and the test data were the 63 tournament games. For development we used cross validation on our training set, splitting the season into ten different buckets and training on nine while holding out the last for testing.

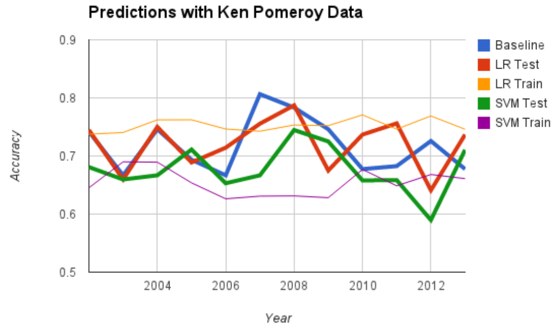
# 5 Results

The structure of the NCAA tournament is a 64 team bracket, split up into 4 groups of 16. Each team is then seeded based on the tournament committee's exhaustive selection process. The committee gauges teams based on wins and losses, success in conference play, success in tournaments, and an overall assessment based on players, coaching, and program tradition, called the eye test. We use a baseline of the percentage of higher seeds (favorites) that win their games. Depending on the year, the higher seeds have won between 65% and 80% of their games, with an average of 72% since 2001. Below we have a graph containing the baseline, and our models using the sports-reference data.



The red line, which is our logistic regression model generally performs better than the baseline, although not by much. We were having some over fitting issues with the SVM model previously, but we adjusted the features and saw a noticeable improvement so that the scores are roughly on par with the baseline.

Below we have our results using the statistics from kenpom.com to train our model:



This graph is similar to the previous one, although we generally perform worse using the SVM which was unexpected. Our understanding of theses statistics is that they would capture more of the "underlying truth" that is a team's skill.

# 6 Discussion

## 6.1 Overall

We ran our algorithm over multiple seasons of data and tried many different features. After initially beginning with 14 features we ran some tests to prune out with ones were either ineffective or adding noise. We narrowed it down to 6 key statistics, but SVM had a lot of variance and turned out to overfit, performing very well on the training data but poorly on the test set. Logistic Regression was more effective and consistent. Using a data aggregation source helped our SVM model, but slightly decreased performance for the logistic regression. We feel confident in the results, and will be constructing our brackets this spring with the help of our program.

These next two tables are our aggregated results over multiple seasons, with both models, data sources, and training/testing data.

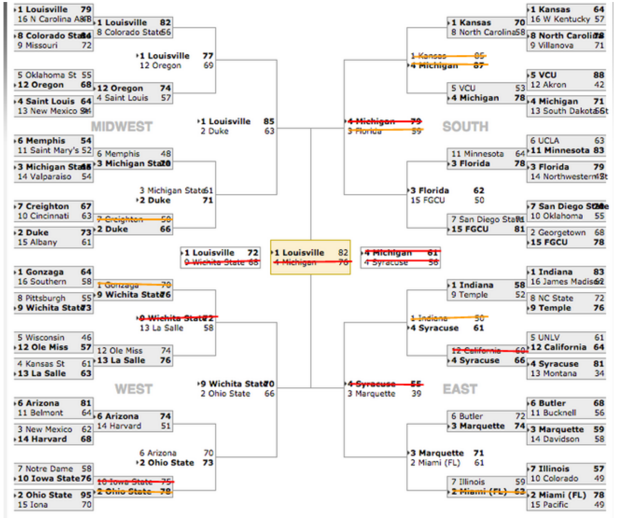
| Sports Reference | LR | SVM |
|---|---|---|
| Training Data | .757 | .722 |
| Test Data | .747 | .738 |

We initally had a problem of overfitting, and our accuracy on the training data scores were much higher than the test set. We corrected for this by emphasizing the strength of schedule feature. Previously the in-season results were skewed by dominant teams in small conferences having trouble when they faced medium skill teams from the bigger conferences.

| Ken Pom | LR | SVM |
|---|---|---|
| Training Data | .752 | .654 |
| Test Data | .721 | .677 |

### 6.2 Example Bracket

We believe our discussion can be enhanced by a specific tournament example.



There are a couple interesting things to note when we looked at where our results were coming from. The bracket has looked like this (64 teams), since 1985, and in those almost 30 years, there has never been a 16 seeded team that won against a 1 seed. However, our model predicts this will happen. In the theme of picking upsets, we also choose three of the four 15 seeds to win their first round game. This has only ever happened 7 times, although one of those upsets did happen in this tournament and we correctly predicted it (Florida Gulf Coast - FGCU). Our picks seemed to get worse as the tournament progressed, which we believe is due to the increased concentration of roughly equal teams. Our model became less confident in the choices as the weaker teams were eliminated.

# 7 Conclusion

Although we did not achieve substantial improvments from the baseline, we were consistently in a similar range, often performing slightly better. In this case, we believe that even getting to the baseline is an accomplishment, because those rankings are choosen by people who watch hundreds of games a season.

# 8 Future Work

There are a couple more things we would have liked to try given more time, mainly testing different features and algorithms. There are novel features such as distance traveled and past season performance, as well as player by player data that we did not have access to. Another big factor in sports is injuries, and right now we have no way to quantify the effect on a team if a player got hurt before the tournament started. We would also like to experiment with regularization and other machine learning libraries.