A Machine Learning Strategy for Predicting March Madness Winners

Jordan Gumm, Andrew Barrett, and Gongzhu Hu

Department of Computer Science, Central Michigan University
Mount Pleasant, Michigan 48859, USA
(gumm1jn, barre2as, hu1g)@cmich.edu

***Abstract*—The Division I NCAA Men's Basketball Tournament is a popular sporting event held annually to determine the leagues National Champion. Over the past several years the betting scene surrounding the tournament has become arguably more popular than the tournament itself, drawing in fans who bet billions overall on its outcome. In this paper, we discuss the statistical challenges in correctly predicting winners in the tournament and present a machine learning strategy for predicting the games. The Kaggle Machine Learning March Mania Competition was used to test the effectiveness of the model by comparing it against other machine-learning-based models submitted to the competition. Overall, the project was considered successful as it scored in the top 15 percentile of all submissions.**

Keywords—predictive modeling, non-linear regression, Kaggle Machine Learning contest, NCAA Men's basketball tournament

I. Introduction

Predicting winners of contests has always been attractive to many people largely due to the challenges presented in the prediction process and the rewards (e.g. pride and/or prizes) for successful predictions. Many events provide opportunities for this kind of prediction "games", including sports, lottery, presidential election, and performing arts. March Madness [1], the annual National College Athletic Association (NCAA) Division I basketball finals tournament, is a typical such event that has become one of the most popular sporting events in the United States. Due to the sheer number of teams involved, it creates a statistical challenge that has come to offer an enticing betting culture for its fans. Recent estimates place gambling on the tournament at around $2.5 billion yearly [1], [9].

The odds of predicting a perfect tournament outcome are 9.2 quintillion to 1 ($9.2e^{18}$: 1) and there has never been a documented perfect bracket [2]. A tournament *bracket* is a "grid of all the teams in the tournament and the path they have to follow to the Final Four and the championship game" [11]. Most tournaments do not have many, if any, brackets that make it beyond the 2nd round of competition in perfect condition.

We developed a model and entered it in the 2014 Kaggle Machine Learning March Mania Competition [5]. The training dataset consisted of every regular season game from 2011-2014 and every tournament game from 2011-2013. The competition used the 2014 tournament as the test dataset. The training process involved examination of individual variables and the correlations between them and the known outcomes. The strengths of these variables were analyzed and those variables with the highest correlations with the winning outcomes were selected to form the final prediction model.

[1]March Madness™is a trademark of NCAA

II. Background

The first NCAA Division I basketball final was held in 1939 with the term March Madness becoming prominent in the mid 1980's. It is so named because most of the games occur at the very end of March with little time between each game for each team, creating an environment where highly unlikely events occur in a very short span of time.

The modern tournament consists of 64 teams grouped into four regions and ranked by the selection committee based on attributes such as wins and strength of schedule. Starting in 2011, four additional teams were added to the field as play-in games, bringing the total number of teams to 68. After the initial play-in games the tournament rounds go 64 teams and 32 games, 32 teams and 16 games, etc. until there is one team left in the tournament. In each round the winning teams advance and the losing teams do not. The lone surviving team is determined to be the winner of the tournament, the National Champion.

Selecting winning teams in the tournament is a statistically difficult problem. While teams are ranked in the tournament based on many attributes, variabilities like player injuries and off-court issues are hard if not impossible to fully account for. Attributes such as the point difference, rebounds, steals, eld goal attempts, among many others will also vary significantly throughout a season due to a variety of factors.

Things such as the point difference of the teams, the rebounds, the steals, the field goal attempts, and many other traits will vary significantly throughout the season. Other factors don't have a statistical value. An example is that of a team that plays better as an "underdog." These types of psychological factors are hard to account for because the mentality of a player or team vary greatly throughout a season and from game to game.

A. Tournament seeding as predictor

The goal of correctly predicting all of the games has become more prominent as of recent, especially when considering that Warren Buffet offered $1 billion prize to the first

978-1-4799-8676-7/15/$31.00 copyright 2015 IEEE

person to achieve such a feat [2]. Sean McCrea examined how individuals make predictions for the tournament pools, in which individuals must correctly predict as many games in the tournament as possible. He demonstrated that people predict upsets (wins by a higher-seeded team) at a rate equal to the past frequency with various statistical results. The actual results of wins for each of the eight first-round games (i.e., 1 seed vs. 16 seed, 2 seed vs. 15 seed, etc.) collapsed across the four groups (i.e., regions) for the years 1985-2005 is shown in Fig. 1 [9]. It is seen in this figure that the tournament seedings are good predictors of the outcome of the games. It is about as good of a predictor as quite few other prediction methods.

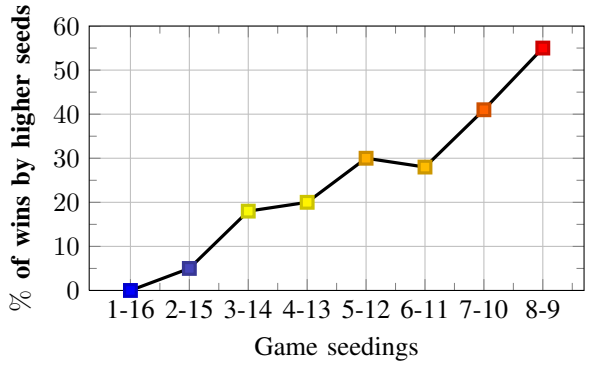


Fig. 1. Aggregate percentage of wins by seeding in round one of past NCAA Tournaments 1985–2005

B. ESPN Tournament Challenge

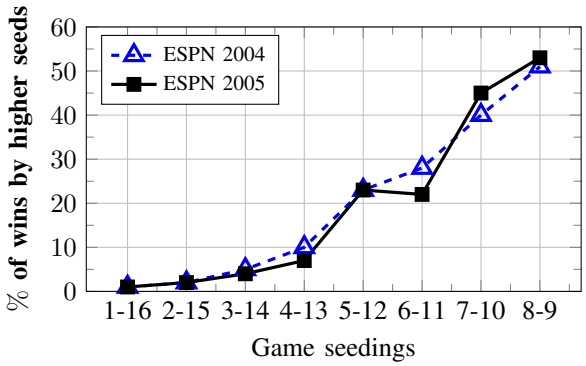
Every year ESPN (originally initials for Entertainment and Sports Programming Network) provides a tournament pool challenge where its users can compete for prizes. The aggregate predictions from 20042005 show that the predictions of ESPN users are less than 50% accurate, or worse than random as shown in Fig. 2(a). One explanation, built off of McCreas thesis, is that the players were stretching to predict all upsets which decreased their overall correct guesses.

The ESPN Tournament Challenge is more or less representative of predictions made by the average person. One would expect that expert predictions might follow a similar trend in an attempt to predict actual upsets. However, Fig. 2(b) shows that this is not the case. The experts rarely pick upsets in games involving the top 4 seeds.

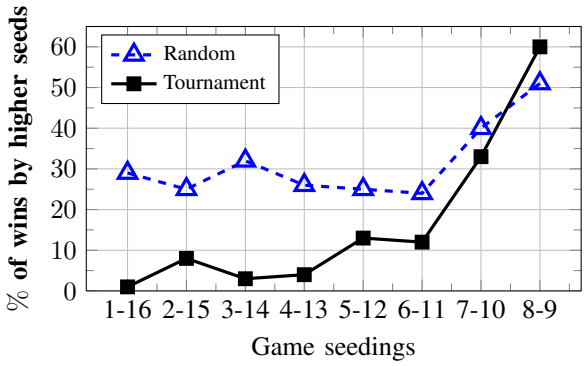
It seems that the rationale behind the methods experts have been using is to increase the changes of having the correct team chosen that is going into the next round. Historically speaking, less than 20% of seeds 1-4 will lose in the first round, so the experts are sacrificing perfection for an overall better result.

C. Other Considerations

People also considered other factors in making predictions, both general predictors (e.g. team ranking) and more specialized predictors. In specialized predictor, every game is judged individually by statistical means. It is an open debate as to which statistics have a highly causal relationship with winning.



(a) Average user predictions, 2004-05



(b) Expert predictions, 2008

Fig. 2. Prediction in round one of NCAA Tournament

III. Our Approach

We used an aggregate model of an ensemble of regression functions derived from the most highly-correlated statistical differentials in regard to wins. The training data included results from the 2011-2014 regular season and 2011-2013 NCAA tournament games including season statistics for each team. It contains 347 college basketball teams with 37 variables, partially shown in Table I.

TABLE I. Variables of Training Data Set

Variable	Description
3FG	Three Point Field Goal
APG	Assist per game
AST	Total Assist
BKPG	Blocks per game
BLKS	Total Blocks
BPI	College Basketball Power Index
FG	Field Goal
FT	Free Throw
L	loses
REB	Total Rebounds
RPG	Rebounds per game
RPI	Rating Percentage Index
PTS	Points
ST	Total Steals
STPG	Steals per game
TO	Turnover
W	Wins
W %	Winning Percentage
...	...

The basic approach we took is to look at individual variables and calculate a variety of basketball statistics, mostly the differences between winning team and losing team, percentages and ratios, such as

- RPI rank difference
- BPI difference
- Point total difference
- Steal total difference
- Block total difference
- Field goal percent difference

We analyzed the correlation between each of these variables and the chance of an individual team beating its opponent. The variables with large win-based correlation coefficients were selected to conduct nonlinear (quadratic) regression analysis regarding the change of winning.

The resulting regression functions for each statistic were used to predict each game by averaging their results, forming an aggregate-based model to calculate the chance of a team winning over another. The algorithm of our approach is given in Algorithm 1 and Algorithm 2.

Algorithm 1: Prediction of Winning

Input: D – Dataset with n data records with m variables, including information of winning W and losing L
Input: V – Set of variables in D
Output: P – Prediction of winning (as $\%$)

begin
  Let V' be the set of k new variables related to V;
  Let D' be the dataset with variables V';
  foreach $v \in V'$ **do**
    Calculate the n values D'_v from D;
    $f_v \leftarrow$ regression$(v, w_v \in W)$;
    $p_v \leftarrow calProbWin(D'_v, w_v, l_v, f_v)$;
  end
  $P \leftarrow \sum_v (p_v)/|V|$;
  return P;
end

Algorithm 2: calProbWin

Input: D'_v – Dataset with n data records under variable v
Input: w_v – winning team statistics on variable v
Input: l_v – losing team statistics on variable v
Input: f_v – regression function on variable v
Output: p – Probability of winning (as $\%$)

begin
  $x \leftarrow |w_v - l_v|$;
  $y \leftarrow f_v(x)$;
  if $(w_v - l_v) > 0$ **then**
    return y;
  else
    return $1 - y$;
end

Of these variables, Rating Percentage Index (RPI) had the largest level of significance per unit of difference. The RPI makes an attempt to take into consideration a teams strength of schedule, making it a valuable feature. Several other ratings indexes have been used to rank teams such as BPI and Sagarin, but RPI still remains in use by the NCAA Tournament selection committee [3].

Statistics such as total 3-point shots and total turnovers contributed insignificantly to estimated chance of winning. This could be due to confounding factors such as the pace a team plays at. A team that plays at a higher pace may have a higher turnover rate, but might also have a higher scoring rate.

The quadratic regression models with these new variables with respect of chance of winning are given below

Variable (x)	**Winning** $\%$ (y)
RPI diff	$y = 6e^{-06}x^2 - 0.0032x + 0.4607$
BPI diff	$y = -0.0002x^2 + 0.0198x + 0.512$
Steals diff	$y = 0.015x^2 - 0.0334x + 0.5556$
Points diff	$y = -2e^{-6}x^2 + 0.0018x + 0.5028$
Blocks diff	$y = 0.0077x^2 + 0.0252x + 0.5465$
Field goal pct	$y = -0.0027x^2 + 0.0698x + 0.4904$

and shown in Fig. 3(a)–3(f).

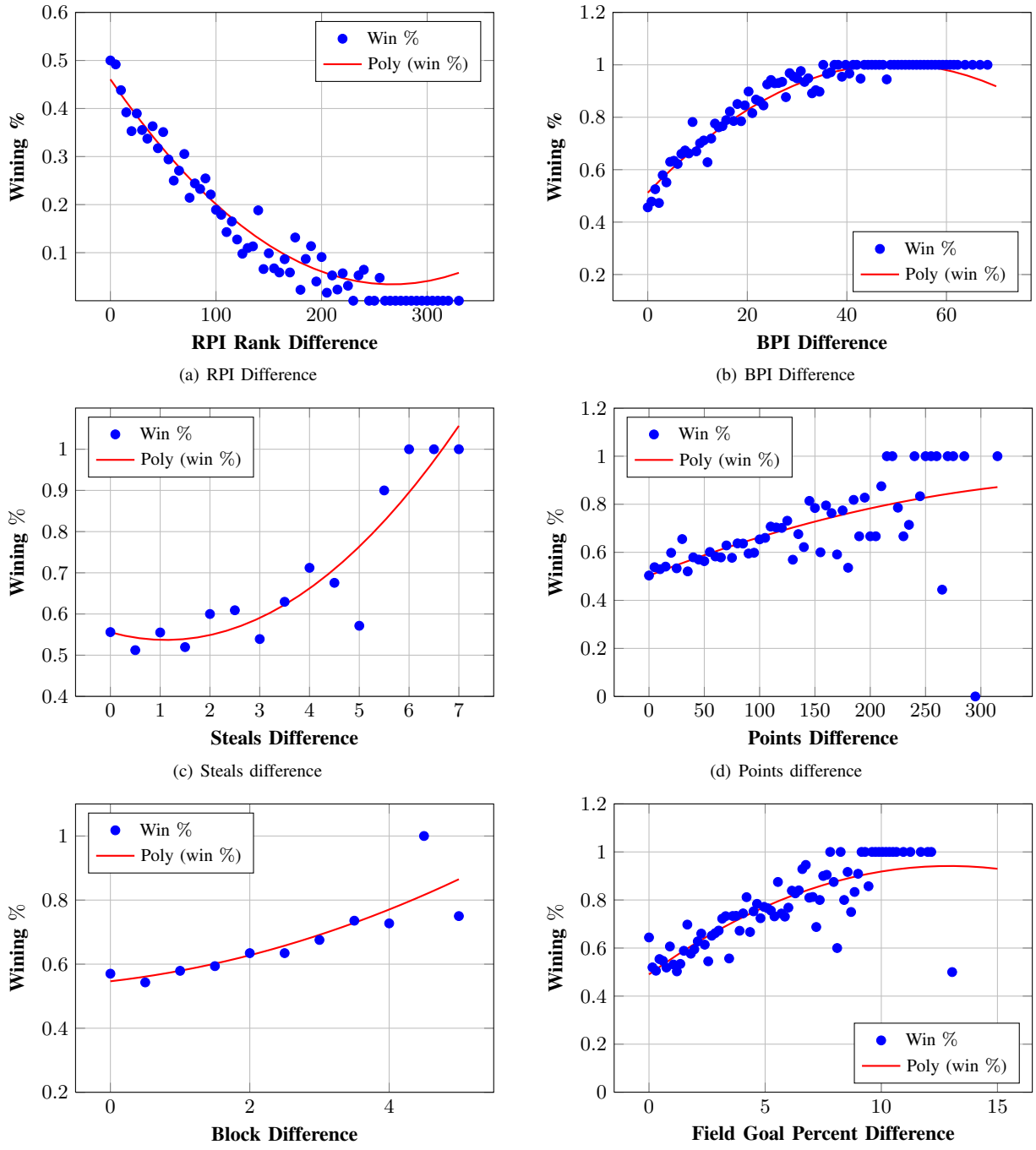
We also analyzed the other variables and combined them into one model that was then executed to produce a statistic analysis of the chances for a team to advance. The result we submitted to the Kaggle competition contains 2,278 entries, each is a team-pair and the predicted winning $\%$ of the first team. Partial results are given in Fig. 4.

```
id,pred
S_Albany_American,0.365
S_Albany_Arizona,0.184
S_Albany_Arizona St.,0.119
S_Albany_Baylor,0.308
S_Albany_BYU,0.181
S_Albany_Cal Poly,0.602
S_Albany_Cincinnati,0.231
S_Albany_Coastal Carolina,0.659
S_Albany_Colorado,0.157
S_Albany_Connecticut,0.183
S_Albany_Creighton,0.083
S_Albany_Dayton,0.369
S_Albany_Delaware,0.343
S_Albany_Duke,0.271
S_Albany_Eastern Kentucky,0.505
S_Albany_Florisa,0.231
.....
S_Wisconsin_Xavier,0.648
S_Wofford_Xavier,0.108
```

Fig. 4. Competitions Results of Model

IV. Evaluation

To do a field test of the accuracy of the model, the resulting probabilities for the 2014 bracket was submitted via the Kaggle March Machine Learning Mania Competition. This competition evaluates the machine learning results of



(a) RPI Difference

(b) BPI Difference

(c) Steals difference

(d) Points difference

(e) Blocks difference

(f) Field goal percent difference

Fig. 3. Quadratic fitting of various variables with winning %

user-submitted model outputs. The submissions are scored against each other via the LogLoss Equation (1).

$$LogLoss = -\frac{1}{n}\sum_{i=1}^{n}(y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)) \quad (1)$$

where n is the number of games played, $\hat{y}_i$ is the probability of team 1 beating team 2 in game i, y_i is the outcome of game i.

Our model peaked at 26th place out of the original 251 submissions but ended up finishing at 37th out of the final 248 submissions (three teams were removed for violating competition rules). This places our model in the top 15th percentile (see the final results in [4]). The Log Loss score of our model was 0.56411, better than both the median (0.65559) and the mean (0.57551) benchmarks of the competition. It easily beat out the seed-based model (0.60021).

As a secondary measure we analyzed the prediction accuracy of our model in regard to the number of total wins it

TABLE II. OUR PREDICTION OF WINNERS (ROUNDS 64 AND 32)

Region/Round	Winner		Loser		Prediction
	seed	team	seed	team	
South First Round	1	Florida	16	Albany	✓
	9	Pittsburgh	8	Colorado	✓
	12	SF Austin	5	VCU	
	4	UCLA	13	Tulsa	✓
	11	Dayton	6	Ohio St.	
	3	Syracuse	14	Western Mi.	✓
	10	Stanford	7	New Mexico	
	2	Kansas	15	E Kentucky	✓
East First Round	1	Virginia	16	Coastal Car.	✓
	8	Memphis	9	George Wash.	
	12	Harvard	5	Cincinnati	✓
	4	MSU	13	Deleware	✓
	6	UNC	11	Providence	✓
	3	Iowa St	14	NC Central	✓
	7	Uconn	10	St Joseph	✓
	2	Villanova	15	Milwaukee	✓
West First Round	1	Arizona	16	Weber St	✓
	8	Gonzaga	9	Oklahoma St	✓
	12	ND St	5	Oklahoma	✓
	4	SD St	13	NM St	
	6	Baylor	11	Nebraska	✓
	3	Creighton	14	La Lafayette	✓
	7	Oregon	10	BYU	✓
	2	Wisconsin	15	American	✓
Midwest First Round	1	Wichita St	16	Cal Poly	✓
	8	Kentucky	9	KSU	✓
	5	St Louis	12	NC State	✓
	4	Louisville	13	Manhattan	✓
	11	Tennessee	6	UMass	✓
	14	Mercer	3	Duke	
	7	Texas	10	Arizona St.	
	2	Michigan	15	Wofford	✓
South Round of 32	1	Florida	9	Pittsburgh	✓
	4	UCLA	12	SF Austin	✓
	11	Dayton	3	Syracuse	✓
	10	Stanford	2	Kansas	
East Round of 32	1	Virginia	8	Memphis	✓
	4	MSU	12	Harvard	
	3	Iowa St	6	UNC	✓
	7	UConn	2	Villanova	
West Round of 32	1	Arizona	8	Gonzaga	✓
	4	SD St	12	ND St	✓
	6	Baylor	7	Creighton	
	2	Wisconsin	7	Oregon	✓
Midwest Round of 32	8	Kentucky	1	Wichita St	
	4	Louisville	5	St Louis	✓
	11	Tennessee	14	Mercer	✓
	2	Michigan	7	Texas	✓

correctly predicted. Taking the teams the model predicted to have the highest probability of winning in each game we came up with 42 correct predictions out of a possible 63, as shown in in Table II (rounds 64 and 32) and Table III (rounds 16, 8, 4, and the championship). Considering that the tournament had both the highest number of seed-based upsets (teams with a higher seed beating teams with a lower seed) in the history of the 64-team tournament [12] and the lowest seed in 29 years to win the tournament [6], we are optimistic that our model could serve as even a better predictor of games in future tournaments. An interesting side-note is that our submission correctly predicted both 12 over 5 upsets and one of the two 11 over 6 upsets (Tennessee over UMass).

TABLE III. OUR PREDICTION OF WINNERS (ROUNDS 16, 8, 4, AND CHAMPIONSHIP)

Region/Round	Winner		Loser		Prediction
	seed	team	seed	team	
South Sweet 16	1	Florida	4	UCLA	✓
	11	Dayton	10	Stanford	
East Sweet 16	4	MSU	1	Virginia	
	7	UConn	3	Iowa St	✓
West Sweet 16	1	Arizona	4	SD St	✓
	2	Wisconsin	6	Baylor	✓
Midwest Sweet 16	8	Kentucky	4	Louisville	
	2	Michigan	11	Tennessee	
Elite 8	1	Florida	11	Dayton	✓
	7	UConn	4	MSU	
	2	Wisconsin	1	Arizona	
	8	Kentucky	2	Michigan	✓
Final Four	7	UConn	1	Florida	
	8	Kentucky	2	Wisconsin	
Final	7	UConn	8	Kentucky	

By these standards of evaluation, our model and algorithms are considered successful.

V. RELATED WORK

The NCCA tournaments are so fascinating that many models have been proposed and developed to predict the outcomes. We shall briefly discuss some of these models in this section.

A. Kaggle

There were 248 final submissios in the 2014 Kaggle contest and each team was allowed two submissions. Many different approaches were presented that had a varying degree of success throughout the tournament. Some of these approaches included deep learning methods.

Many previous Kaggle contests have been won with the use of a more sophisticated approach called Deep Learning to derive their conclusions. These deep learning methods can involve many different characteristics, but many of them use artificial neural networks. An example of this deep learning concept with neural networks was created by Google Brain in a project that learned what a cat was by watching YouTube videos.

The winner of the 2014 Kaggle contest was one of the two entries submitted by Lopez and Matthews. Their model was built using logistic regression that the maximum likelihood estimates derived from the model are based on maximizing a function that was equivalent to the scoring function used in the contest [8].

B. Harvard Sports

It is argued by Harvard Sports that there is significant evidence that tournament models don't reflect regular seasons models [12]. Reasons that they listed included that the "NCAA tournament games carry much more added pressure and added attention," as well as the fact that for a lot of the teams there

are more fans present than any games that they had played previously throughout their regular season (Harvard Sports).

They recommend weighing certain attributes, such as the consistency of a team, to see beyond the faults of normal variability. The variables that are deemed to be highly significant focus primarily on the previously stated consistency, the strength of a schedule, the experiences of the players on the team, and the amount of time that the team had played in previous tournament games. It is argued that a team even getting a bid into the tournament should be considered with a higher sense of prestige than teams who had not previously made it into the tournament. Their model is an attempt at determining the survivability of a team.

C. Ken Pom Model

A highly successful predictive model for basketball analysis was created by Ken Pomeroy. This model is purely predictive of how well a team would do against the average team. It does not consider injuries or emotional factors of a team for a game. This model has become very popular, requiring a subscription to to view a majority of the advanced statistics.

This model uses a Pythagorean Expectation approach, shown in Equation (2). It was initially created to be used for predictive baseball modeling, but was adapted by Ken Pomeroy to be used with variables pertaining to basketball.

$$Win = \frac{s^2}{s^2 + a^2} = \frac{1}{1 + (a/s)^2} \tag{2}$$

where s is number of runs scored and a is number of runs runs allowed.

This model is represented in such a way that it is easy to observe and analyze previous seasons. It creates an adjusted offensive and defensive rating based on the Strength of Schedule and the opponents offensive vs defensive skills.

D. Academic Work

Researchers in the academic world have also studied the NCCA tournaments data and developed various models for the prediction of winners. Schwertman et al [10] built linear and logistic regression models using seed position for predicting the probability of each of the 16 seeds winning the regional tournament. Kvam and Sokol [7] presented a combined logistic regression/Markov chain (LRMC) model for predicting the outcome of NCAA tournament games. Zimmermann and his colleagues evaluated various machine learning techniques applied to predicting NCCA basketball matches and concluded that attributes seem to be more important than models, and there seems to be an upper limit to predictive quality [13]. Fuqua attempted to accurately predict the four semi-finalists of the NCCA Division I men's basketball tournament using a binary choice logit model to predict and claimed that the model does better than any current rating system at predicting the final four teams [3].

VI. Conclusion

We applied data mining methods to the NCAA Division I Men's Basketball Tournament and built a machine learning prediction model that was submitted to the 2014 Kaggle March Mania competition. Although our entry did not win, it beat out all benchmarks while achieving better results than the majority of competitors, finishing in the top 15 percentile at the end of the tournament.

Our predictions certainly could have benefited from a more sophisticated model, such as one created by a deep neural network (DNN) or deep belief network (DBN). It would be ideal to gain a grasp on deep learning to further the development of algorithms for this March Madness problem.

It is important that the information used to derive these machine learning approaches utilize many years worth of data to account for the variability of these factors. The purpose of this solution is not necessarily to solve the tournament every time, but rather to improve the odds of predicting the tournament correctly through continuous development.

Thus it is important that the information used to derive these machine learning approaches utilize many years worth of data to account for the variability of these factors. The purpose of this solution is not necessarily to solve the tournament every time, but rather to reduce the odds of predicting a tournament correctly through continuous development.

References

[1] Liz Clarke. March madness turns gamblers into basket cases. *Washington Post*, page A1, 2005.

[2] Sydney Ember. $1 billion for a perfect N.C.A.A. bracket, courtesy of Warren Buffett. *New York Times*, 2014.

[3] Cameron Fuqua. The final four formula: A binary choice logit model to predict the semi-finalists of the NCAA Division I mens basketball tournament. *Major Themes in Economics*, pages 31–49, 2014.

[4] Kaggle. March machine learning mania. https://www.kaggle.com/c/march-machine-learning-mania/leaderboard.

[5] Kaggle. March machine learning mania 2014. http://www.kaggle.com/c/march-machine-learning-mania-2014.

[6] Howie Kussoy. UConn defeats Kentucky to win NCAA championship. *New York Post*, 2014.

[7] Paul Kvam and Joel S Sokol. A logistic regression/markov chain model for ncaa basketball. *Naval research Logistics (NrL)*, 53(8):788–803, 2006.

[8] Michael J Lopez and Gregory Matthews. Building an NCAA mens basketball predictive model and quantifying its success. *arXiv preprint arXiv:1412.0248*, 2014.

[9] Sean M McCrea and Edward R Hirt. Match madness: Probability matching in prediction of the NCAA basketball tournament1. *Journal of Applied Social Psychology*, 39(12):2809–2839, 2009.

[10] Neil C Schwertman, Kathryn L Schenk, and Brett C Holbrook. More probability models for the NCAA regional basketball tournaments. *The American Statistician*, 50(1):34–38, 1996.

[11] Wikipedia. March machine pools. http://en.wikipedia.org/wiki/March_Madness_pools.

[12] Ben Zauzmer. NCAA tournament 2014: The most upsets ever. *Harvard Sports Analysis*, 2014.

[13] Albrecht Zimmermann, Sruthi Moorthy, and Zifan Shi. Predicting college basketball match outcomes using machine learning techniques: some results and lessons learned. *arXiv preprint arXiv:1310.3607*, abs/1310.3607, 2013.