Predicting March Madness via Classification

A Data Mining Approach to Bracketology

Jonathan Pfeil

April 28, 2014

Contents

Abstract

The annual 64-team NCAA Men's Basketball tournament, known as "March Madness", is one of the most popular sporting events of the year both in terms of viewership and gambling. Fans and analysts use various strategies while filling out their brackets, ranging from "gut feelings" to advanced ranking metrics such as Kenpom and Saragrin ratings. This purpose of this study was to evaluate the effectiveness of classification algorithms in predicting winners of individual basketball games, as well as entire tournament brackets. The motivating hypothesis was that differences in team's play styles may be able to predict the outcome of individual matchups better than just ranking teams according to a statistical model, and that classification algorithms would be able to learn these matchup-specific patterns.

1 Overview and Related Work

1.1 Background: March Madness

The NCAA Men's Division I Basketball Championship, commonly known as "March Madness", is a 64-team Men's College basketball tournament held every March. March Madness is one of the most popular sports events of the year, both in terms of viewership and gambling. 52 million viewers tuned in for the tournament last year [1], and Las Vegas bets exceeded $100 million [2]. Fans and analysts pride themselves in their abilities to accurately predict the outcomes of games. This has led the massive popularity of "bracket pools", where fans submit a bracket predicting the outcome of every one of the 63 games. In 2013, ESPN alone had 8.15 million bracket submissions to its online competition [3].

Basketball has an enormous amount of individual and team statistics collected and recorded every game. With years of historical NCAA data available, Data Mining techniques can be used to provide quantitative analysis on which teams have the highest probability of success. By applying these principles to each game of the tournament, it is possible to fill out an entire 63-game bracket.

1.2 Background: Statistical Predictions

The idea of leveraging data to predict outcomes of sporting events is a subset of a popular field called Sports Data Analysis [4]. Professional sports teams now employ teams of data scientists to help them make decisions at all levels of the game. Predictive models are now used in choosing when to make player substitutions, surfacing hidden talent in scouting reports, and finding optimal strategies to play against opponents [4].

Many attempts have been made to apply Sports Data Analysis principles to NCAA basketball games.

Studies such as Schwertman et al. [5] and Boulier and Stekler [6] built models using only team seedings, and found that seeding is a statistically significant predictor of success. More advanced ranking mechanisms popularized by Jeff Sagarin, Ken Pomeroy, Kenneth Massey, and others are derived from a combination of attributes such as home and away wins, winning margin, strength of schedule, and performance against high-ranked teams [8]. Nate Silver has experimented with using a linear combination of multiple ranking systems to generate a new, composite ranking system [9]. After a ranking system has been constructed, winners are predicted by just choosing the higher ranked team in a given matchup.

A slightly more complicated system can be used to generate probabilities of each team winning a given game. Instead of just ranking teams, a statistical model can be built to assign each team what is called a "power rating". A power rating is a system in which the difference of two teams' ratings yields the point-spread, or expected margin of victory, if the two teams were to play each other [11]. Research has been done by Jess Sonas showing that ordinal ranks can be transformed to power ratings via a logarithmic function if a given rating is not a power rating [13]. Given a matchup, the power ratings are then subtracted to obtain the point spread. Stern showed that point spreads are normally distributed with $\mu = 0, \sigma = 10.75$ and that the probability of a team winning a matchup is given by the cumulative distribution function $\Phi(Y/\sigma)$ [12]. Therefore, given a rating or ranking system, we can generate the probability of victory for each team in a given matchup.

Even though these probabilities are more descriptive than the original models, they suffer from the same basic flaw: in using a well-ordered ranking or rating system to choose the winner, we have ignored matchup-specific information. Differences in teams' play styles may give a worse overall team an advantage over a better overall team when they face each other. For example, a team which has dominated the league through its strong interior defense may not be favored against a small, three-point shooting team, even though they would perform better than their opponent against the rest of the league. With this in mind, the hypothesis of this paper is that more advanced data mining techniques may be able to learn matchup-specific patterns and outperform traditional statistical prediction methods.

1.3 Background: Data Mining Predictions

Although classification and regression typically take a back seat to statistical methods in mainstream sports data analysis, research has still been done on the subject. Dragan Miljkovic et al. treated NBA game-outcome prediction as a classification problem, in which each game could be labeled as "Home Team Victorious" or "Away Team Victorious" [10]. By applying a decision tree and a Naive Bayes' classifier, they were able to

predict the outcome of NBA games with 67% success [10].

Research done by Zifan Shi et al. looked at using Machine Learning techniques to predict the outcome of individual NCAA basketball games – a similar goal to this project. This research used various classification algorithms to try to predict the outcome of individual games throughout a season. The features used in this research fell into two categories: base stats and derived stats. The base stats were just aggregated stats throughout a season (Points Scored, Points Allowed, etc.), while the derived statistics relied on a wealth of statistical basketball knowledge from analysts such as Dean Oliver and Ken Pomery. There were two major takeaways from this research, both of which influenced the way I conducted my research:

1. The complexity of the classifier used did not have a major impact on the predictive power. Simple classifiers such as Naive Bayes performed as well as advanced methods such as Artificial Neural Networks. Rather, the success of classifiers was almost entirely dependent on the quality of the features used [14]. Because of this, I spent most of my effort on this project researching advanced basketball metrics that could be useful as features.

2. There appears to be a glass ceiling in predictive power in sports data mining. Zifan Shi et al. estimated that the inherent volatility on live sports made it impossible to consistently predict more than 74-75% of NCAAB games correctly [14]. Similar findings have been reported in sports data mining applied to soccer, American football, NCAA football, and NBA basketball [14]. With this in mind, any results showing successful predictions of $> 80\%$ should be a red flag of either over training or of leaking future data into my datasets.

2 Methodology

Below is a general outline of the methodology used in this project. The details of each step will be expanded upon in the sections afterwards.

1. Obtain box-score data for each game played in the 2010-2014 seasons and save it in a postgresql database.

2. Calculate aggregated and derived stats for each team after each game. These will be used as features in the classification of their next game.

3. Train game-prediction classifiers by treating each game as a classification problem which gets labeled as “Upset” or “No Upset”, where “No Upset” predicts the higher ranked team to win (according to

NCAA's RPI rating) and "Upset" predicts the lower ranked team to win.

4. Recursively apply the trained classifiers to NCAA brackets to make a prediction for each of the 63 games.

5. Evaluate the success of each classifier.

3 Data

3.1 Data Acquisition

The NCAA Archive has box-score data available for each of the 27,080 Division-I NCAAB games played in the 2010-2014 seasons, available at `http://stats.ncaa.org/team/inst_team_list?division=1&sport_code=MBB`. An example box-score of a game played between Pittsburgh and North Carolina is shown in Figure 1. Highlighted in red shows the stats for each team that would be scraped and saved to a postgresql database.

3.2 Base Stats

For each game, the stats shown in Table 1 are saved for each team.

Stat Name	Abbreviated Name
Points	PTS
Field Goals	FG
Field Goal Attempts	FGA
Three Field Goals	3FG
Three Field Goal Attempts	3FGA
Free Throws	FT
Free Throw Attempts	FTA
Offensive Rebounds	OREB
Defensive Rebounds	DREB
Total Rebounds	REB
Assists	AST
Turnovers	TOV
Steals	STL
Blocks	BLK
Fouls	FOUL

Table 1: Table showing team's base stats

Additionally, each team's opponents stats are stored. The base stats of a team's opponent are notated with a superscript as shown in Table 2.

Box Score | Play by Play | Summary by Halves

Pittsburgh

Player	Pos	MP	FGM	FGA	3FG	3FGA	FT	FTA	PTS	ORebs	DRebs	Tot Reb	AST	TO	STL	BLK	Fouls
Patterson, Lamar	F	35:00	4	11	3	8	1	5	12		5	5	2	6			4
Robinson, James	G	35:00	5	7	1	1	8	12	19		6	6	3	2	4		3
Wright, Cameron	G	34:00	4	10			3	8	11		2	2	2	2	2		2
Young, Michael	F	20:00	2	5		2	1	1	5	1	2	3				1	4
Zanna, Talib	F	34:00	8	13			3	3	19	10	11	21	3	2	1	1	5
Artis, Jamel		12:00	2	3	1	2			5		2	2		1			4
Newkirk, Josh		28:00	2	3			5	10	9		1	1	2	2			2
Randall, Derrick		2:00						2									2
TEAM										1	2	3					
Totals		200:00	27	52	5	13	21	41	80	12	31	43	12	15	7	2	26

North Carolina

Player	Pos	MP	FGM	FGA	3FG	3FGA	FT	FTA	PTS	ORebs	DRebs	Tot Reb	AST	TO	STL	BLK	Fouls
Tokoto, J.P.	F	31:00	2	8		2	1	5	5	3	4	7	4	1	1		5
McAdoo, James Michael	F	36:00	4	13		1	7	9	15	1	6	7	2	1			2
Meeks, Kennedy	F	8:00		2			1	2	1	2	1	3		1			2
McDonald, Leslie	G	16:00		2		1				1	1	2		3			5
Paige, Marcus	G	35:00	9	20	4	12	5	5	27		4	4	1	2			5
Johnson, Brice		21:00	8	9					16	3	3	6			1	1	3
Hubert, Desmond		7:00									1	1				1	
Hicks, Isaiah		10:00					1	2	1		1	1					4
Simmons, Jackson		5:00					1	2	1								1
James, Joel		3:00						1			2	2		1			1
Davis, Luke		1:00															
Britt, Nate		26:00	2	7			5	6	9	1		1	5	3	3		4
Moody, Wade		1:00															
TEAM											1	1					
Totals		200:00	25	61	4	16	21	32	75	11	24	35	12	12	5	2	32

Figure 1: Example box-score with team stats highlighted

3.3 Basic Derived Stats

Given the previous Base Stats, we can define the following derived stats for a team and its opponent:

$$FG\% = \frac{FG}{FGA} \tag{1}$$

$$3FG\% = \frac{3FG}{3FGA} \tag{2}$$

$$FT\% = \frac{FT}{FTA} \tag{3}$$

Stat Name	Abbreviated Name
Points	PTS^{OPP}
Field Goals	FG^{OPP}
Field Goal Attempts	FGA^{OPP}
Three Field Goals	$3FG^{OPP}$
Three Field Goal Attempts	$3FGA^{OPP}$
Free Throws	FT^{OPP}
Free Throw Attempts	FTA^{OPP}
Offensive Rebounds	$OREB^{OPP}$
Defensive Rebounds	$DREB^{OPP}$
Total Rebounds	REB^{OPP}
Assists	AST^{OPP}
Turnovers	TOV^{OPP}
Steals	STL^{OPP}
Blocks	BLK^{OPP}
Fouls	$FOUL^{OPP}$

Table 2: Table showing opponent's base stats

3.4 Dean Oliver Derived Stats

In *Basketball on Paper*, author Dean Oliver gave a thorough statistical analysis of NBA basketball teams in an attempt to provide quantitative explanations for team success [15]. My research relies heavily on his work in order to provide predictive features for use in classification; all of the equations below come from *Basketball on Paper*.

In order to normalize statistics w.r.t the "pace of the game", we must know the number of possessions that a team had in a game. A possession represents one team controlling the ball; by the rules of basketball, teams necessarily alternate possessions, so a team and its opponent will have the same number of possessions in a given game. Although the number of possessions is not an officially recorded stat, we can estimate it because we know that a possession can only end in one of three ways: a field goal attempt that is not rebounded by the offense, a turnover, or a set of free throws [15]. With this in mind, Dean Oliver has defined the following:

$$Possessions = FGA - OREB + TOV + 0.4 * FTA \tag{4}$$

where the 0.4 coefficient represents the proportion of free throws that end a possession.

Given the number of possessions in a game, we can define a team's offensive and defensive efficiencies as

the number of points scored and allowed per 100 possessions respectively.

$$OE = \frac{100 * PTS}{Possessions} \tag{5}$$

$$DE = \frac{100 * PTS^{\text{OPP}}}{Possessions} \tag{6}$$

We can then derive stats that give us an idea of *how* teams score their points. This will help capture a team's style of play, which we have hypothesized may have strong predictive power in determining the winner a given matchup. We define the following [15]:

Scoring Possessions - Number of possessions in which at least one point is scored

$$ScoringPossessions = FG + \left[1 - (1 - FT\%)^2\right] * 0.4 * FTA \tag{7}$$

Plays - Number of times a team has the ball (unlike possessions, an offensive rebound marks a new play)

$$Plays = Possessions + OREB \tag{8}$$

Floor Percentage - Percentage of team's possessions in which at least one point is scored

$$Floor\% = \frac{ScoringPossessions}{Possessions} \tag{9}$$

Field Percentage - Percentage of team's non-foul shot possessions in which at least one point is scored

$$Field\% = \frac{FG}{FGA - 1.07 * \frac{OREB}{OREB + DREB^{\text{OPP}}} * (FGA - FG) + TOV} \tag{10}$$

Play Percentage - Percentage of team's plays in which at least on point is scored

$$Play\% = \frac{ScoringPossessions}{Plays} \tag{11}$$

Dean Oliver provides a few rules of thumb regarding what these derived stats imply about a team's style of play [15]:

1. Teams with high floor percentages and high offensive ratings score very well from the two-point area

2. Teams with low floor percentages and high offensive ratings score their points by shooting three-pointers
3. Teams with high field percentages commit few turnovers, get offensive rebounds, and shoot well
4. Teams with high play percentages and high offensive ratings score well on initial shots, not relying on offensive rebounds to generate points

In addition to these, Dean Oliver relies heavily on the "Four Factors" – four derived stats that he believes captures how effective a team is. They are defined and calculated as follows:

Effective Field Goal % - Field goal % adjusted for the importance of three-pointers

$$eFG\% = \frac{FG = 0.5 * 3FG}{FGA} \tag{12}$$

Offensive Rebound % - Percentage of offensive rebound opportunities which result in a rebound

$$OREB\% = \frac{OREB}{OREB + DREB^{\text{OPP}}} \tag{13}$$

Turnover Rate - Percentage of possessions in which a turnover is committed

$$TOV\% = \frac{TOV}{Possessions} \tag{14}$$

Free Throw Rate - Percentage of field goal attempts in which a team draws a foul

$$FTR = \frac{FTA}{FGA} \tag{15}$$

Finally, Dean Oliver provides a way to estimate the win percentage of a team given its statistics. This stat relies on the effect that volatility has on a team's likelihood to win a game. If a team is better than their opponents on average, then being inconsistent will cause them to lose games that they should have won. However, if a team is worse on average, being inconsistent will cause them to win games that they should have lost. This means that the having a large variance in points scored and points allowed brings teams closer to .500, no matter how good they are. With this knowledge, Dean Oliver uses the cumulative distribution function of the standard normal distribution to estimate win percentage as follows:

$$Win\% = \Phi\left(\frac{FG_{\text{mean}} - FG^{\text{OPP}}_{\text{mean}}}{FG_{\text{variance}} + FG^{\text{OPP}}_{\text{variance}} - 2 * COV(FG, FG^{\text{OPP}})}\right) \tag{16}$$

where $COV(x, y)$ is the covariance of x and y

3.5 KenPom Derived Attributes

Ken Pomeroy, author of `http://www.kenpom.com`, has expanded upon the work of Dean Oliver in advanced basketball analysis. The equations in this section come from the work he has published on his blog, `kenpom.com` [16].

Pomeroy has advanced the idea of offensive and defensive efficiency under the notion that scoring points against a good defense should be considered more impressive than scoring points against a bad defense, and stopping a good offense should be considered more impressive than stopping a bad offense. This leads to the idea of adjusted offensive and defensive efficiencies. We define them as follows:

Adjusted Offensive Efficiency - An estimate of the offensive efficiency a team would have against the average D-I defense

$$AdjOE = \frac{OE * OE^{\text{LeagueAvg}}}{AdjDE^{\text{OPP}}} \tag{17}$$

Adjusted Defensive Efficiency - An estimate of the defensive efficiency a team would have against the average D-I offense

$$AdjDE = \frac{DE * DE^{\text{LeagueAvg}}}{AdjOE^{\text{OPP}}} \tag{18}$$

To avoid the recursive nature of these equations, I used the mean of the opponent's adjusted offensive and defensive efficiencies as of the $(t-1)^{st}$ game of the season to calculate the adjusted offensive and defensive efficiencies in the t^{th} game. This gives the following equations:

$$AdjOE^t = \frac{OE * OE^{\text{LeagueAvg}}}{(AdjDE^{\text{OPP}})^{t-1}} \tag{19}$$

$$AdjDE^t = \frac{DE * DE^{\text{LeagueAvg}}}{(AdjOE^{\text{OPP}})^{t-1}} \tag{20}$$

Since kenpom only makes end-of-season ratings available, I used the above equations to calculate these ratings myself at every point in the season. My end-of-season results matched kenpom's within a constant factor, indicating that the only difference in our calculations were the values used for $OE^{\text{LeagueAvg}}$ and $DE^{\text{LeagueAvg}}$.

Finally, kenpom has created a statistic that uses a teams adjusted offensive and defensive ratings to generate an expected winning percentage:

Pythagorean Winning % - A team's expected winning percentage against an average D-I team

$$Pyth = \frac{AdjOE^{\gamma}}{AdjOE^{\gamma} + AdjDE^{\gamma}} \tag{21}$$

with an empirically derived value of $\gamma = 10.25$

3.6 Cumulative Stats

Given the base stats and derived stats for each team in each game, I then calculate a set of Cumulative Stats at each point in the season to be used as features in predicting the result of the next game. For each base stat and derived stat (both of a team and its opponent), I keep track of four new cumulative stats:

1. Minimum value in a game up to this point in the season
2. Maximum value in a game up to this point in the season
3. Mean value across all games up to this point in the season
4. Variance of value across all games up to this point in the season

The following formulas are used to update these values iteratively over the course of a season, where V is an arbitrary stat and V^t is the value of that stat in the team's t^{th} game of the season:

$$V^{\text{t}}_{\text{min}} = min(V^{\text{t-1}}_{\text{min}}, V^{\text{t}}) \tag{22}$$

$$V^{\text{t}}_{\text{max}} = max(V^{\text{t-1}}_{\text{max}}, V^{\text{t}}) \tag{23}$$

$$V^{\text{t}}_{\text{mean}} = \frac{V^{\text{t}} + (t-1)V^{\text{t-1}}_{\text{mean}}}{t} \tag{24}$$

$$V^{\text{t}}_{\text{variance}} = \left(\frac{t-2}{t-1}\right) V^{\text{t-1}}_{\text{variance}} + \frac{\left(V^{\text{t}} - V^{\text{t-1}}_{\text{mean}}\right)^2}{t} \tag{25}$$

The following base cases are used to complete these recursive definitions:

$$V^{1}_{\text{min}} = V^{1} \tag{26}$$

$$V^1_{\max} = V^1 \tag{27}$$

$$V^1_{\text{mean}} = V^1 \tag{28}$$

$$V^1_{\text{variance}} = 0 \tag{29}$$

3.7 Rating Systems

In addition to the Base Stats and Derived Stats listed above, there exists a wealth of knowledge about how good teams are at every point in a season in the form of rating systems. ESPN, sports writers, newspapers, bloggers, etc., all have various statistical methods that they use to generate rating systems, as discussed in Section 1.2. Although these may not perform well as individual predictors of success, they may be valuable features in a more complex classification model.

Kenneth Massey, an American sports statistician has compiled and maintained weekly historical team ratings for forty different rating systems [17]. The ratings used are shown below in Table 3. These ratings are available publicly at `http://www.masseyratings.com/cb/compare.htm`.

3.8 Team Feature Summary

In the above sections, we have looked at 15 base stats, 3 basic derived stats, 13 Dean Oliver stats, and 5 kenpom stats. For each of these[1] we also record V^{OPP}, the opponent's value for the stat (or the amount of that stat that the team has "allowed"). We then calculate 4 cumulative stats for each of the stats we have at this point: $V_{\min}, V_{\max}, V_{\text{mean}}, V_{\text{variance}}$. Finally, we add in the 40 ratings obtained from Kenneth Massey. In total this gives us 316 features that describe a team at a given point in the season.

This can be viewed as a vector $\mathbf{x} = \langle x_1, x_2, ..., x_n \rangle$, where $|\mathbf{x}| = 316$

3.9 Game Features

All of the features discussed so far have been with respect to a team; however, we ultimately want to run our classification algorithm on games, not teams. Therefore we need a way to generate a feature vector for a

[1] Excluding OE, DE, $AdjOE$, and $AdjDE$, since $OE \equiv DE^{\text{OPP}}$

System Name	Abbrev.	System Name	Abbrev.
Associated Press	AP	Pugh Power	PPR
Baker Bradley-Terry	BBT	Rewards	REW
Bihl	BIH	RPI	RPI
BVL Least Squares	BLS	RoundTable	RT
Bobcat	BOB	Roundtable BCS	RTB
Colley	COL	Rothman	RTH
D1A Sports	D1A	RT Power	RTP
DeSimone	DES	Sagarin	SAG
Donchess Inference	DII	Sagarin-Elo	SE
Dokter Entropy	DOK	Self	SEL
Dolphin	DOL	Sagarin Predictor	SP
ESPN BPI	EBP	Snapper's World	SPW
Kirkpatrick	KPK	Stanford	STF
Krach	KRA	The Power Rank	TPR
Massey	MAS	TeamRankings Pred	TRP
Moore	MOR	USA Today Coaches	USA
Nolan	NOL	Wilson	WIL
Pugh	PGH	Wiemeyer	WMR
Pigskin	PIG	Wobus	WOB
Pomeroy	POM	Wolfe	WOL

Table 3: Summary of Rating Systems

game given the two teams participating in it. The strategy taken for this project was to use the difference of the two teams' feature vectors – specifically, the high-seed's feature vector minus low-seed's feature vector.

$$\mathbf{g} = \langle g_1, g_2, ..., g_n \rangle = \mathbf{x}^{\text{high-seed}} - \mathbf{x}^{\text{low-seed}} \tag{30}$$

Since all of the features are positive, we have not lost information about which team's feature value was larger; this is now indicated by the sign of the game feature. Additionally, since the original features were approximately normally distributed, and we know that the sum/difference of normally distributed RVs is normally distributed, the game features are normally distributed as well.

This representation is not without weaknesses, however. It is quite possible that information regarding a team's style of play has been lost in this transformation – a team shooting more three-pointers than its opponent does not necessarily mean that team is a deep threat. Alternative representations of a game is a potential area for future research.

4 Preprocessing

4.1 Normalization

Many of the classification algorithms that will be used require that the features roughly follow a standard normal distribution. The game features are already roughly normally distributed; therefore, to transform our features, we only have to remove the mean and scale by the standard deviation.

Let $\mathbf{g}_\mu$ be the mean feature vector over all games

Let $\mathbf{g}_\sigma$ be a vector whose i^{th} entry is the standard deviation of the i^{th} game feature

Then we can define $\mathbf{g_c}$, the normalized version of a given game vector $\mathbf{g}$, elementwise as:

$$g_{c_i} = \frac{g_i - g_{\mu_i}}{g_{\sigma_i}} \tag{31}$$

4.2 Feature Selection

27,080 game vectors of length 316 each is a large amount of data – more data than I hoped to use in classification. Since a large number of games provides helpful training data, I chose to prune the set of features associated with a game. Ideally this would increase the speed of computation by reducing the dimensionality and improve the accuracy by removing features that only serve as noise.

An ANOVA F-test was performed on the set of normalized game vectors, $\left\{\mathbf{g}_\mathbf{c}^1, \mathbf{g}_\mathbf{c}^2, ..., \mathbf{g}_\mathbf{c}^\mathbf{N}\right\}$, against the set of game labels $\left\{l^1, l^2, ..., l^N\right\}$, $l^i \in \{0, 1\}$ where 0 represents "No Upset" and 1 represents "Upset", in order to remove features with little-to-no correlation with the result of a game.

During classification, the top k features according to the ANOVA test were used, with k determined experimentally. The top 51 features are listed in Table 4.

As we would expect, popular ranking systems and derived possession-based statistics are the most predictive stats. The least predictive stats are all variances of base or derived stats.

5 Classification

5.1 5-fold cross validation methodology

The methodology used for testing the classification algorithms is as follows:

1. Let $\mathbf{G}$ be the set of game vectors $\mathbf{g}$, each of which is composed of the top k selected features. Let

Rank	Feature	Rank	Feature	Rank	Feature
1	$PYTH_{\text{mean}}$	18	$FG\%_{\text{mean}}$	35	$AdjOE_{\text{min}}$
2	POM	19	PTS_{mean}	36	$FG\%_{\text{mean}}^{\text{OPP}}$
3	$Win\%_{\text{mean}}$	20	$ScoringPossessions_{\text{mean}}$	37	$eFG\%_{\text{mean}}^{\text{OPP}}$
4	MOR	21	$Field\%_{\text{max}}$	38	$Play\%_{\text{max}}$
5	SAG	22	$eFG\%_{\text{mean}}$	39	$REB_{\text{mean}}^{\text{OPP}}$
6	$Floor\%_{\text{mean}}$	23	OE_{max}	40	$Field\%_{\text{min}}^{\text{OPP}}$
7	OE_{mean}	24	$ScoringPossessions_{\text{mean}}^{\text{OPP}}$	41	$FG_{\text{mean}}^{\text{OPP}}$
8	$Field\%_{\text{mean}}$	25	$PTS_{\text{mean}}^{\text{OPP}}$	42	$Field\%_{\text{min}}$
9	RPI	26	FG_{mean}	43	$PTS_{\text{min}}^{\text{OPP}}$
10	$Play\%_{\text{mean}}$	27	DE_{min}	44	$DREB_{\text{mean}}^{\text{OPP}}$
11	$AdjOE_{\text{mean}}$	28	$AST_{\text{mean}}^{\text{OPP}}$	45	$Play\%_{\text{min}}$
12	$Floor\%_{\text{mean}}^{\text{OPP}}$	29	OE_{min}	46	$ScoringPossessions_{\text{min}}^{\text{OPP}}$
13	DE_{mean}	30	$Ply\%_{\text{min}}^{\text{OPP}}$	47	$OREB\%_{\text{mean}}^{\text{OPP}}$
14	$Play\%_{\text{mean}}^{\text{OPP}}$	31	$Floor\%_{\text{min}}^{\text{OPP}}$	48	PTS_{min}
15	$AdjDE_{\text{mean}}$	32	AST_{mean}	49	$Field\%_{\text{max}}^{\text{OPP}}$
16	$Field\%_{\text{mean}}^{\text{OPP}}$	33	$AdjDE_{\text{min}}$	50	$Pay\%_{\text{max}}^{\text{OPP}}$
17	$Floor\%_{\text{max}}$	34	$Floor\%_{\text{min}}$	51	AST_{max}

Table 4: Summary of Feature Selection

$\mathbf{l} = \{l_1, l_2, ..., l_N\}$ be the set of labels corresponding to game outcomes

2. Transform each $\mathbf{g} \in \mathbf{G}$ to a corresponding $\mathbf{g_c}$ via the normalization procedure described in Section 4.1, generating a new set $\mathbf{G_c}$

3. Partition $\mathbf{G_c}$ into 5 subsets $\mathbf{G_c} = \{\mathbf{G_{c_1}}, \mathbf{G_{c_2}}, \mathbf{G_{c_3}}, \mathbf{G_{c_4}}, \mathbf{G_{c_5}}\}$ for use in 5-fold cross validation.

4. For each $\mathbf{G_{c_i}} \in \mathbf{G_c}$

   (a) Train the classifier with $\mathbf{G_c} - \mathbf{G_{c_i}}$ as the training set

   (b) Predict labels $\mathbf{m} = \{m_1, m_2, ..., m_N\}$ for the test data set $\mathbf{G_{c_i}}$

   (c) Calculate the classification success via

$$success = \frac{\#correctpredictions}{\#predictions} \tag{32}$$

$$success = \frac{|\{m_i | l_i = m_i\}|}{|\mathbf{m}|} \tag{33}$$

5. Take the weighted mean of each partition's success, according to how many elements were in it

5.2 Classifiers

The following classifiers were tested:

1. Baseline classifier – Always picks the favorite to win
2. Naive Bayes classifier
3. Decision Tree
4. Random Forest
5. Support Vector Machine – Radial Basis Function
6. Support Vector Machine – Linear
7. Artificial Neural Network
8. Linear Regression – Predict point differential of high-seed score - low-seed score. If positive return "No Upset" else return "Upset"

5.3 Full Season Results

The results in this section come from are the result of trying to classify every game[2] in the data set via the methodology discussed in Section 5.1, and a k-value of 50.

Classifier	Accuracy
Baseline	66.30%
Naive Bayes	65.92%
Decision Tree	70.93%
SVM – RBF	71.39%
SVM – Linear	68.92%
Neural Net	70.96%
Linear Regression	64.92%

Table 5: Full Season Classification Accuracies

5.4 Piecewise Classification

Intuitively, it makes sense that a feature's predictive power may not remain constant throughout a season. We would expect that early in the season, rankings would be more important (since there only a few games

[2]This excludes the first 2 games played by each game each season. This is because cumulative stats don't make sense this early in the season.

of cumulative stats), but by the end of the season the more advanced computed statistics would be powerful predictors. With this hypothesis, I tried training classifiers on small windows throughout the season, rather than once over the entire season. The results of this are shown in Figures 2 & 3 with the average accuracies shown in Table 6.

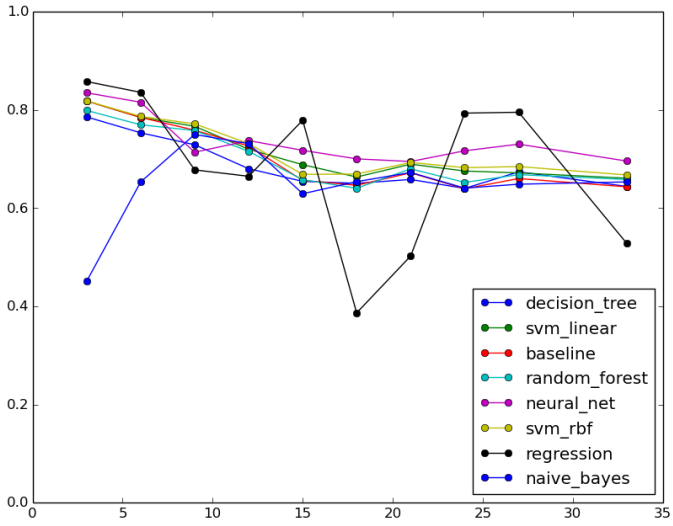

Figure 2: Piecewise Classifier

Classifier	Accuracy
Baseline	66.30%
Naive Bayes	64.83%
Decision Tree	68.71%
SVM – RBF	71.73%
SVM – Linear	71.40%
Neural Net	72.37%
Linear Regression	68.21%

Table 6: Piecewise Classification Accuracies

Although the performance of Naive Bayes and the Decision Tree both dropped due to weak early season results, the classifier accuracies went up in general. Most notably, the neural networks were able to achieve a season-wide accuracy of 72.37%, approaching the glass ceiling estimated in [14].

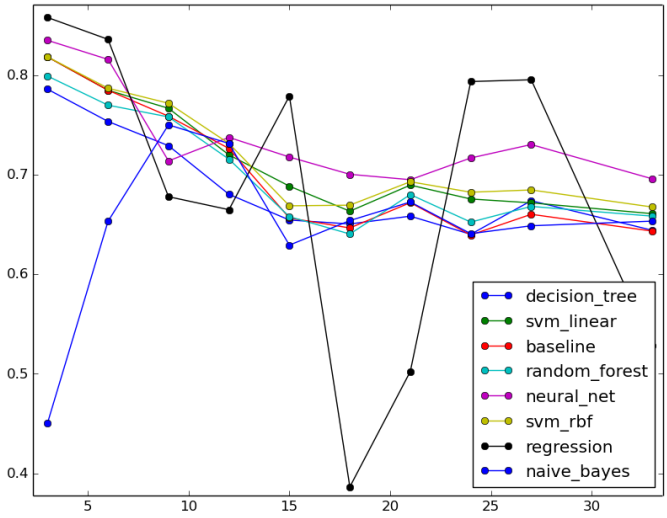

Figure 3: Piecewise Classifier Zoomed

5.5 Bracket Prediction

The results in Section 5.4 support my hypothesis that the importance of a given feature varies over the course of a season. Because my goal is to predict tournament success (and tournament's occur at the end of a season), the training data will only come from the end of seasons; specifically, I will exclude everything before a team's q^{th} game of the season, where q is an experimentally derived integer. To predict a given season y's bracket, all games played in $\{2010, 2011, 2012, 2013, 2014\}$ - y that were at least the q^{th} game of the season for both teams involved were used as a training set. The success of a bracket prediction was measured in two ways:

1. The accuracy formula presented in Section 5.1

2. ESPN's bracket scoring formula shown below, where $\mathbf{m}$ is the set of predicted labels, $\mathbf{l}$ is the set of correct labels, and $Round(m_i)$ is the round of the tournament that the predicted game was played in (first round = 1, second round = 2, ..., championship round = 6). This formula gives a fixed amount of 320 points to each round, with the points distributed evenly across the games played in that round.

$$u(\mathbf{m}, \mathbf{l}) = 10 * \sum_{m_i = l_i} 2^{Round(m_i)-1} \tag{34}$$

The bracket-prediction results are summarized in 7. The full season-by-season results are shown in Table 8. The propagated error of predicting incorrect winners in early rounds necessarily means that the accuracy percentages will drop from when we were only looking at individual games. This is reflected in the baseline falling from 66.30% to 40.00%.

Classifier	Accuracy	ESPN Score
Baseline	40.00%	368
Naive Bayes	32.70%	260
Decision Tree	39.05%	340
Random Forest	44.45%	410
SVM – RBF	45.40%	370
SVM – Linear	45.40%	456
Lin. Reg.	40.00%	368
Neural Net	48.57%	474

Table 7: Average Bracket Prediction Results

6 Conclusion

6.1 Analysis of Results

Looking at the results listed in Tables 7 & 8 we see results that are approximately in line with our regular season predictions. The Naive Bayes classifier still performs worse than the baseline, and the Decision Tree and Linear Regression perform about the same as the baseline. The Random Forest, SVM – RBF, SVM – Linear, and Artificial Neural Network classifiers all performed substantially better than the baseline both in terms of the percentage of games predicted correctly and the ESPN bracket score.

Unlike in the work done by [14], the more advanced classification algorithms vastly outperformed the more simplistic models. Perhaps this is because the strong linear dependence between many of the derived attributes; applying principal component analysis before classifying may be able to improve the performance of the simpler models.

6.2 Future Work

The hypothesis inspiring this research was that differences in team's play styles may be able to predict the outcome of individual matchups better than just ranking teams according to a statistical model. While the above results show that more advanced classification algorithms were able to substantially outperform the baseline prediction method of always choosing the high-seed, there is still room for improvement. It is

unclear whether or not accumulated end-of-game stats have a fine enough granularity to truly capture *how* a team plays. For example, the 2014 Creighton Bluejays ran almost all of their offense through small forward Doug McDermott; McDermott took 37% of the teams shots [16]. This is a defining characteristic of the Creighton offense, and yet it wouldn't be captured in the box-score data that was used for this project.

A future project then, is to look at a more granular data source. In addition to the box-scores used for this project, the NCAA Archive hosts play-by-play data describing each possession in a given game. By applying data mining techniques to this data set, we may be able to create a more informative description of *how* a given team plays.

Time	Pittsburgh	Score	North Carolina
19:45		0-0	PAIGE,MARCUS missed Three Point Jumper
19:42		0-0	MEEKS,KENNEDY Offensive Rebound
19:40		0-0	PAIGE,MARCUS missed Three Point Jumper
19:36	ZANNA,TALIB Defensive Rebound	0-0	
19:13	**ZANNA,TALIB made Dunk**	2-0	
19:13	PATTERSON,LAMAR Assist	2-0	
18:58		2-0	TOKOTO,JP missed Two Point Jumper
18:55	ROBINSON,JAMES Defensive Rebound	2-0	
18:43		2-0	TOKOTO,JP Commits Foul
18:32	**ROBINSON,JAMES made Two Point Jumper**	4-0	
18:32	ZANNA,TALIB Assist	4-0	
18:03		4-0	MCADOO,JAMES M missed Two Point Jumper
18:00		4-0	MEEKS,KENNEDY Offensive Rebound
17:59		4-0	MEEKS,KENNEDY missed Two Point Jumper
17:58		4-0	MCADOO,JAMES M Offensive Rebound
17:57		4-0	MCADOO,JAMES M missed Two Point Jumper
17:56	PATTERSON,LAMAR Defensive Rebound	4-0	
17:55	WRIGHT,CAMERON Commits Foul	4-0	
17:55	WRIGHT,CAMERON Turnover	4-0	

Figure 4: Example play-by-play data

Year	Classifier	Accuracy	ESPN Score
2010	Baseline	46.03%	540
	Naive Bayes	26.98%	180
	Decision Tree	39.38%	350
	Random Forest	55.56%	650
	SVM – RBF	49.20%	380
	SVM – Linear	53.97%	580
	Lin. Reg.	46.03%	540
	Neural Net	57.14%	670
2011	Baseline	30.16%	200
	Naive Bayes	31.75%	210
	Decision Tree	38.10%	310
	Random Forest	41.27%	310
	SVM – RBF	39.68%	280
	SVM – Linear	39.68%	290
	Lin. Reg.	30.16%	200
	Neural Net	47.62%	390
2012	Baseline	26.98%	180
	Naive Bayes	34.92%	250
	Decision Tree	38.10%	390
	Random Forest	41.27%	330
	SVM – RBF	38.10%	310
	SVM – Linear	33.33%	400
	Lin. Reg.	26.98%	180
	Neural Net	44.44%	440
2013	Baseline	47.62%	410
	Naive Bayes	34.92%	380
	Decision Tree	33.33%	260
	Random Forest	36.51%	350
	SVM – RBF	46.03%	410
	SVM – Linear	46.03%	450
	Lin. Reg.	47.62%	410
	Neural Net	42.86%	430
2014	Baseline	49.21%	510
	Naive Bayes	34.92%	280
	Decision Tree	46.03%	390
	Random Forest	47.62%	410
	SVM – RBF	53.97%	470
	SVM – Linear	52.97%	560
	Lin. Reg.	49.21%	510
	Neural Net	50.79%	440

Table 8: Full Bracket Prediction Results

7 Bibliography

References

[1] L. Petrecca, "March Madness in the office: Work comes in second," USA TODAY, 3/15/2012.[Online]. http://usatoday30.usatoday.com/money/workplace/story/2012-03-09/march-madness-ncaa-workplace/53538598/1.

[2] Denny Lee."Betting on March's Madness." New York Times: F.1. 2005. Print.

[3] D. Haar, "How Many ESPN.com Brackets Got All Four Right? Very Few," Hartford Courant, 4/1/2013.[Online]. http://courantblogs.com/dan-haar/how-many-espn-com-brackets-got-all-four-right-very-few/.

[4] Schumaker, R., Solieman, O., Chen, H. "Sports data mining. Springer", 2010.

[5] Schwertman, Neil C., Kathryn L. Schenk, and Brett C. Holbrook. "More Probability Models for the NCAA Regional Basketball Tournaments." The American Statistician 50.1 (1996): 34-8. Print.

[6] Boulier, Bryan L., and H. O. Stekler."Are Sports Seedings Good Predictors?: An Evaluation." International Journal of Forecasting 15.1 (1999): 83-91. Print.

[7] Jacobson, S. H., and King, D.M. (2009)."Seeding in the NCAA Mens Basketball Tournament: When is a Higher Seed Better", The Journal of Gambling Business and Economics, 3, 63-87.

[8] Toutkoushian, Emily. "Predicting March Madness: A Statistical Evaluation of the Men's NCAA Basketball Tournament." 2011. Print.

[9] N. Silver, "How We Made Our N.C.A.A. Picks," FiveThirtyEight – The New York Times, 3/14/2011.[Online]. http://fivethirtyeight.blogs.nytimes.com/2011/03/14/how-we-made-our-n-c-a-a-picks/.

[10] Miljkovic, D., Gajic, L., Kovacevic, A., Konjovic, Z.,"The use of data mining for basketball matches outcomes prediction", IEEE 8th International Symposium on intelligent and informatics, Subotica, Serbia, 2010, pp.309-312.

[11] Sorensen, S. P., "An Overview of Some Methods For Ranking Sports Teams", http://www.phys.utk.edu/sorensen/ranking/

[12] Stern H. (1991) "On the Probability of Winning a Football Game", The American Statistician, 45, 179-183.

[13] Jess Sonas, "EXTRA DATA - Sagarin Predictive Ratings", Kaggle, 2/24/14. [Online]. http://www.kaggle.com/c/march-machine-learning-mania/forums/t/6767/extra-data-sagarin-predictive-ratings

[14] Shi, Zifan et al. "Predicting NCAAB match outcomes using ML techniques - some results and lessons learned". 14 Oct, 2013. arxiv.org.

[15] Dean Oliver. *Basketball on Paper*. Brasseys, Inc., 2002.

[16] Ken Pomery, "the kenpom.com blog" [Online]. http://www.kenpom.com

[17] Kenneth Massey, "College Basketball Ranking Composite", Massey Ratings [Online]. http://www.masseyratings.com/cb/compare