Matthew Boutot and Tobias Lochner

Malinda Iluppangama

STA 4102

December 9,2024

Modelling March Madness from 2010-2014 using Logistic Regression

1. Abstract

March Madness is among one of the most difficult tournaments to predict the outcome. The annual basketball tournament held in March is a 68-team, single-game elimination tournament where qualification for March Madness is determined by regular season results. The factor contributing most to the unpredictability of the tournament is the single-game elimination format, which is subject to high amounts of in-game variance due to the complexity of basketball as a sport. However, because there are so many variables within a basketball game, March Madness can be a laboratory of sorts to test the efficacy of certain in-game descriptors, like 3-point percentage or rebounding, on predicting the outcome of a game.

As such, in this report, a logistic regression model was implemented to test if the model could accurately predict the outcome of a March Madness tournament. The model used March Madness results from 2010-2013 as training data sets and then used the 2014 March Madness tournament as a testing data set. The model itself originally incorporated 20 in-game statistics as independent variables to predict the probability of the game outcome, (either win or loss), given the opponent.

After applying backward AIC to the model, the independent variables were narrowed down to 11 from 20. In the end, the model predicted the results of the training datasets with about 80 percent accuracy and predicted the accuracy of the testing dataset with 73 percent accuracy. The model itself did not predict upsets (defined as a difference in seeding above 5) well but was able to quite accurately predict the results of certain matchups.

Despite the model being relatively accurate considering the format of the tournament, the results could be skewed by the model violating normality and only having 4 tournaments as training data.

2. Introduction

Prediction of game outcomes has always been a challenging prospect for most sports, but it has been particularly challenging with basketball especially. In the context of the highest caliber of professional basketball in the NBA, some predictors can allow for a better understanding of which teams are likely to win in the playoffs like net rating (the difference between a team's offensive and defensive rating) and fundamental factors like player matchups or defensive schemes. However, there is a much larger sample size in the NBA with teams playing best of 7 series against each other as opposed to the single-game elimination format found in March Madness.

Having a single game be the sole decider of a team’s elimination makes the result of the tournament incredibly difficult to predict. This is because basketball as a sport is already subject to large swaths of variance, even over multiple games. After the emergence of the three-point revolution,

where teams have started taking a larger number of threes, the variance of in-game results has increased even further. This is because as the distance of a shot increases, there is more variability. This combined with how wildly the outcome of a basketball game can vary in a single-game setting, contributes to the erratic results that March Madness can produce. It is not unheard of for a first seed with the best record to be upset by the 16th seed in the tournament.

If the model accurately can predict the outcomes of March Madness tournaments, then there can be a better understanding of how to model basketball. Basketball is subject to variance as previously mentioned, and because of that variance, modeling what leads to the highest likelihood of winning a game is incredibly challenging. Taking in-game information like rebounds, free throws, and three-pointers gives a semblance of what could occur, but especially during a single game setting there are a variety of confounding factors like whether a game is played at home, or defensive scheme that can wildly swing the outcome of a game. If the model can identify significant parameters, then with further research, there is a possibility that a more advanced model can be created to better account for the variance that makes basketball so unpredictable.

3. Data Description

The data we used for this project was a collection of various important statistics regarding a given team's performance throughout the regular season. This data was taken from various websites including Sports Reference, Ken Pom, and ESPN. From these sites, we extracted 20 of what we deemed the most relevant statistics to a team's success in the postseason. These were based solely on opinions using the knowledge of basketball that we have.

An example of the raw data is shown below, we transformed this data which will be explained later.

	2014 Teams	Seed	Age Diff	BPI Diff	ORtg Diff	DRtg Diff	AdjT Diff	TFGA Diff (Team
Virginia	Virginia	*NA*	*NA*	5	114.8	88.3	59.5	
Coastal Carolina	Coastal Carolina	*NA*	*NA*	197	98.2	103.2	65.6	
Memphis	Memphis	*NA*	*NA*	35	112.4	97.1	67.5	
George Washington	George Washington	*NA*	*NA*	44	110.8	97.0	65.3	
Cincinnati	Cincinnati	*NA*	*NA*	26	109.7	90.4	61.1	
Harvard	Harvard	*NA*	*NA*	28	112.9	95.3	64.1	
Michigan State	Michigan State	*NA*	*NA*	11	118.6	94.7	64.4	
Delaware	Delaware	*NA*	*NA*	103	112.0	106.2	70.4	
UNC	UNC	*NA*	*NA*	27	112.6	94.2	69.2	
Providence	Providence	*NA*	*NA*	48	114.6	100.8	62.9	
Iowa State	Iowa State	*NA*	*NA*	23	118.9	97.8	70.2	
North Carolina Central	North Carolina Central	*NA*	*NA*	68	108.6	100.7	62.3	
UConn	UConn	*NA*	*NA*	25	113.6	91.5	63.2	
St. Joseph's	St. Joseph's	*NA*	*NA*	40	112.3	99.4	64.3	
Villanova	Villanova	*NA*	*NA*	6	116.3	92.6	66.0	
Milwaukee	Milwaukee	*NA*	*NA*	158	104.7	104.6	65.8	
Wichita State	Wichita State	*NA*	*NA*	14	117.8	92.4	63.5	
Cal Poly	Cal Poly	*NA*	*NA*	208	103.6	104.4	58.8	
Kentucky	Kentucky	*NA*	*NA*	12	118.3	95.8	64.0	
Kansas State	Kansas State	*NA*	*NA*	52	107.8	93.7	63.3	
Saint Louis	Saint Louis	*NA*	*NA*	37	105.7	90.0	65.6	
NC State	NC State	*NA*	*NA*	77	114.0	102.4	64.7	
Louisville	Louisville	*NA*	*NA*	1	118.9	88.5	67.2	
Manhattan	Manhattan	*NA*	*NA*	69	107.2	97.0	67.7	
UMass	UMass	*NA*	*NA*	47	110.0	96.5	70.3	
Tennessee	Tennessee	*NA*	*NA*	15	117.9	94.2	61.6	
Duke	Duke	*NA*	*NA*	3	124.7	100.4	64.6	

Figure 1: Snapshot of the data showing what the raw data looks like

Shown below is a key showing what all of the abbreviations mean for each statistic, with a brief explanation of what it means.

Statistic Abbreviation	Meaning	Description
BPI	Basketball Power Index	A rating used by ESPN to rank how good each team is
ORtg	Offensive Rating	An adjusted rating from Ken Pom, that describes how good an offense is
DRtg	Defensive Rating	An adjusted rating from Ken Pom, that described how good a defense is
AdjT	Adjusted Tempo	An adjusted rating from Ken Pom, that describes how fast a team plays, i.e. how many possessions they have in a game
TFGA	Team Field Goals Attempted	The average amount of field goals (Shots) a team attempts during a game
TFG%	Team Field Goal Percentage	The percentage of field goals that the team makes
OFG%	Opponent Field Goal Percentage	The percentage of field goals that the team allows their opponent to make
T3PA	Team 3-Point Attempts	The average number of 3-pointers a team attempts
T3P%	Team 3-Point Percentage	The percentage of 3-pointers a team makes
O3P%	Opponent 3-Point Percentage	The percentage of 3-pointers a team allows their opponent to make
TFTA	Team Free Throw Attempts	The average number of free

		throws a team attempts per game
TFT%	Team Free Throw Percentage	The percentage of free throws a team makes
OFTA	Opponent Free Throw Attempts	The average number of free throws a team allows their opponent to attempt
TORB	Team Offensive Rebounds	The average number of rebounds a team gets per game when they have the ball
TDRB	Team Defensive Rebounds	The average number of rebounds a team gets per game when they don't have the ball
TTOV	Team Turnovers	The average amount of times a team turns the ball over
OTOV	Opponent Turnovers	The average amount of times a team causes their opponent to turn the ball over
TPts	Team Points	The average number of points a team scores per game
OPts	Opponent Points	The average number of points a team allows per game
Seed	Tournament Ranking	A number to represent how the team ranks in the tournament and is used to determine who they play

Figure 2: Key describing the statistics we used

We handpicked data from many different websites to create our dataset, so there was no need to clean the data as we manually did this when creating the above spreadsheet. Along with this, our data did not have any limitations because we purposely selected the exact data to match our needs, and only used relevant statistics. So there were no missing statistics or other similar problems. Since we later transformed our data there wasn't a need to check for outliers or any other distribution problems.

As we've stated above, the goal of this model was to create a game-by-game predictor that would give a percentage probability p, that a certain team had to win vs. another team. With this we wanted the opponent's probability to be 1-p. To do this we had to transform the data to create the exact opposite statistics, so that when the model was run if it gave Team A p=0.67 then Team B (their opponent) would have p=0.33. To do this we transformed all of the statistics to differences based on the opponent. So for every instance of a team playing a game, the statistics entered into the model would be Team A Stat - Team B Stat, and then Team B - Team A for the opponent. This way one team would have a -15 advantage in BPI, and their opponent would have +15. Therefore, our model when running would spit out two probabilities that add to 1.

Before we started our regression model we wanted to see if there were any interesting correlations to note that could help us develop our model. A correlation plot we developed is shown below. As shown there is very little correlation between any of the variables especially for the W/L (Win or Lose) variable, which is our response. This lack of correlation influenced the creation of our model which is discussed in the next section.

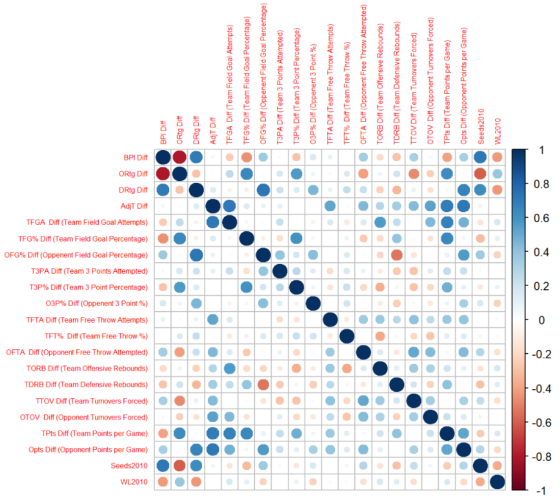


Figure 3: Correlation Plot for all of our variables

4. Methodology

Logistic regression was chosen for this model as the response variable we hoped to predict was win or loss, which we coded as a binary variable 1 or 0 (1 being a win and 0 being a loss). To set up this data we used the difference in statistics between the two schools. We coded this in R by manually inputting the raw statistics from each team into a data frame in R. Once this data frame was in there, we created a function that would take the two team names as inputs and subtract the team from their opponent and store it as a vector. For every instance of a team playing a game, we created a vector. For example, if Team A played Team B in the Round of 64 then we created a vector for Team A and Team B. If Team A won and moved on to play Team C in the Round of 32 another vector was created

for Team A but this time with the opponent Team C, and a vector was created for Team C with Team A as their opponent.

We created these vectors for every instance of a team playing across the March Madness tournaments from 2010-2013, we then used rbind() to create a dataframe of our training data. We then repeated this process for the tournament from 2014 and used that as our test data. Following the suggested 80-20 rule (80% of the data for training and 20% for testing). Once we had the two dataframes created we were able to create our model.

As explained in Section 3, Data Description, none of the variables had a strong correlation to W/L, so we decided to create our model using backward Step AIC. This means that we started with all of our independent variables and removed them stepwise until we arrived at the most significant model. We did this in R using the stepAIC() function. This took our original model with all 20 variables in it, down to a model with just 11 independent variables as shown below.

```
MMModel3 <- glm(formula = WL2010 ~ `BPI Diff` + `ORtg Diff` +
                        `DRtg Diff` +
    `AdjT Diff` + `T3P% Diff (Team 3 Point Percentage)` + `O3P%
                  Diff (Oppenent 3 Point %)` +
        `TFT%  Diff (Team Free Throw %)` + `TORB Diff (Team
                     Offensive Rebounds)` +
        `TDRB Diff (Team Defensive Rebounds)` + `OTOV  Diff
                (Opponent Turnovers Forced)` +
    `TPts Diff (Team Points per Game)`, family = binomial, data
                        = MMTrainData)
```

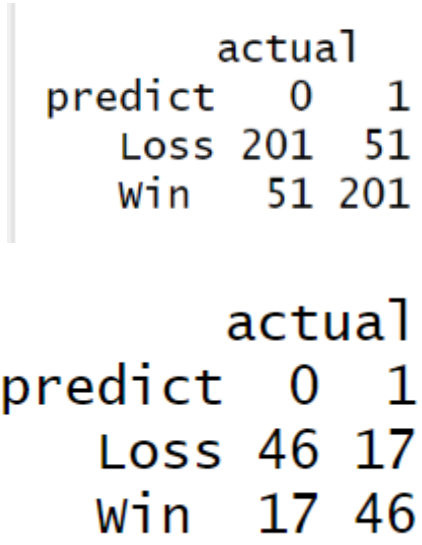
Figure 4: Code for the Final Model we used

This model was found to have the lowest AIC, however, one of the variables was not significant at an alpha of 0.05. Since the p-value is only slightly higher than 0.05, we decided to keep it in our model, since removing it would increase the AIC. Once the model was complete, we created the predictions for both the training and test dataframes and created a confusion matrix to check the accuracy of our model.

5. Results

The results show that overall, our model performed fairly well considering we were asking it to predict the impossible. As shown in the confusion matrices below, our model was between 70% to 80% accurate between the predictions for the training and test data. Given how hard it is to predict March Madness I would say that our model was a moderate success, though it may lack practical applications as we'll explain later. Another thing to note is that our calculated AUC or Area Under the

Curves was about 0.85, which again indicates that our model does a fairly decent job at predicting the outcomes on a game-to-game basis.



	actual	
predict	0	1
Loss	201	51
Win	51	201

	actual	
predict	0	1
Loss	46	17
Win	17	46

Figure 5: Confusion Matrices for our training(above) and test(below) data

Despite the success described above our model can not be considered reliable in a statistical sense. For a logistic regression model to be valid it must pass certain assumptions that our model does not. The logistic regression model must have independence of observations, linearity, absence of multicollinearity, and no outliers. But based on the very nature of our study these assumptions could never be met. For example, while every game in March Madness is independent, since we consider every instance of a team playing there will be by nature a lack of independence between the observations. If Team A won, then by default their opponent must have lost, so therefore our data is more akin to paired data than independent data. Also because of the paired nature of our data, our data was unlikely to be linear. Multicollinearity has to do with interactions between independent variables in a model and again by the nature of the sport of basketball, almost every statistic is going to have some interaction with any other statistic. So based on this, while our model may have some application in the practical sense, it lacks credibility in the statistical sense.

6. Discussion

Based on our results we can say that we did a fairly good job of accomplishing our goal of creating a model to predict the outcome of March Madness games. Overall creating a model that can accurately predict March Madness games 70% to 80% of the time is incredibly good, and thus has some practical value. However, it is important to note some of the limitations when it comes to our model both practically and statistically.

First, our model was built and tested on March Madness tournaments that had already occurred and because of this, we knew which team was going to win any given game and move on to the next round. Typically, when one thinks about filling out a March Madness Bracket, you do it

before any games are played. This affects the results because if you were to pick a given team to win the National Championship, but they lost in the first round your bracket would be completely busted. With that being said our model just looked at predicting the result of a game knowing the two teams that are in that game. But if from the beginning we had gone round by round and predicted the winners of each game and then used the predicted winners the next round, and not reset it to the real winners, our model would have been significantly less accurate. However, given this our model 4 out of the 5 times correctly predicted the national champion, meaning it predicted the real national champion to win every game from the outset, however, this may have changed if they faced a stronger opponent that the model likes, but didn't win in the real tournament.

In the most practical sense, our model does a fairly decent job of predicting a single game, given it knows which two teams are playing. So with that, this model may have some practical applications to sports betting, in terms of picking a winner of a game. One drawback of this is that the model often picks the favorite and struggles to predict big upsets, which would have big money winnings, so you may not make a huge profit. But I imagine if you were to bet on all the games and were to get between 70% to 80% of them right using the model you would probably make some money.

In a statistical sense, as highlighted in Section 5, Results, our model assumptions are not met which means this model cannot be seen as reliable in a statistical sense. Since, this was applied to a real-world problem, statistical reliability may be seen as not that important, as the model was shown to be accurate, which is what matters. However, what this statistical reliability could point to is that, if this model were to be expanded to include more tournaments, or we were to test the data on different tournaments our results may be less accurate. We used tournaments from 2010-2014, which means this may be less accurate for newer tournaments.

Overall, this model showed some definite promise and could have some practical use for us as an individual game predictor. However, this model lacked the volume of data or statistical backing to concretely say that this model is a reliable predictor to put any real value into. And given the structure of how March Madness predictions work this model may be even less accurate. This model can be used as a fun tool to try and predict games and maybe put some money into it to see if you win overall, but it should be used for nothing more. Next year's March Madness I will try and predict a bracket using the model, but I doubt it will do any better than the best bracket someone can come up with on their own, but we will see.

7. Conclusion

Overall, the model we created was a good first step in trying to find a way to predict the outcomes of individual games, given the immense amount of variance found in March Madness. Despite there being several limitations to the model including: already having the results and the tournaments and the violations of the logistic regression assumptions, the model provided relatively

accurate results. The question then becomes would the model's accuracy be sustained if it was tweaked to follow the assumptions of logistic regression and whether the model would be accurate without knowing the two given teams in the tournament.

An expansion of the model would have to change such a wide variety of factors that I would be very surprised if the 70-80 percent accuracy from our model held. Considering that the data we used was closer to paired than independent and there is a violation for nearly all assumptions of logistic regression, there would likely have to be a different framework for the model altogether.

However, even with the given limitations of our model, it still serves as a relatively accurate predictor of a single game outcome, which has some value in showing that modeling the results of March Madness can be done, even with the aforementioned limitations. If the framework of the model was changed to better fit with the dataset, then there could be a more applicable way to model the outcome of March Madness, which could be used as a test case to better understand predicting outcomes with small samples and large variances.

8. References

2010 men's NCAA tournament summary: College basketball at sports. Reference.com. (2024a). https://www.sports-reference.com/cbb/postseason/men/2010-ncaa.html

2011 Men's NCAA Tournament Summary: College Basketball at Sports. Reference.com. (2024b). https://www.sports-reference.com/cbb/postseason/men/2011-ncaa.html

2012 Men's NCAA Tournament Summary: College Basketball at Sports. Reference.com. (2024c). https://www.sports-reference.com/cbb/postseason/men/2012-ncaa.html

2013 Men's NCAA tournament summary: College basketball at sports. Reference.com. (2024d). https://www.sports-reference.com/cbb/postseason/men/2013-ncaa.html

2014 Men's NCAA tournament summary: College basketball at sports. Reference.com. (2024e). https://www.sports-reference.com/cbb/postseason/men/2014-ncaa.html

ESPN Internet Ventures. (2024). *2024-25 Men's College Basketball Power Index*. ESPN. https://www.espn.com/mens-college-basketball/bpi

2014 Pomeroy College Basketball Ratings. (2024). https://kenpom.com/index.php?y=2014