**How to not Blow a 3-1 Lead in this Year’s Bracket Predictions**

**Damir M. ZhaksilikovJason Frazier**

Department of Computer Science Department of Computer Science

University of Washington University of Washington

Seattle, WA 98105 Seattle, WA 98105

*damir@cs.washington.edu* *jasonf56@cs.washington.edu*

**Abstract (Damir)**

As of recent, the NCAA March Madness tournament has piqued the interests of the computing and machine learning community. The tournament’s seemingly unpredictable nature questions whether tournament brackets and the associated upsets can be accurately predicted. In this document, we will explore the existence of upset indicators, success of various models in predicting brackets, and impact of mixing raw team statistics with various measures of team rankings. First, we describe the results of regular season data manipulation and model performance of Logistic Regression, Random Forest, Multilayer Perceptron, and K-Nearest Neighbors models, of which Random Forest consistently outperformed others. Next, we outline and model improvement to cater towards the nature of March Madness Tournament play. Our model’s success in prediction is then quantified – we estimate that it has an accuracy of about 70 and 60 percent accuracy in predicting regular and tournament games respectively. Our techniques in tuning our model provides a 25% increase in correctness from predictions made by non-expert humans [1].

**1 Project Outline (Damir)**

**1.1 NCAA March Madness Tournament**

The NCAA Division I Men’s Basketball (nicknamed March Madness) is a single elimination basketball tournament played by 68 Division I collegiate basketball teams. The tournament consists of 32 teams who win Division I conferences and 36 teams awarded positions via committee choice. It has become tradition for fans to predict the tournament bracket given initial team slots. Due to the nature of single elimination, game outcomes are highly unpredictable. Additionally, there are 9.2 quintillion potential bracket combinations. Thus, the prediction of the correct bracket combination is highly improbable.

**1.2 Workflow**

Our first goal was to determine the best model for predicting NCAA basketball game results. Although there are differences between tournament and regular season play, we operated under the assumption that model success would reflect across all forms of NCAA basketball games. Using regular season data from 2003 – 2016 we tested the success of Logistic Regression, Random Forest, Multilayer Perceptron, and K Nearest Neighbors models. Once Random Forest was determined as the most successful model, we adjusted our tournament data set to reflect teams’ yearly performance, tournament seeding, and Las Vegas point spreads for each game and weighted upsets. Our models were then run on tournament data on a March Madness tournament simulator which provided the most probable bracket based on our models’ predictions.

**1.3 Key focuses**

The following questions were central to formulating our workflow and model/data manipulation techniques:

* Do predictable indicators of an upset exist?
* Is there a correlation between tournament seeding and the Las Vegas Point Spread?
* Is it possible to combine team statistics with seeding and the point spread to predict more accurately than the spread/seeding on their own?
* How does performance during the regular season transfer to tournament performance?

**1.4 Hypothesis**

As discussed by Lopez and Matthews, the “rules of efficient gambling markets” state that it is impossible to beat the spread of the Las Vegas sports books in the long run [2]. We believe that it is possible to train our model to predict upsets more accurately than the spread through weighting upsets more heavily. The negative aspect of this, though, is that our ability to predict “straightforward” results will suffer. Additionally, we expect that our model prediction performance will drop from regular season to tournament games as tournament games are considered to be more unpredictable.

**2 Data Overview (Damir)**

**1.1 Original dataset**

Kaggle’s March Machine Learning Mania 2017 challenge provided data for all NCAA basketball regular season and tournament games from 2003 – 2016. Each data point contained a game’s statistics pertaining to game information such as the year, location, and participating team IDs. Additionally, each point provided game statistics for both participating teams (e.g. points scored, field goals made, steals, etc.). Teams for each game were marked as team1 and team2 where team1 was the winner of each matchup.

The Kaggle challenge also provided a table mapping team IDs to the school name. Other data was provided, but not used. We used Python’s SQLite library to format and modify data as necessary.

**1.2 Regular Season Games**

To experiment with various models on the success of predicting regular season games we needed to reformat point features to reflect the strength of participating teams rather than their actual performance during that game. Obviously, in its original format the model would simply choose the winner of the game as whoever scored more points, this was not the function of our intended model.

To reformat data, we replaced each team’s statistic for a game with the average of that statistic for an entire year. This way, the averages reflected the team’s average strength in that statistic and the model could use that as a feature impacting game evaluation. Additionally, we added an indicator value to classify the games, where 1 meant team1 won and 0 meant otherwise. Since team1 won each game in the original dataset, each classifier value was 1 for the original data. In this given format, the model would simply predict that team1 for any game. Thus, for each data point, we created a copy with a team1 and team2 flipped and an indicator value of 0. We had hoped, as well, that this would completely remove any bias towards weights in team1 over team2 and that those weights would ultimately have an inverse relationship.

**1.3 March Madness Games**

The March Madness games were reformatted contained the same averages found in the regular season for each team during a given year. This made sense to us, as a team’s performance during the regular season would be an accurate reflection of the same team playing in the tournament. In addition, we added the tournament seed of each team and the Las Vegas point spread for the game.

**1.3.1 Seeding**

Tournament seeds are preliminary rankings, released by the NCAA committee for March Madness purposes. For each region, there are seeds given 1-16, but the NCAA also releases an all-inclusive 1-68 ranking of all participants. For game in the March Madness dataset we added participating teams’ tournament seed. The intuition is that the highest seeded teams consistently (for the most part) make it to the final rounds of the tournament. Additionally, although raw statistics provide insight on individual strength of each team, the seed provides an all-encompassing team ranking which we believed would holistically strengthen our model. These values were entered manually and due to time constraints, we entered values for 2012-2016 all other values were 0.

**1.3.2 Las Vegas Point Spread**

The Las Vegas Point Spread “provides the predicted difference in total points scored between the visiting and the home team; a spread of – 5.5, for example, implies that the home team is expected to win by 5.5 points” [1]. In our case, team1 was considered the home team as all games were in neutral locations. Our intuition for adding these values comes from Lopez and Matthews who argue that it is impossible to beat the Las Vegas Point Spread in the long run due to normalization associated with the large number of bets made [1]. Lopez and Matthews cite that “should act as the standards on which to judge any pre-game predictions” [1].

We ran into numerous issues with Odds Shark. Firstly, the values were entered manually and we simply did not have the resources to gather data for every game (naming conventions between our name-to-ID and Odds Shark were different therefore data scraping was ruled out). Additionally, Vegas odds did not exist for all games between 2003-2016 and no odds exist for this year as the bracket has not been finalized. Therefore, we a built a L2 Normalized Linear Regression model to find relationship between tournament seeding and Las Vegas odds. For games with seeds, but no Las Vegas Odds we used the relationship to guess the Las Vegas odds. Our intuition was that both values are based on overall perceived team strength and thus must be at least loosely correlated. Our model found the following weights:

0.29599983 for the higher team seed

-0.37982249 for the lower team seed

These values are consistent and the Las Vegas odds are negative towards the better (lower seeded) team. But, since the higher seed values are larger than the lower seed values, the weight must be lower for higher seeds.

**1.4 Feature selection results**

The tournament data originally used contained 35 features including both raw statistics as well as seeding and Las Vegas odds. We ran a L1 Normalized Linear Regression feature selection to determine the most impactful features. After running feature selection, our model contained between 17 and 20 features depending on the run. The results are as follows:

Figure 1.2: Scores for Features after Feature Selection on Regular Season Data

Figure 1.1: Scores for Features after Feature Selection on Regular Season Data

Figure 1.2: Scores for Features after Feature Selection on Tournament Data

Regular Season Data

* Location was the most impactful feature for determining the winner of regular season games. All March Madness games are played from a neutral location and could account for some of the unpredictability in games results.
* As expected, average score and score difference were the most impactful raw team statistics.

Tournament Data

* Average score difference was the most important feature for prediction.
* Vegas odds were the second most important feature and this supports our claim that Vegas odds should be a standard for pre-game indication.
* Seeding plays very little role in the result of the tournament games. This may be because we did not have seeds for all games, but if this was the case our Vegas odds would have a similar score (as we have the same number of Vegas odds and seeds statistics). Thus, this implied that our assumption that Las Vegas odds and tournament seeds are correlated is incorrect.

**3 Model Selection (Jason)**

**4 Tournament Adjustments (Damir)**

**4.1 Feature Additions**

**4.2 Upset Weight**

**4.3 Simulated Tournament Play**

**5 Results (Jason)**

**6 Conclusions (Jason)**

**References**

[1] <http://stephenpettigrew.kinja.com/11-million-brackets-vs-espn-cbs-and-fox-experts-who-1561354312/1562354592>

[2] Lopez, M. J., & Matthews, G. J. (2015). Building an NCAA men’s basketball predictive model and quantifying 131 its success. Journal of Quantitative Analysis in Sports. Retrieved February 20, 2017

[3] March Machine Learning Mania 2017 | Kaggle. (n.d.). Retrieved February 20, 2017, from https://www.kaggle.com/c/march-machine-learning-mania-2017