**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 Regular Season Games**

**1.2 March Madness Games**

**1.2.1 Seeding**

**1.2.2 Las Vegas Point Spread**

**1.3 Feature selection results**

**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