UNITED STATES MILITARY ACADEMY

WEST POINT, NEW YORK

HONORS THESIS

MODELING MADNESS

by

Cadet Gino Nicosia

13 May 2019

Thesis Advisor:	Dr. William Pulleyblank
Thesis Advisor:	Major James Pleuss
Thesis Advisor:	Major Dusty Turner
Second Reader:	Major Daniel Baller

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE		Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE 13 May 2019	3. REPORT TYPE AND DATES COVERED

4. TITLE AND SUBTITLE Modeling Madness 6. AUTHOR(S) Cadet Gino Nicosia	5. FUNDING NUMBERS
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) United States Military Academy West Point, NY 10996	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) NA	10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited	12b. DISTRIBUTION CODE

13. ABSTRACT (maximum 200 words)

The NCAA Mens Division I Basketball Championship Tournament is an annual collegiate competition of some of the best 68 teams across the nation. The nature of the competition is deemed unpredictable by the masses and, therefore every year, millions of people attempt to create the perfect bracket. For every correct game prediction, the bracket earns a certain number of points which is determined by different submission platforms. This project used a series of modeling techniques and statistics to determine the winner of each game. By using game statistics from 2003 until 2018, a linear model was trained to predict the number of points that a given team scores against another opponent. These predictions were based upon selective offensive and defensive variables from regular season play. From the regressed model, a game probability matrix was created containing a certain probability for each team to defeat every other team in the tournament against every other team in the nation. These single game probabilities are inputted into a singular Monte Carlo simulation to create 23 brackets for submission to ESPN's Bracket Challenge. They also are the basis for a more probabilistic tournament competition through Kaggle.com. This paper discusses the process of attempting to correctly predict the 2019 NCAA Mens Basketball Championship Tournament. Let the Madness begin!

Keywords: Linear Model, Probability Matrix, Monte Carlo Simulation, March Madness, Basketball

14. SUBJECT TERMS	15. NUMBER OF PAGES
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT OF
Unclassified	Unclassified	Unclassified	UL

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

MODELING MADNESS

Cadet Gino Nicosia
Cadet, Armor
B.S., United States Military Academy, 2019

Submitted in partial fulfillment of the
requirements for the degree of
BACHELOR OF SCIENCE
in **MATHEMATICAL SCIENCES**
with Honors
from the
UNITED STATES MILITARY ACADEMY
13 May 2019

Author: Cadet Gino Nicosia

Advisory Team: Dr. William Pulleyblank
Thesis Advisor

Major James Pleuss
Thesis Advisor

Major Dusty Turner
Thesis Advisor

Major Daniel Baller
Second Reader

Colonel Tina Hartley
Chair, Department of Mathematical Sciences

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The NCAA Mens Division I Basketball Championship Tournament is an annual collegiate competition of some of the best 68 teams across the nation. The nature of the competition is deemed unpredictable by the masses and, therefore every year, millions of people attempt to create the perfect bracket. For every correct game prediction, the bracket earns a certain number of points which is determined by different submission platforms. This project used a series of modeling techniques and statistics to determine the winner of each game. By using game statistics from 2003 until 2018, a linear model was trained to predict the number of points that a given team scores against another opponent. These predictions were based upon selective offensive and defensive variables from regular season play. From the regressed model, a game probability matrix was created containing a certain probability for each team to defeat every other team in the tournament against every other team in the nation. These single game probabilities are inputted into a singular Monte Carlo simulation to create 23 brackets for submission to ESPN's Bracket Challenge. They also are the basis for a more probabilistic tournament competition through Kaggle.com. This paper discusses the process of attempting to correctly predict the 2019 NCAA Mens Basketball Championship Tournament. Let the Madness begin!

Keywords: Linear Model, Probability Matrix, Monte Carlo Simulation, March Madness, Basketball

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

I would like to acknowledge and thank my advisor team: Dr. William Pulleyblank, Major James Pleuss and Major Dusty Turner for their hard work and dedication in assisting me to complete this project. Without these individuals this project never would have been possible.

Additionally, I would like to thank the Second Reader of this paper, Major Daniel Baller. Thank you for taking time out of your busy schedule to make my project better.

Thank you Retired Lieutenant Colonel Shaw Yoshitani, a fellow at the Mounger Writing Center. He worked with me to help polish this paper into something presentable.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. OVERVIEW

Every March, as the college basketball regular season comes to a close, the annual Division I Men's Championship Basketball Tournament begins. The tournament is sanctioned by the National Collegiate Athletic Association (NCAA) and the Selection Committee chooses 68 teams to compete for the National Championship. The tournament, known as "March Madness," is infamous for the unpredictable nature of each of the games. Such a tournament creates an excitement that grows beyond the basketball faithful and the United States. Millions of individuals from across the world turn to their televisions and computer screens to watch Madness unfold.

March Madness is a single elimination basketball tournament that starts with 68 teams and finishes with one champion. A single elimination tournament removes a team from the competition after only one loss and advances the team winning each game to the subsequent round. March Madness consists of six rounds of games. The winner of each game advances to the next round.

The first games played are four play-in games. The last eight teams to qualify for the tournament play an extra, pre-tournament play-in game to determine which four teams should continue to the 64-team competition which forms a balanced tree.

Millions of fans make their predictions for every game in the tournament and enter them into a variety of competitions in hope of being the most accurate. Due to the large number of different platforms of competition, this naturally creates several different scoring methods to determine the winner. The most common submission approach is creating a bracket, as seen in Figure 1.

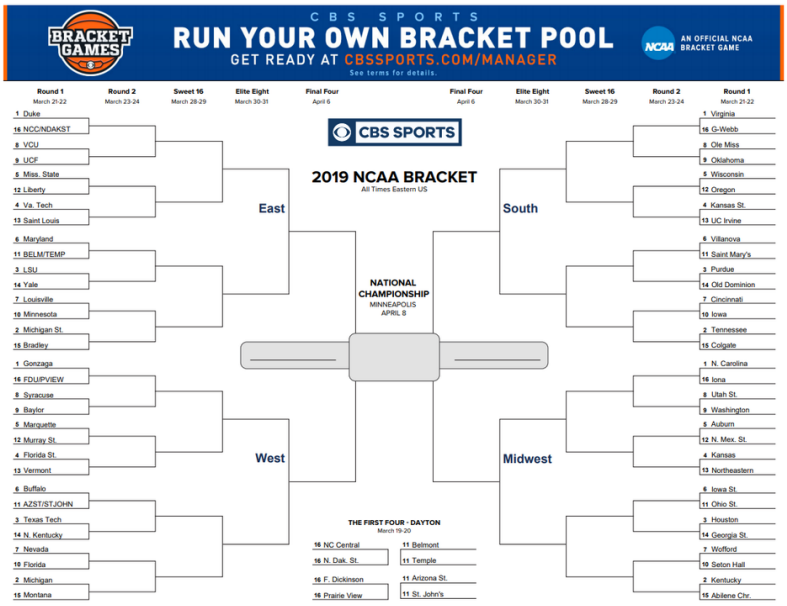

Figure 1. 2019 Blank Bracket for March Madness

Some scoring platforms do not use the common bracket submission approach, but require that a contestant estimate the probability of a team winning against each other team. Each scoring platform has a limited number of entries that an individual can enter to prevent a person from submitting every possible outcome. The probability of creating a perfect bracket - one which correctly predicts the winner of every game (63 games) - is minuscule, which allows billionaire Warren Buffet to temptingly offer a million dollar prize a year for life to anyone who can create a perfect bracket without facing any significant financial risk. [1]

The goal of our project was to correctly predict the winner of each game of the 2019 NCAA tournament via the ESPN and Kaggle scoring platforms. However, our project would be viewed as successful if it scored highly in the annual March Madness competition.

B. SCORING ALGORITHMS

There are different criteria for scoring tournament submissions in accordance with a specific competition. Different competitions prioritize different aspects of the tournament. For example, some scoring methods give a higher percentage of points to a bracket that predicts the tournament champion correctly. Others methods award higher performance based upon confidence of estimated win probabilities. Evaluation of the results of the predictive win/loss model for various scoring systems helped determine the success of the model. Two submission platforms are utilized to score the output of this model: ESPN and Kaggle.[2]

C. DATA COLLECTION

The data collected from various sources was analyzed to create a model that predicts the number of points a team will score against an opponent. The data determined which variables to include in this final model.

A main data source for this model was Kaggle.com which hosts hundreds of competitions. The website has accessible, relevant of data sets for making predictions. These data sets include information from each game: *Play-by-play* commentary, general statistics, the names of the team and mascots, different ranking system of teams at different points in the season, and much more.

A kaggle.com file with a highly specific set of variables for every game from the 2003 season until 2018 was the basis of the model building process. Each row details a singular game containing offensive and defense statistics that can predict how many points each team scores. Some examples of offensive statistics include: points scored, offensive rebounds, assists, and other performance measures. Some defensive statistics are also included in the data: defensive rebounds, points allowed, steals, blocks, and other defensive measures. The developed model uses some offensive and defensive variables to predict the number of points a team will score against a certain opponent in any given game.[3][4][5]

D. METHODOLOGY

The selected data sets provided certain variables, discussed more in depth in later sections, that developed the model. Data exploration commenced with a variable selection process to determine which variables that significantly influence the predictor variable of

points scored in a game. Bottom-up variable selection is the manner in which the model becomes more complex by gradually adding more variables through an iterative process.

The overall process used to make predictions for the winner of every game was intensive. Using a simple linear regression model, a distribution for the number of points both teams should score in a given game was determined. This model provides an easy interpretation of the coefficients' effect on the predicted amount of points scored for each team. Using the difference in the normal distributions of the two teams' predicted score, an estimated probability that Team A could win against Team B and vice versa is created for every combination of teams in NCAA Division I. Because the distribution of scores for both teams is normally distributed, the difference in scores can also be represented with a normal distribution. Such probabilities filled a 350 by 350 matrix that contains all estimated probabilities for a specific Division I team to beat another team based upon the proposed model.

Due to the fact that the March Madness Tournament is an iterative process, to extrapolate these winning probabilities, conditional probabilities were applied. Every team has a chance to win the tournament and if a team is to win the championship, it must beat a series of six teams along the way. After the 68 teams were selected, a new matrix was created specifying the probabilities of each of the 68 teams advancing to every round of the tournament.. In order for a team to win in the second round, it needs to beat one of the two possible opponents that it could face if it beat its first round opponent. Therefore, the probability of a team winning the second game must account for both possible outcomes. This process carries through the entire six rounds.

The probabilities that every team will win the championship allowed for the creation of brackets. After solidifying the champion and through a singular Monte Carlo Simulation, a bracket is developed based upon the probability matrix and random number generation.

The model was tested by applying it to previous NCAA Tournaments and refined to improve its accuracy. A set of offensive and defensive variables, as well as outside ranking systems, were used to develop a linear regression model.

E. RESULTS AND CONCLUSIONS

At the conclusion of the 2019 NCAA March Madness Tournament, analysis of the model's performance showed that the model objectively failed to correctly predict the outcomes of the tournament. The model predicted less than 25 percent of the games correctly

and scored poorly in both the ESPN and Kaggle competitions.

THIS PAGE INTENTIONALLY LEFT BLANK

II. LITERATURE REVIEW

A. INTRODUCTION

Over recent years, mathematics has seen tremendous growth and widespread acceptance in the field of sports. More specifically, the interest of mathematics in the NCAA Basketball Championship Tournament or more colloquially, March Madness has ballooned in recent years. According to ESPN, there were over 17 million brackets submitted last year in ESPN's bracket challenge. Scholars have attempted to model the complexity of the unpredictable nature of the Tournament, fueled by the increasing popularity from not only college basketball enthusiasts, but also those interested in the cash prizes. Many have tried to model this real world complex problem and published their work to help future statisticians. The large monetary prizes associated with the tournament is that successful models are rarely to disclosed. Therefore, research on public forums are limited to failed attempts at modeling this situation or academic work.[1]

B. HOBLIN AND KOCHER

Two college students from Ball State, Tim Hoblin and Cody Kocher, used predictive models on the 2017 NCAA Basketball Tournament. Their model was mainly based upon a logistic regression model and randomization. The model included many offensive and defensive statistics, to include seed, shooting percentages, and turnover ratios to determine the winner of a particular game. [6]

They used data from 2006 to 2016, but some rule changes over the ten year span added the need for some modifications to the data. In 2009, the three-point arc was moved back a foot and, in 2016, the shot clock limit was shortened from thirty-five seconds to thirty seconds. The change in the three-point arc distance from the hoop decreases a team's overall shooting percent of three-pointers. The change in the shot clock led to more points, due to the quickened play of offenses. These key rule changes had major effects on the models. [7]

The models competed very well, seventeen out of twenty one scored in the top fifty percent of all submissions to ESPN's bracket challenge. The success of the brackets was attributed to the model and the alterations in the data due to rule changes. [7]

Although the team used a logistic model, the work inspired some ideas for this

project's model. For example, Hoblin and Kocher calculated shooting percentages and used them as a predictor variable. This is similar to the model seen in this paper. Other variables, such as turnover ratios, were tested in this model, but were determined not statistically significant.[7]

C. NATE SILVER

In an article from fivethirtyeight.com, Nate Silver, a well-known statistician, attempted to predict the NCAA Tournament over recent years. Silver's model consists of eight general team ratings. Six of the rankings are computer-based and two are founded in subjective rankings. The computer-based ratings are similar, using statistics such as the number of wins and losses, the strength of a team's schedule, margin of victory, and other common offensive and defensive statistics. The reason for incorporating a multitude of ranking systems is to clean up and correct biases that some contain. Some ranking systems are biased or heavily weigh certain statistics, and the incorporation of multiple systems removes the impact any on variables. One ranking system in particular is the ELO, created by fivethrityeight.com, in which the final score of the game, the location of the game (and thus the travel distance for each team), and the status of being home or away are the main factors. [8]

In addition to the six computer ratings, Silver's system includes two human factored rankings. The first is the NCAA selection of the 68 teams with seeds. This committee attempts to provide an order for the teams based on regular season performance. The second is the Associated Press' preseason rankings. The preseason rankings are based upon the talent prior to the start of the season, team chemistry, and experience of the players and coaches. These rankings benefits talented teams that can still make an impact in the tournament but were plagued by injuries, needed time to build team chemistry, and generally could not meet lofty expectations of the preseason. There is some subjectivity to Nate Silver's model, however the majority is based on objective ranking systems. In aggregate this helps mitigate biases and over-emphasis on certain factors.[8]

Silver uses ranking systems to better to assign strengths to teams. Although he uses both preseason and regular season rankings, only regular season rankings were incorporated into this paper's model. In general, the ranking systems update a weekly basis and are generally similar to another. These ranks were incorporated to remedy the problem of mixing teams from different conferences. [8]

D. JOHN EZEKOWITZ

John Ezekowitz attempted to model the Tournament as well, however he decided to use factors that were structurally different from most. He made the assumption that the games played in the post-season are principally different than the regular season. Ezekowitz created his own factors that he deemed important for teams to have in the tournament, such as a team's confidence and experience in the tournament. [9]

In creating a variable for the confidence of a team, the main assumption is that a team that beats more teams that will be playing in the tournament is more confident. This is based upon the logic that if a team has already played and beaten multiple opponents in the tournament, they will enter the tournament with a higher confidence in their ability to beat other teams. In an attempt to capture a team's tournament experience, he took the team's success in past tournaments coupled with the players from those post-season games. The assumption is that the teams with experienced coaches and players will play better and advance to the later rounds of the tournament. [9]

E. STEKLER AND KLEIN

Two students from George Washington University, H.O. Stekler and Andrew Klein built upon some earlier work on ranking systems. The previous work was only able to predict the first four of the six rounds of the NCAA Basketball Tournament, whereas the work by Stekler and Klein is able to predict through the final round. This is the main limitation of this model, the lack of ability to model the fifth and sixth rounds. [10]

A combination of over thirty ranking systems, from the years of 2003 to 2010, was compared to the tournament's seeding. Their combined ranking system with the tournament seeding was consistent. However, the combined ranking predicted better than the tournament seeding ranking but could only predict the earlier rounds well. [10]

F. NEW IDEAS

In addition to incorporating ideas from published work, we brought some new ideas to this project. Nate Silver uses an aspect of Bayesian statistics, where a previous state is updated and a new, more accurate state is created. Nate Silver uses this by incorporating a ranking system which includes preseason rankings. The model that is discussed in this paper uses a rolling mean. A rolling mean is a technique to harness a team's recent game

play. It incorporates the impact of key players being injured, suspensions, or even hot or cold streaks. As previously stated, the two undergraduates from Ball State, Hoblin and Kocher, used logistic regression. This is a more difficult model to interpret compared to the linear model used in this paper. [8][7]

G. CONCLUSION OF LITERATURE REVIEW

There are a number of published techniques for modeling and predicting the NCAA Division I College Basketball Tournament, including those outlined in this chapter. The methods used in this paper build upon some of the existing strategies with the same goal - predicting the winner of March Madness Games.

III. MODEL

A. INTRODUCTION

The end goal of this project is to correctly predict the winners of games in the NCAA Tournament, To accomplish this, a technique was needed to accurately model the number of points a team will scored. The method used in this project is a linear regression model. There are a variety of models in the world of mathematics; however, the most simple and intuitive is the general linear regression model. Linear models show relationships between certain factors and their influence upon a predicted value.

B. THE LINEAR MODEL

1. Understanding Linear Models

A linear model has two parts: a response variable and predictor variables. A predictor variable models the response variable. The model built in this paper attempts to predicts the number of points a team will score against another team based upon a number of factors. Equation III-1 shows an example multiple linear regression in which the estimated response variable, $\hat{y}$, is modeled by the predictor variables x_i and their coefficients β_i. For every one unit increase in x_i, $\hat{y}$ is subsequently increased by the corresponding β_i.

$$\hat{y} : \beta_0 + \beta_{11} + \beta_{22} + ... + \beta_{nn} \tag{III-1}$$

where

$\hat{y}$: Estimated number of points scored by a team

n : Total variables in the model

$x_1, x_2, x_3, ...x_n$: Each variables in the model

$\beta_1, \beta_2, \beta_3, ...\beta_n$: Each coefficient in the model

β_0 : Intercept of the model

Every statistical model, including a linear model, uses the Model Utility Test (MUC) in order to determine which factors are statistically significant in modeling the response variable. The MUC test has a Null Hypothesis, or the statement that is being tested and always states that each of the coefficients equal zero, in which the corresponding variables have no influence on $\hat{y}$. In other words, there is a lack of a linear relationship. This can be seen in the following equation:

$$H_o : \beta_1 = ... = \beta_n = 0$$

The opposite of the Null Hypothesis is the Alternative Hypothesis, in which, for the MUC test is that at least one variable is linearly related. In other words, at least one variable affects the response variable. The Alternative Hypothesis for a linear model is:

$$H_a : \text{At Least One } \beta \neq 0$$

The p-values allow statisticians to either reject the Null Hypothesis or fail to reject the Null Hypothesis. In a linear model, rejecting the Null Hypothesis is verification that there is a linear relationship between predictor variable(s) and the response. The significance of predictor variables occurs when the p-value is smaller than 0.05. When a p-value is larger than 0.05, the associated variable is not statistically significant, and therefore not linearly related to the response. This results in a failure to reject the Null Hypothesis, and hence, due to the lack of statistical evidence, therefore it is not included in the model.

In the case of modeling the NCAA Championship Tournament, the linear statistical model was applied in order to predict the number of points that Team A and Team B scored. It was based upon the variables seen in the following section.

2. Determining Factors

To build a model that best predicts the number of points scored in a NCAA game, some 15 factors were selected. The idea of "rolling statistics" is introduced into this model to capture the most recent strength and performance of a team. Teams that have star players get injured, suspended or benched could have a decrease in points scored. This directly influences the number of points scored. This model calculated the average of specific statistics over the last three games to provide the rolling means.

To ensure only quality variables were being added to the proposed model, a potential factor was included into the model and analyzed one at a time. To identify significant variables, the p-value, a measurement of statistical significance, was compared to the significance level of 0.05. If the factor produced a p-value less than 0.05, it remained in the model.

Analyzing the Akaike Information Criterion (AIC) occurred after the addition of a new variable. The smaller value the AIC, the model is determined to be of higher quality. Additional factors were added one at a time until the AIC began to increase. The final model had an AIC value of 35299.51. The AIC of the final model is half of the AIC value for the preliminary model development phase. The final model can be seen below:

$$\hat{y} = 0.81 \times x_1 + 0.43 \times x_2 + 0.05 \times x_3 - 0.04 \times x_4 + 0.05 \times x_5 + 0.13 \times x_6 + 1.00 \times x_7 + 1.15 \times x_8 - 0.72 \times x_9 + 0.01 \times x_{10} + 0.02 \times x_{11} - 0.01_{12} + 0.002_{13} + \beta_{1414} + \beta_{15} \times x_{15} - 42$$

where:

x_1 : Variable of Rolling Mean of Points Scored from Last Three Games
x_2 : Variable of Rolling Mean of Points Allowed from the Last Three Games
x_3 : Variable of Rolling Mean of Personal Rebounds Grabbed Team A
x_4 : Variable of Rolling Mean of Personal Rebounds Grabbed Team B
x_5 : Variable of Rolling Mean of Personal Fouls Committed Team A
x_6 : Variable of Rolling Mean of Personal Fouls Committed Team B
x_7 : Variable of Rolling Mean of Field Goal Percentage
x_8 : Variable of Rolling Mean of 3-pointer Percentage
x_9 : Variable of Rolling Mean of Free Throw Percentage
x_{10} : Variable of Kenpom Ranking Team A
x_{11} : Variable of Kenpom Ranking Team B

x_{12} : Variable of Massay Ordinals Ranking Team A
x_{13} : Variable of Massay Ordinals Ranking Team B
x_{14} : Variable of Conference of Team A
β_{15} :Coefficient for x_{14}
x_{15} : Variable of Conference of Team B
β_{15} :Coefficient for x_{15}

In the final model, each variable has a coefficient term that affects the response variable by the value of β multiplied by the quantity of the variable x. The β value has a linear relationship with the predicted number of points scored. For every one unit increase in a specific variable, the predictor variable increases by the value of the coefficients (β terms). For example, for every one unit increase in x_1, there is a increase in $\hat{y}$ by β_1.

These 15 predictor variables and their associated values generate the final model that predicts the number of points a team scores. All of the x terms represent one of the predictor variables of the model. There are two β terms in the linear model written above. Each of the conferences was a factor for both Team A and Team B. This list of conferences and their impact on the model can be seen in Appendices I and II.

The model states that every team, if all of the variables are zero, will score -42 points. This is the model's intercept, however does not make pragmatic sense due to the fact that when a team finishes a game without any statistic or any other variable, they cannot score a negative number of points. This is acceptable because it is almost impossible for a team not to register statistics for a game and therefore it can be assumed that a team would be able to score some points in every game. The remaining β values seem to make more intuitive sense. These coeficents are found in the linear model previously mentioned or in Table III. 1.

3. Assumptions of a Linear Model

A linear model requires four assumptions: linearity, normality of the error of the residuals, homoscedasticity, and independence. Linearity assumes a relationship between the predictor variables and response variable. Normality is the distribution of the residuals; usually verified by a Quantile-Quantile Plot. Homoscedasticity is the presence of constant variance of the residuals (error of the model): we must assume there is minimal. and is needed to allow a linear model to be accepted. The final assumption is independence. This

is the lack of an impact of one obseveration to the next observation.

a. Linearity

In a linear model, each of the variables is compared to the response variable to prove the presence of linearity of the model. Table III.1 describes the variables and the associated coefficients and p-values.

Variable	Coefficient	p-Value
Rolling Mean of Last 3 Games: Points Scored	0.81	$2 \times e^{-16}$
Rolling Mean of Last 3 Games: Points Allowed	0.43	$2 \times e^{-16}$
Rolling Mean of Last 3 Games: Rebounds Grabbed	0.05	$33 \times e^{-2}$
Rolling Mean of Last 3 Games: Rebounds Allowed	-0.04	$32 \times e^{-2}$
Rolling Mean of Last 3 Games: Fouls Committed	0.05	$27 \times e^{-2}$
Rolling Mean of Last 3 Games: Fouls Received	0.13	$1 \times e^{-2}$
Rolling Mean of Last 3 Games: Field Goal Percentage Shot	1.00	$82 \times e^{-2}$
Rolling Mean of Last 3 Games: 3-pointer Percentage Shot	1.15	$59 \times e^{-2}$
Rolling Mean of Last 3 Games: Free Throw Percentage Shot	-0.72	$60 \times e^{-2}$
Kempom Rank	0.01	$13 \times e^{-2}$
Opponent Kempom Rank	0.02	$3 \times e^{-5}$
Massey Rank	-0.01	$2 \times e^{-16}$
Opponent Massey Rank	-0.002	$62 \times e^{-16}$
Conference	Located in Appendix A	

Table III.1 Summary of the linear model

The results of the model can be seen in Table III.1. The table is the output of the fitted linear model based upon the data from 2019 and variables discussed earlier. The level of statistical significance is found under p-values. Due to the complexity of modeling a basketball games, variables with higher p-values were included. These variables were included to decrease the AIC of the model. This decision allows a larger number of variables to be included. The intuitive part of this table is the Coefficient column. This shows that with every one unit increase of a certain variable, there is a change of the total points scored by the value of Coefficient. For example, for every one unit increase in the "Rolling Mean of Last 3 Games: Points Scored", there is a consequent increase of 0.81 in the predicted points scored for Team A.

Linearity can also be inspected visually. Plotting the predictor variables of the model against the standard residuals (error) provides the ability to analyze the linearity

of the model. The following two figures: Figure 2 and Figure 3 are two examples of the predictor variables.

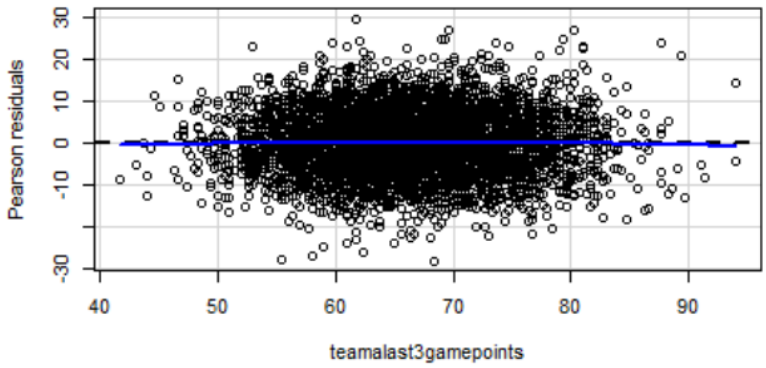

Figure 2. Standardized Residuals against Rolling Mean of Points Scored of Team A

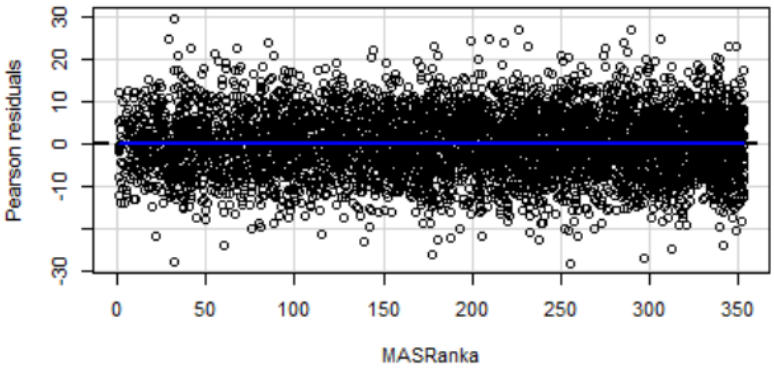

Figure 3. Standardized Residuals against Massey Ranking of Team A

Figure 2 and Figure 3 show the two types of comparisons between the predictor variables of this model and the standardized residuals. All 15 variables followed on of these two patterns. Figure 2 shows some data that seems a little concerning with the consolidating mass in the middle of the graph. Both figures show standardized results that have error terms more than +/- 2. The data should be inside of this range 96 percent of the time.

b. Normality

The normality of the distribution for the error term needs to be assumed for a linear model to be applied. This error (represented by ϵ) is written mathematically as:

$$\epsilon \sim \text{Norm}(0, \sigma) \tag{III-2}$$

Normality can be assumed by of viewing the normality of a data set is through the Quantile-Quantile Plot (qq-plot). The plot attempts to show the difference between the theoretical normal quantiles and the sample quantiles. The closer the plot is to the normal line (represented by the positive diagonal line), the more likely the distribution is normal. This is represented in Figure 4.

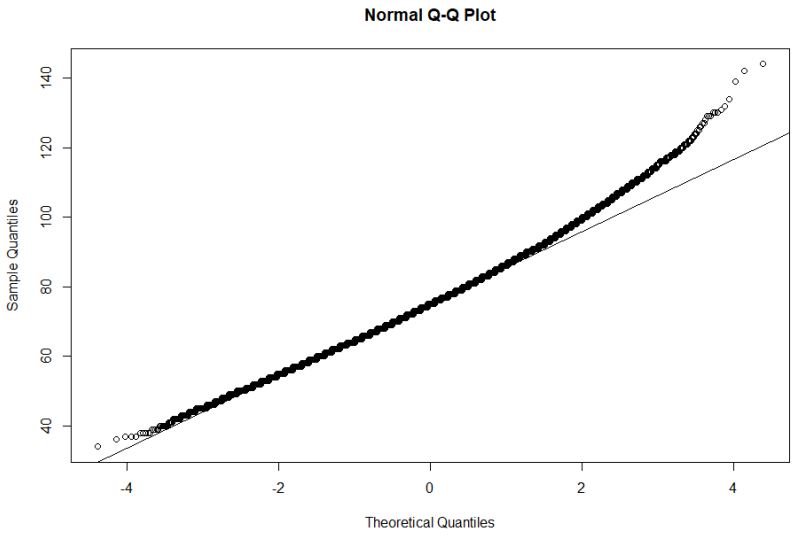

Figure 4. Quantile-Quantile Plot to Measure Normality of Points Scored by Team A

Figure 4 shows that the points scored data is fairly normally distributed. There are some slight "tails" to at the extreme ends of the qq-plot, however the expected for real-life data passes the assumption of normality.

c. Homoscedasticity

Constant variance, also known as homoscedasticity, is important because it shows that the model predicts equally well for all values of the factors in the model. Figures 2 and Figure 3 show that as x increases (this is the predictor variable) the error does not increase or decrease. Fanning, or irregular variance is a key indicator that error term is not normally distributed. Additionally, large variance does not coincide with a normally distribution, the majority of the data (ninety-six percent) does not reside within two standard deviations from zero. The data passes the constant variance assumption.

d. Independence

Independence of the model is the final assumption that is validated in order to accurately fit a linear model. The majority of the games are independent, however there are some in-conference games that are dependent. The outcome of the previous encounters create a relationship between these events. For the majority of the games, the following is true: game between teams A and B are not affected by the game between teams C and D. However, the first encounter between team A and B affects the outcome of the second game between team A and B. This assumption is valid.

e. Conclusion

Sufficient evidence has been shown for all four of the assumptions needed for a linear model to be valid. Therefore the points scored can be modeled using a linear regression model. The linear assumption is verified through the individual relationships between the response variable and each predictor variables. The normality of the model is validated by the qq-plot that shows the residuals are normally distributed with slight deviations in the tails. The presence of homoscedasticity is found by looking at the spread of the residuals. The lack of fanning and the constant variance proves this assumption. The final assumption, independence is assumed due to the overall lack of dependency of one games outcome on any other game. The validation of each of these four assumptions allow the general linear model to be used.

IV. GAME MATRIX

A. INTRODUCTION

The model described in chapter 3 was used to construct submissions for two scoring platforms, the first of which is a competition hosted by Kaggle. Submissions to this competition require an estimated probability for every team to win against all other possible match-ups in the tournament. The game matrix from this chapter provides the probability of each NCAA Division I team beating every other team in the division and is calculated using the difference in projected points scored for every pair of teams in the tournament.

B. DIFFERENCE IN POINT DISTRIBUTION

The model in chapter 3 predicts the number of points a team (Team A) will score against another team (Team B) using the necessary model variables for the two teams. Because the model's residuals follow a normal distribution sufficiently, the points scored for Team A against Team B is also distributed normally (*Section 3.4*) and vice versa albeit with differing means (μ) and variance (σ^2). An example of the two distributions for points scored when Team A plays Team B can been seen in Figure 5.

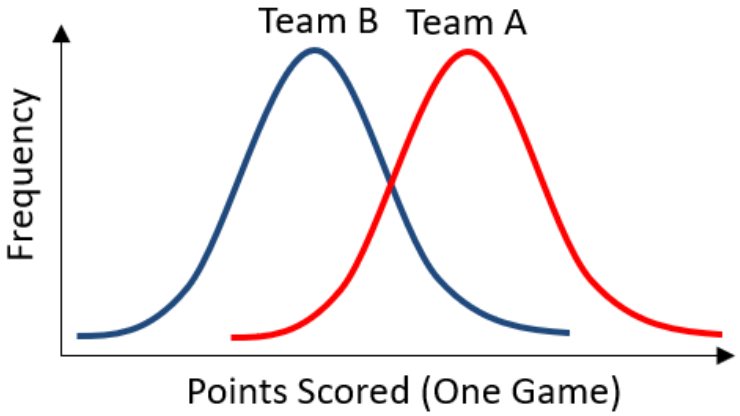

Figure 5. Graphical Representation of the Distribution of Points Scored for Teams A and B

These distributions are defined in mathematical notation below.

$$\text{AvB} \sim \text{Norm}(\mu_{AvB}, \sigma^2_{AvB}|\vec{x}) \tag{IV-1}$$

$$\text{BvA} \sim \text{Norm}(\mu_{BvA}, \sigma^2_{BvA}|\vec{x}) \tag{IV-2}$$

where

AvB : Number of Points Scored by Team A against Team B

BvA : Number of Points Scored by Team B against Team A

μ_{AvB} : the population mean of the points scored by team A against Team B

σ^2_{AvB} : the population variance of the points scored by team A against Team B

$\vec{x}$: Model input data

The difference between two normally distributed variables is also normally distributed as depicted in Equation IV-4 and Figure 6.

$$\text{AvB} - \text{BvA} \sim \text{Norm}(\mu_{AvB} - \mu_{BvA}, \sqrt{\sigma^2_{AvB} + \sigma^2_{BvA}}|\vec{x}) \tag{IV-3}$$

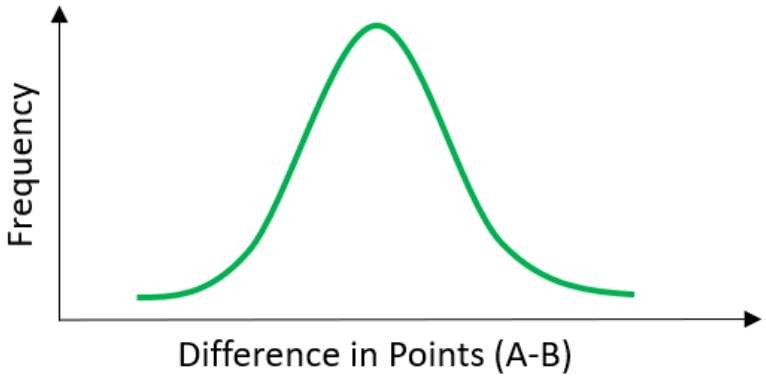

Figure 6. Difference of Points Scored between Team A and B (Normally Distributed)

This combined distribution is the difference of the normal distributions and is visually represented in Figure 6. When the points scored by Team A is greater than the points by Team B (AvB - BvA > 0), this represents a victory for Team A. When the points scored by Team B is greater than the points from Team A (BvA - AvB > 0), Team B wins. Calculating the probability of Team A scoring more than Team B (P(AvB - BvA < 0) give the probability of Team A beating Team B.

C. FINAL GAME MATRIX

This process continues for all combinations of teams to develop the game matrix. The goal of the game matrix is to give a specified probability that a team will win a particular game over another team. The game matrix is broken into two complement portions: the upper and lower halves. Each team has a probability of beating another team with the process just defined. The middle diagonal is blacked-out because teams do not play themselves. An excerpt from of the game matrix is shown in Figure 7. For example, Abilene Christian, written as Abilene Chr, has an estimated 0.90 probability to win against Air Force.

	Abilene Chr	Air Force	Akron	Alabama	Alabama A&M	Alabama St
Abilene Chr		0.90102415	0.782959469	0.835275815	0.9997752	0.9972276
Air Force	0.0989758492		0.306696067	0.377487128	0.9868658	0.9314352
Akron	0.2170405314	0.69330393		0.576559241	0.9968079	0.9767979
Alabama	0.1647241849	0.62251287	0.423440759		0.9943620	0.9639461
Alabama A&M	0.0002247577	0.01313423	0.003192064	0.005637983		0.2310322
Alabama St	0.0027724442	0.06856484	0.023202071	0.036053862	0.7689678	

Figure 7. Exerpt from the 350 x 350 Game Matrix

D. KAGGLE: CONFIDENCE WEIGHTED

One of the platforms used to evaluate the proposed model is Kaggle.com, a data science competition submission site. For entry into this competition, a probability of every team winning against every other team is obligatory. These entries are produced by pre-

dicting the win probability for each of the sixty-eight teams playing in the tournament and are taken directly from the game matrix. The scoring is determined by a log loss equation, which rewards submissions for high probabilities associated with the wining team and penalizes submissions with lower probabilities for the losing team. The equation for the log loss of a single game is shown in IV-1.

$$\text{log loss} : -2(y_{ij} \log p_{ij} + y_{ij} \log p_{ji})[11] \tag{IV-4}$$

where:
p_{ij} is the predicted probability of team i beating team j
y_{ij} is 1 if Team i wins, 0 if Team j wins
$\log()$is the natural logarithm (base e)

The smaller the value of the log loss, the better the model's prediction. The purpose of the log loss formula is to penalize probabilities that are both overconfident and incorrect. Submission's final score is the average log loss across all 63 games in the tournament.[12]

In this scoring method, every game matters and every probability is important. Although extreme probabilities add more risk to the total log loss calculation, this fact did not affect the modeling approach. In other words, the model was not modified in order to make probabilities less extreme.

V. PROBABILITY MATRIX

A. INTRODUCTION

Every team has a chance of being crowned the NCAA Tournament Champion, and therefore each team has a nonzero probability to advance through each round of the tournament. Here, a matrix is created with the conditional probabilities for every tournament team to advance to every round in the tournament.

B. CONDITIONAL PROBABILITY

Conditional Probability is a term in statistics that refers to the probability of Event A occurring if Event B occurs. It is denoted below in mathematical notation.

$$P(A|B) \tag{V-1}$$

where
A : Event that has a probability of occurring
B : Event that has a probability of occurring

An example involves the weather. Event B may be the probability that it rains, while Event A is the probability that you will eat lunch outside. The probability that you will choose A is dependent and influenced on the probability of Event B.

In this model, Event A is the probability that a specific team wins against all possible opponents and Event B is the probability that they won the previous game. Event A becomes more and more complex as the rounds advance and the increased number of teams that could be playing in that particular game. Conditional probabilities were the basis of creating the Probability Matrix to determine which team will win its games.

C. BUILDING THE TOURNAMENT PROBABILITY MATRIX

Figure 8 depicts the probability of every team in the East Region to advance to every round of the tournament. A team has a higher chance of advancing into the later rounds if the other teams in that region are weaker. For Duke to make it to the Sweet Sixteen, they must beat North Dakota State, and the winner of the Virginia Commonwealth Uni-

versity and University of Central Florida game. Duke, has a 0.61 estimated probability of winning against North Dakota State (NDU) and advancing to the next round. In the subsequent round, Duke's estimated winning probability decreases to 0.41. This probability is the conditional probability that Duke will beat Virginia Commonwealth University (VCU) or University of Central Florida (UCF) given that Duke won its first game and is calculated as follows:

$$\begin{aligned}
\text{P(Duke wins Round 2)} &= \text{P(Duke Beats NDU)} \times (\text{P(Duke beats VCU)} \times \text{P(VCU beast UCF)} + \\
&\qquad \text{P(Duke beats UCF)} \times \text{P(UCF beats VCU)} \\
\text{P(Duke wins Round 2)} &= 0.61 \times (0.597x0.316) + (0.861x0.684) \\
\text{P(Duke wins Round 2)} &= 0.406
\end{aligned}$$

The probability matrix includes all of the possible probabilities for every team to advance to all six round. The championship round is the most complex conditional probability statement that includes every teams' probabilities in playing in the final championship game. The model produced that the University of Buffalo has a thirty percent chance of winning the tournament. The following figure is an excerpt of entire probability matrix that demonstrates the probabilities of each team advancing to every round.

TeamName	Round1	Round2	Round3	Round4	Round5	Round6
Duke	0.61084599	0.406207145	0.2963350156	1.641733e-01	6.412715e-02	1.335001e-02
N Dakota St	0.38915401	0.217945092	0.1382721371	6.195296e-02	1.812260e-02	2.561488e-03
VA Commonwealth	0.31562158	0.081328469	0.0369938177	1.030339e-02	1.619548e-03	1.042401e-04
UCF	0.68437842	0.294519294	0.1863573263	8.315125e-02	2.418227e-02	3.390547e-03
Mississippi St	0.48043475	0.079836985	0.0086739295	9.212930e-04	4.363120e-05	6.825523e-07
Liberty	0.51956525	0.092880556	0.0109527471	1.259091e-03	6.568759e-05	1.147960e-06
Virginia Tech	0.57913138	0.491654526	0.2064262816	8.244844e-02	2.068972e-02	2.394913e-03
St Louis	0.42086862	0.335627932	0.1159887451	3.814888e-02	7.423522e-03	6.221390e-04
Maryland	0.73047590	0.613951137	0.4269425405	2.869047e-01	1.290035e-01	3.264159e-02
Belmont	0.26952410	0.176953731	0.0832689653	3.670576e-02	8.748947e-03	9.482063e-04
LSU	0.74444718	0.184844252	0.0695191896	2.408925e-02	4.104073e-03	2.918589e-04
Yale	0.25555282	0.024250880	0.0042199320	6.419107e-04	3.698773e-05	7.239892e-07
Louisville	0.69987370	0.277638869	0.0967844066	3.883275e-02	8.098708e-03	7.404298e-04
Minnesota	0.30012630	0.067384041	0.0129988400	2.928682e-03	2.800583e-04	9.893493e-06
Michigan St	0.22207030	0.096287065	0.0236519589	6.761603e-03	8.863576e-04	4.567527e-05
Bradley	0.77792970	0.558690025	0.2826141670	1.607768e-01	5.576052e-02	9.881632e-03

Figure 8. Excerpt from the Probability Matrix: Eastern Region

D. MONTE CARLO SIMULATION

A Monte Carlo Simulation used a random number generator to simulate random outputs. In this case, a bracket is simulated based upon the probabilities from the probability matrix. 23 brackets are created fro submission to the ESPN Bracket Challenge. An example output of this Monte Carlo Simulation is seen in the Figure 9. In this iteration, the winner of the tournament is Gonzaga University.

E. ESPN: ROUND WEIGHTED

There are a series of rounds throughout the March Madness Tournament. The first round contains 64 teams and after six iterative rounds, the Champion is crowned. ESPN is the most traditional and well-known NCAA March Madness submission platform. For the ESPN bracket competition, the objective is to have the most amount of points accumulated at the end of the competition. ESPN's scoring algorithm has a weighted algorithm that emphasizes the correct prediction of the winners of the later rounds and therefore the champion. See the equation below:

Roundof32	Roundof16	Roundof8	Roundof4	Roundof2	Championship
Duke	Duke	Duke	Duke	Gonzaga	Gonzaga
UCF	St Louis	Maryland	Gonzaga	Cincinnati	
Liberty	Maryland	Gonzaga	Utah St		
St Louis	Bradley	N Kentucky	Cincinnati		
Maryland	Gonzaga	Utah St			
LSU	Vermont	Houston			
Louisville	N Kentucky	Oklahoma			
Bradley	Florida	Cincinnati			
Gonzaga	Utah St				
Syracuse	New Mexico St				
Murray St	Houston				
Vermont	Kentucky				
St John's	Oklahoma				
N Kentucky	Oregon				
Florida	Old Dominion				
Michigan	Cincinnati				

Figure 9. Example Bracket Produced by Monte Carlo Simulation

Equation that calculates the total points scored for a bracket:

$$y = 10 \times x_1 + 20 \times x_2 + 40 \times x_3 + 80 \times x_4 + 160 \times x_5 + 320 \times x_6 \qquad \text{(V-2)}$$

where

y : Number of Points Scored

i : Round$\{1, 2...6\}$

x_i : Number of games whose outcome is correctly predicted in Round *i*

Note that the number of points awarded is doubled with each consecutive round, emphasizing the need to correctly predict the winner of the later rounds. In terms of the tournament, single game upsets, when a team upsets a higher ranked team, but loses in the second round, have little effect on the total points of the bracket. For example, if the University of Baylor is ranked 9, and beats the University of Syracuse, ranked 8, but loses

in the next round, there is little impact on the number of potential points that could be lost if the upset is not predicted. However, picking teams that advance further and make it farther in the tournament will be paramount in obtaining a high score. ESPN's bracket competition has a high level of prestige (to the 17.3 million participants)[1]. It is the most well known submission platform and by performing well, the model can gain much credibility.[2] The ESPN bracket is only concerned with which team is victorious; this is different from the other submission platform Kaggle.

To score the largest amount of points in the ESPN competition a bracket must correctly predict the champion. This will be the strategy for this project's bracket submissions. Only brackets that contain the top 12 teams that have the most likely chance of winning will be submitted to ESPN submission platform. The assumption that the actual winner of the tournament was in this top 12. To achieve the total number of brackets (23), the top 11 winners were chosen again. If the winner is correctly predicted, the bracket will earn at least 630 points. This is thirty percent of the total points available. The nature of the scoring algorithm, which emphasizes predicting the later rounds correctly, makes this strategy more appealing for this project.

THIS PAGE INTENTIONALLY LEFT BLANK

VI. MODEL VALIDATION

Since the goal of this analysis is to perform as well as possible in both the Kaggle and ESPN competitions, the model was tested in both scoring scenarios.

A. KAGGLE

Using data from the last five tournaments (2014 to 2018), individuals can make submissions to this pre-competition model validation. The benefit is to understand how effective one's model predicts the winner of games from past tournaments. Using the same scoring method as seen in *Chapter IV, Equation 4*, each submission receives a score; the lower the score, the better the model predicted the winners. All submissions are compared to other individuals that have also submitted their models.

The probabilities from the game matrix were implemented to predict each outcome. Using the actual results from 2014 to 2018 March Madness tournament, the model did a mediocre job of predicting the outcome of each game. The model scored a 1.3, where random guesses are scored at 0.6. This unfortunately shows that the model is far from successful due to some limitations.

B. ESPN

To validate the model for use in the ESPN Tournament Challenge, it used data from 2016 by seeing how often the winner for every game was correctly selected. 57.1 percent of the games were correctly selected ($\frac{36}{63}$ games). The main limitation of the model was the lack of strength of schedule factor. This is shown when Hampton (the sixteenth seed from the Mid-West Region) was expected to beat Virginia (the number one seed from the region). Such an output does not make common sense. This, among other examples, shows the model allows very weak teams to have higher than fifty percent probability of beating basketball powerhouses.

THIS PAGE INTENTIONALLY LEFT BLANK

VII. THE TEST: MARCH MADNESS 2019

A. NCAA TOURNAMENT 2019

1. Selection Sunday to First Round

On Selection Sunday, the 17th of March, 2019 the selection committee picked the 68 teams. This officially opened the ESPN and Kaggle Competitions for submission. The Kaggle submission was completed the following morning and included all 68 teams. Due to the fact that the original model had some of the play-in game team moving deep into the tournament, the creation of the brackets waited until the end of all of the play-in games. This gave the model some help in alleviating any error in the elimination of four teams from 68 to 64 teams.

2. Bracket Selection

For ESPN, 23 brackets out of the 25 brackets submitted were based upon the statistical model created. Using the output of the Monte Carlo Simulation, the 12 teams with the highest probability of winning the Champion were selected. These top 12 teams and their associated probability of winning the Championship game are seen in the following Figure 10. The top 11 teams were selected again to create the remainder of the possible brackets.

The figure below shows the top teams that the model predicted to win the championship based upon the conditional probabilities of the winning of the championship round. All of the other games, minus the champion, were selected based off a Monte Carlo Simulation and random number generation.

Team	Probability to Win Championship
Cincinnati	0.319117198
Houston	0.274518287
Gonzaga	0.106683961
New Mexico St	0.063585216
Old Dominion	0.048954970
St John's	0.039919241
Michigan	0.036009731
Maryland	0.032641590
Florida St	0.013466076
Duke	0.013350013
Bradley	0.009881632
N Kentucky	0.005442658

Figure 10. Model's Top Predictions to Win

B. RESULTS

1. ESPN

The model performed poorly and incorrectly predicted the outcomes of the majority of the games played in the tournament. The submissions for ESPN performed in the bottom of all brackets submitted. Of the 23 brackets, the highest performing brackets fell into the fourth percentile, in other words, such brackets beat only four percent of all of the millions of brackets submitted. The average placing of the brackets was 17 millionth place. These are clear indicators that the model struggled to find a successful manner to predict the correct outcome of each game.

2. Math Department Pool

The Department of Mathematical Science created a group on ESPN to compare the results of the project's brackets against a smaller population. Including the 23 submission

based upon the model described in this paper. The submissions created using the model finished last in the pool. The submitted brackets populated the bottom of the group in terms of performance. Of the 127 submissions, the 23 brackets based off of this model were beaten by all other 104 bracket submissions. The conclusions drawn from this is that the model performed worse than a causal fan's intuition.

3. Kaggle

The scoring platform of Kaggle received 866 submissions. Using the calculation of Log Loss, as previously mentioned in Chapter Four, the competition analyzes the correctness and the confidence of each prediction. From the 866 submissions, the model arrived at 816th place. The one highlight from the Kaggle submission is that the model's performance was less than 100 places behind 50 percent probabilities, or flipping a coin. A brief summary of the results of the competition can be found in the table below.

Place	Individual	Score
1	Winner	0.41
729	Random Prediction	0.69
816	Model	1.28
866	Last Place	20.81

Table VII.1 Summary of the Kaggle Competition

C. CONCLUSION

The aforementioned information presented display the unsuccessful performance of the model. The model's brackets was bettered by 16.5 million other submissions across the world in the ESPN competition. In the Kaggle submission, the model secured 816th place out of 866 total competitors and performed worse than a random selection (all teams have a 0.50 probability of winning every game). While it is quite possible the the model performed poorly by chance based off the circumstances of how the tournament played out, there is much improvement for future years.

THIS PAGE INTENTIONALLY LEFT BLANK

VIII. FUTURE WORK/CONCLUSION

A. INTRODUCTION

The performance of the model was somewhat disappointing, but there is potential for improvement. There are certain revisions and future work will allow for large improvements. There are many other of methodologies or techniques to model this scenario.

B. MODEL LIMITATIONS AND IMPROVEMENTS

The model has much room for improvements, evident by the poor performance in both Kaggle and ESPN submission competitions. The lack of success is attributed to the limitations of the data and the model. In terms of the data, no individual player data was available, restricting the prediction to be based upon team statistics. The strength of schedule, the lack of defensive statistics, the lack of shooting percentages for the opposing teams, and the overall type of model all could be improved to make a better product.

Players have intangibles. Some players have a natural ability to make important plays that help their teams win. Including player data could be interesting to see what factors: height, speed, or points per game affect the model. Coach and player experience in important games or past tournaments has a potential to impact the model.

Although the model accounted for strength of the team, there possibly could be more ranking systems in order to better mask the effect of weaker teams having better statistics due to their weaker conferences. This lack of in-depth modeling of the strength of schedule dramatically affected the model. For example, the model did not allow Virginia, the winner of the 2019 tournament to win against Gardner-Webb, a 16th seed, in any of the 23 brackets submitted. The effect was detrimental to the model, in which Duke was the only 1-seed to make an appearance in Figure 10 of Chapter VII.

Secondly, the defensive statistics are limited in the model. The only statistic that has some defensive nature is the number of rebounds and the opponents' points scored. There are more defensive statistics that could be attributed to winning games and scoring points. Turnover ratio is a statistics that encompassed many defensive statistics.

Additionally, the lack of shooting percentages of the opponent could also have bettered the model. The completion rates of each of the different opponent shots: free throws,

field goals and three point shots, have a relationship with the number of points Team A allowed.

Finally, one last improvement could be the type of model. A logistic model could be more applicable because of its binary nature (a team wins or loses). This could possibly assist in the prediction power of the model.

C. CONCLUSION

Although, the model performed poorly, much was learned and can be refined. Tournaments that are based off of single elimination games have a large variance associated with the outcomes. A strong team could beat a weaker opponent nine times out of ten games played, however if the weaker team wins that one time, the favored team is out of the competition. This high level of chance makes these situations difficult to model. Looking forward, the model can continued to be refined and improved to create a more holistic predictive model of the Division I Men's NCAA College Basketball Tournament. The tournament is full of uncertainty and lives up to its name, March Madness.

IX. APPENDIX

A. TEAM A CONFERENCE (LINEAR RELATIONSHIP)

Conference Name	Coefficient	p-Value
American Athletic Conference	3.79	$2 \times e^{-3}$
American Ten Conference	2.45	$2 \times e^{-2}$
American Sun Conference	1.08	$3.6 \times e^{-1}$
Big 12 Conference	4.63	$4 \times e^{-4}$
Big East Conference	2.62	$4.2 \times e^{-3}$
Big Sky Conference	-0.053	$9.6 \times e^{-1}$
Big South Conference	1.47	$2.2 \times e^{-1}$
Big Ten Conference	6.23	$4.9 \times e^{-7}$
Big West Conference	2.51	$4 \times e^{-2}$
Colonial Athl. Athletic Conference	0.54	$6.4 \times e^{-1}$
Conference USA Conference	0.82	$4.7 \times e^{-1}$
Horizon League Conference	1.4	$2.4 \times e^{-1}$
Metro Atlantic Athletic Conference	0.58	$5.8 \times e^{-1}$
Mid-American Conference	1.5	$2.1 \times e^{-1}$
Mid-Eastern Athletic Conference	0.48	$6.6 \times e^{-1}$
Missouri Valley Conference	0.87	$4.8 \times e^{-1}$
Mountain West Conference	1.88	$1.2 \times e^{-1}$
Northeast Conference	-0.16	$9.9 \times e^{-1}$
Ohio Valley Conference	0.66	$5.9 \times e^{-1}$
Pacific-12 Conference	2.03	$9.7 \times e^{-2}$
Patriot League Conference	1.16	$3.1 \times e^{-1}$
Southeastern Conference	4.04	$1 \times e^{-3}$
Southern Conference	1.317120	$3.1 \times e^{-1}$
Southland Conference	0.76	$5.3 \times e^{-1}$
Southwest Athletic Conference	1.317120	$3.1 \times e^{-1}$
Summit League	1.58	$2 \times e^{-1}$
Sun Belt Conference	1.45	$2.2 \times e^{-1}$
West Coast Conference	3.1	$1 \times e^{-2}$
Western Athletic Conference	0.85	$5 \times e^{-1}$

Table IX.1 Summary of Team A Conference Affected Points Scored

B. TEAM B CONFERENCE (LINEAR RELATIONSHIP)

Conference Name	Coefficient	p-Value
American Athletic Conference	-5.71	$5.4 \times e^{-7}$
American Ten Conference	-1.61	$1.5 \times e^{-1}$
American Sun Conference	-5.71	$5.4 \times e^{-7}$
Big 12 Conference	-1.32	$3 \times e^{-1}$
Big East Conference	-1.96	$1.1 \times e^{-1}$
Big Sky Conference	-1.70	$2.3 \times e^{-1}$
Big South Conference	-1.92	$1.3 \times e^{-1}$
Big Ten Conference	-4.92	$3.7 \times e^{-5}$
Big West Conference	-2.96	$2 \times e^{-2}$
Colonial Athl. Athletic Conference	0.04	$9.7 \times e^{-1}$
Conference USA Conference	-0.55	$6.4 \times e^{-1}$
Horizon League Conference	-1.13	$3.8 \times e^{-1}$
Ivy League Conference	-0.41	$7.3 \times e^{-1}$
Metro Atlantic Athletic Conference	-1.12	$3.4 \times e^{-1}$
Mid-American Conference	-1.59	$1.9 \times e^{-1}$
Mid-Eastern Athletic Conference	-1.9	$1.4 \times e^{-1}$
Missouri Valley Conference	-0.11	$9.3 \times e^{-1}$
Mountain West Conference	-2.69	$3 \times e^{-2}$
Northeast Conference	-1.2	$3.2 \times e^{-1}$
Ohio Valley Conference	-1.9	$1.2 \times e^{-1}$
Pacific-12 Conference	-1.9	$13. \times e^{-1}$
Patriot League Conference	-1.68	$1.6 \times e^{-1}$
Southeastern Conference	-3.6	$2 \times e^{-3}$
Southern Conference	-2.7	$4 \times e^{-2}$
Southland Conference	-2.6	$5 \times e^{-2}$
Southwest Athletic Conference	-1.67	$2.4 \times e^{-1}$
Summit League	-2.7	$5 \times e^{-2}$
Sun Belt Conference	-2.2	$7 \times e^{-2}$
West Coast Conference	-3.8	$1 \times e^{-3}$
Western Athletic Conference	-1.6	$2 \times e^{-1}$

Table IX.2 Summary of Team B Conference Affected Points Scored

LIST OF REFERENCES

[1] March madness 2019 dates and schedule. NCAA, 2018.

[2] Kaggle. Scoring your march madness bracket.

[3] Kaggle. Regularseasondetailedresults.csv.

[4] Kaggle. Ncaatourneyseeds.csv.

[5] Kaggle. Teams.csv.

[6] Curtis Gary Dean. Predictive modeling and march madness. Actuarial Expertise, 2017.

[7] Cody Kocher and Tim Hoblin. Predictive Model for the NCAA Men's Basketball Tournament. Ball State University, Muncie, Indianapolis, 2017.

[8] Nate Silver. Building a bracket is hard this year, but we'll help you play the odds. FiveThirtyEight, 2014.

[9] John Ezekowitz. Quantifying intangibles: A new way to predict the ncaa tournament. Harvard Sports Analytics Collective, 2011.

[10] A. Klein and Stekler. Predicting the outcomes of ncaa championship basketball games. research program on forecasting. 2011.

[11] Making sense of logarithmic loss. Data Wookie, 2015.

[12] Daniel Becker. What is log loss?