The Efficacy of Modern Basketball Analytics in College Basketball  
Predictive Win-Probability Algorithms

by

Robert Mepham

A thesis presented to the Honors College of Middle Tennessee State University in Partial Fulfillment of the requirements for graduation from the University Honors College.

Fall 2022

Thesis Committee:

Dr. Ramchandra Rimal, Thesis Director

Dr. John R. Vile, Thesis Committee Chair

The Efficacy of Modern Basketball Analytics in College Basketball   
Predictive Win-Probability Algorithms

by Robert Mepham

APPROVED:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dr. Ramchandra Rimal, Thesis Director

Assistant Professor, College of Basic and Applied Sciences

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dr. John R. Vile, Thesis Committee Chair

Dean, University Honors College

AKNOWLEDGEMENTS

My sincerest thanks go to my girlfriend, Kylie Dykgraaf, for supporting me throughout the entire thesis process, including serving as a mentor and editor. I would like to recognize my parents, Bob & Dinah Mepham, for their unconditional love and invaluable advice, not only for this endeavor, but throughout life as well. Finally, I would like to acknowledge my thesis committee. Thesis Director Dr. Ramchandra Rimal, Assistant Professor, College of Basic and Applied Sciences, Middle Tennessee State University - for providing vision, awareness, and countless insight during the entire course of my research. Thesis Committee Chair, Dr. John R. Vile, Dean, University Honors College, Middle Tennessee State University for his vital critiques, guidance, feedback, and encouragement for this project. Every one of you has played a role to make this opportunity possible, and for that I am forever grateful.

ABSTRACT

This study examines the efficacy of different types of machine learning algorithms, as well as different types of basketball statistics, in predicting both the season-long win/loss percentage (W-L%) of a team, as well as the individual game win-probability for every Division 1 Men’s College Basketball Game in the 2022 season. Types of algorithms examined include LASSO regression, logistic regression, Random Forest regressors and classifiers, and AdaBoost regressors and classifiers. Data were also collected to compare the optimized models against commonly accepted models used by sportsbooks, against which the models developed from this research hold a 6.69% edge. The null hypothesis, that modern (“advanced”) basketball statistics better predict both the season-long W-L% and the outcomes of individual games, was supported by the optimized machine learning algorithms having higher accuracy scores and lower error values when trained on advanced statistics compared to when trained on traditional box-score measures. It was also found that the optimal number of previous games to consider when predicting a team’s next outcome is only a team’s single previous game (excluding variance from injury).

TABLE OF CONTENTS

Page

LIST OF TABLES AND FIGURES ………………………………………………......... **vii**

LIST OF APPENDICES ……………………………………………………………….... **xi**

I. INTRODUCTION ……………………………………………………………………... **1**

Statement of Purpose …………………………………………………………...... **5**

Research Questions ……………………………………………………………..... **6**

Definition of Terms ……………………………………………………………..... **7**

Significance of Study ………………………………………………………….... **10**

II. REVIEW OF LITERATURE ………………………………………………………... **11**

The History of Statistics in Basketball ………………………………………….. **11**

How are Statistics Leverages By Coaches? ……………………………………... **12**

The Difference Between Division 1 Collegiate Basketball and the NBA ………. **13**

Modern Factors Impacting the Division 1 Landscape …………………………... **15**

Machine Learning Algorithms Used for Regression ……………………………. **16**

Machine Learning Algorithms Used for Classification ………………………… **20**

Data and Output Validation in Machine Learning Algorithms …………………. **23**

Techniques To Optimize Model Performance ………………………………….. **24**

III. METHODOLOGY …………………………………………………………………. **28**

Measures ………………………………………………………………………... **29**

Procedures …………………………………………………………………….... **29**

Data Analysis ………………………………………………………………….... **31**

IV. RESULTS …………………………………………………………………………... **33**

Stage 1 – Determining Significant Predictors of Winning Basketball ………….. **33**

LASSO Regression Models …………………………………………….. **35**

Random Forest Regressor Models …………………………………….... **38**

AdaBoost Models ………………………………………………………. **42**

Stage 1 Conclusion ……………………………………………………... **46**

Page

Stage 2 – Evaluating Significant Predictor’s Ability to Classify The ……………….  
Outcome of a Division 1 College Basketball Game.…………………………...... **48**

Optimal Lag Term ……………………………………………………..... **48**

Reduced Predictor Set …………………………………………………... **53**

Basic vs. Advanced Statistics …………………………………………… **58**

Optimal Machine Learning Algorithm Approach …………………...….. **61**

Finalized Models and Evaluation ……...……………………………....... **62**

V. DISCUSSION ……………………………………………………………………….. **67**

Findings ……………………………………………………………………….... **67**

Outlier Analysis: 2022 Season ………………………………………………….. **68**

Optimal W-L% Model ………………………………………………….. **68**

Optimal Game Simulation Model ………………………………………. **73**

In Depth team Analysis: MTSU Blue Raiders Men’s Basketball, 2022 Season ... **78**

Game Prediction Model ………………………………………………… **78**

Post-Game Outcome Analysis Model …………………………………... **81**

Conclusion and Future Direction ……………………………………...………... **85**

Conclusion ……………………………………………………………… **85**

Future Direction ……………………………………………………….... **87**

REFERENCES …………………………………………………………………………. **88**

APPENDICES ……………………………………………………………….…………. **96**

LIST OF TABLES AND FIGURES

Page

Table 1: Data Dictionary, Basic Team Statistics ……………………………………….... **7**

Table 2: Data Dictionary, Advanced Team Statistics ……………………………………. **8**

Table 3: Stage 1 Significant Predictor Summary By Model Type ……………………… **46**

Table 4: Greatest Predicted W-L%, 2022 Season, Optimal W-L% Model ……………... **69**

Table 5: Lowest Predicted W-L%, 2022 Season, Optimal W-L% Model ……………… **70**

Table 6: Highest W-L% Residual, Optimal W-L% Model ……………………………... **71**

Table 7: Lowest W-L% Residual, Optimal W-L% Model ……………………………... **72**

Table 8: Highest % of Games Correctly Predicted, 2022 Season, Optimal …………………   
Game Prediction Model …………………………………………………………...……. **73**

Table 9: Lowest % of Games Correctly Predicted, 2022 Season, Optimal Game …………  
Prediction Model ……………………………………………………………………….. **74**

Table 10: Highest Residual of Predicted Total Wins, 2022 Season, Optimal Game ……… Prediction Model ……………………………………………………………………...... **75**

Table 11: Lowest Residual of Predicted Total Wins, 2022 Season, Optimal Game ………... Prediction Model …………………………………...…………………………………... **76**

Table 12: MTSU Men’s Basketball, 2022 Season, Predicted W/L Totals By Location… **78**

Table 16: MTSU Men’s Basketball, 2022 Season, Predicted W/L Totals ……………… **79**

Table 17: Incorrectly Predicted Games, MTSU Men’s Basketball, 2022 Season ………. **80**

Table 18: Post-Game Analysis Output, MTSU Men’s Basketball, 2022 Season ……….. **82**

Table 19: Incorrectly Classified Games, 2022 Season, Post-Game Classification …………  
Algorithm ………………………………………………………………………………. **84**

Page

Figure 1: Sample “Basic” Box Score for a Basketball Game ……………………………. **3**

Figure 2: Sample “Advanced” Box Score for a Basketball Game ……………………….. **4**

Figure 3: Sample of BasStat Dataset ……………………………………………………... **9**

Figure 4: Sample of AdvStat Dataset …………………………………………………...... **9**

Figure 5: Trade-off Between Variance Decrease and Bias Increase, LASSO Regressor . **17**

Figure 6: Sample Random Forest Model Diagram, Regressor …………………………. **18**

Figure 7: Sample AdaBoost Regressor Diagram …………………………………........... **19**

Figure 8: Sample Logistic Regression Equation Graph ………………………………… **20**

Figure 9: Sample Random Forest Classifier Diagram ………………………………....... **21**

Figure 10: AdaBoost Regressor Decision Boundary Example ………………………….. **23**

Figure 11: Example of Classification Model Output Bins ………………………............ **25**

Figure 12: Example of Receiver Operator Characteristic Curve (ROC) and Area ………….  
Under the Curve (AUC) …….......……………………………..………………………... **26**

Figure 13: LASSO Regression Feature Importance, BasStats ………………………….. **35**

Figure 14: LASSO Regression Feature Importance, AdvStats …………………………. **36**

Figure 15: Root Mean Square Error Comparison, LASSO Regressors …………...…….. **37**

Figure 16: Random Forest Regressor Feature Importance, BasStats ………………....... **38**

Figure 17: Random Forest Regressor Feature Importance, AdvStats …………………... **39**

Figure 18: Root Mean Square Error Comparison, Random Forest Regressor ………….. **40**

Figure 19: AdaBoost Regressor Feature Importance, BasStats ………………………..... **42**

Figure 20: AdaBoost Regressor Feature Importance, AdvStats ……………………........ **43**

Figure 21: Root Mean Square Error Comparison, AdaBoost Regressor ……………....... **45**

Figure 22: Logistic Regression Accuracy by Lag Depth ……………………………….. **49**

Figure 23: Random Forest Accuracy by Lag Depth …………………………………….. **50**

Figure 24: AdaBoost Accuracy by Lag Depth ………………………………………….. **51**

Figure 25: Logistic Regression AUC Analysis, Optimal Lag Depth ………………….... **52**

Page

Figure 26: Logistic Regression Test Accuracy, Full vs. Reduced Predictor Sets ……..... **54**

Figure 27: Random Forest Test Accuracy, Full vs. Reduced Predictor Sets …………..... **55**

Figure 28: AdaBoost Test Accuracy, Full vs. Reduced Predictor Sets …………………. **56**

Figure 29: Logistic Regression AUC Analysis, Optimal Lag Depth ………………….... **57**

Figure 30: Optimal Predictor Set, Logistic Regression Approach …………………........ **58**

Figure 31: Optimal Predictor Set, Random Forest Approach ………………………….... **59**

Figure 32: Optimal Predictor Set, AdaBoost Approach ……………………………….... **60**

Figure 33: AUC Analysis, Optimal Machine Learning Algorithm Decision ………….... **61**

Figure 34: Confusion Matrix, Optimized Game Prediction Model ……………………... **63**

Figure 35: AUC Analysis, Optimized Game Prediction Model ……………………….... **64**

Figure 36: Confusion Matrix, Optimized Game Recap Model …………………………. **65**

Figure 37: AUC Analysis, Optimized Game Recap Model …………………………….. **66**

Page

Figure A-1: Sample Basic Team Stats ……………………………………………........... **96**

Figure A-2: Sample Basic Opponent Stats …………………………………………........ **96**

Figure A-3: Sample Advanced Team Stats ………………………………………..……. **97**

Figure A-4: Sample Advanced Opponent Stats ………………………………...…...…... **97**

Figure A-5: Sample Basic Game Log Data ………………………………………...….... **97**

Figure A-6: Sample Advanced Game Log Data ……………………………………........ **98**

Figure A-7: Sample NCAA Team Sheet (University of Virginia, 2019 – Partial) …........ **98**

LIST OF APPENDICES

Page

APPENDIX A: Additional Figures ……….…………………………………………….. **96**

**Introduction**

Ever since Dr. James Naismith hung up two peach baskets at his local YMCA in 1891, many coaches, players, statisticians, and the like have all debated about what the “right” way to play basketball is (Lunardi, 2022). The sport that Naismith invented over 130 years ago is a far cry from the sport that modern-day basketball players like LeBron James play today.

The first basketball game ever contested saw 18 players on the court at one time and finished with the awe-inspiring final score of 1-0. Today, it is not uncommon to see NBA games won by a team scoring more than 100 points. Even the losing team in most NBA contests in the past decade have scored over 100-points.

The sport of organized basketball is now played across the world by kids as young as four years old. In the United States, there are well-defined age groups that play basketball, ranging from youth to professional. This paper will primarily focus on the men’s college basketball played as a part of Division 1 of the National Collegiate Athletic Association (NCAA).

The game has changed drastically since its inception. Many coaches have introduced their own take on analytics throughout the years. John Wooden brought his UCLA dynasty to the forefront of college basketball based on implementing practice plans tailored for his team based on opponent’s statistics (Wooden & Jamison, 1997). Syracuse’s Jim Boeheim took these defensive schemes a step further by pioneering the zone defense in the collegiate game. Gregg Popovich of the NBA’s San Antonio Spurs revolutionized the value of corner three-pointers in his coaching system (Staffo & Bing, 2018). However, for the longest time, an overwhelming majority of coaches have believed that there is something missing from the traditional statistical view of basketball. Many coaches hone in on the individual statistics that a player can output, but this is not a perfect system. A basketball team is a perfect example of a whole that is greater than the sum of its parts. If basketball is reduced down to simply scoring baskets, it would be a disservice to the complexity of the game. There are 10 players on the court, but only one ball. John Calipari, formerly of the University of Memphis, the University of Massachusetts, and currently the Head Coach of the University of Kentucky, sums it up by stating that “[I] love when a player can show me how he’s helping us win the game when [he’s] not scoring baskets” (Calipari, 2015).

These coaches were operating in a basketball era where traditional “box-score” statistics were invented. This era did not apply the concept of efficiency outside of the context of shooting efficiency. Field Goal Percentage (FG%), 3-Point Percentage (3P%), and Free Throw Percentage (FT%) are the only basic box-score statistics that account for a team’s (or player’s) efficiency on the basketball court. These original statistics that have been recorded for over 50 years are often referred to as “basic” statistics in the game of basketball.

**Figure 1: Sample “Basic” Box Score for a Basketball Game**

Table

Description automatically generated

In a more modern sense, the game of basketball has become revolutionized thanks to an increased interest from the statistics community. Concepts such as performance variance, deep-level efficiencies, home-court advantage, the “hot hand”, and many more are now regularly factored into modern statistical models in the game of basketball. The concept of efficiency is now being applied to all facets of the game, not just shooting efficiency (Zak et al., 1979).

These metrics are so beneficial to teams, many modern-day coaches utilize these metrics to formulate practice plans, on-court tactics, and make in-game decisions (Shea & Baker, 2013). These new-age metrics are referred to by the basketball community as “advanced” metrics.

**Figure 2: Sample “Advanced” Box Score for a Basketball Game**

Table

Description automatically generated

Many coaches praise the “Four-Factors” that statistician Dean Oliver coined in the early 2000’s. A more modern view on the game of basketball is understanding that a team that can score, protect, crash, and attack on the basketball court will win more basketball games than a team simply focused on scoring more points and stopping their opponent from scoring (Oliver, 2011).

Over the second half of the century that has passed since Dr. Naismith first hung that peach basket on the wall in Springfield, MA, a new concept has been born in the world of statistics and computer science called Machine Learning (ML). ML has evolved the ways in which people are able to solve complex tasks such as speech recognition and object detection among others. ML has its applications across millions of disciplines. The sports industry has been primed with an influx of ML algorithms and the real-world advantages they can give to those who leverage them.

The game of college basketball is no stranger to the expansion of ML algorithms from the tactical side of the game to the recruiting cycle. One of the most popular uses for ML in sports is to determine the outcome of a given contest. In other words, “Who is going to win?”. This question can be asked of any game at any level, but the data is not readily available and interpretable at every level. College basketball is the first level of basketball where the kind of data that is needed to make these ML algorithms is readily available for all teams across a given league.

Furthermore, the evolution of ML into mainstream practice within the sport of college basketball has some potentially fatal flaws. Many coaches and staff aren’t properly trained in the use of ML algorithms. This can lead to results being misconstrued, and even entire models being built on biased interpretations of basketball concepts. Coaches and their staff are privy to dozens of “advanced” datapoints that seem to deliver easier ways to evaluate player performance. The issue comes into play when those metrics are misconstrued as providing windows into the answer to that “Who is going to win?” question that is so often asked.

**Statement of Purpose**

The purpose of this research is to quantitatively determine whether advanced statistics are more apt for predicting the outcome of a given collegiate basketball game at the Division 1 level as compared to traditional “basic” game-level data. This will be accomplished by developing and testing a series of machine learning algorithms and comparing their resulting accuracy and bias to assess the true predictive power of each set of statistics. These algorithms will be classification-based, and their dependent variable will be a binary classification of whether a team won a given game. This research will also preface these machine learning algorithms with a set of variable selection algorithms to determine which features from the available datasets are deemed to be significant predictors of winning in college basketball. These variable selection algorithms will have the dependent variable of a team’s W-L% from a particular season.

**Research Questions**

The main research questions of this paper are:

* Is there an ideal lag term to consider when factoring in a team’s past performances in predicting future outcomes?
* Is there an ideal subset of predictors that can cross over from season-long W-L% calculations to individual game simulation?
* Is it better to utilize basic box score output statistics or advanced efficiency metrics when predicting the outcome of a team’s future contests?
* Is there an optimal machine learning algorithm approach to predicting the outcome of future Division 1 Men’s Basketball games?

**Definition of Terms**

**Table 1: Data Dictionary, Basic Team Statistics**

|  |  |  |
| --- | --- | --- |
| Variable | Description | Formula (If Applicable) |
| W-L% | Win Loss Percentage; A measure of the percentage of games a team wins | Wins / (Wins + Losses) |
| Points Tm | Points per Game; A measure of how many points a team scores per game |  |
| FGM | Field Goal Makes; A measure of how many Field Goals a team makes per game |  |
| FGA | Field Goal Attempts; A measure of how many Field Goals a team attempts per game |  |
| FG% | Field Goal Percentage; A measure of what percentage of a team’s Field Goals are made | FGM / FGA |
| 3PM | Three Point Makes; A measure of how many Three Pointers a team makes per game |  |
| 3PA | Three Point Attempts; A measure of how many Three Point Attempts a team takes per game |  |
| 3P% | Three Point Percentage; A measure of what percentage of a team’s Three Pointers are made | 3PM / 3PA |
| FTM | Free Throw Makes; A measure of how many Free Throws a team makes per game |  |
| FTA | Free Throw Attempts; A measure of how many attempted Free Throws a team gets per game |  |
| FT% | Free Throw Percentage; A measure of what percentage of a team’s Free Throws are made | FTM / FTA |
| ORB | Offensive Rebounds; A measure of how many offensive rebounds a team grabs in a game |  |
| TRB | Total Rebounds; A measure of how many total rebounds a team grabs in a game |  |
| AST | Assists; A measure of how many assists a team earns in a game |  |
| STL | Steals; A measure of how many steals a team gets in a game |  |
| BLK | Blocks; A measure of how many opponent's shots are blocked in a game |  |
| TOV | Turnovers; A measure of how many turnovers a team gives up during a game |  |
| PF | Personal Fouls; A measure of how many fouls a team commits in a game |  |

Sources: <https://www.sports-reference.com/cbb/>, (Oliver, 2011)

**Table 2: Data Dictionary, Advanced Team Statistics**

|  |  |  |
| --- | --- | --- |
| Variable | Description | Formula (If Applicable) |
| W-L% | Win Loss Percentage; A measure of the percentage of games a team wins | *Wins / (Wins + Losses)* |
| Possession | A game action that starts when a player gets the ball, and eds only in one of four ways (FGM, FTM, DREBOpp, TOV) | *FGA + 0.44\*FTA – OREB + TOV* |
| Pace | A measure of a team's total possessions in a single game | *FGA + 0.44\*FTA – 1.07 (OR / OR+DREBopp) \* (FGA-FGM) + TOV* |
| Ortg | Offensive Rating; A measure of points scored per 100 possessions | *(Total Pts Scored / Possessions) \* 100* |
| FTr | Free Throw Rate; A measure of how often a team is getting Free Throw Attempts | *FTA / FGA* |
| 3PAr | Three Point Attempt Rate; A measure of what percentage of a team's total FGA are 3PA | *3PA / FGA* |
| TS% | True Shooting Percentage; A measure of shooting efficiency that adjusts for differing attempt rates | *PTS / (2 \* TSA); TSA = FGA + 0.44 \* FTA* |
| TRB% | Rebounding Percentage; A measure of what percentage of total available rebounds a team grabs | *Rebounds / [(FGA - FGM) + (FGAOpp - FGMOpp)]* |
| AST% | Assist Percentage; A measure of what percentage of FGM that are Assisted | *AST / FGM* |
| STL% | Steal Percentage; A measure of what percentage of opponent possessions end in a steal | *STL / Possessions* |
| BLK% | Block Percentage; A measure of what percentage of opponent FGA are blocked | *BLK / FGAOpp* |
| eFG% | Effective Field Goal Percentage; A measure of Field Goal efficiency that adjusts for shot distance | *(FGM + 0.5 \* 3PM) / FGA* |
| TOV% | Turnover Percentage; A measure of what percentage of offensive possessions end in a turnover | *Turnovers / Possessions* |
| ORB% | Offensive Rebounding Percentage; A measure of what percentage of available offensive rebounds a team is able to secure | *Off. Rebounds / (FGA - FGM)* |
| FT/FGA | Free Throw Efficiency; A measure of how efficiently a team is getting to the line | *FTM / FGA* |
| DRtg | Defensive Rating; A measure of points allowed per 100 possessions | *(Total Pts Allowed / Possessions) \* 100* |

Sources: <https://www.sports-reference.com/cbb/>, (Oliver, 2011)

Basic Team Statistics (BasStat): Refers to the dataset comprised of values in Table 1. A sample of the BasStat dataset is provided in Figure 3.

Advanced Team Statistics (AdvStat): Refers to the dataset comprised of values in Table 2. A sample of the AdvStat dataset is provided in Figure 4.

**Figure 3: Sample of BasStat Dataset**

**Calendar

Description automatically generated**

**Figure 4: Sample of AdvStat Dataset**

**Table

Description automatically generated**

**Significance of Study**

The purpose of this study is to better understand the efficacy of advanced statistics when it comes to predicting the outcome of a Division 1 collegiate basketball game. Since the prevalence of advanced basketball analytics has grown ten-fold since the turn of the 21st century (Bechtold, 2021), it is necessary to understand which, if any, of these statistical measures of efficiency are significant predictors of winning the game, or rather better utilized with a great deal of skepticism and potentially nothing more than metrics for evaluating individual players. This study has the potential to set in stone the legitimacy or illegitimacy of the “four-factors” as metrics for evaluating team performance at the collegiate level.

**Review of Literature**

**The History of Statistics in Basketball**

When James Naismith first hung a peach basket on a wall in 1891, the sport of basketball played on that court is unrecognizable to the modern eye. Shortly after the invention of competitive basketball, statistics like points, rebounds, assists, and steals were tabulated into a readable format called a “box score” (TeamMom, 2017). Many analysts, including Joe Lunardi (the father of NCAA Bracketology) note that production statistics like those found in a box score are not the most indicative of team success (Lunardi, 2022). This problem with box score statistics was known to coaches and analysts in the mid-20th century, leading to hall of fame coach Dean Smith’s introduction of a possession-based statistics system in the 1950s (Smith et al., 2002). The invention of a possession as a base unit of play in basketball, like an “at bat” in baseball, allowed for a new set of statistics to be recorded. In the late 1970s, researchers found that using efficiency statistics like field goal percentage (FG%) and free throw percentage (FT%) were much better indicators of a player’s value to his team than traditional “counting” statistics like total field goals made (FGM) or free throws made (FTM) (Zak et. Al, 1979).

The modern era of basketball statistics, sometimes referred to as the “basketball analytics revolution”, had its beginnings in the early 21st century (Bechtold & Streetwood, 2021). During those years, statistician Dean Oliver published a plethora of research concluding that there are four main factors (the Four Factors) that explain how a team wins a basketball game (Oliver, 2011). Those Four Factors include scoring, crashing, protecting, and attacking. Basing his work off of baseball guru Bill James (his work popularized in the film *Moneyball*), Ken Pomeroy introduced a series of tempo-based metrics in the early 2000’s that went on to change the way in which Las Vegas predicted the moneyline and O/U of games moving forward (Bechtold & Streetwood, 2021).

Further research concluded that these principles that Oliver defined, and the metrics that can be derived from them explain significantly more variability in the outcome of an NBA basketball game than the more “basic” statistics available via the box score (Shea, 2014). The analytics revolution evolved one step further in the 2010s as spatial analytics are popularized to add value to analytics-based decision making that is often omitted when simply looking at production and efficiency statistics (Shea & Baker, 2013). These spatial analysis methods also allow for the contextualization of many previously used statistics such as the innovation of “clutch” and “garbage” time.

**How Statistics Are Leveraged By Coaches**

The sport of basketball, specifically at the college level, is uniquely different than other levels. This is because of the nature of the players coming in from high school, many of which are looking to turn professional after one or more seasons. For this reason, significant investments are made to not only maximize a player’s on-court contribution to his current team, but also teach him both the schematics of the game and important life lessons for him to succeed at that next level (Calipari, 2015). Coaches will sometimes pioneer their own evaluation systems in order to place the highest value on players that fit their system (Staffo & Bing, 2018).

Some coaches even suggest the minimization of statistics in the success of a program, going as far to say that it’s solely the fight in the players and the staff that deliver the product on the court (Krzyzewski & Hill, 2001). However, despite the common rhetoric from coaches that it’s all about the players and not about the scheme, that is simply untrue. In creating and sustaining his dynasty at UCLA from 1964-1975, in which his Bruins won 10 out of 12 national championships, coach John Wooden credited his team’s success on the court to his “pyramid of success”. This pyramid consists of 15 key components to any championship player/team. Organized as a pyramid, the concept suggests certain competencies are foundational in nature, and must be mastered before more abstract ideals can be truly realized. At the core of this pyramid is the “skill” concept, in which Wooden claims his players must master every minute detail when it comes to his offensive and defensive scheme, and only when they have mastered that can they perform at a championship level (Wooden & Jamison, 1997). This insistence that the scheme that Wooden implemented during his time as the head coach of UCLA was core to the program’s success refutes any idea that coaches don’t pay attention to statistics. In fact, it is commonly accepted that coaches, players, and front office staff not only see the potential in leveraging the most cutting-edge analytics, but they also often seek out analytics as a mode of competitive advantage over other teams (Demenius & Kreivyte, 2017).

**The Difference Between Division 1 Collegiate Basketball and the NBA**

There are many slight differences between the basketball played in the NBA and at the Division 1 level in college. Outside of the obvious differences in rules and length of play (e.g., four 12-minute quarters in the NBA vs. two 20-minute halves in college), there are several factors that impact analysts’ abilities to predict the outcome of each game at the different competition levels. In the NBA, there is a built-in incentive system to lose games. This phenomenon was brought to light in the 2010s by the Philadelphia 76ers and is often referred to as “tanking” (Weitzman, 2020). Tanking for higher draft picks means that the fundamental assumption of perfect competition, that each team is playing in a way that they believe maximizes their winning chances, is violated quite often at the NBA level (Beck, 2021). The lack of a draft system in college basketball eliminates the incentive to tank for higher draft stock, and thus the purity of competition is still sacred at the collegiate level.

Other analysts have noted that the accessibility of cutting-edge data is often a limiting factor of improvements for lower-tier programs (Mudric, 2020). This discrepancy exists at the NBA level between organizations that are willing to spend more money on analytics departments than others but is much more pronounced at the collegiate level when programs such as Duke and Kentucky can have operating budgets over $20 million per year, where other programs such as Western Carolina only have operating budgets of ~$1 million per annum (Root, 2020).

Another key difference in assessing the relative strength between teams at the NBA level and at the Division 1 level is the size of the leagues. In the NBA, each team plays every other team, making comparing records much easier. For this reason, it makes sense that the NBA uses W-L% as the primary ranking factor when determining playoff positions (Lunardi, 2022). However, in division 1 college basketball, there are over 350 teams, meaning that each team cannot play every other team. This makes direct comparisons much harder. Many systems have been invented to try and adjust performance based on strength of opponent such as the infamous KenPom rating system that adjusts offensive and defensive performance for both strength of opponent and pace (Pomeroy, 2006). These types of metrics have led to the official decision structure for college basketball, the NCAA Selection Committee, utilizing a system known as “team sheets” to determine a team’s post-season fate. An example of these team sheets, and the metrics that they provide can be found in Figure A-7 in the appendix.

**Modern Factors Impacting the Division 1 Landscape**

The game of college basketball is ever changing. Several fundamental factors surrounding the sport have changed in recent years, and these factors could have a foreseeable impact on the continued validity of any past research. The continued expansion of legalized sports betting has led to the increased investment in private-sector analytics to model the outcome of collegiate basketball games more accurately (White, 2022). However, as of this point it is inconclusive as to whether the models used to predict games pre-sports gambling era have seen significant changes in their accuracy due to the legalization of sports gambling.

A similar paradigm shift is occurring with regards to the players themselves. In the summer of 2021, the NCAA lost its appeal on a Supreme Court case (*Alston v. NCAA*), which opened the doors for players to directly profit off their own name, image, and likeness (NIL) (Osterman, 2022). This decision changes the definition of amateurism that the NCAA held sacred for so many years, effectively leveling the playing field for collegiate athletes when it comes to making money outside of their NCAA-sanctioned athletics scholarships (Dodd, 2021). The most direct impact that NIL has had on college basketball up until this point is allowing many players to return to college for an additional season(s) because it is now a financially lucrative decision (Dellenger, 2022).

**Machine Learning Algorithms Used for Regression**

*LASSO Regression*

The Least Squares regression model is the most basic quantitative predictive model in machine learning. The model takes in a set of parameters and yields coefficient estimates that when combined with the intercept and input parameter instance values yield the predicted independent variable value. However, this basic machine learning model has its flaws, and can be adjusted in slight ways to improve its deployability in most circumstances.

A picture containing text, watch, clock

Description automatically generated

(1)

The above figure is a sample LASSO loss function error where the coefficient estimates are scaled by lambda (lambda). A LASSO model is a particularly good substitute for large input sets where variable selection through traditional least squares linear regression would be tedious and unreasonable (Hans, 2009).

**Figure 5: Trade-off Between Variance Decrease and Bias Increase, LASSO Regressor**

Chart, line chart

Description automatically generated

The lambda parameter for the LASSO regression allows the model to scale the coefficients of variables to decrease the variance. As this lambda coefficient increases, the variance of the model decreases. However, the bias of the model simultaneously increases. Therefore, to optimize a LASSO regression algorithm, the value of lambda for which the minimal Mean Squared Error (MSE) is realized is the optimal value for lambda for that LASSO model (Zhao & Yu, 2006).

*Random Forest Regressor*

The decision tree model represents the most simplistic form of decision-based machine learning algorithm (Rokach & Maimon, 2005). By assembling an ensemble of pruned decision trees, each making their best estimate of the quantitative independent variable, the accuracy of said prediction increases dramatically (Biau et al., 2018). This concept is referred to as a random forest. Given a set of inputs, the random forest runs that input set against its n-regressors and averages all quantitative estimates of the independent decision tree regressors to form a finalized prediction. These independent estimates are sometimes weighted by the inverse of the MSE for each member of the ensemble, thus making the most accurate trees have a greater weight in the overall end estimation (Tuv et al., 2009).

**Figure 6: Sample Random Forest Model Diagram, Regressor**

Diagram

Description automatically generated

The diagram above details the process flow of a random forest regressor from the input set to the final prediction for that input set. The nature of a random forest regressor is such that each tree is generated not by analyzing which factor and cutoff value reduces error by the greatest amount, but rather a truly randomized selection of variables. This approach seems counterintuitive on the surface, but at the aggregate, the prediction error decreases dramatically due to the ensemble’s ability to leverage value out of predictors that don’t decrease error when used as lone criteria (Zhang et al., 2006).

*AdaBoost Regressor*

The concept of a random forest ensemble regressor is taken to its extrema via the AdaBoost model. This method of boosting leverages an ensemble of n-decision tree “stumps” (trees with a depth of 1) to serve as the predictors of the independent variable (Esposito, 1997).

**Figure 7: Sample AdaBoost Regressor Diagram**

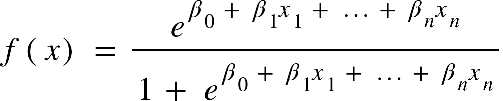
*Diagram

Description automatically generated*

The figure above shows the process flow of the AdaBoost regressor ensemble, which is very similar to that of a random forest. Although the process and structure are similar, it has been found that AdaBoost models can reduce prediction error compared to random forests trained on similar datasets, but also have to deal with the problem of overfitting of data at a much higher rate (Chengsheng et al., 2017).

**Machine Learning Algorithms Used for Classification**

*Logistic Regression*

The simplest machine learning algorithm to classify objects into classes is the logistic regression model.

(3)

Given above, the logistic regression equation takes predictors x subscript 1 comma space... comma space x subscript n as inputs, and yields a series of coefficients beta subscript 1 comma space... comma space beta subscript n as outputs (Lemeshow, 2008). These coefficients plug into the model above to yield a logistic curve seen below.

**Figure 8: Graph of Logistic Function**

Shape

Description automatically generated with medium confidence

**Probability of Positive Class**

One advantage of a logistic regression model for classifying discrete events is that it limits the bias present due to overfitting and overparameterization of some of the more complex ensemble algorithms (Rocks & Mehta, 2020). Furthermore, logistic regression models take up much less space computationally and is much leaner compared to their peer classification models (Dreiseitl & Ohno-Machado, 2002).

*Random Forest Classifier*

The decision tree is arguably one of the most human readable and understandable classification models. With it’s “if-then” structure, it applies Bayesian logic to dissecting a group of records down until they are homogenous lead nodes (Rokach & Maimon, 2005). However, there is a major drawback to these models as they often introduce bias in the form of overfitting. The random forest model alleviates that bias and increases classification accuracy by making an ensemble of pruned predictors that are individually less accurate than an individual optimized decision tree, but at the aggregate level provide much more insight and accurate classifications than a single decision tree (Biau et al., 2018).

**Figure 9: Sample Random Forest Classifier Diagram**

Chart, radar chart

Description automatically generated

The diagram for a random forest classifier demonstrates the decision path that each tree in the ensemble takes to come to their independent classification “votes”. Those “votes” are then weighted by the individual tree’s accuracy and summed up to see the majority class for each instance, thus providing the final classification prediction for each instance fed into a random forest ensemble.

*AdaBoost Classifier*

The AdaBoost classifier builds on the methodology of the random forest by taking the ensemble idea to its extreme. Since a Random Forest derives more test-set prediction accuracy by making individual trees that are less accurate than the optimal tree, the AdaBoost algorithm takes that ensemble method to its extreme by making the individual members of the ensemble decision trees with a node-depth of 1 (Esposito, 1997). These “weak” learners (aka “stumps”) are fed the instance and the predicted classes are summed on the same voting system as the random forest. AdaBoost models have been proven to provide better classification rates than random forest models, but often have to deal with overfitting bias more than the random forest algorithm does (Chengsheng et al., 2017).

**Figure 10: AdaBoost Classifier Decision Boundary Example**

*Diagram

Description automatically generated*

The figure above shows the AdaBoost ensemble’s ability to establish non-linear classification boundaries through an aggregation of weak linear classifiers.

**Data and Output Validation in Machine Learning Algorithms**

When considering the model inputs for a machine learning algorithm, the problem of covariance often emerges as a primary blocker for deploying an effective and interpretable model. Covariance is when two or more input variables are significantly correlated with each other, thus taking away from the effect that each of them are supposed to have on the explained variance of the independent variable (Chen et al., 2021).

Another consideration that must be made when dealing with regression-based algorithms specifically is the concept of heteroskedasticity. This is the idea that the prediction errors should be consistent and randomly scattered around the N-Dimensional line of best fit. If this condition is not met and the residuals are not homoscedastic (i.e. they are heteroskedastic), then the model is exhibiting some implicit bias in its predictions in the sense that an input variable can tell you how “accurate” the prediction of that model will be (Long & Ervin, 2000). Both covariance and heteroskedasticity are serious issues with data integrity and must be at the forefront of data preprocessing during the construction of all models.

**Techniques To Optimize Model Performance**

Every model discussed so far has so-called hyperparameters that are separate from the coefficients specifically assigned to the input factors. These hyperparameters are function level coefficients that adjust for things like ensemble size, learning rate, and loss-bias. By optimizing these hyperparameters, higher rates of classification accuracy, and lower rates of prediction error can be realized when the models are evaluated on the test set (Yu & Zhu, 2020). By leveraging a grid search technique the optimal combination of n-hyperparameters can be found for each modeling solution. This process can be very time consuming and resource intensive, which will drive up the runtime of any model creation script (Jebara et al., 2004) The optimization of these hyperparameters also prevents the overfitting of each model to the training set, and instead optimizes each model to be deployed on the test set and future prediction sets not included in the training data for the algorithm (Cawley & Talbot, 2007).

The hyperparameter conversation intensifies when dealing with ensemble classifiers and regressors such as Random Forest and AdaBoost models. This is because the ensemble size itself is a hyperparameter. The tuning of this specific hyperparameter, as well as the pruning of decision trees to constrain overfitting of the ensemble members to the training dataset will lead to a direct increase in prediction accuracy once the algorithm is deployed on the test set (Niculescu et al., 2006). Furthermore, boosting models need to have their base predictor set as its own hyperparameter. While it is typically assumed that the depth of this base “weak” learner is supposed to be 1, that is not a fixed constant, and must also be fit to provide the optimal output (Bennet & Parrado-Hernandez, 2006).

One drawback of reinforcement models is that their hyperparameters are considered fixed throughout deployment. This means that if future data is needed to retrain the model, the entire time-consuming process of grid search optimization of the hyperparameters must be repeated for each subsequent retraining and redeployment of the model (Singh, 1992).

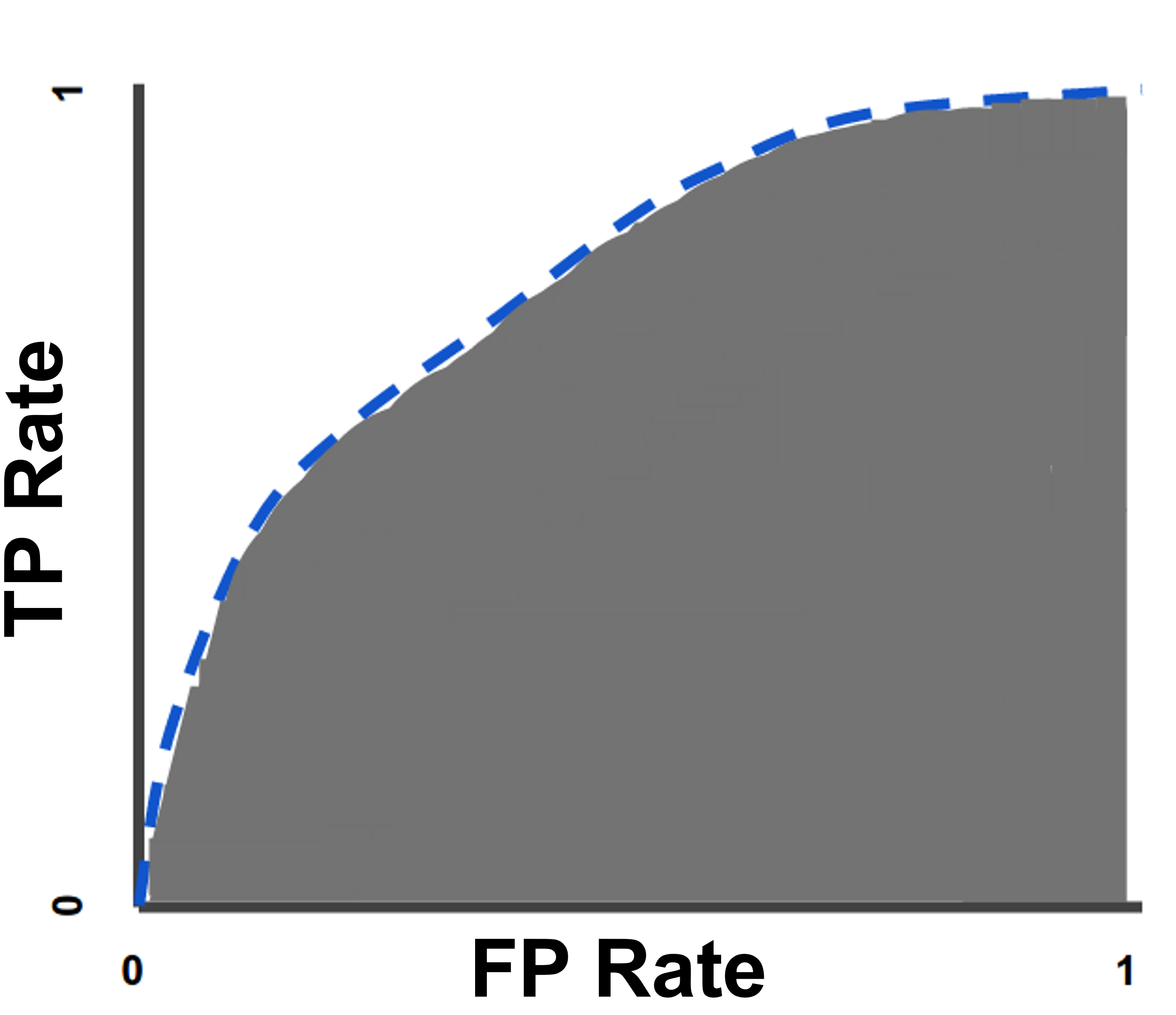
A picture containing diagram

Description automatically generated**Figure 11: Example of Classification Model Output Bins**

*Source: (Google, 2019)*

Classification models must be optimized in ways that are slightly more specific than regression models. The classification model’s predictions can be split into four types: True Positives, True Negatives, False Positives, and False Negatives. The goal of any classification model is to maximize the proportion of predictions that fall into the “true” categories, meaning that the model correctly predicted their actual state. A classification model achieves higher true positive and true negative rates by going through standard process of hyperparameter optimization (Google, 2019).

**Figure 12:** **Example of Receiver Operator Characteristic Curve (ROC) and Area Under the Curve (AUC)**



The above figure outlines a hypothetical Receiver Operating Characteristic (ROC) Curve for a classification model. The blue-dashed line shows the trade-off in true positive rate and false positive rate as the decision threshold varies from 0 to 1. In a typical classification model, the threshold as to whether a record has a predicted state of 0 or 1 is whether the probability of the positive (1) outcome is greater than 0.5 (Malato, 2022). By varying the threshold, it can be determined which model correctly balances the trade-off between correctly identifying all true positives and avoiding the misclassification of the true negatives. The method most used for this is called AUC analysis, in which the area under the ROC curve is used in a pseudo-integral form to compare classification models on a level deeper than simply evaluating the test-set accuracy (Google, 2019).

**Methodology**

The current study attempts to observe the difference in efficacy between two competing sets of predictors in classifying the result of a Division 1 College Basketball game. This study contains data as outlined in Table 1 and Table 2.

The study proposes multiple machine learning algorithms to be fit and tested on each predictor set to determine which predictors are significant, and which overall set of predictors are more accurate in their classification abilities. This research set out under the working hypothesis that the advanced efficiency metrics would be better overall predictors of both season-long W-L% and individual game prediction probabilities.

For variable selection, two LASSO Regression models, one for each predictor set, were used to scale the coefficients associated with each predictor in each set. In addition, two random forest regressor models were fit to the data to determine their own set of significant predictors. Furthermore, an AdaBoost model was fit to each dataset to form a third opinion on the significant predictors from each dataset. These sets of significant predictors were cross-referenced and used in stage two of the research.

For game prediction, two Logistic Regression models, one for each reduced predictor set, were used to classify each game as either a Win (1) or a Loss (0). In addition to these two models, two Random Forest Classifier models were fit to the data, along with two AdaBoost classification models to provide three total predicted classifications of each game.

**Measures**

There were two distinct stages for this research. The primary stage was the variable selection stage and the secondary stage was the game prediction stage. These two stages have different dependent variables, predictor sets, and units of study.

Each stage used separate units of study. The variable selection stage used the team as the unit of study. The measures in this stage were season-long averages for the metrics laid out in Table 1 and Table 2. Each NCAA Division 1 program has a record in the dataset for each year that they competed at the Division 1 level.

The two stages of research used separate dependent variables for their respective machine learning algorithms. The dependent variable used in the variable selection stage of research was a team’s W-L%. This was selected as the dependent variable because the primary goal of this stage was to determine which predictor variables in each dataset were significant contributors to a team winning more games. With this goal in mind, it stands to reason that the variables that can significantly predict a team’s W-L%, would be variables of interest for the second stage. The second stage used the result of each game as a dependent variable. This measure is a classifier with values of either W or L and was re-coded as a binary classifier with values of either 1 or 0.

**Procedures**

The data were retrieved from their various sources and compiled into merged datasets via Python. Python scripting was used to extract, clean, merge, and analyze the data for each stage of this research.[[1]](#footnote-1) After the full datasets were assembled, a series of machine learning algorithms were trained on data from the 2010-2011 season through the 2020-2021 season. The models were then compared and evaluated on their error rate(s) to answer the key research questions set forth at the start of the project. The questions were answered using scores obtained via data from the 2021-2022 season. This separation of training and testing data into two separate datasets ensures that overfitting bias is limited.

To compare the result of the final model to a similar model that is currently deployed in the industry, data were gathered from Caesar’s Sportsbook on every money line offer they booked for the 2019-2020 NCAA Basketball Season. The data was then processed in combination with the actual results of each game to determine whether the team that was implied to win by Caesar’s Sportsbook’s odds won that particular game. This process was aggregated across all games for which data was available (sportsbooks often don’t book extremely lopsided games for profit reasons).

The strengths of this research are two-fold. First, the completeness of the datasets from both a team-level and game-level perspective allow for the full training and testing of a series of complex machine learning algorithms. This research would not be possible in its current format if the data were not so complete and accurate. The second main strength of this research are its adherence to the highest levels of proper procedure and evaluation. On top of simply training and testing these series of algorithms, this research has ensured to control for covariance and heteroskedasticity, as well as propose two research-level parameters to test while evaluating these models: The number of games to consider for rolling average calculations, and whether reducing the dataset by heeding the results from stage one is appropriate.

Instead of making assumptions about whether it is appropriate to use a reduced dataset, this research empirically tests it by a comparison of the accuracies of identically trained models, one on the full dataset, the other on the reduced. This research also utilizes the optimal hyperparameters for each individual model. These hyperparameters were obtained through extensive K-Fold Cross Validation, as well as implementing a well-structured Grid Search to determine which exact combination of hyperparameters yields the highest level of accuracy for each model.

The weaknesses of this research include the fact that the data only goes back to the 2010-2011 NCAA season. This restriction on the dataset is in place due to the fact that gamelog data are not readily available for seasons prior to the 2010-2011 season. This is a key weakness because the dataset used in the first stage of the research has the capability to extend back to the 1999-2000 season. The size of the data is adequately large enough to draw statistical conclusions on the efficacy of each ML algorithm, however, since more data is available for one set than another, predictive power was omitted from this research on the grounds of not having full data for each model until the 2010-2011 season.

**Data Analysis**

After constructing the models on data from the 2010-2011 season through the 2020-2021 season, each model was tested for its efficacy on data from the 2021-2022 NCAA Division 1 season. For the variable selection stage, feature importance was used as the selection criteria to evaluate each individual predictor’s ability to predict the dependent variable. To evaluate the models for the variable selection stage, Root Mean Squared Error (RMSE) was used to define which model was best at predicting W-L%. For the game prediction stage, two different evaluation methods were utilized to compare the models. First was the model’s overall classification accuracy (i.e., what percentage of games did the model correctly predict as either a win or a loss). The second evaluation method that was used to compare the models was the Receiver Operator Characteristic curve (ROC). More specifically, the Area Under the ROC curve (AUC) was used to balance both the overall accuracy of the models with their recall ability when determining which model was the best at classifying the results of individual games.

Each research question was answered independently, and the conclusions drawn from each individual comparison were synthesized into one ‘optimal’ model at the end of the research. This optimal model was then used to demonstrate several use-cases that the model derived from this research can have to different user bases including coaches, fans, and analysts.

**Results**

This research was conducted in two primary stages, the first of which consisted of the development and evaluation of a series of competing machine learning algorithms with the sole purpose of determining which variables from each of the datasets are significant predictors of a team’s ability to win a basketball game. The second stage focused on leveraging the significant predictors yielded from stage one into several optimized algorithms to predict the outcome of individual games between Division 1 opponents. These stage two models were evaluated against each other on both precision and accuracy for the 2022 NCAA Division 1 Men’s College Basketball season, in order to determine which algorithm type, depth of rolling average, and predictor set would yield the most accurate predictive results. Therefore, stage one will only yield results pertaining to the significance of predictors in each dataset, and stage two will yield the results to the main hypothesis of this research.

**Stage 1 – Determining Significant Predictors of Winning Basketball**

For the first stage, each record was an individual team in an individual season. Therefore, each Division 1 program would have multiple records used to train the models if they competed at the D1 level for more than one season during the timeframe of this study (2011-2022). The dependent variable for each model in this stage is a team’s W-L% for a given season. Therefore, each model will use the data associated with a team in each season to predict that team’s W-L%. In the training of these models, variable importance was assessed to determine which individual variables from the datasets pulled the most weight when it came to predicting a team’s W-L%. These variables that were determined to be significant for each type of ML model were fed into the models in stage two to reduce the dimensionality of the datasets.

This research did not attempt to determine which machine learning algorithm type was the best at predicting a team’s W-L% because the main goal of stage one was to determine significant predictors of winning college basketball games at the Division 1 level. Each type of algorithm was used in stage two as well, so the significant predictors for each algorithm type were fed into the corresponding algorithm in stage two as a reduced dimension predictor set.

Each model discussed in stage one was trained on the respective dataset (BasStat or AdvStat) for the 2011-2021 NCAA Division 1 Men’s College Basketball season. The feature importances were determined from these trained models, and then the models were also evaluated for fit via comparison of test set root mean squared error (RMSE). The models that have hyperparameters were optimized via an exhaustive grid search to ensure the optimal set of hyperparameters was chosen. A significance cutoff of 0.05 was established to determine which variables would be counted as ‘important’ predictors. This value was chosen after looking at the datasets as it allows the reduced predictor set to be smaller than the full dataset without being too small that it doesn’t take advantage of the thoroughness of the dataset.

**LASSO Regression Models**

**Figure 13: LASSO Regression Feature Importance, BasStats**

**Graphical user interface, application

Description automatically generated**

From figure 13, it was determined that only three variables from a school’s basic team stats were significant predictors of a team’s W-L%:

* A Team’s Defensive FG%
* A Team’s Offensive FG%
* A Team’s Offensive FT%

These results are interesting in the sense that the data suggest that any statistic outside of these three key predictors doesn’t do much at all to drive a team’s W-L% at a college level. This is not to say that those facets of the game are not important, but it does suggest that a significant majority of the variance in W-L% between teams at the D1 level comes from their ability to shoot the ball at a high level of efficiency, and their ability to force opponents to shoot the ball at low levels of efficiency.

**Figure 14: LASSO Regression Feature Importance, AdvStats**

Graphical user interface

Description automatically generated

When looking at the advanced team statistics, more variables emerge as significant predictors of a team’s W-L%. These predictors include:

* A Team’s Offensive True Shooting % (TS%)
* A Team’s Defensive Effective Field Goal % (eFG%)
* A Team’s Defensive Free Throw Efficiency (FT/FGA)
* A Team’s Offensive Effective Field Goal % (eFG%)
* A Team’s Offensive Free Throw Efficiency (FT/FGA)
* A Team’s Defensive Free Throw Rate (FTr)
* A Team’s Defensive True Shooting % (TS%)

All these indicators correlate with the significant predictors from the BasStats dataset. TS% and eFG%, both on the offensive and defensive side of the ball, are more advanced metrics that weight a team’s offensive and defensive FG% according to shot difficulty. FT/FGA is a measure of how efficient a team is at making free throws, which is a slightly different angle on the FT% metric.

Moreover, all but one of these predictors fall into the Four Factors, which have largely been accepted as the main drivers of winning a basketball game during the 21st century basketball analytics revolution (Oliver, 2011).

**Figure 15: Root Mean Square Error Comparison, LASSO Regressors**

Chart, bar chart

Description automatically generated

Figure 15 shows the BasStat model yields a slightly more accurate model when fit to the 2022 season data as the RMSE for the test set has a slight difference in favor of the BasStat model (0.001). Therefore, for the Logistic Regression approach to predicting a team’s W-L% at the college level, using the basic box score statistics is slightly more accurate than using advanced stats to predict a season-long W-L%.

**Random Forest Regressor Models**

**Figure 16: Random Forest Regressor Feature Importance, BasStats**

**Graphical user interface

Description automatically generated**

When a random forest regressor model was fit to the data, four predictors from the BasStat dataset emerged as significant predictors of a team’s W-L%:

* A Team’s Offensive FG%
* A Team’s Defensive FG%
* A Team’s Points Scored Per Game (PPG)
* A Team’s Points Allowed Per Game (PPG)

The first two predictors above are similar to the significant predictors from the LASSO Regression model that was fit to the same data. The two FG% metrics make sense from a simple eye test perspective, as the higher percentage of shots a team makes and the lower percentage of shots their opponent makes, the more likely they are to win more basketball games. The second two predictors in the above list, PPG and opponent PPG, were not close to the top of the feature importance in the LASSO regression model. This means that the pruned random forest model has leveraged these two predictors in a higher capacity than the LASSO model was able to. Again, from a logical standpoint, it reasons that a team that scores more points and holds their opponents to less points per game will win more basketball games than a team that fails to do so.

**Figure 17: Random Forest Regressor Feature Importance, AdvStats**

A picture containing graphical user interface

Description automatically generated

The Random Forest regressor has only identified two significant predictors of a team’s W-L% from the AdvStats section.

* A Teams Offensive Efficiency (ORtg)
* A Team’s Defensive Efficiency (DRtg)

These features fall in line with the Four Factors mentioned earlier and make sense from a purely superficial standpoint as well. This random forest model seems to have yielded results that line up with even the most amateur understanding of what it takes to win the game of basketball: *Score more points than your opponent in less tries*.

**Figure 18: Root Mean Square Error Comparison, Random Forest Regressor**

**Chart, bar chart

Description automatically generated**

The above figure shines a light on the strength, and comparative weakness of a Random Forest regressor for this type of quantitative predictive problem. The training RMSE for both the BasStat dataset and the AdvStat dataset are immensely better than the corresponding RMSE for the logistic regression approach. This shows that the expansive ensemble approach of a Random Forest when fit to this problem yields significantly more accurate training results. However, the test RMSE for both datasets are slightly higher than the Logistic Regression approach introduced earlier, and therefore the Random Forest model is a slightly worse predictive model than the Logistic Regression approach for a season-long W-L% model.

As it pertains specifically to the efficacy of the BasStat dataset vs. the AdvStat dataset in predicting season-long W-L% for a Division 1 Men’s College Basketball team, the test RMSE in the figure above shows that there is an advantage in favor of the AdvStat set of predictors. This contradicts the absolute value of the test RMSE figures from the logistic regression models, and suggests that if using solely advanced team stats to make a prediction at a team’s season-long W-L%, a Random Forest regressor approach will yield better relative results than a logistic regression model.

**AdaBoost Models**

**Figure 19: AdaBoost Regressor Feature Importance, BasStats**

Chart

Description automatically generated with low confidence

The AdaBoost regressor, when fit to the BasStat dataset, returns a more balanced feature importance distribution. This model, using a cutoff of 0.05 as the value of significance (consistent with previous models), yields 10 significant predictors from the BasStat dataset:

* A Team’s Offensive FG%
* A Team’s Defensive FG%
* A Team’s Turnovers Per Game (TOV)
* A Team’s Points Scored Per Game (PPG)
* A Team’s Defensive 3-Point Percentage (3P%)
* A Team’s Points Allowed Per Game (PPG)
* A Team’s Assists Per Game (AST)
* A Team’s Steals Per Game (STL)
* A Team’s Total Rebounds Lost Per Game (TRB)
* A Team’s Defensive Assists Allowed Per Game (AST)

These predictors that the AdaBoost model has given significant feature importance to do not overlap nearly as much as the previous two models. FG%, both on the offensive and defensive side of the ball are still the two most significant predictors in this model, but apart from those, most of the other predictors were deemed insignificant by the other models. The AdaBoost regressor, and its ensemble nature of weak-classifiers allows for it to derive predictive value from predictors that the other algorithms overlook, such as STL, AST, TRB, etc.

**Figure 20: AdaBoost Regressor Feature Importance, AdvStats**

Graphical user interface

Description automatically generated

In contrast to the AdaBoost’s feature importance distribution from the BasStat dataset, its distribution of feature importance from the AdvStat dataset is much more top-heavy, like the LASSO regression and the Random Forest regressor. There are three predictors that break the plane of significance, although the top two are magnitudes more important than the third:

* A Team’s Offensive Efficiency (ORtg)
* A Team’s Defensive Efficiency (DRtg)
* A Team’s Defensive Rate of Rebounds Lost (TRB%)

The ORtg and DRtg metrics are a common thread among the algorithm’s important variables derived from the AdvStat dataset. This passes the naïve logic that has been applied to “eye test” the results of these algorithms. As has been the movement in basketball analytics in the past two decades, a more efficient team seems to win more games than a less efficient team. It is also significant to note here that the “weak” classifier model has yielded the TRB% metric as a significant predictor, as this is the one Four Factor metric that the other models failed to flag as significant (eFG%/TS%, TOV%, FT/FGA being the other three).

**Figure 21: Root Mean Square Error Comparison, AdaBoost Regressor**

Chart, bar chart

Description automatically generated

The above plot shows the most significant edge yet in favor of the AdvStat dataset in both the training and testing set RMSE values for the ensemble AdaBoost models assembled in stage 1. With a RMSE delta of 0.015 in favor of the AdvStat dataset for both the training and testing RMSE values, this analysis provides the most concrete evidence yet in favor of the AdvStats providing the most accurate predictive power for predicting a team’s W-L% in a given season.

**Stage 1 Conclusion**

The following table recaps the significant predictors that each machine learning algorithm selected from each dataset.

**Table 3: Stage 1 Significant Predictor Summary by Model Type**

|  |  |  |
| --- | --- | --- |
| **Model** | **BasStat Predictors** | **AdvStat Predictors** |
| LASSO | FG%\_Opp,  FG%,  FT%, | TS%,  eFG%\_Opp,  FT/FGA\_Opp,  eFG%,  FT/FGA,  FTr\_Opp,  TS%\_Opp, |
| Random Forest | FG%,  FG%\_Opp,  PPG,  PPG\_Opp | ORtg,  DRtg |
| AdaBoost | FG%,  FG%\_Opp,  TOV,  PPG,  3P%,  PPG\_Opp,  AST,  STL,  TRB\_Opp, AST\_Opp | ORtg,  DRtg,  TRB%\_Opp |

To recap where the research went from here, each model type (or it’s classification-based equivalent) was leveraged in a stage two model with the goal of correctly classifying the most D1 Men’s College Basketball games as either a Win or a Loss. The significant predictors for each model type in stage one were fed into the models in stage two as a reduced dimension dataset and compared against the full datasets (BasStat and AdvStat) to evaluate the efficacy of performing variable selection in this two-stage manner.

**Stage 2 – Individual Game Analysis**

Stage two of this research study developed, trained, and evaluated a series of machine learning algorithms against each other to determine which combination of factors yields the most accurate predictive model for classifying the outcome of a college basketball game. The types of machine learning algorithms that will be compared against each other are a Logistic Regression based approach, Random Forest classifier, and an AdaBoost classifier. Each model type will be fit to a series of training datasets to determine the optimal data structure and dataset to predict the outcome of a Division 1 Men’s College Basketball game.

**Optimal Lag Term**

To assess whether there is an optimal number of previous games to consider when predicting the outcome of a future game, each of the candidate machine learning algorithm types (Logistic Regression, Random Forest Classifier, and AdaBoost Classifier) were fit considering a variable number of previous games. The range of previous games considered ranged from 1 to 9. These fits were done on both the BasStat and AdvStat set of predictors since it had not yet been concluded which of these sets were optimal for this classification problem.

**Figure 22: Logistic Regression Accuracy by Lag Depth**

Chart, shape

Description automatically generated

There is a significant drop off in test accuracy between different lag depths. There appears to be diminishing decrease in accuracy as the lag term increases, with there not being a significant difference in test set accuracy when considering say a 6-game lag term versus a 9-game lag term. However, for the Logistic regression approach, it is clear that the test set accuracy is highest when only a team’s prior game statistics are considered in predicting the outcome of the upcoming game.

Now would be a good time to disclose that this study did not account for injuries. Therefore, if a team’s composition heading into their next game is significantly different, or their opponents’ lineup is significantly different due to injury or a player returning from injury, the optimal model that will be developed by the end of this research does not specifically account for that case. However, this study didn’t omit those games from its training or testing sets, so in some sense, the model developed from this research is a “general case” model that can be deployed regardless of injury status because it was trained with those injury games as a part of its training set. From a coach’s standpoint, they might have just slightly less confidence in the prediction if they know a significant contributor is not being accounted for with the previous single-game lag term.

**Figure 23: Random Forest Accuracy by Lag Depth**

Chart, shape

Description automatically generated

A similar accuracy trend is observed for the Random Forest model as well, as there is clearly a downward sloping accuracy curve as lag depth increases. The same diminishing returns can be seen around the 5-9 game depth levels, but again, there is still a clear frontrunner in that the single-game lag term has the highest test set accuracy of any lag depth (avg of 0.954 between BasStat and AdvStat) by a margin of over 15% compared to the next most accurate lag term.

**Figure 24: AdaBoost Accuracy by Lag Depth**

Chart, shape

Description automatically generated with medium confidence

The same trend of decreasing accuracy with increasing lag term is observed for the AdaBoost approach as well. The single-game lag term again is the optimal lag term for this machine learning algorithm. Something of note that can be seen in this particular case as compared to the logistic regression and random forest approaches is that the difference in test accuracy between AdvStat and BasStat is muted as the single-game lag depth, whereas in the other cases there was a visible difference (albeit small) in test set accuracy between BasStat and AdvStat even at the single-game level.

**Figure 25: Logistic Regression AUC Analysis, Optimal Lag Depth**

**Chart

Description automatically generated**

Based on the conclusions from the previous three accuracy curves, only the lag terms from 1-game to 4-games were included in the AUC analysis as any further lag depth seemed to have both diminishing returns and be less accurate than the lag depths included in the AUC analysis above. It is clear from the observed test set AUC values for each lag depth (averaged across each ML algorithm type at each lag depth), that the single-game lag depth performs significantly better at predicting the outcome of Division 1 Men’s Basketball games than any other lag depth. In fact, it’s AUC-edge over the next closest lag term (2-games) is roughly 0.05. From both the perspective of sports betting and from a coach’s standpoint, a 5% edge over the next best model is invaluable.

**Reduced Predictor Set**

The next research question that was assessed for the game outcome classification problem was the efficacy of using a reduced predictor set versus a full predictor set for each of the two candidate datasets (basic box score stats vs. advanced efficiency metrics). The two datasets had significant predictors selected via the season-long W-L% models in stage one of this research. The goal of this particular sub-question is to determine whether the same factors that predict season-long W-L% at the highest levels of accuracy are also the same factors that predict the outcomes of individual games with the highest levels of accuracy.

Using this methodology, each candidate model approach was fit with both the full and reduced predictor sets to determine whether or not the same significant features can be focused on to strengthen a team’s season-long W-L% and a team’s probability to win an individual game.

**Figure 26: Logistic Regression Test Accuracy, Full vs. Reduced Predictor Sets**

Chart, bar chart

Description automatically generated

The figure above shows that for each dataset, the logistic regression models obtained higher levels of test set accuracy when leveraging the full slate of predictors available to the model instead of using the reduced set determined by the season-long W-L% models.

**Figure 27: Random Forest Test Accuracy, Full vs. Reduced Predictor Sets**

Chart, bar chart

Description automatically generated

This figure shows slightly contradictory evidence compared to the Logistic Regression approach. For the BasStat predictor set, there was a slight (0.02%) advantage in test accuracy for the reduced predictor set, but there was a 0.2% advantage for the AdvStat dataset when utilizing the full predictor set. This means that for the Random Forest approach, the conclusion on whether or not the same predictor set that optimally predicts season-long W-L% can be leveraged to predict single-game outcomes is inconclusive without first selecting the optimal predictor set (BasStat vs AdvStat).

**Figure 28: AdaBoost Test Accuracy, Full vs. Reduced Predictor Sets**

Chart, bar chart

Description automatically generated

For the AdaBoost candidate algorithms, the test accuracy was again very close between the reduced and full predictor sets. For both the BasStat and AdvStat predictor sets, there was a 0.1% advantage in test set classification accuracy when utilizing the reduced predictor set compared to the full predictor set. This means that utilizing the predictors that the AdaBoost model was able to determine as significant predictors of season-long W-L% is also a better approach to predicting outcomes of a single game compared to using the full set of available predictors.

While the results remain inconclusive as to whether the reduced or full predictor set is optimal based on what type of algorithm is at play, using results from a future section of this conclusion can help achieve the final optimal decision on this sub-point. Further on in the conclusion of this research, it will be found that the Logistic Regression approach to this classification problem yields both the best test set accuracy and the best test AUC value. Therefore, if the Logistic Regression approach is deemed to be the optimal model, looking at the AUC analysis for reduced versus full predictor set for that particular model type can help cement the answer to this sub-research question as it pertains to selecting the optimal structure for the final game prediction model.

**Figure 29: Logistic Regression AUC Analysis, Optimal Lag Depth**

**A picture containing chart

Description automatically generated**

The figure above shows a significant AUC-advantage for the model that leverages the full predictor set compared to the one leveraging the reduced predictor set from stage one. Therefore, in the final optimized game prediction model, the full predictor set will be leveraged when training the model.

**Basic vs. Advanced Statistics**

The next sub-optimization that will be determined for the game predictor classification portion of this research is whether the BasStat set of predictors or the AdvStat set of predictors is better at predicting the outcome of a Division 1 Men’s College Basketball game. This analysis will be shown for each of the three candidate model types, as that determination will be the last conclusion before the optimized model is presented.

**Figure 30: Optimal Predictor Set, Logistic Regression Approach**

**A picture containing graphical user interface

Description automatically generated**

For the logistic regression approach, there is a slight (0.005) advantage in AUC in favor of using the AdvStat predictor set to predict the outcome of each division 1 Men’s Basketball game of the 2022 season. This advantage might seem slim, but on the scale of ~4500 games contested at the D1 level per season, that 0.5% advantage for the AdvStat predictor set corresponds to approximately 23 extra games correctly predicted. This edge might not be significant in the statistical sense, but it is an observable advantage that a coach, analyst, or sportsbook would take note of.

**Figure 31: Optimal Predictor Set, Random Forest Approach**

Graphical user interface

Description automatically generated with low confidence

A similar edge in AUC is observed for the AdvStat predictor set in the Random Forest model as the Logistic Regression model. This 0.005 edge in AUC does mean that from an objective standpoint, the AdvStat predictor set is again the superior predictor set when it comes to predicting the outcome of Division 1 Men’s Basketball contests.

**Figure 32: Optimal Predictor Set, AdaBoost Approach**

A picture containing graphical user interface

Description automatically generated

The AdaBoost model bucks the trend when it comes to a difference in AUC between the BasStat and AdvStat predictor sets. The AdaBoost model has no observed difference in AUC for the competing models used to predict games from the 2022 NCAA Men’s Basketball season. This could be due to the ensemble weak-learner nature of the AdaBoost approach to this classification problem. However, the key takeaway from this model is that there is not an advantage to using the BasStat predictor set. This means that across all candidate models there is either an advantage, or at least not a discernable disadvantage to using the AdvStat predictor set to predict the outcomes of D1 Men’s College Basketball games.

**Optimal Machine Learning Algorithm Approach**

To fully fit and present the optimized game prediction model, it must first be determined which of the three candidate classification algorithms best fits this classification problem. The three classification algorithm candidates are a logistic regression model, a random forest classifier, and an AdaBoost classifier approach. Each model type was fit to their optimized hyperparameter values determined from the previous three sub-optimization tasks (lag depth, breadth of predictor set, specific predictor set). Then each model’s respective test AUC values compared to determine which candidate model type is the best algorithmic approach to this classification problem.

**Figure 33: AUC Analysis, Optimal Machine Learning Algorithm Decision**

A picture containing graphical user interface

Description automatically generated

According to the above figure, there is a noticeable difference in the predictive quality of each candidate model type. The rankings of the candidate models in their ability to predict the outcome of Division 1 Men’s Basketball games in the 2022 season from most powerful to least powerful are as follows:

1. Logistic Regression Model
2. Random Forest Classifier
3. AdaBoost Classifier

Therefore, the best machine learning algorithm to approach this classification problem with is the logistic regression model, trained on AdvStat data from 2011-2021, using all predictors in the AdvStat dataset, while using only a team’s previous 1-game to predict the outcome of the next contest. This algorithm combines the optimization of each of the four sub-research questions that this classification model was developed to answer.

**Finalized Models & Evaluation**

*Game Predictor Algorithm*

Using the prescribed features for the logistic regression game predictor, a model was constructed using data from the 2011-2021 seasons and tested on the games from the 2022 season. This model is the key model to come out of this research as it will directly demonstrate the optimal approach to predicting the outcome of a Division 1 Men’s College Basketball game given the constraints of the candidate models selected and the data available to the model to train on.

**Figure 34: Confusion Matrix, Optimized Game Prediction Model**

Chart, treemap chart

Description automatically generated

The above confusion matrix demonstrates that the overall classification accuracy of this game predictor algorithm is 76.91%. Roughly 3 in 4 games had their outcomes correctly predicted in the 2022 Men’s Division 1 College Basketball season. This accuracy will be compared to a model that is currently deployed by one of America’s largest sportsbooks during further discussion of these results.

**Figure 35: AUC Analysis, Optimized Game Prediction Model**

*A picture containing chart

Description automatically generated*

The AUC curve above shows the predictive power of this model is high with a test set AUC = 0.867. This high AUC value signifies that the optimized game predictor model holds significant predictive power above that of a simple “coin-flip” model that guesses at the outcome of each game.

*Game Recap Algorithm*

In a similar vein as the game predictor, another model was fit to similar data. This model was specifically fit on box score data from each individual game, with the intent of classifying the outcome of that game as either a win or a loss depending on the box score data from that game. Certain data points that were considered “deterministic”, that is, they directly either state or can be used to calculate the outcome of a game were omitted from the input set for this model, as the intent for this model is to try and classify the outcome of a game after it happened using data from that games box score, and/or relevant advanced efficiency metrics.

**Figure 36: Confusion Matrix, Optimized Game Recap Model**

Chart, treemap chart

Description automatically generated

The game recap model only misclassifies a total of 62 out of the over 11,000 games played at the division 1 level during the 2022 season. This means that only 0.54% of all games were unable to be correctly classified as a win or a loss given data from that contest after it concluded. This highly accurate model has use-cases for coaches at the Division 1 level that will be explained in further detail at a later point.

**Figure 37: AUC Analysis, Optimized Game Recap Model**

A picture containing chart

Description automatically generated

By evaluating the model and its test set AUC, it can be seen that this model is highly effective at classifying the outcomes of Division 1 Men’s College Basketball games given their statistics. A test set AUC that rounds to 1.000 means that this is almost a perfect classifier, with only 0.54% of the test set being misclassified. With only 27 false positives at the 0.5 probability threshold, it can be concluded that if the model predicts a game was a win given its statistics, that there is only a 0.48% chance that the prediction is incorrect (28 false positives out of 5,612 total predicted positives).

**Discussion**

**Findings**

This research has concluded that the best type of model to predict a team’s season-long W-L% is a LASSO Regression model fit with alpha = 1\*10-6. This LASSO model was trained on the AdvStat set of predictors. Moreover, the best model to deploy to predict the outcome of a single game or a series of games is a Logistic Regression model trained on the full predictor set from the AdvStat set of predictors, utilizing only a team’s most recent game as the input layer.

With this pair of optimal models, it can be concluded that the AdvStat set of predictors is indeed the better set of predictors to use to predict both a season-long metric such as W-L%, as well as individualized game metrics like the classification of W/L.

This research also found that when evaluating a team’s performance in an upcoming game, apart from injuries or other external factors, it is best to only consider the statistical output from a team’s most-recent game to define the predicted statistical output for the upcoming game. This finding is essential in understanding how to deploy a game predictor model in Division 1 Men’s College Basketball, as the volatility of team’s and their performances can often skew other predictive models. The optimal Logistic Regression model proposed from this research is as resistant to such volatility as it possibly can be by being built directly on this 1-game optimal lag depth.

This research has also found that the features that are deemed most important in the prediction of a team’s season-long W-L% are not necessarily the same features that are best suited for predicting the outcome of individual games. This was concluded from the fact that at no point during the A/B testing for reduction of predictor set was there an advantage in either test set prediction accuracy or test set AUC for the reduced predictor set. There were instances where there was no delta between the reduced and the full predictor set, but there was not a single model that derived higher results from the reduced sets than the full sets. Therefore, it cannot be stated that using the significant predictors from the season-long success metric prediction model is an effective variable selection process for predicting individual game outcomes.

**Outlier Analysis: 2022 Season**

In order to fully contextualize the value of the optimal models that this research has developed, the following tables are presented as a sample of the analysis that both of the optimized models are capable of facilitating.

*Optimal W-L% Model*

The W-L% model can be leveraged in a multitude of ways. The most forthcoming of those ways is to use the model to analyze every team’s predicted vs. actual W-L% for the 2022 Men’s Basketball season (the test set). Analyzing the outliers in the prediction set of any machine learning algorithm can help to identify any unintentional bias that the model may have. In the case of college basketball, unintentional model bias might take the form of the consistent over/under ranking of a particular conference or teams with a particular play style.

**Table 4: Greatest Predicted W-L%, 2022 Season, Optimal W-L% Model**

Table

Description automatically generated with medium confidence

From a season-long perspective, Table 4 offers the W-L% model’s predictions for the Top 25 highest predicted W-L% from the 2022 Men’s Basketball season along with their actual W-L%, and the residual between the model’s prediction and the actual realized W-L% for the 2022 season. Teams that achieved Top-2 seeds in March Madness such as Gonzaga, Houston, Arizona, and Duke all populate the top of this table. Mid-Major powerhouses such as Vermont, Murray St., and South Dakota St., all of whom had program-defining post-season runs also sit at the top of the W-L% model’s predictions. The fact that the model’s predictions match so well with the “eye-test” speaks to the power of the model to provide a rigorous scientific backing for analysts to back up what they see with their eyes.

**Table 5: Lowest Predicted W-L%, 2022 Season, Optimal W-L% Model**

Table

Description automatically generated

Like the previous table, the lowest W-L% predictions from the 2022 Men’s Basketball season also match up with the eye test. Programs that underwent huge turmoil during the 2022 season find themselves at the lowest end of the model’s predictions.

**Table 6: Highest W-L% Residual, Optimal W-L% Model**

Table

Description automatically generated with medium confidence

Analyzing the Top 25 programs with the highest W-L% residual from the optimal W-L% model shows a potential bias of the model. Many Power-6 programs (members of the ACC, Big Ten, Big 12, Big East, Pac-12, or SEC conferences) that did not fall at the top of their conference, like Wisconsin, Providence, and Arkansas, all had much lower predicted W-L% than they realized in the 2022 season. This is evidence that strength of opponent is not being properly captured for the middle of the pack teams like it is for topflight programs like Duke, Kentucky, and Kansas.

**Table 7: Lowest W-L% Residual, Optimal W-L% Model**

Table

Description automatically generated

The reverse end of the residual table shows a plethora of Mid Major programs that missed out on the NCAA Tournament, but still had spectacular seasons possessing the most under-predicted W-L% based on their statistics. The fact that very few Power 6 programs populate the bottom of the list suggests the potential use-case for separate models to predict W-L% for Power 6 programs compared to Mid and Low Major schools.

*Optimal Game Prediction Model*

The game prediction model has its own unique use-cases for both coaches, analysts, and fans of the game. The use-case that this paper will demonstrate is using the game-by-game Win% predictions to tabulate a team’s predicted records over the course of a season.

The model outputs a percentage for each game input that indicates a team’s propensity to win that matchup. By treating the probability prediction of the model as an expected wins variable, one can sum those expected wins over the course of a season and determine a team’s predicted record. This can be subdivided to analyze more specific scenarios by month, conference vs. non-conference, home court games, neutral site games, or true road games.

**Table 8: Highest % of Games Correctly Predicted, 2022 Season, Optimal Game Prediction Model**

Graphical user interface

Description automatically generated with medium confidence

Looking at the Top 15 teams for which the optimal game prediction model predicted the highest percentage of their games correctly, there appears to be a nice mixture of teams from both Power 6 and Mid Major conferences. There also appears to be a mixture of teams that had stellar records in 2022 along with teams that had subpar records. This mixture of programs at the top of the prediction accuracy table suggests that there is no overt bias being demonstrated by the classification model on the top end.

**Table 9: Lowest % of Games Correctly Predicted, 2022 Season, Optimal Game Prediction Model**

A picture containing graphical user interface

Description automatically generated

Looking at the bottom end of prediction accuracy, another story presents itself. Most of the teams for which the model had the lowest prediction accuracy in the 2022 Men’s College Basketball season are lower major programs (programs outside the top ~12 conferences). The only Power 6 program to make the list are St. John’s, which only had 53.1% of its games correctly predicted over the course of the season.

The model has a range in accuracy for different teams. The good news is that the floor for accuracy is 50%. This means that for all 350+ programs, this model is at least as good as a coin flip at predicting the outcome of their games. However, this research didn’t set out to develop a model with a coin flip’s chance at predicting the outcome of a game. The ceiling for prediction accuracy is 89.7%, with a mean prediction accuracy of 78.4%.

Therefore, the typical Division 1 program can expect this model to correctly predict 78.4% of their games before they are even played. The value that such a tool can have to coaches is apparent.

**Table 10: Highest Residual of Predicted Total Wins, 2022 Season, Optimal Game Prediction Model**

Table

Description automatically generated with medium confidence

By leveraging the probability to win each of a team’s games as an expected wins metric, the model can yield predicted win totals for a team as well. The table above shows the Top 15 teams that won more games this season than their predicted win totals. For example, UNC Wilmington had a record of 27-9 for the 2022 NCAA Men’s Basketball Season, but the optimal game prediction model shows their record should have been 18.9 – 15.1 based on their probability to win each individual game. This reflects positively on both the players and the coaches as it suggests they were able to win roughly 8.1 games that they were not supposed to win based on the model’s predictions.

Comparing the list of teams in the Top 15 for predicted wins residual and the teams presented in Table 6 with the highest W-L% residual, many of the same teams can be found on both lists. This suggests that although the models are independent of each other, the numbers yield similar results when it comes to teams that outperformed their statistics for the 2022 season. This is a positive for the models as it suggests that both are operating within the same plane of prediction, where the highest season-long W-L% residuals also have the highest Win Total residuals.

**Table 11: Lowest Residual of Predicted Total Wins, 2022 Season, Optimal Game Prediction Model**

Table

Description automatically generated

The reverse end of the residual spectrum for teams that won less games than the model predicted based on the summation of individual game win probabilities yields many of the same programs that had the lowest W-L% residuals. Only four programs had their season win totals differ by more than 4 games. This figure is much greater for teams that won significantly more games than the model predicted with 32 teams falling into the bin with more than four additional wins over the model’s prediction.

This suggests that the model tends to underestimate a team’s win totals, which checks out as the model tends to classify win probability on a normal distribution, thus many of the individual game win probabilities will be farther away in magnitude from the absolute outcomes of 1/0. This discrepancy in win probability and the binary outcome of a basketball game leads to the under-prediction of the season-long win totals that the game prediction model is being leveraged to create. However, despite the inability of the model to provide more extreme predictions, the season win totals fall completely within the realm of possibility and are very interpretable from a coaching perspective.

A coach could even leverage this data on a game-by-game basis to identify games for which the real outcome of a contest and the predicted outcome did not match. On the opposite end of the spectrum, a coach could also query which games the team won where their win probability prediction was < 50%. This type of analysis can and will lead to more in-depth understanding of what drives the actual outcome of a basketball game at the collegiate level. Coaches, particularly those who leverage these types of analytics at the highest level, will find greater success for their program.

**In-Depth Team Analysis: MTSU Blue Raiders Men’s Basketball, 2022 Season**

*Optimal Game Prediction Model*

Another use-case for the models generated from this research is to analyze the performance of an individual team against expectations. The game prediction model, as previously shown, can yield expected win totals based on a team’s probability to win each individual game. Breaking these numbers down for an individual team across an entire season can lead to interesting insights for coaches, analysts, and fans alike.

**Table 12: MTSU Men’s Basketball, 2022 Season, Predicted W/L Totals By Location**

Text

Description automatically generated

Comparing the actual record for the MTSU Men’s Basketball program in the 2022 season overall, the model only has a residual of about 0.5 games on each side of their actual W/L record. Given an actual W/L record of 26-11 during the 2022 campaign, the optimal game prediction model says that their expected record based on the opponents that they played and their level of play at the time of playing them, the Blue Raiders should have ended the season with a record of 23.05 – 11.95. Given that partial wins and losses aren’t interpretable at a macro level, rounding off these predictions to the nearest whole number makes comparison between the actual record and the predicted record much easier.

Rounding MTSU’s predicted overall record, the model gave them a predicted record of 23-12, roughly two games worse than they ended up finishing during their actual season. Now is a good time to note that because the optimal prediction model leverages the results from a team’s previous three games to use as inputs to predict the upcoming contest, predictions from the first three games of the season are not based on the same set of inputs (3-game average), and therefore are excluded from any analyses on the basis that they are fundamentally different predictions.

Dividing the games based on the venue at which they were played, more value can be derived on where MTSU over/under achieved on those expectations. MTSU went 15-0 at home during the 2022 season, a record that the model says should have been 10-3 (omitting 2 home games due to the inability to predict). This means that MTSU outperformed expectations at home by at least a 3-game margin. MTSU’s record on the road for the 2022 season was 5-9, which the model suggests should have been above .500 at 8-6. This underperformance on the road roughly cancels out their overperformance at home, which explains why their predicted record and their actual record are much more similar than these 3-game margins put together. Considering games played on a neutral court, MTSU had a predicted record of 5-3 in those contests this year, while attaining an actual neutral site record of 6-2.

**Table 13: MTSU Men’s Basketball, 2022 Season, Predicted W/L Totals**

A picture containing chart

Description automatically generated

The data analyzed from the previous table can be summed up at a macro level in the table above. MTSU ended up winning 3 more games than predicted this year, while only losing 1 more game than predicted. This 2-game discrepancy in delta values is again because the first few games were omitted from predictions as their inputs were unable to be based off the required 3-game rolling average. These first two games of the season were played at home against Brescia University and Bethune-Cookman University, both of which resulted in wins for the Blue Raiders.

**Table 14: Incorrectly Predicted Games, MTSU Men’s Basketball, 2022 Season**

Table

Description automatically generated

The above table shows the 9 games that were incorrectly predicted by the optimal game predictor model from the Blue Raider’s 2022 campaign. Leveraging the summation of win-probability across these 9 games, the Blue Raiders had an expected record of 5.49 – 3.51. MTSU’s actual record across these nine games was 4-5. These incorrectly predicted games include Wins by the Blue Raiders where their predicted win probability was 0.497, 0.483, 0.410, and 0.461 respectively. The incorrectly predicted games also included losses by the Blue Raiders where their win probability was 0.696, 0.811, 0.676, 0.609, and 0.849 respectively. The magnitudes of these win probabilities being starkly different between the wins and losses that were incorrectly predicted suggests that MTSU won some very tight games that they had a slightly higher chance of losing than winning, while also losing games where they were the overwhelming favorite heading into said game.

A pattern among the win-probabilities of incorrectly classified game outcomes like these suggest that a team is streaky, thus coming into some games with overinflated win-probability based on their previous games. This also could be an artifact of scheduling bias. For example, a team might play their toughest three opponents in a row, so their win probability for that next game will be drastically understated as they are coming off their toughest stretch against a lesser opponent. The same could be said conversely about coming off a stretch of “easy” opponents into an unfavorable matchup against a strong opponent. The win probability would be inflated due to the inflated statistics achieved against lesser competition being used as the input vectors for the game prediction model.

*Post-Game Outcome Analysis Model*

Another team-level use-case that this research provides is the leveraging of the game prediction model that was able to 100% accurately classify the outcome of a game after given its non-deterministic statistics after the fact. Non-deterministic statistics mean that the model was fed every output statistic from a game (e.g. REB, AST, TRB, eFG%) that is not directly able to determine and/or calculate the outcome of that game. Therefore, statistics like PTS, FGM, 3PM, and FTM were omitted from the predictor set as those can be used to directly determine the final score of a game, and thus the outcome of said game. The model that successfully identifies ~100% (99.85%) of games as either wins or losses has a use-case that can be leveraged by coaches of a team after a game has concluded.

**Table 15: Post-Game Analysis Output, MTSU Men’s Basketball, 2022 Season**

Table

Description automatically generated with medium confidence

The table above shows each game that the MTSU Men’s Basketball program played in the 2022 season, including the two early season games that had to be omitted from the optimal game prediction model due to not having the proper rolling average for the input layer. The table also gives the actual outcome of each game, the predicted outcome, and the “true” win probability of a game given the actual statistics put up by each team during that game.

The Blue Raiders had 0 games for which the predicted outcome and the actual outcome differed in the 2022 season. Therefore, it can be concluded that based on the games that MTSU played, they walked away with the result that they deserved on each night. A coach can use these calculations to identify what are commonly referred to as “bad losses”. These are losses where a team “should” have won, but for one reason or another is walking back to the locker room with a loss. These games are especially important for a coach as it allows them to avoid overreacting to a game where a team “should” have won given the game that they played. This model can be a deployable tool for coaching staffs to determine whether the result of a game is worth changing the team’s strategy for. If a game is incorrectly classified by this model (0.142% of all games), it is not worth changing a team’s composition or strategy as a result of that game because the game can be considered a “fluke” from an outcome perspective, meaning the box score and advanced statistics from that game suggest that the actual outcome of the game is opposite of what it “should” have been.

On the other hand, the model could also identify “fluke” games that went the other way. Where a team ends up winning a game that the statistics indicate they “should” have lost. The acceptance of this win as deserved might cause a coaching staff to minimize the number and/or magnitude of changes that need to be implemented before a team’s next match.

**Table 16: Incorrectly Classified Games, 2022 Season, Post-Game Classification Algorithm**

Table

Description automatically generated

Out of the 5,976 independent Division 1 Men’s College Basketball games played during the 2021-22 season, this post-match analysis model incorrectly classified the outcomes of 31 games. This means that for 99.49% of all contests, this model can accurately help explain the “why” regarding the outcome of a game. Furthermore, only one contest that was incorrectly classified had a predicted probability over 60%, meaning that the model still correctly classified the games as close contests, thus even further explaining the incorrect nature of the prediction.

This tool, which is distinctly separate from the “optimal” models discussed at length already, can provide post-match analysis value to a Men’s College Basketball program, and it only exists as a by-product of the main models generated from this research.

**Conclusion and Future Direction**

*Conclusion*

This research set out with the objective of determining a few specific criteria for optimizing both the season-long W-L% prediction and the game-by-game outcome prediction of Division 1 Men’s College Basketball. The results of this research have unearthed so much more value and understanding than purely concluding on that original goal.

Three models have been developed as a result of this research that can be deployed to deliver real-world value for a wide range of end-users. The process of developing these models also led to the uncovering of other optimal parameters for evaluating Men’s College Basketball games such as only leveraging a team’s past three games in predicting the outcome of a future game, as the test set accuracy maximizes itself at this depth.

Comparing the optimized game prediction model developed in this research to the Las Vegas Sportsbook average for percent of all basketball games correctly predicted (66.9%), shows that the model (76.91%) developed through this research has a 10.01% outright advantage over the typical basketball model currently employed by several sportsbooks (Admin., 2022). Using data more specific to the sport of college basketball, it was determined that Caesar’s Sportsbook correctly predicted the results of 70.22% of college basketball games during the 2019-20 season. As mentioned earlier, this data only represents games for which they accepted money line bets. Therefore, the optimized model that this research yielded has a 6.69% edge over Vegas in this use-case. The caveat to this edge is that Vegas Sportsbooks often set odds at sub-optimal levels to ensure profitability based on the wagers currently being taken on any given matchup (Fanson, 2020). Therefore, without knowing the internal financials for a sportsbook, it is impossible to know if this 6.69% edge that is observed is due to superior prediction capability, or sub-optimal prediction quality from Vegas in exchange for profits.

The need for further research in this field cannot be understated. While this research has unearthed three valuable tools to assist in the analysis of Division 1 Men’s College Basketball games, there is undoubtedly more in-depth models that could be developed with more domain knowledge and more specific criteria. For example, many analysts suggest that the college basketball season is really comprised of two seasons: the non-conference schedule from November through December, and the conference schedule that begins in January. This domain knowledge might lead to the investigation of whether there are significantly different winning drivers of games between the two sub-seasons.

The most popular aspect of college basketball is the end-of-season March Madness tournament. This is when the sport gets most of its national attention, and thus models that specifically explain and predict the outcomes of these games would provide the most tangible value to a national audience.

*Future Direction*

This research only proposed a few alternative models to be leveraged and deployed in each scenario. There are a plethora of algorithms and techniques that were not tested as solutions to the two models, such as Principal Component Analysis to reduce dimensionality, Support Vector Machines to leverage that PCA in a classification context, or even the leveraging of a feed forward neural network to train models beyond the limits of traditional supervised learning techniques. Therefore, the findings of this research, although significant and deployable, suggest that further research be conducted on the topic of game prediction in the college basketball industry.

REFERENCES

Admin. (2022, May 12). *How often does the favorite win in NBA - Underdog Winners*. Offshoresportsbookfact. Retrieved August 11, 2022, from https://offshoresportsbookfact.net/nba/how-often-does-the-favorite-win-in-nba/

Bechtold, T., & Sweetwood, J. (2021, April 16). *How the 'idiots who believe' in the analytics movement have forever changed basketball*. The Analyst. Retrieved July 19, 2022, from https://theanalyst.com/na/2021/04/how-advanced-analytics-have-changed-basketball/

Beck, H. (2021, April 26). *The tanking era as we know it is over*. Sports Illustrated. Retrieved August 15, 2022, from https://www.si.com/nba/2021/04/26/tanking-era-is-dead-play-in-tournament

Bennett, K. P., & Parrado-Hernandez, E. (2006). The Interplay of Optimization and Machine Learning Research. *Journal of Machine Learning Research*, *7*, 1265–1281.

Biau, G., Scornet, E., & Welbl, J. (2018). Neural random forests. *Sankhya A*, *81*(2), 347–386. https://doi.org/10.1007/s13171-018-0133-y

Bilogur, (2018) Missingno: a missing data visualization suite. Journal of Open Source Software, 3(22), 547, https://doi.org/10.21105/joss.00547

Caesars Entertainment. (n.d.). Retrieved September 14, 2022, from https://www.caesars.com/sportsbook-and-casino/tn/

Calipari, J. (2015). *Players first: Coaching from the inside out*. PENGUIN Books.

Cawley, G. C., & Talbot, N. L. C. (2007). Preventing Over-Fitting during Model Selection via Bayesian Regularisation of the Hyper-Parameters. *Journal of Machine Learning Research*, *8*, 841–861.

Chen, A. A., Beer, J. C., Tustison, N. J., Cook, P. A., Shinohara, R. T., & Shou, H. (2021). Mitigating site effects in covariance for machine learning in Neuroimaging Data. *Human Brain Mapping*, *43*(4), 1179–1195. https://doi.org/10.1002/hbm.25688

Chengsheng, T., Huacheng, L., & Bing, X. (2017). AdaBoost typical algorithm and its application research. *MATEC Web of Conferences*, *139*, 00222. https://doi.org/10.1051/matecconf/201713900222

*College Basketball Statistics and history: College basketball at sports*. sports-reference.com. (n.d.). Retrieved July 19, 2022, from https://www.sports-reference.com/cbb/

Dellenger, R. (2022, May 2). *Nil, 'Booster Banks' and recruiting wars: For some, it doesn't add up*. Sports Illustrated. Retrieved August 15, 2022, from https://www.si.com/college/2022/05/02/nil-name-image-likeness-experts-divided-over-boosters-laws-recruiting

Demenius, J., & Kreivytė, R. (2017). The benefits of advanced data analytics in basketball: Approach of managers and coaches of Lithuanian Basketball League teams. *Baltic Journal of Sport and Health Sciences*, *1*(104), 8–13. https://doi.org/10.33607/bjshs.v1i104.11

Dodd, D. (2021, June 21). *NCAA amateurism is effectively dead, and the Association itself may not be far behind*. CBSSports.com. Retrieved August 15, 2022, from https://www.cbssports.com/college-football/news/ncaa-amateurism-is-effectively-dead-and-the-association-itself-may-not-be-far-behind/

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and Artificial Neural Network Classification models: A methodology review. *Journal of Biomedical Informatics*, *35*(5-6), 352–359. https://doi.org/10.1016/s1532-0464(03)00034-0

Esposito, F., Malerba, D., Semeraro, G., & Kay, J. (1997). A comparative analysis of methods for pruning decision trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(5), 476–493. https://doi.org/10.1109/34.589207

Fanson, P. (2020, September 29). *Vegas always knows? A mathematical deep dive*. The Only Colors. Retrieved August 11, 2022, from https://www.theonlycolors.com/2020/9/29/21492301/vegas-always-knows-a-mathematical-deep-dive

Google. (2019). *Classification: Precision and recall  |  machine learning  |  google developers*. Google. Retrieved July 28, 2022, from https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall

Google. (2019). *Classification: Roc curve and AUC  |  machine learning  |  google developers*. Google. Retrieved July 28, 2022, from https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, *96*(4), 835–845. https://doi.org/10.1093/biomet/asp047

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in science & engineering, 9(03), 90-95.

Jebara, T., Kondor, R., & Howard, A. (2004). Probability Product Kernels. *Journal of Machine Learning Research*, *5*, 819–844.

Krzyzewski, M., & Hill, G. (2001). *Leading With the Heart: Coach K's Successful Strategies for Basketball, Business, and Life*. GRAND CENTRAL PUB.

Lemeshow, S., & Hosmer, D. W. (2008). Logistic regression. *Wiley Encyclopedia of Clinical Trials*. https://doi.org/10.1002/9780471462422.eoct964

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, *54*(3), 217. https://doi.org/10.2307/2685594

Lunardi, J. (2022). *Bracketology: March madness, college basketball, and the creation of a national obsession*. TRIUMPH BOOKS.

Malato, G. (2022, May 27). *Are you still using 0.5 as a threshold?* Your Data Teacher. Retrieved August 15, 2022, from https://www.yourdatateacher.com/2021/06/14/are-you-still-using-0-5-as-a-threshold/#:~:text=The%20reason%20behind%200.5&text=In%20fact%2C%20if%20the%20probability,the%20probability%20of%20having%200.

McKinney, W., & others. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56).

Mudric, M. (2020, November 22). *How the NBA data and Analytics Revolution has changed the game*. SmartData Collective. Retrieved August 15, 2022, from https://www.smartdatacollective.com/how-nba-data-analytics-revolution-has-changed-game/

Niculescu, R. S., Mitchell, T. M., & Rao, R. B. (2006). Bayesian Network Learning with Parameter Constraints. *Journal of Machine Learning Research*, *7*, 1357–1383.

Oliver, D. (2011). *Basketball on paper: Rules and tools for performance analysis*. Potomac Books, Inc.

Osterman, Z. (2022, July 19). *How nil is reshaping college basketball: 'everything that was a sin is now saintly.'*. The Indianapolis Star. Retrieved August 15, 2022, from https://www.indystar.com/story/sports/college/2022/07/19/ncaa-basketball-nil-reshaping-college-sonny-vaccaro-march-madness-transfer-portal-nba-draft/65373592007/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of machine Learning research, 12, 2825-2830.

Pomeroy, K. (2006, November 29). *Ratings explanation: The Kenpom.com blog*. The KenPom.com Blog. Retrieved August 15, 2022, from https://kenpom.com/blog/ratings-explanation/

Pomeroy, K. (n.d.). *2022 Pomeroy College Basketball Ratings*. KenPom College Basketball Ratings. Retrieved July 19, 2022, from https://kenpom.com/

Rocks, J. W., & Mehta, P. (2020). Memorizing without overfitting: Bias, variance, and interpolation in over-parameterized models. *Phys. Rev. Research*, *4*. https://doi.org/ https://doi.org/10.48550/arXiv.2010.13933

Rokach, L., & Maimon, O. (2005). Decision trees. *Data Mining and Knowledge Discovery Handbook*, 165–192. https://doi.org/10.1007/0-387-25465-x\_9

Root, J. (2020, April 17). *Fun with finances: Basketball budgets - three-man-weave*. Three Man Weave. Retrieved August 15, 2022, from https://www.three-man-weave.com/3mw/college-basketball-budgets-2020

Shea, S. M. (2014). *Basketball Analytics: Spatial Tracking*. Createspace Independent Publishing Platform.

Shea, S. M., & Baker, C. E. (2013). *Basketball analytics: Objective and efficient strategies for understanding how teams win*. Advanced Metrics.

Singh, S. P. (1992). Scaling reinforcement learning algorithms by Learning Variable Temporal Resolution Models. *Machine Learning Proceedings*, 406–415. https://doi.org/10.1016/b978-1-55860-247-2.50058-9

Smith, D., Kilgo, J., & Jenkins, S. (2002). *A Coach's Life: My 40 Years in College Basketball*. Random House.

Staffo, D. F., & Bing, D. (2018). *Jim Boeheim and Syracuse Basketball: In the zone*. Sports Publishing.

TeamMom. (2017, January 19). *The story in a box score*. Blazer's Edge. Retrieved August 15, 2022, from https://www.blazersedge.com/2017/1/19/13755264/nba-box-score-history

Tuv, E., Borisov, A., Runger, G., & Torkkola, K. (2009). Feature Selection with Ensembles, Artificial Variables, and Redundancy Elimination. *Journal of Machine Learning Research*, *10*, 1341–1366.

Van Rossum, G. (2020). *The Python Library Reference, release 3.8.2*. Python Software Foundation.

Waskom, M., Botvinnik, Olga, O&#x27; Kane, Drew, Hobson, Paul, Lukauskas, Saulius, Gemperline, David C, … Qalieh, Adel. (2017). *mwaskom/seaborn: v0.8.1 (September 2017)*. Zenodo. https://doi.org/10.5281/zenodo.883859

Weitzman, Y. (2020). *Tanking To The Top: The Philadelphia 76ers and the Most Audacious Process in the History of Professional Sports*. Grand Central Publishing.

White, J. (2022, May 3). *How continued legalization of sports betting affects college sports*. GMTM. Retrieved August 15, 2022, from https://gmtm.com/articles/how-does-the-continue-legalization-of-sports-betting-effect-college-sports

Wooden, C. J., & Jamison, S. (1997). *Wooden*. Contemporary Books.

Yu, T., & Zhu, H. (2020). Hyper-Parameter Optimization: A Review of Algorithms and Applications. *Journal of Machine Learning Research*.

Zak, T. A., Huang, C. J., & Siegfried, J. J. (1979). Production Efficiency: The Case of Professional Basketball. *The Journal of Business*, *52*(3), 379–392. http://www.jstor.org/stable/2352368

Zhang, Y., Burer, S., & Street, N. W. (2006). Ensemble Pruning Via Semi-Definite Programming. *Journal of Machine Learning Research*, *7*, 1315–1338.

Zhao, P., & Yu, B. (2006). On Model Selection Consistency of Lasso. *Journal of Machine Learning Research*, *7*, 2541–2563.

APPENDIX A

**Figure A-1: Sample Basic Team Stats**

Calendar

Description automatically generated

*Source: https://www.sports-reference.com/cbb/seasons/2022-school-stats.html*

**Figure A-2: Sample Basic Opponent Stats**

Calendar

Description automatically generated

*Source: https://www.sports-reference.com/cbb/seasons/2022-opponent-stats.html*

**Figure A-3: Sample Advanced Team Stats**

Table

Description automatically generated

*Source: https://www.sports-reference.com/cbb/seasons/2022-advanced-school-stats.html*

**Figure A-4: Sample Advanced Opponent Stats**

Calendar

Description automatically generated

*Source: https://www.sports-reference.com/cbb/seasons/2022-advanced-opponent-stats.html*

**Figure A-5: Sample Basic Game Log Data**

A screenshot of a computer

Description automatically generated

*Source: https://www.sports-reference.com/cbb/schools/middle-tennessee/2022-gamelogs.html*

**Figure A-6: Sample Advanced Game Log Data**

Graphical user interface, application, table

Description automatically generated

*Source: https://www.sports-reference.com/cbb/schools/middle-tennessee/2022-gamelogs-advanced.html*

**Figure A-7: Sample NCAA Team Sheet (University of Virginia, 2019 – Partial)**

*Diagram

Description automatically generated with medium confidence*

1. Full Python package citations can be found in the references section. [↑](#footnote-ref-1)