A Machine Learning Approach to Predicting National Champions in Division 1 Women's Basketball

D'Anna Cvetnich

April 20, 2023

1 Problem

Each March, the top-ranked basketball teams in the first division of the National Collegiate Athletic Association, commonly referred to as the NCAA, engage in a single-elimination tournament to become the National Champion of the season. This tournament draws national attention, even from those who are not typically basketball fans. However, the focus during this time of year is primarily on the men's road to the national championship. The women's side of NCAA Division 1 basketball also competes in a similar tournament, however, it is often overlooked. One of the significant draws of "March Madness" is the accompanying bracket challenge, where millions of people try to predict the tournament's outcome before it begins with the chance to win substantial rewards for a perfect bracket. While much analysis has been done on the men's side of the tournament, the women's side remains ignored. The student-athletes participating in this tournament deserve recognition for their hard work and effort. Playing on the female side of the game should not diminish this deserved recognition in the media and sports analytics.

Despite the lack of recognition and media coverage, the women's side of NCAA Division 1 basketball has consistently produced some of the most talented and skilled players in the sport. These student-athletes face the same level of competition and pressure as the men's teams, and their hard work and dedication to their sport should not be overlooked. The success of these women's teams can also serve as an inspiration to young girls and women everywhere, showing them that they too can achieve their dreams and succeed in a male-dominated field. Additionally, the use of data and analytics in this project can help to shed light on the talent and abilities of the women's teams and give them the recognition they deserve. This project aims to not only predict the outcomes of the women's national championship tournament but also to bring more attention to the women's side of the sport and give credit where credit is due.

The structure of this paper begins with a look at previous work that has been completed on related problems as well as a discussion of what these previous studies used to complete their machine learning. From there, the third section regards the material that was new and necessary to learn in order to complete the project. Furthermore, the next section goes into detail on the steps taken and the methodology followed throughout the completion of the project. Following, the fifth section describes the results gathered via the completion of the machine learning model, as well as some examples showcasing the accuracy of the predictions and a bracket containing the predicted data. Finally, section six describes the possible future work that could be completed on this project.

2 Literature Review

Machine learning algorithms can process and analyze vast amounts of data much faster and more efficiently than traditional methods, making them ideal for sports game outcome prediction [7]. There are different types of machine learning to choose from when approaching such a problem. Machine learning is divided into three main categories: supervised learning, semi-supervised learning, and unsupervised learning [9]. The majority of these broad categorizations have more specific sub-categories. The category of supervised learning can be further divided into three sub-categories: classification, regression, and forecasting. The category of unsupervised learning can be further divided into two sub-categories: clustering and dimension reduction [9]. With machine learning, we can not only make predictions about future games but also identify key variables that have the greatest impact on the outcome [5]. This information can be used to improve player performance, optimize team strategies, and make more informed decisions about player management. Additionally, machine learning algorithms can continuously learn and adapt to new data, making predictions more accurate over time [8].

As machine learning has shown itself to be well-suited to this area of sports analytics, there have been several who have taken on this task with none going about in exactly the same ways. One of the more popular approaches is discussed in Bunker and Thabtah's "A Machine Learning Framework for Sports Result Prediction". Bunker and Thabtah focus on the use of Neural Networks, mainly focusing on Artificial Neural Networks (ANN), in the realm of sports predictions. The framework to be followed in the creation of the models is SRP-CRISP-DM, or Sport Result Prediction CRISP-DM framework, and is an extension of the standard CRISP-DM framework. This approach follows six main steps: Domain Understanding, Data Understanding, Data Preparation and Feature Extraction, Modelling, Model Evaluation, and Deployment of the Model [5]. The paper reviews previous studies that have used artificial neural networks exclusively and provides a critical discussion and observations on prior work in this field, in the context of the SRP-CRISP-DM framework and conventional measures of model performance. ANNs have been used in the prediction of outcomes for several different sports and sports organizations including the National Football

League (NFL), Australian Rules Football League (AFL), horse races, and many others [5]. The section describing the creation of the horse race prediction model references several different training algorithms that were used to create the most accurate model possible for their purposes. These training algorithms included gradient-descent (BP), gradient-descent with a momentum parameter (BPM), Levenberg-Marquadt (LM), and conjugate gradient-descent (CGD). BP had the highest accuracy, with a rate of 77%, however, BP has the drawback of requiring a longer completion time for training [5]. Bunker and Thabtah conclude that machine learning is a suitable methodology for this problem and stress the need for more accurate models despite the increasing popularity of using machine learning in this field.

While a main focus of the proposed solution to this problem is the creation of machine learning models to predict the outcomes of games, in order for these models to be of use they must be validated. This is the primary emphasis of Tomislav Horvat, Ladislav Havaš, and Dunja Srpak in "The Impact of Selecting a Validation Method in Machine Learning Predicting Basketball Game Outcomes". The models developed for this study utilized supervised learning techniques, with a concentration on the classification sub-class. Seven classification machine learning algorithms were applied and their results were validated using two methods: Train&Test validation and cross-validation. The purpose of the research was to determine the best combination of algorithm, validation method, and data preparation method for better prediction results. The research also explores the impact of different validation methods on prediction accuracy when using different machine learning algorithms. Cross-validation was shown to be a superior validation technique, but it relies on future event data that cannot be predicted. The Train&Test validation method, however, still produced acceptable results [7]. In situations that include the inability to predict accurate future event data Horvat, Havaš, and Srpak actually recommend the use of the Train&Test validation method, along with the use of current data. They also concluded that the machine learning algorithms used performed quite similarly. However, the front runner in accuracy was the nearest-neighbors algorithm.

The research by Chenjie Cao is focused on building a model to accurately predict the outcomes of National Basketball Association (NBA) games. Cao adopted a mostly supervised learning approach, focusing on the classification of wins and losses in their predictions. The algorithms fitting under this umbrella used in this study were Simple Logistics Classifier, Support Vector Machine algorithm, and Naïve Bayes. Cao also used reinforcement learning in the form of artificial neural networks [6]. Cao tested several different machine learning algorithms in order to strive towards choosing the correct algorithm for their situation. After creating and testing their models, it was shown that the Simple Logistics Classifier attained the highest accuracy in its predictions, with 69.67% [6].

Based on the information provided in the resources above, it is clear that supervised learning is the most promising method for creating high-quality machine learning models for sports outcome prediction. Specifically, the classification sub-class within supervised learning is highlighted as the best approach.

Furthermore, the algorithms described in these resources are touted as having the greatest potential for producing accurate models to address the issue at hand.

3 Technical Material

This project did not require an abundance of technical material to be learned for its completion. The largest learning curve was during the web scraping phase. An understanding of the intricacies of the different packages that were attempted to be used in Python to get the data needed was necessary to complete the tasks at hand. The recalling of how machine learning is to be implemented and how the data must be formatted to ensure the best results was also a stepping stone in this project.

3.1 Web Scraping

Web scraping is the process of extracting data from a web page, usually through the use of automation tools. This section of the project presented the largest learning curve, as I had never been required to go through the process in the past. In order to complete this step, several methods were tried. The first method attempted was to collect data via the use of the Selenium package in Python. I was somewhat familiar with this package and some of its capabilities, including manipulating and gathering information from web pages. However, this package proved not to be the optimum method; as it took a long time to complete and required a large number of server calls. Due to this, another package, Beautiful Soup was researched and was attempted to be implemented. This package is specifically made for web scraping data, therefore, it vastly outperformed the code written with Selenium. The package has readily available documentation that made it easier to understand and apply the package to the project's specific use case. As a semi-working automation suite had been created and I was originally unfamiliar with the Beautiful Soup package, ChatGPT was initially used to translate the code written for the Selenium package into code that worked with Beautiful Soup. This new package was used to navigate to the desired URLs, extract table data from the page, and export it, with the help of Pandas' data frames, into .CSV files for future use.

3.2 Machine Learning & Feature Manipulation

While familiar with machine learning from past coursework, I was most definitely rusty with the skills associated. In order to begin the machine learning phase, the data had to first be cleaned and formatted in a useful way; this step came in the form of removing any games that contained any null values. After that, the features were manipulated to only contain the statistics believed to be most helpful in determining the winner of a basketball game. These statistics mostly stayed true to the basic statistics gathered at almost any game, e.g.

percentage of made shots, assists, turnovers, etc.

4 Methodology

4.1 Data Gathering

The first task of this project was to gather the data needed. This data will take the form of common basketball-centric statistics, i.e. assists, points, turnovers, blocks, fouls, etc., and will be both the averages over the entire team for that particular game as well as the maximum of that individual statistic during that game. This data was acquired from https://www.sports-reference.com/cbb/seasons/ [7] [3], an open data source in beta for college basketball statistics under the umbrella of the Sports-Reference family of websites. This data was obtained via web scraping. This data was collected via the use of the Beautiful Soup Python library [1] and then converted into Pandas [2] data frames for easier manipulation.

4.2 Feature Manipulation

After the data was collected, the next logical step was to begin the feature engineering process. Feature engineering is an important step in the data science process that involves creating new features from existing data to improve the accuracy and performance of machine learning algorithms. This feature engineering adhered to the standard feature engineering process of data cleansing, data transformation, feature extraction, feature selection, and feature iteration.

The data was cleaned by removing extraneous columns, ensuring the data was in a usable form, and condensing each game to a single row. When the data was collected, each game's statistics consisted of four separate tables. The data in these tables were then evaluated for their importance to the the result of team in the game, i.e. determining if the data held any bearing on if the team would win or not. Data that simply gave the result of the game, other than the boolean 'WIN' column, was removed. This included any column from which the total number of points could be extrapolated, for example, the 'POINTS' column would directly tell the model who the winner was and any of the field goal (or 'FG') columns. However, this information is too important to remove completely from the data set, this was amended by including the percentage made of each field goal. This was completed by dividing the total made field goals by the number attempted in the game. This was repeated for two-point, three-point, and free-throw field goals. While the data collected was from completed games, these statistics are also available before games. All of the statistics used in this model are readily recorded for teams' past games. The data from these statistics can be averaged over the past few games a team has played or over the season thus far.

New data columns were also created, such as the assist-to-turnover ratio, created by dividing the 'AST' and 'TOV' columns, if a null value was calculated

due to a turnover value of zero, the value was converted to zero. The team that the 'WIN' column pertained to was represented by the default column names, data pertaining to the opposing team was labeled by attaching '_OPP' to the end of the column name. The maximum values of each statistic for a team were labeled with '_MAX'.

4.3 Machine Learning

After this, the data was ready to be input into the machine learning algorithms. This project took a supervised machine learning approach, focusing on the classification sub-type. Classification machine learning has shown itself to be useful for prediction and forecasting situations such as these due to the fact that the program must estimate the relationships present in the given data. Several machine learning algorithms were tested to see which would provide the highest results. The algorithms explored included Ridge Regression, Linear Regression, Logistic Regression, SVM or support vector machine, k-nearest neighbors classifier, decision tree classifier, random forest classifier, Gaussian Naive Bayes, and Multi-Layer Perceptron (MLP) classification.

4.4 Testing and Validation

Out of the ten tournament seasons of data collected, three seasons (2018, 2021, and 2022) were set aside for testing and validation of the created model. The tournament data set aside for testing was taken from some of the more recent tournament years to ensure the model can reliably predict wins and losses in games currently being played and that will be in the future. In the testing data, the 'WIN' column was separated into a separate file so that it can be input without worry of giving the model the answers. These games were fed into the model to produce the predictions.

The produced predictions were then evaluated by using cross-validation. SciKit Learn's cross_val_predict and cross_val_score methods were used. The validation was completed on a split of size three on the testing data. The resultant score of the cross-validation was on average 0.611, with a maximum of 0.648. Therefore, this model was able to predict the correct winner of the game approximately 61% of the time.

4.5 Bracket Creation

With the model completed, a program was created to log the win/loss predictions of a tournament and output them into a bracket format, for easier readability for the user. This output bracket was created in HTML, with the predicted winners of each game input into designated spots in the HTML for display. The program is able to output this bracket in both image and pdf format.

5 Results

This model is most certainly not the most accurate game prediction model, as it was only able to predict correctly approximately 61% of the time. However, these more accurate models often have much larger data sets than the one that was used in this project. The data set utilized here was limited to women's NCAA division one basketball tournament games. Each tournament season has a total of 67 games played, and only data from the past thirteen seasons were available. This gives a total of 871 games played. When researching for this project, models were found that were trained on datasets upwards of fifty times larger than our dataset. As there are many more games to learn from in bigger datasets, it is only logical that these models are more likely to have a higher accuracy rate than those trained with smaller datasets.

5.1 2022 Season

An example of a bracket completely made from predictions made based on data from the 2022 tournament season is shown in Appendix 7.1. This was created by first gathering games from the testing data that were played in 2022, this was then further shortened by only selecting the games played in the first round of the tournament (Round of 64). This data set was then given to the model to make predictions. This produced a binary list designating whether the model predicted the first team named in the row to win or lose.

From this list, the data pertaining to the winning team was extracted and then arranged so that the next match-up, according to the bracket, could be 'played.' This then continued for the remaining rounds of the tournament, until a predicted national champion was achieved. The actual bracket results from 2022 are shown in Appendix 7.2. The model predicted South Carolina to win the national championship based on the data presented to it. This matches the actual outcome of the tournament. The model was able to correctly guess 39 games out of 63, or 62% of games played.

Bracket creation for March Madness most often occurs through the ESPN Tournament Challenge website and/or application. This program allows the user to make their own predictions of how the tournament will play out before it begins. The created brackets are then scored using a system of rules in order to compare the user's bracket to others, either in a created group or against the entire player base. The point system for the ESPN Tournament Challenge is outlined in Table 1 [4]. Using this scoring system, the predicted bracket attains a score of 1190 out of a total of 1920 possible points.

Table 1. Point system for created brackets according to the ESPN Tournament Challenge.

ROUND	GAMES	POINTS (ea.)	POINTS (pred.)
First Round	32	10	230
Second Round	16	20	160
Sweet Sixteen	8	40	160
Elite Eight	4	80	160
Final Four	2	160	160
National Championship	1	320	320
Totals	63	–	1190

6 Future Direction

This project can be continued by adding more seasons of data to the training if or when they come available. It would be ideal to essentially rotate out the oldest of the testing set and place it into the training set as newer seasons can be added to the testing data. More advanced statistics could also be added to possibly increase the accuracy of the model. Implementation of the model and bracket producer would be of more use within a webpage or phone application, so another future step could be implementing this.

All of the code created and used in this project can be found on Git-Hub at the following link: https://github.com/deeCvetnich/WBBTournamentPredictions

References

[1] Beautiful soup documentation. https://beautiful-soup-4.readthedocs.io/en/latest/. Accessed: 2023-04-11.

[2] Pandas documentation. https://pandas.pydata.org/docs/. Accessed: 2023-04-11.

[3] Sports-reference college basketball. https://www.sports-reference.com/cbb/seasons/. Accessed: 2023-04-11.

[4] Tournament challenge - how to play. https://fantasy.espn.com/tournament-challenge-bracket/2023/en/story?pageName=tcmen%5Chowtoplay. Accessed: 2023-04-17.

[5] Rory P. Bunker and Fadi Thabtah. A machine learning framework for sport result prediction. *Applied Computing and Informatics*, 15(1):27–33, 2019.

[6] Chenjie Cao. Sports data mining technology used in basketball outcome prediction. *School of Computer Sciences at ARROW@TU Dublin Dissertations*, Sep 2012.

[7] Tomislav Horvat, Ladislav Havaš, and Dunja Srpak. The impact of selecting a validation method in machine learning on predicting basketball game outcomes. *Symmetry*, 12(3):431, 2020.

[8] Herman O. Stekler and Andrew Klein. Predicting the outcomes of ncaa basketball championship games. *Journal of Quantitative Analysis in Sports*, 8(1), 2012.

[9] Katrina Wakefield. A guide to the types of machine learning algorithms. https://www.sas.com/en_gb/insights/articles/analytics/machine-learning-algorithms.html. Accessed: 2023-02-11.

7 Appendix

7.1 2022 Prediction Bracket

Below is a completed bracket based on the predictions made by the model based on the data collected from the 2022 women's national tournament.

7.2 2022 Actual Bracket

Below is a completed bracket based on the actual outcomes of games played in the 2022 women's national tournament. Teams that are in red indicate that the prediction model did not predict this team would be in that position in the bracket.