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

D'Anna Cvetnich

February 11, 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.

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 [4]. 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 [6]. 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 [6]. 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 [2]. 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 [5].

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 [2]. 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 [2]. 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 [2]. 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 [4]. 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 [3]. 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% [3].

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 Methodology

The first task of this project will be 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 as well as the individual statistics for each player, with the added metric of time played per game added for individual statistics. The seed of the team, i.e. the rank of the team within the bracket, in the tournament will also be given. This data will be acquired from various sources including the NCAA official Statistics website [1], ESPN, and, if needed, the websites of individual participating colleges. Other sources of data may be used if these sources do not see themselves useful, such as `https://www.sports-reference.com/cbb/` [4], an open data source in beta for college basketball statistics under the umbrella of the Sports-Reference family of websites. This data will either be obtained manually or via web scraping, depending on the format of the particular data source.

After the data is collected, the project will move on to 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 will adhere to the standard feature engineering process of data cleansing, data transformation, feature extraction, feature selection, and feature iteration. In completing this feature engineering, we will gain a further understanding of the dataset, identify its relevant features, apply appropriate transformations, and engineer new features that will add value to the analysis.

After this, the creation of the model can begin. This project will take a supervised machine learning approach, most likely focusing on the regression and or classification sub-types. Regression machine learning has shown itself to be useful for prediction and forecasting due to the fact that the program must estimate the relationships present in the given data. Classification machine learning has shown very similar uses to regression; therefore it may also be of use in the model creation process. This phase of the project will also include the testing of different regression and or classification machine learning algorithms such as linear regression, decision trees, and random forests [6].

The dependent variable the model will be predicting is if a game between two teams will result in a win or a loss for each team, with one team receiving the result of a loss and the other a win as a game cannot result in two wins or two losses. Predictions will be made in succession, according to the tournament year's bracket, with the winning team advancing into the next round and the losing team being taken out of consideration. Approximately 20-25% of the data collected will be set aside for testing. The tournament data set aside for testing will be taken from a spread of tournament years to ensure the model can reliably predict wins and losses through different eras of play types. The last three tournament seasons will also be set aside for testing. Once the model is complete, a program will be 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 will most likely take the form of a webpage. This webpage will allow the user to observe the seed of the teams playing in each game and whether that team is predicted to win or lose. The page will also allow the user to export the bracket as different file types.

A preliminary model will be created by March 16^{th} in order to create a bracket for the upcoming tournament season. This year's tournament begins on March 17^{th}, thus no completely predictory brackets will be able to be created after this date. This will allow the testing of the model with real time results, as well as add entertainment value to the project.

4 Timeline

Week Of	**Planned Task**
January 15	Researching Possible Topics
January 22	Researching Possible Topics
January 29	Understanding Chosen Topic
February 5	Written Proposal
February 12	Research Models and Methods
February 19	Gather Data
February 26	Test Models/Choose Training Model
March 5	Training Model
March 12	Training Model, 2023 Tournament Begins
March 19	Testing Model
March 26	Validate Model
April 2	Finalize Model, Draw Conclusions, 2023 Tournament Ends
April 9	Report Draft Submission
April 16	Finalize Draft and Report
April 23	Final Oral Presentation

References

[1] NCAA statistics. `https://stats.ncaa.org`. Accessed: 2023-02-11.

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

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

[4] 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.

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

[6] 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.