Mining NCAA basketball data to inform March Madness predictions

Ethan Meyer
CSCI 4502
Ethan.Meyer@Colorado.edu

Ishika Patel
CSCI 5502
Ishika.Patel@Colorado.edu

1 INTRODUCTION

Over the last decade, sports betting has experienced a meteoric rise in popularity. In 2021 alone, March Madness—the annual NCAA Basketball Tournament—garnered approximately $ 3.1 billion in bets on platforms such as FanDuel, DraftKings, and BetMGM (Fortune). With this incredible amount of participation, increasing rates of sports betting legality, and the ample amount of historical data available on NCAA teams, predicting the winner NCAA basketball matches is a prime subject for investigation.

March Madness matches are renowned for their unpredictability. In fact, it is one of the main reasons the tournament ganrers so much attention from fans. Factors such as volatile player contribution on the court, streaks in game-play, buzzer beaters, and more all come together to form "Cinderella Stories" - situations in which teams achieve far greater successes than spectators could have reasonably expected. These dynamic aspects of the game engage fans and present a rewarding challenge—that of predicting game winners—for data miners to approach.

2 RELATED WORK

There exists a myriad of related work in the NCAA data mining genre from fanatics who are tracking their college teams to research papers on predicting game outcomes based on past team performance.

Contributions to the current literature on NCAA tournament prediction focuses primarily on model and feature selection, with an emphasis on mining historical data for predictive features. These predictive features are leveraged for the outcome of classification which has several methodologies available in the literature for this purpose.

With reference to model selection, Yuan, et al. recently published with The 5th International Conference on Big Data Research (ICBDR) on different classification models such as decision tree, random forest, Linear Discriminant Analysis, QDA, Support Vector Machines, and Naïve bayes were used to predict the result of a game.

Variable/feature selection is varied among the models available because NCAA data is in abundance. However, Lo-Hua Yuan et al. found that parsimonious feature sets and relatively simple algorithms tend to outperform more complicated models with numerous features which will be taken into consideration.

Many models use averages over seasons to calculate feature metrics. Given the spontaneous nature of NCAA seasons, we wish to bring new team insights based on sequential game performance. Student researcher Bryce Brown leverages 5-game averages in on logistic regression model to perform NCAA predictions.

In this paper, we branch out from the existing classification literature to investigate how such models can be leveraged to place moneyline bets on NCAA games by accounting for recent team performance over game averages.

3 PROPOSED WORK

3.1 Data

For the purpose of training and testing our models, we will use the dataset from NCAA 2022 machine learning competition on Kaggle.com. The dataset includes historical performance metrics (statistics), game-by-game stats at a team level (free throws attempted, defensive rebounds, turnovers, etc.) for all regular season, conference tournament, and NCAA tournament games since the 2002-03 season. A few data points we intend to include when training our model are:

- WFGM3 - three pointers made (by the winning team)
- WFGA3 - three pointers attempted (by the winning team)
- WFTM - free throws made (by the winning team)
- WFTA - free throws attempted (by the winning team)
- WOR - offensive rebounds (pulled by the winning team)
- WDR - defensive rebounds (pulled by the winning team)
- WAst - assists (by the winning team)
- WTO - turnovers committed (by the winning team)

As part of evaluating the effectiveness of our model, we will also be using historical NCAA betting odds data. Specifically, we will be gathering the Vegas moneyline odds for each match. In sports betting, a moneyline bet is a bet on the winner of the game. For each bet, the casino/bookmaker releases odds for users to bet on. Moneyline odds can be positive or negative and have different implications. For positive odds, the odds number represents how much a user could win off a $100 bet. For example, a moneyline bet on team A with odds of +150 indicate that a user would win $150 off a $100 bet if team A won. For negative odds, the odds number represents how much a user would need to bet in order to win $100. For example, a moneyline bet on team B with odds of -150 indicate that a user would win $100 off a $150 bet if team B won.

From these Vegas moneyline odds, we can compute the bet's implied probability of a team winning via the following formula:

$$P(win) = \begin{cases} \frac{|odds|}{|odds|+100} & \text{if } odds < 0 \\ \frac{100}{odds+100} & \text{if } odds < 0 \end{cases} \tag{1}$$

From our examples earlier, a bet having odds of +150 has implied probability of $\frac{100}{250}$ = 40% and a bet having odds of -150 has implied probability of $\frac{150}{250}$ = 60%. We will use the implied probability from each match to inform our betting strategy.

3.2 Approach

We have planned to implement a selection of models in order to compare them against one another to see which one performs best. These include:

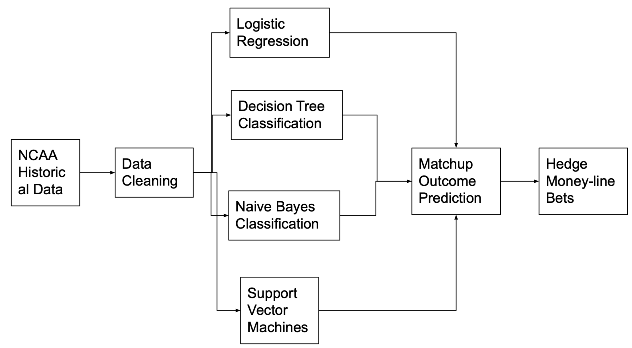
- Logistic Regression
- Decision Trees
- Support Vector Machines
- Bayesian Classifiers

After training our models, we will compare the computed probability of winning to the implied probability from the moneyline Vegas odds. The difference between these two figures—a.k.a. our edge—will determine whether or not we bet, and if so, how much. For example, if we predict the probability of team A winning a match against team B to be 45% but derive that the implied probability from the Vegas moneyline odds to be only 40%, we would want to make a bet as our computed probability is greater than the bets probability.

One of the challenging aspects of this project will be fine tuning our betting strategy based on our edge. Intuitively, we know that the higher our edge, and thus greater difference between our computed probability and Vegas probability, the higher we should bet. However, determining the most effective relationship between edge and bet amount will require extensive testing.

3.3 Process

To outline the general process we plan to follow, we have provided the following diagram:



3.4 Feasibility

Seeing as our team only has two members (both being undergraduate students), we recognize this is an ambitious project. However, we believe that we will be able to implement and compare the results of at least two models and can take on additional work as time permits.

4 EVALUATION

Our models will be evaluated on two metrics. First, we will evaluate our prediction accuracy on correctly identifying wins and losses for specific NCAA tournament matchups. This includes the optimal selection of features to include in the model to confidently predict game outcomes.

Second, we will test our model through a sports betting application. Specifically, we will leverage our model's predictions and confidence levels to inform our betting strategy. In the end, the total amount of money won/lost will be reflective of the success of the model. We see this as more relevant test for the model and believe it will provide a meaningful contribution to the space.

5 MILESTONES

(1) **Data Cleaning & Integration (Oct. 13):** Importing data into local environment and joining from multiple sources. Due to the sheer amount of data, we will be splitting data sets up by seasons. A training and test set are needed for the implementation of an accurate model meaning we are able to use all years of data if need be. The money line data is limited to odds from the 2007 season on with intention of testing betting practices on the most recent seasons.
(2) **Exploratory Data Analysis (Oct. 20):** Exploring data to identify relationships between team statistics (specifically recent performance) and tournament game outcomes
(3) **Attribute Subset Selection (Nov. 10):** Narrowing in on informative features from large dataset. As mentioned, smaller feature subset selection has demonstrated better model performance.
(4) **Model Creation, Examination, and Revision (Nov. 17):** Implementing, testing, and comparing multiple classification models (Logistic Regression, Bayesian Classification, Decision Trees, Random Forests, Support Vector Machine). Fine-tuning models to increase performance. Model selection is flexible and scope to the models is bigger than anticipated output.
(5) **Synthesizing Results (12/6):** Compiling results from work into written project report and video describing which features and methods created the most profitable prediction in betting.

6 REFERENCES

(1) Brown, Bryce, "Predictive Analytics for College Basketball: Using Logistic Regression for Determining the Outcome of a Game" (2019). Honors Theses and Capstones. 475.
(2) University, Yuan Liu Georgetown, et al. “Prediction for NCAA Championship: 2021 the 5th International Conference on Big Data Research (ICBDR).” ACM Other Conferences, 1 Sept. 2021
(3) Yuan, Lo-Hua, et al. A Mixture-of-Modelers Approach to Forecasting NCAA Tournament Outcomes.

7 HONOR CODE PLEDGE

On my honor, as a University of Colorado Boulder student, I have neither given nor received unauthorized assistance.