Mining NCAA basketball data to inform March Madness predictions

Team Members

Ishika Patel

ispa0196@colorado.edu
CSCI 5502

Ethan Meyer

etme9299@colorado.edu
CSCI 4502

Team Members

Ethan Meyer

President

[redacted]

Ashley Jung

VP of Finance

Introduction

- In 2021, March Madness generated about **$3.1 billion in bets** for companies like FanDuel, DraftKings, and BetMGM | Fortune. This amount of revenue, recent looser betting restrictions, and the ample amount of data available on NCAA teams makes this topic fruitful to investigate.

- We wish to bring new team insights based on **sequential game performance** a opposed to overall performance over the season.

- **Moneyline bet**: A college basketball moneyline bet is a wager on which team will win an upcoming game regardless of points scored

Proposed Work

Given NCAA team specific statistics we wish to confidently **predict which bets to place given the game-specific data and moneyline odds**. We would like to place emphasis on how recent team performance (e.g. team/player streaks, upsets in game-play, general team stats) can inform on future game outcomes.

Proposed Work

Data Sets

- Comprehensive NCAA Kaggle Dataset: https://www.kaggle.com/competitions/mens-march-mania-2022/data
    - Tournament-level data (seeds, outcomes, etc.) since 1984 season
    - Team-specific aggregated season stats (Points per game, rebounds per game, etc.)
- Vegas NCAA betting odds: https://www.sportsbookreviewsonline.com/scoresoddsarchives/ncaabasketball/ncaabasketballoddsarchives.htm
    - Game-specific moneyline odds for all NCAA games (regular and postseason) 2007 - present

Milestones

Subtasks and Milestones

1. **Data Cleaning & Integration:** Importing data into local environment and joining from multiple sources
2. **Exploratory Data Analysis:** Exploring data to identify relationships between team statistics (specifically recent performance) and tournament game outcomes
3. **Attribute Subset Selection:** Narrowing in on informative features from large dataset
4. **Model Creation, Examination, and Revision:** Implementing, testing, and comparing multiple classification models (Logistic Regression, Bayesian Classification, Decision Trees, Random Forests, Support Vector Machine). Fine-tuning models to increase performance.
5. **Synthesizing Results:** Compiling results from work into written project report and video

Feasibility of Work

Seeing as our team only has two members, 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.

Evaluation

Model Evaluation

Our models will be evaluated on two metrics. First, we will evaluate our **prediction accuracy on identifying wins and losses** for specific NCAA tournament matchups. Second, we will test a theoretical application of each model through calculating the **amount of money we would have won/loss** were we to have bet in alignment with its prediction. This believe will add a more practical metric for evaluating a model's performance.

Potential Tools

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

Related Work

- Prediction for NCAA championship | 2021 the 5th International Conference on Big Data Research (ICBDR)
    - Implementation of models to predict game winner, no sports-betting application
- A mixture-of-modelers approach to forecasting NCAA tournament outcomes
    - Multiple model development, minimize log-loss, no sports-betting application
- Predictive Analytics for College Basketball: Using Logistic Regression for Determining the Outcome of a Game
    - Computing running averages as features for logistic regression model, no implementation of any other models (SVM)