Asher Erickson
Elise Longenecker
Colin Truran
Yvan Longin

Applied Analytics Project Week 1 Assignment

Problem Statement:

The goal of this project is to evaluate whether historical team data can be used to accurately predict outcomes in the NCAA Men's Basketball March Madness tournament. The project aims to predict winners of each game to improve bracket accuracy, identify games with elevated upset potential, and assess whether these predictions could theoretically produce a positive return when used in sports betting markets. March Madness is known for being unpredictable, as noone has ever predicted the entire bracket before. This unique unpredictability makes the tournament a great challenge for data modeling.

Value: The primary value of this project lies in turning an otherwise unpredictable March Madness bracket into a decision-making model. By using season-long team metrics, statistics, and qualities and turning them into probabilities, we can make consistent decisions and potentially identify matchups that the markets may be under/overvaluing.

Calculation of economic Value:

The economic value of this project is evaluated by testing whether the model consistently outperforms commercial betting odds when its predictions are translated into a simple, transparent betting rule.

For each March Madness game, the model produces a predicted probability of winning for each team based on historical team metrics from our dataset. These probabilities will be compared to the implied probabilities from commercial betting odds, which represent the market's consensus expectation for each game's outcome.

A bet is placed only when the model indicates that a team's predicted probability of winning is higher than the probability implied by the betting odds. This difference suggests that the market may be underpricing that team's chances. To keep the analysis simple and comparable across games, we assume a fixed dollar wager on every qualifying bet.

Economic value is then measured using two straightforward metrics:

- **Total profit**: the sum of wins minus losses across all bets.
- **Return on Investment (ROI)**: total profit divided by total amount wagered.

If the model-generated betting strategy produces a positive profit and positive ROI over historical tournaments, this suggests that historical team data contains exploitable information and that the model adds value beyond random guessing, seed-based picks, or market expectations alone.

As the project progresses, this framework can be extended to more advanced analyses, such as expected value optimization, confidence thresholds, upset-focused strategies, and comparisons across different model types. For this report, however, the goal is to establish a clear, intuitive link between predictive accuracy and real economic value.

Dataset Discussion:

The dataset we chose for this project is a dataset available on Kaggle by Jonathan Pilafas titled "March Madness Historical Dataset (2002 to 2025). This dataset is a collection of historical NCAA March Madness statistics spanning from the years 2002 to 2025. It contains team-level performance metrics from regular season games and NCAA tournament games. The dataset contains traditional metrics, as well as more advanced statistics, which makes it ideal for bracket analysis, machine learning models, and historical trend analyses.

The dataset contains extensive team performance metrics, organized by season and team. Some of these statistics are wins, losses, points scored, points allowed, scoring margin, effective field goal percentage, turnover percentage, rebounding percentage, free throw rate, possessions per game, and strength of schedule. The dataset also contains physical statistics on players, such as average player height and bench minutes. Some of the more uncommon and advanced statistics included in the dataset are adjusted offensive efficiency, adjusted defensive efficiency, and adjusted efficiency margin.

Aside from in-season statistics, the dataset also has historical NCAA tournament results. Some of these statistics are winning and losing teams, score differentials, seed, and round advancement.

All of the statistics in this dataset will provide us with the ability to build machine learning models to try to predict outcomes of the NCAA March Madness tournament.

Type of Modeling:

The model used for the project will employ supervised learning, since the outcomes of the NCAA tournaments are available. Every row in the dataset represents a team that has entered the tournament within the last 25 years, with a long list of rankings in offense, defense, tempo, and many other factors. Because this dataset has labeled inputs and outcomes, the model can infer patterns that map how specific indicators factor into a team's performance.

This model will be using binary classification. In March Madness, there are no ties; it is either a win or a loss. In this dataset, indicators show how far each team has advanced in the tournament. To predict matchup winners, the team's metrics for the year will be compared against their opponent's metrics for the season. For every game, this model will take into account the metrics, seeding, conference, and more to create inputs. Using previous tournament games, the model will learn which features matter most, then predict the probability of a team winning a single matchup.

Project Plan:

Weeks	Weekly Plan
1 - The Problem and Dataset	● Define the problem of predicting March Madness game outcomes using historical team data ● Identify a dataset containing over 20 years of NCAA Division I men's basketball data ● Create a GitHub repo and format
2 - Explore the Dataset	● Define the primary prediction targets (game outcome, point differential, total points scored) ● Perform initial exploration of the dataset to understand structure, data types, missing values, etc. ● Create matchup data by pairing team stats for each game
3 - EDA	● Split dataset into training, validation, and test sets using year by year data ● Conduct EDA to find relationships between team statistics and outcomes ● Analyze upset pattern and determine preprocessing needs
4 - Preprocessing	● Clean and preprocess across all datasets based on findings from EDA ● Encode any necessary categorical variables ● Prepare datasets for different model strategies (classification for winners, regression for point totals)
5 - Feature Engineering	● Create tournament specific features and stat differentials between opposing teams ● Reduce dimensionality by using feature selection or correlation analysis

	● Finalize training, validation, and test sets for modeling
6 - 1st Model	● Develop simple baseline models, such as logistic regression for game outcomes and linear regression for point totals/differentials ● Tune hyperparameters and establish benchmark performance levels ● Compare baseline model performance against naive strategies such as always picking higher seed
7 - 2nd Model	● Build more advanced models such as random forests or gradient boosting models ● Tune hyperparameters and evaluate models compared to baselines ● Analyze overfitting and bias variance tradeoffs
8 - 3rd Model	● Develop additional models such as XGBoost, neural networks, or ensemble models ● Perform advanced hyperparameter tuning and cross validation ● Compare model performance across multiple prediction targets (winner, point margin, upset probability)
9 - Winning Model	● Evaluate all possible models using consistent metrics ● Select the best performing model for each task and calculate final performance on the test dataset ● Interpret results in the context of real world tournament prediction
10 - Data Centric AI	● Improve model performance by refining features and correcting data issues ● Revisit feature engineering choices for better matchup data ● Analyze errors to identify data gaps

11 - Model Explanation and Ethics	● Explore feature importance to understand what drives predictions ● Explain individual game predictions, especially upsets ● Discuss ethical considerations related to sports betting and responsible model use
12 - Model Monitoring	● Save and serialize the final model for deployment ● Outline hypothetical deployment scenario (pre-tournament prediction tool) ● Develop monitoring plan and discuss data drift in college basketball
13 - Assemble Everything	● Integrate all project components into a complete machine learning workflow ● Finalize documentation, code, and results in GitHub ● Summarize findings, limitations, and future improvements, including implications for bracket prediction and betting strategies