# STA 561 Final Project

**Yuxuan Zhang, Qinzhi Peng, Kewei Xu, Huanli Gong, Jianyang Zhou**

**Date: April 26**

# Contents

# 1 Press Release

Durham, NC – April 26, 2024 – In the midst of the fervor surrounding collegiate basketball tournaments, an innovative team composed of Yuxuan Zhang, Huanli Gong, Qinzhi Peng, Kewei Xu, and Jianyang Zhou from Duke University introduces their groundbreaking initiative, ”March Madness Mastery.” Utilizing the capabilities of data analytics and machine learning, this project seeks to redefine the anticipation and comprehension of the 2024 NCAA basketball tournaments across both men's and women's divisions.

”With countless fans eagerly anticipating each matchup, the excitement is palpable,” remarks Yuxuan Zhang, a master's candidate in Statistical Science. ”We identified an opportunity to enrich the tournament experience by providing fans with unprecedented insights into potential outcomes and intriguing plot twists.”

Over an intense final week period, the team immersed themselves in an extensive collection of historical NCAA games, meticulously scrutinizing gameplay, and team statistics. Leveraging advanced predictive modeling techniques, they have developed algorithms capable of producing a diverse range of tournament scenarios, spanning from anticipated favorites to captivating underdog narratives.

”Our models offer a nuanced comprehension of team strengths, weaknesses, and strategic maneuvers,” explains Huanli Gong, a master's candidate in Computer Science. ”By simplifying intricate data into actionable insights, we empower fans to make informed predictions and enhance their engagement with the tournaments.”

Initial testing of the predictive models has yielded promising results, with early users expressing enthusiasm for the project's potential to elevate their tournament experience. ”March Madness Mastery not only predicts outcomes but also deepens appreciation for the complexities of the game,” remarks Qinzhi Peng, a pivotal contributor to the endeavor.

As excitement mounts for the impending tournaments, ”March Madness Mastery” aims to democratize access to predictive analytics, granting fans a glimpse into the future of collegiate basketball. ”Our vision transcends mere predictions; we aim to enrich the fan experience and ignite a passion for the sport,” adds Jianyang Zhou, a master's candidate in Electrical and Computer Engineering.

With ”March Madness Mastery,” the team envisions a future where basketball enthusiasts are equipped with the tools to navigate the thrill of tournament season with confidence and insight. As the countdown to tip-off commences, anticipation surges for the dawn of a new era in collegiate basketball analysis.

March Madness Mastery has achieved a milestone, and now it's your turn to immerse yourself in the excitement of predictive prowess with March Madness Mastery. Let the games commence!

## 2 FAQ

1. What data will you collect in training your predictive model?

   Our model utilizes 4 datasets (specified in the Dataset Introduction section) which are posted on the Kaggle competition platform. In summary, these datasets encompass information concerning the histories of past regular season games and MCNNA tournaments from 2003 to 2023 for both men's and women's divisions, facilitating the determination of the winning rate for each team per game. Additionally, we integrated an external dataset of 538 predictions into our features to enhance the accuracy of win/loss predictions.

   For further insights into the technical details, simulations, and preliminary results, please refer to the corresponding sections.

2. How does the predictive model work?

   Our predictive model utilizes a stack of predictive models with binary outcomes (logistic regression, random forest, SVM, etcs).

   Finally, we also build a Bayesian logistic regression model. It will help with hype tuning all the parameters for the model and possibly increase the accuracy of our base model.

   After the model is built, we use it to predict the 2024 regular season games as our preliminary results.

3. What measures are in place to prevent overfitting or underfitting of the predictive model?

   To prevent overfitting or underfitting of our predictive model, we implement several key measures. Firstly, we utilize cross-validation techniques like k-fold cross-validation to evaluate the model's performance on unseen data, ensuring its ability to generalize effectively. Ensemble methods like bagging is also employed to combine multiple models and reduce the risk of overfitting by averaging individual predictions. Finally, we balance model complexity and performance to strike an optimal balance, avoiding overly complex or overly simplistic models that may lead to either overfitting or underfitting. Through these measures, we aim to ensure that our predictive model delivers accurate and reliable results across various scenarios.

4. How do you handle uncertainty and variability in your predictions?

   To handle uncertainty and variability in bracket predictions, we employ advanced statistical techniques and machine learning algorithms that can adapt to changing circumstances. Additionally, we continually refine our model based on real-time data and user feedback, allowing us to account for unforeseen events and adjust predictions accordingly. While no model can predict with absolute certainty, our approach minimizes uncertainty and provides informed predictions for users.

5. Why does it really matter?

Forecasting tournament results is crucial as it boosts fan involvement, aids in bracket competitions, offers insightful analytics, fuels media attention, affects betting trends, and influences revenue. Such predictions drive innovation, investment, and expansion in broadcasting, merchandise, sponsorships, and ticketing within the sports sector. Precise forecasts draw larger audiences, enhance fan engagement, and enable stakeholders to leverage collegiate basketball tournaments' popularity, collectively adding to their thrill and significance.

6. Can users provide feedback on the accuracy of your predictions?

   Absolutely! We value user feedback immensely in refining and improving our predictive model. After each game or tournament outcome, users can rate the accuracy of our predictions based on their own observations and experiences. Additionally, we welcome users to share any discrepancies they may have noticed between our predictions and the actual outcomes. This feedback loop is crucial for the continuous improvement and refinement of our predictive model. By incorporating user insights and observations, we can adjust and fine-tune our algorithms to better reflect the dynamic nature of collegiate basketball tournaments. Ultimately, user feedback plays a vital role in ensuring that our predictive model remains accurate, reliable, and aligned with the expectations of our user community.

7. Can users interact with your model to explore ”what-if” scenarios?

   Indeed, our model possesses the capability to predict the outcome of any matchup between teams within a gameplay. While this feature is hypothetical in nature, it serves as a valuable reference point derived from the insights of our model. Users can rely on these predictions for guidance, understanding that they represent informed estimations rather than definitive outcomes.

8. Can users access the methodology and technical details behind your predictive model?

   Yes, users have full access to the methodology and technical details behind our predictive model. We believe in transparency and provide comprehensive information about our approach, including data sources, statistical techniques, and algorithms employed. Users can review this information to understand how our predictions are generated and make informed decisions based on their knowledge of our methodology.

9. Are there any limitations or constraints to the accuracy of your model?

   While our model excels in predicting tournament outcomes based on historical data, it's important to note some limitations. These include the dynamic nature of collegiate basketball, which can introduce variability due to factors like player injuries and coaching strategies, and star player absences. Additionally, unforeseen events can impact game results. Despite these constraints, our model provides valuable insights for understanding tournament dynamics.

10. How frequently is your model updated with new data?

    Our model is updated with new data on a continuous basis throughout the 2024 NCAA tournaments. As each round of games concludes for all pairs of teams, new data becomes available. Our model is designed to automatically incorporate this new data into a streaming version, allowing it to continually refine and update its accuracy predictions in real-time. This ensures that our predictions remain as current and precise as possible throughout the tournament proceedings.

11. How do you ensure transparency and accountability in your predictions?

    Ensuring transparency and accountability in our predictions is paramount. We achieve this through several measures. First, we disclose our methodology in detail, outlining the data sources, statistical techniques, and algorithms employed. We also track the performance of our predictions over time, providing users with transparent updates for accuracy. Furthermore, we actively encourage user feedback to address concerns and enhance our model's precision continuously. Maintaining open communication channels with our users and stakeholders allows for ongoing dialogue and updates on our predictive modeling process. Finally, we remain open to external review by independent experts or auditors, ensuring the integrity and reliability of our forecasts. Overall, our commitment to transparency and accountability fosters trust in our predictions, enabling informed decision-making for our users.

12. Are there plans to expand your predictive modeling beyond basketball tournaments?

    While our primary focus lies in predictive modeling for collegiate basketball tournaments through our project "March Madness Mastery," we are certainly open to extending our modeling techniques to other similar activities in the future. The principles of data analytics and machine learning that we employ are adaptable across various domains, including other sports tournaments and similar competitive events. Expanding our predictive modeling beyond basketball tournaments could involve adapting our algorithms and methodologies to analyze data from different activities such as tennis, soccer, chess, or esports competitions. By doing so, we aim to provide valuable insights and predictions to enhance the anticipation and understanding of a wider range of competitive events. Our team is committed to leveraging advanced analytics to unlock new possibilities and contribute to advancements in predictive modeling across diverse fields, ultimately empowering individuals to make informed decisions and engage more deeply with their favorite activities and events.

# 3 Dataset Overview

## 3.1 Dataset Introduction

In the March Machine Learning Mania 2024 competition hosted on Kaggle, multiple datasets are provided for the participants to construct features to build models. The datasets available for this project include:

- **MNCAATourneySeeds.csv**:

We use this file to obtain seed data for the teams participating in the tournaments. We also uses seeds as one of the predictors as they represent a team's ranking and expected performance in the tournament. Later in this project, we also leverage seed data to create a baseline prediction model. This model simply compares the seeds of the competing teams to generate a forecast. Essentially, this metric represents what might be considered an "educated" guess by someone who relies exclusively on seed rankings to predict outcomes.

- **M/WRegularSeasonDetailedResults.csv**

  These files contain detailed game-by-game results for the men's and women's regular seasons, respectively. They provide comprehensive statistics such as points scored, assists, rebounds, three-pointers made, and other metrics that are essential for creating detailed models to predict tournament outcomes. We uses most of the information in this data to construct the features for our predictive models.

- **M/WNCAATourneyCompactResults.csv**

  These files provide compact results of tournament games, specifically focusing on which team won or lost, which is used as the target variable in our prediction models. The simplicity of the win/loss outcome helps streamline the modeling process by focusing on the binary outcome of each match.

For detailed data frame construction, see Python codes in attachment.

At the conclusion of this project, we have also utilized the prediction results shared by a famous bracketing enthusiast to evaluate the value of our model. During our study of the NCAA bracketing culture, we noted that the former US. President Barack Obama is an enthusiastic participant who began sharing his bracket predictions annually after his first year in office in 2009. Considering his interest and engagement in bracket forecasting, we view him as a potential client of our product, under the premises that our model will demonstrate higher-than-average accuracy. This acknowledgment of his involvement highlights his relevance as a prospective end-user for our predictive tools.

## 3.2 Feature Construction

We construted our dataset in the following way: each row represent a tournament game, with "Season", "League", "TeamID", "OppTeamID" as the identifier. The column "Win" indicates if the team won or loss the game, which is our target variable. The remaining columns are quantitative features of both teams. These columns includes:

- **TeamScore, OppScore**:

  Points scored by the team and its opponent.

- **AvgScoreDiff, MedianScoreDiff, MinScoreDiff, MaxScoreDiff**:

  Statistics that reflect the team's scoring performance relative to their opponent over the season.

- **AvgThreePointerMade, AvgFreeThrowsAtp, AvgFreeThrowHitRate, etc**:

  Statistics related to the shots made and shots attempted that reflects both teams scoring ability and shots accuracies.

- **SeedNum, OppSeedNum, SeedNumDiff**:

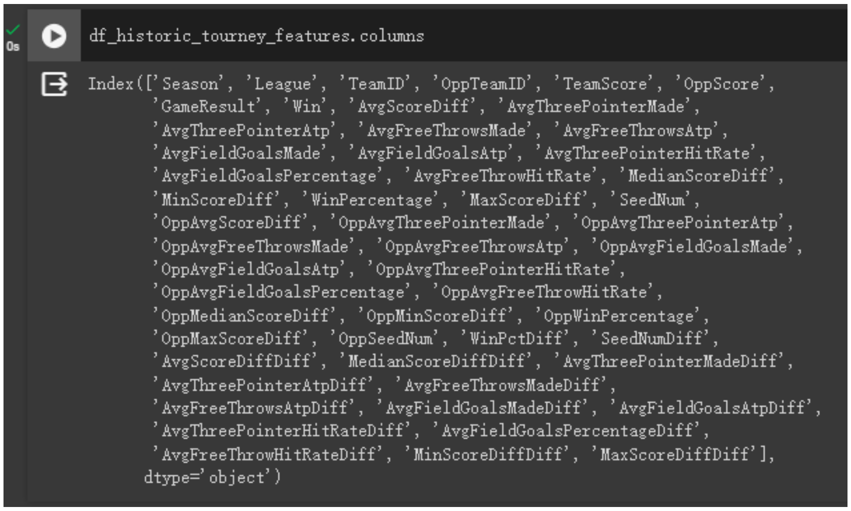
  Seeding numbers for the team and their opponents.

Taking the curse of dimensionality in consideration, we uses a selected set of the features stated above to build our machine learning model. We will now present a brief analysis of the selected features, which is particularly useful for the construction of our Bayesian prior distributions. The code for data extraction and feature construction is in the attachment.

# 4 Data Analysis

## 4.1 Feature Selection

After conducting an initial round of feature engineering, our dataset expanded to include over 40 features.



```
df_historic_tourney_features.columns
```

```
Index(['Season', 'League', 'TeamID', 'OppTeamID', 'TeamScore', 'OppScore',
       'GameResult', 'Win', 'AvgScoreDiff', 'AvgThreePointerMade',
       'AvgThreePointerAtp', 'AvgFreeThrowsMade', 'AvgFreeThrowsAtp',
       'AvgFieldGoalsMade', 'AvgFieldGoalsAtp', 'AvgThreePointerHitRate',
       'AvgFieldGoalsPercentage', 'AvgFreeThrowHitRate', 'MedianScoreDiff',
       'MinScoreDiff', 'WinPercentage', 'MaxScoreDiff', 'SeedNum',
       'OppAvgScoreDiff', 'OppAvgThreePointerMade', 'OppAvgThreePointerAtp',
       'OppAvgFreeThrowsMade', 'OppAvgFreeThrowsAtp', 'OppAvgFieldGoalsMade',
       'OppAvgFieldGoalsAtp', 'OppAvgThreePointerHitRate',
       'OppAvgFieldGoalsPercentage', 'OppAvgFreeThrowHitRate',
       'OppMedianScoreDiff', 'OppMinScoreDiff', 'OppWinPercentage',
       'OppMaxScoreDiff', 'OppSeedNum', 'WinPctDiff', 'SeedNumDiff',
       'AvgScoreDiffDiff', 'MedianScoreDiffDiff', 'AvgThreePointerMadeDiff',
       'AvgThreePointerAtpDiff', 'AvgFreeThrowsMadeDiff',
       'AvgFreeThrowsAtpDiff', 'AvgFieldGoalsMadeDiff', 'AvgFieldGoalsAtpDiff',
       'AvgThreePointerHitRateDiff', 'AvgFieldGoalsPercentageDiff',
       'AvgFreeThrowHitRateDiff', 'MinScoreDiffDiff', 'MaxScoreDiffDiff'],
      dtype='object')
```

Figure 1: Features Overview

However, it became apparent that not all these features contribute unique information due to interactions, as some were derived from others. For instance, the 'Three-Pointer Hit Rate' is a direct calculation from ’Three-Pointers Made’ divided by ’Three-Pointers Attempted’. In such cases, the hit rate alone may be enough to encapsulate the shot accuracy information of the team. Similarly, while metrics like ’Median Score Difference’, ’Maximum Score Difference’, and ’Minimum Score Difference’ provide insights into a team’s scoring variability, the ’Average Score Difference’ is a comprehensive feature that effectively summarizes a team’s overall scoring performance relative to their opponents. Therefore, we focused on features that offer distinct and direct insight into the outcome of games, streamlining our feature set to those that best capture the nuances of team performance and game dynamics. The features selected for our following analysis are: Win Percentage Difference, Seed Number Difference, Average Score Difference Difference, Average Three Pointer Hit Rate Difference[1], Average Field Goals Percentage Difference, and Average Free Throws Hit Rate Difference.

## 4.2 Visualization

We can construct a pair plot with all the variables of interest against each other.

In this plot, orange dots represent win, and blue dots represent loss. From the pair plot, there are several clear trends observed:

- Both the win percentage difference and seed number difference show a clear separation between wins and losses. This suggests that these features could be significant indicators of the game’s outcome.
- For average score difference difference, we see a small region of overlap, but there is a visible trend where a higher average score difference difference tends to be won games.
- For the scoring metrics like the average three-pointer hit rate difference, we do see more overlaps than the other features, as the separation between win and loss isn’t as clear as the others. However, the plots show distinct clusters and some plots show trends where higher-scoring performance correlates to more won games.

Overall, these visual patterns suggest that the features we have selected could be strong predictors of the outcome of a game.

[1] With one team member being an alumnus of Stephen Curry, we believe this statistic is indispensable.

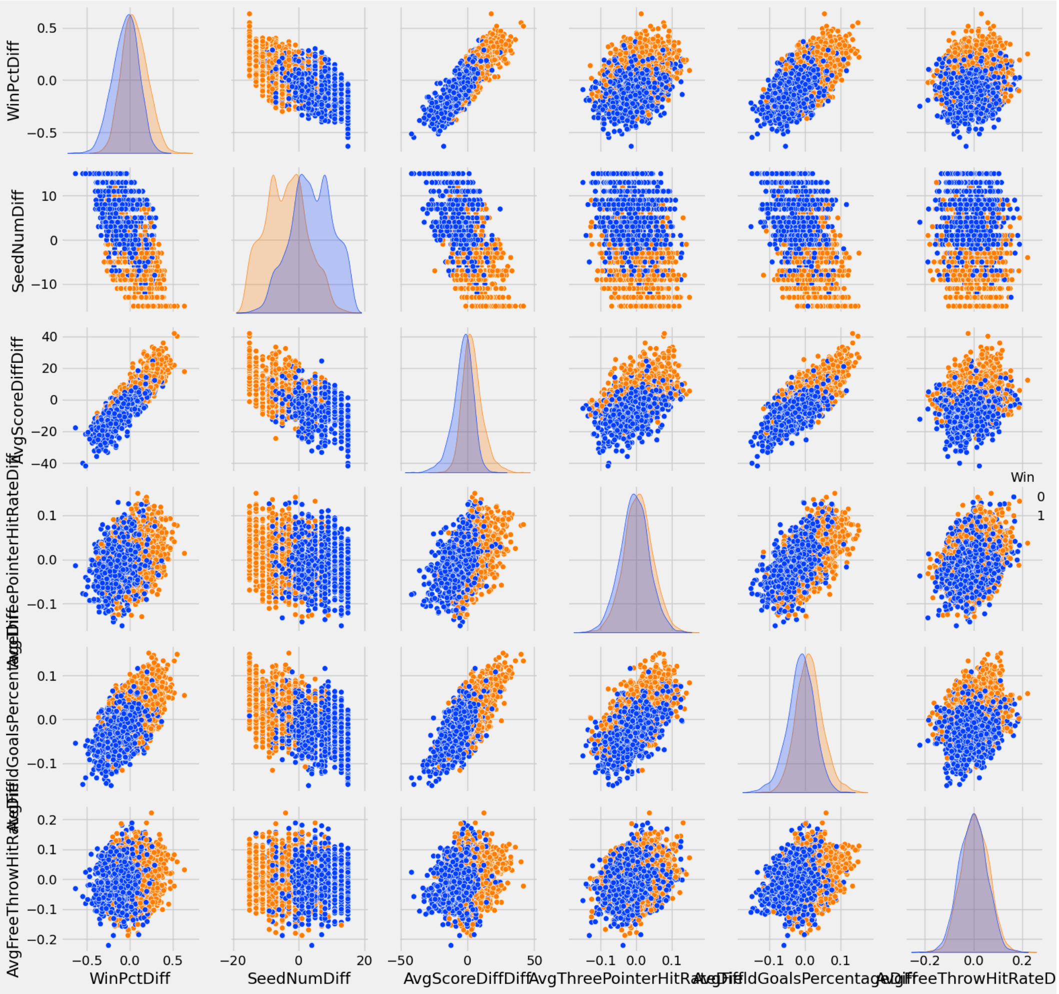


Figure 2: Pairs Plot

# 5 Modeling

## 5.1 Classical Predictive Models

The models were trained using a cross-validation "leave one season out," which is a variation of K-Fold cross-validation. In this method, the data for one entire season is held out as the test set while the models are trained on the remaining seasons. This process is repeated such that each season serves as the test set once. We can use this method to choose the important features and compare the model performance.

We first implement various machine learning algorithms, including general linear models, support vector machines, and ensemble models like XGBoost and random forests. As shown in the Table 1, most of machine learning methods have 0.734 accuracy.

Table 1: Model CV Accuracies

| Model | Average CV Accuracy |
|---|---:|
| LinearReg | 0.7317 |
| LassoReg | 0.7345 |
| RidgeReg | 0.7350 |
| LogitReg | 0.7352 |
| LinearSVC | 0.7353 |
| PolySVC | 0.7353 |
| RBFSVC | 0.7352 |
| XGBoosting | 0.7348 |
| Random Forest | 0.7326 |

## 5.2 Bayesian Logistic Regression

Bayesian logistic regression is favored in predictive modeling for its ability to incorporate prior knowledge and provide a probabilistic interpretation of model parameters. Unlike traditional logistic regression, which offers point estimates, Bayesian logistic regression calculates a full posterior distribution for each parameter. Additionally, Bayesian logistic regression can adeptly handle missing data and complex hierarchical structures, making it suitable for various levels of data aggregation.

In our case, the logistic regression formula is as follows:

$$\begin{aligned}\text{logit} = \beta_0 + &\beta_1\text{WinPctDiff} + \beta_2\text{SeedNumDiff} + \beta_3\text{AvgScoreDiffDiff} + \beta_4\text{AvgThreePointerHitRateDiff} \\ &+ \beta_5\text{AvgFieldGoalsPercentageDiff} + \beta_6\text{AvgFieldGoalsPercentageDiff}.\end{aligned}$$

Then, the probability estimate of the game outcome is:

$$P = \frac{1}{1 + e^{-\text{logit}}}.$$

From a Bayesian standpoint, our focus is to determine the probability of a team winning a game given the empirical information of the features, which is the conditional probability $P(\text{Win}|\text{Features})$. Instead of obtaining the maximum likelihood estimation, the Bayesian method finds appropriate prior distributions of the model's parameters $\boldsymbol{\beta}$.

In this project, we are dealing with binary win/loss outcomes. Hence, the sample likelihood function follows a Bernoulli distribution:

$$P(Win|Features) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i},$$

where $y_i = 1$ if the outcome is win and 0 is it is a loss, and $p_i$ is the probability of winning as previously defined.

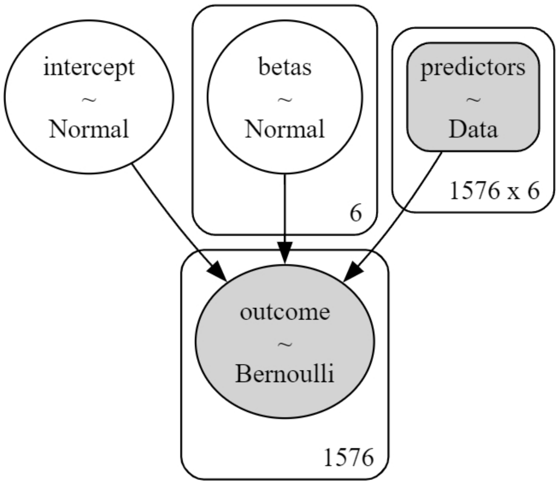


Figure 3: Generalized Linear Model

After we have the prior distribution of the parameters and the sampling likelihood function, our objective is to find the posterior probability distribution given the input features and observed outcomes.

$$P(\boldsymbol{\beta}|\boldsymbol{X}, y) = \frac{P(y|\boldsymbol{\beta}, \boldsymbol{X})P(\boldsymbol{\beta})}{P(y)}.$$

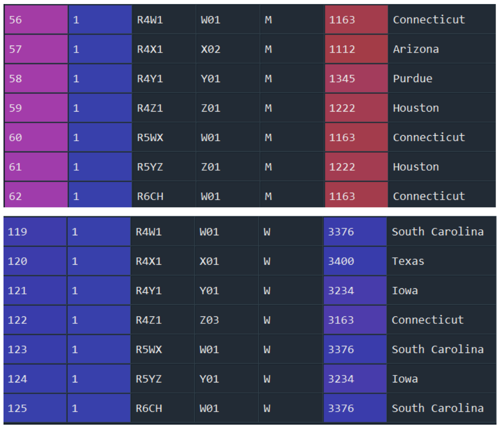
Using this framework, we construct and refine a Bayesian logistic regression model, which allows us to integrate prior knowledge and update our beliefs about the parameter's distribution as we receive more data from future games.

# 6 Model Simulation

Given the model above, we are able to simulate a bracket of game results for the 2024 NCAA tournament. Note that we have a lot of missing values in the feature SeedNumDiff (since most teams don't have seed numbers).

The file "2024 tourney seed.csv" includes the game schedule, and we output a table that includes the winner for each game (each row) in the column "TeamName".

The following table is a bracket of output of predictive result from our classical predictive models. It includes the predicted winner for each game.



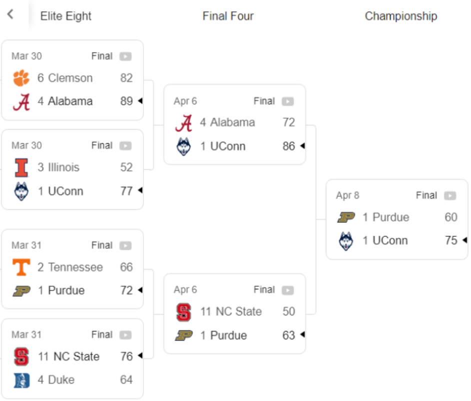
| | | | | | | |
|---|---|---|---|---|---|---|
| 56 | 1 | R4W1 | W01 | M | 1163 | Connecticut |
| 57 | 1 | R4X1 | X02 | M | 1112 | Arizona |
| 58 | 1 | R4Y1 | Y01 | M | 1345 | Purdue |
| 59 | 1 | R4Z1 | Z01 | M | 1222 | Houston |
| 60 | 1 | R5WX | W01 | M | 1163 | Connecticut |
| 61 | 1 | R5YZ | Z01 | M | 1222 | Houston |
| 62 | 1 | R6CH | W01 | M | 1163 | Connecticut |

| | | | | | | |
|---|---|---|---|---|---|---|
| 119 | 1 | R4W1 | W01 | W | 3376 | South Carolina |
| 120 | 1 | R4X1 | X01 | W | 3400 | Texas |
| 121 | 1 | R4Y1 | Y01 | W | 3234 | Iowa |
| 122 | 1 | R4Z1 | Z03 | W | 3163 | Connecticut |
| 123 | 1 | R5WX | W01 | W | 3376 | South Carolina |
| 124 | 1 | R5YZ | Y01 | W | 3234 | Iowa |
| 125 | 1 | R6CH | W01 | W | 3376 | South Carolina |

Figure 4: Simulation Results of Quarterfinals, Semifinals, and Final (Men on the top; Women at the bottom)

Note that we just include the simulated winners for Quarterfinals, Semifinals, and the Final (top for men, bottom for women). The column "Slots" is identified by a four-character string of the form R (round) (region) (chalk seed). For example, "R6CH" represents the final game for the champion. The column "TeamName" represents the winner for that game.

According to the classical predictive models, for men, it predicts that the first place is Connecticut, the second place is Houston, and the semifinalists are Purdue and Arizona. For women, it predicts that the first place is South Carolina, the second place is Iowa, and the semifinalists are Texas and Connecticut.



Elite Eight

Final Four

Championship

Mar 30 Final
6 Clemson 82
4 Alabama 89

Mar 30 Final
3 Illinois 52
1 UConn 77

Mar 31 Final
2 Tennessee 66
1 Purdue 72

Mar 31 Final
11 NC State 76
4 Duke 64

Apr 6 Final
4 Alabama 72
1 UConn 86

Apr 6 Final
11 NC State 50
1 Purdue 63

Apr 8 Final
1 Purdue 60
1 UConn 75

Figure 5: 2024 NCAA Men's Tournament Results

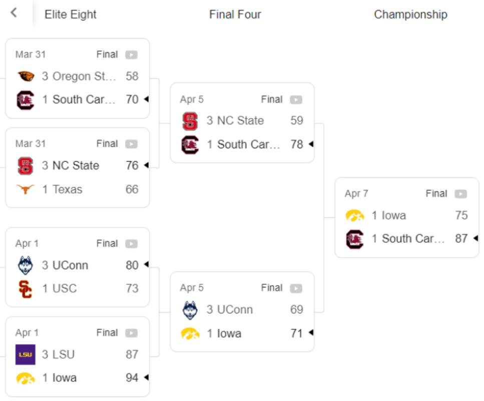


Figure 6: 2024 NCAA Women's Tournament Results

The above figures are the results of the 2024 NCAA Men's and Women's tournaments, respectively. The champions for the 2024 NCAA are Connecticut and South Carolina, respectively. We can see that at least our predictions for the champions in both the men's and women's divisions are correct!

Note that this is just 1 bracket of the simulation. We could add more brackets of predictions in the future (maybe 100,000 brackets).

# 7 Prediction

## 7.1 2024 NCAA Tournament Prediction

The complete dataset of the NCAA Tournament results is not yet available online. We employed the most sophisticated biological neural network known to science for visual data scraping from the NCAA website: our eyes and brains. To test the performance of our Bayesian model, we give a first-round prediction where the accuracy is about 0.71 for 2024 NCAA Tournament data.

## 7.2 Obama was pretty accurate, but not enough...

As previously mentioned, the former U.S. President, Barack Obama, is an enthusiast of the NCAA tournament and shares his bracket forecast on his website every year. This year, we took the opportunity to examine his prediction results in detail and calculate his accuracy.

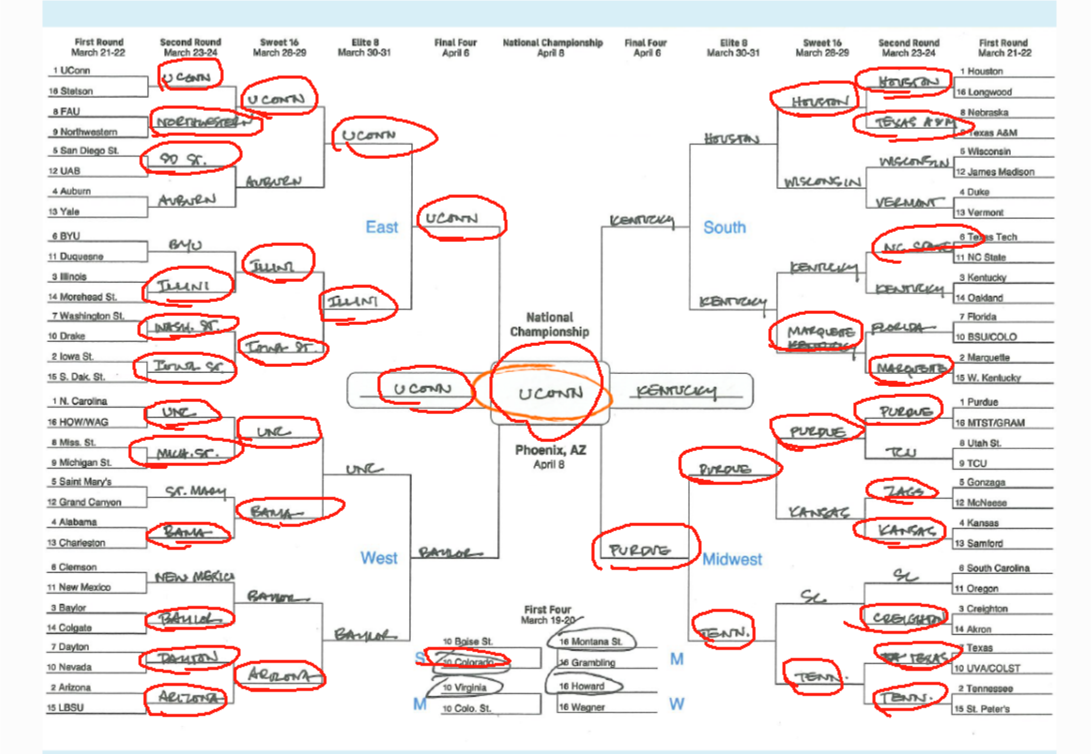


Figure 7: Obama's Prediction of 2024 Men's NCAA Tournament

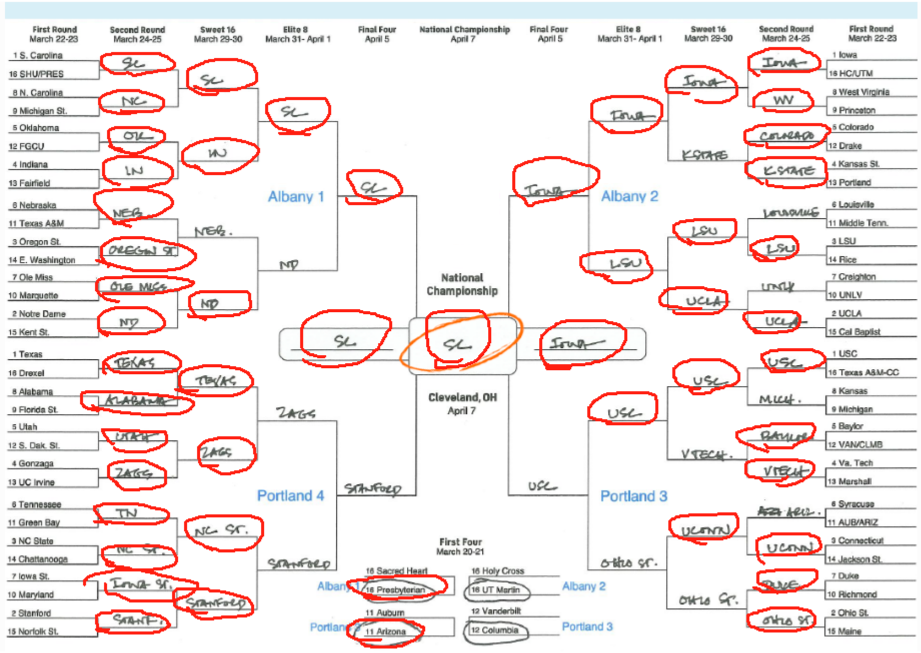


Figure 8: Obama's Prediction of 2024 Women's NCAA Tournament

After manually labeling the correct forecasts, we found out that President Obama's overall forecast precision this year stood at approximately 0.694, 6.8% lower than the baseline forecast, which relies solely on seed numbers. His forecast accuracy for the men's NCAA tournament stood at approximately 0.627, while his predictions for the women's NCAA tournament were significantly more precise, at around 0.761 this year. Given the strengths of our predictive model, which outperformed the baseline, we would be honored to see how it might enhance President Obama's bracket selections. With this in mind, we extend an invitation to President Obama to consider our model for his forecasts in the 2025 NCAA tournament.

# 8 Bibliography

1. March machine learning mania 2024. Kaggle. (n.d.). https://www.kaggle.com/competitions/march-machine-learning-mania-2024

2. Robikscube. (2024, March 10). machine learning bracket - GPU powered. Kaggle. https://www.kaggle.com/code/robikscube/machine-learning-bracket-gpu-powered

3. WillKoehrsen. (n.d.). Data-analysis/bayesian_log_reg/bayesian-logistic-regression.ipynb at master · Willkoehrsen/Data-analysis. GitHub. https://github.com/WillKoehrsen/Data Analysis/blob/master/bayesian_log_reg/Bayesian-Logistic-Regression.ipynb

4. NCAA.com. (2021, January 5). NCAA bracket for March madness. NCAA.com. https://www.ncaa.com/march-madness-live/bracket

5. Obama Foundation. (n.d.). President Obama's March Madness picks. https://www.obama.org/stories/march-madness-2024/