**Written Report on Predicting NCAA Tournament Success**

**Mathletes Consulting Group**

**Team Members:** Austin Lien, Sneh Patel, Reid Erickson, & Jake Burger

**1. Executive Summary**

This project aimed to develop a predictive model to forecast the success of the NCAA Division I Men's Basketball teams in the March Madness tournament. Success was defined as advancing to the Sweet Sixteen. By analyzing historical game data, team statistics, and performance metrics, we sought to identify patterns and factors significantly influencing team performance.

After finding data about March Madness from 2008-2023 and preparing the data for the models, we built a logistic regression, gradient boosting, and random forest models. Logistic regression, with balanced precision and recall metrics, was chosen as the best-performing model. It showed high accuracy in predicting both true negative and true positive occurrences while both gradient boosting and random forest showed signs of possible overfitting.

Exploring key performance metrics like Offensive and Defensive Efficiency, and Efficiency Margin we found they significantly influence Sweet Sixteen success. Less impactful factors include Adjusted Tempo and Defensive Rebounds. Our findings suggest prioritizing core efficiency metrics can enhance March Madness betting strategies. However, external factors like injuries and coaching changes may impact predictions.

Future work may include incorporating additional data sources and segmenting data for more accurate predictions. In conclusion, our project provides valuable insights for sportsbook companies aiming to optimize March Madness betting strategies, enhancing operational efficiency and profitability.

**2. Problem Description**

March Madness, officially known as the NCAA Division I Men’s Basketball Tournament, is an annual sporting event that captivates millions of fans through its unpredictable nature and thrilling matchups. Each spring, 68 college basketball teams from across the United States compete in a single elimination tournament to determine the national champion. The complexity and high stakes of the tournament provided grounds for developing sophisticated models that can predict various outcomes, crucial for stakeholders such as fans, analysts, and sports betting companies.

**2.1 Business Goal**

The primary business goal of this project is to enhance the decision-making process for sportsbook companies by providing them with accurate predictions for March Madness outcomes. This can significantly improve the formulation of betting odds, enhance strategic planning for engagements and promotions around the tournament and increase overall customer satisfaction with betting platforms.

**2.2 Data Mining Goal**

The data mining goal is to develop a robust classification model capable of predicting whether NCAA basketball teams will advance to the Sweet Sixteen phase of the tournament- the final 16 teams. By leveraging historical data and detailed team performance statistics, the model aims to identify that most influential factors that contribute to a team’s tournament success. This predictive insight will enable our clients to make more informed decisions, backed by data-driven analyses, thereby potentially increasing their operational efficiency and profitability in the highly competitive sports betting market.

**3. Data Description**

**3.1 Data**

Our dataset, obtained from Kaggle (<https://www.kaggle.com/datasets/danieljohnsen/college-basketball-torvik-and-kenpom-data?select=Torvik.csv> ), includes 4159 rows (individual team seasons) and 30 columns (features) for NCAA basketball teams participating in March Madness tournaments from 2008 to 2023. Features encompass key performance metrics like Wins Above Bubble (WAB), field goal percentages, and advanced statistics like Strength of Schedule (SOS) with KenPom stats. Additionally, it delves into opponent-related factors (e.g., opponent field goal efficiency). Our target variable for this dataset was the placement of the team in the Sweet Sixteen of the tournaments, 1 for they made it to the Sweet Sixteen or farther and 0 for they did not make the Sweet Sixteen.

**3.2 Exploratory Data Analysis (EDA)**

Our Exploratory Data Analysis (EDA) aimed to identify factors influencing a team's success in the March Madness tournament, specifically focusing on their likelihood of reaching the Sweet Sixteen.

A key finding was the significant impact of core efficiency metrics on a team's success. Teams with higher core efficiency metrics (Offensive Efficiency, Defensive Efficiency, and Efficiency Margin) exhibited a significantly higher frequency of reaching the Sweet Sixteen (as shown in Figures 2 & 4). This suggests that these metrics hold significant weight in predicting success and should be highly considered when setting betting lines.

We also investigated factors with a potentially lower influence on success. Analyses of Adjusted Tempo (Figure 3) and Defensive Rebounds (Figure 5) revealed minimal differences between teams that made the Sweet Sixteen and those that did not. These findings suggest that Tempo and Defensive Rebounds may not be strong differentiating factors for predicting Sweet Sixteen appearances and may not warrant significant influence on betting lines.

**3.3 Data Pre-processing**

A crucial step in our data preparation was converting the target variable, "Sweet Sixteen," into a binary format (0 or 1). This transformation facilitates the use of logistic regression modeling for our analysis. We also removed variables using the select columns widget in Orange such as Year, Unnamed and Luck. Finally, we removed Seed as a feature to avoid any data leakage into our modeling as having a seed number indicates a team has already made it into the tournament.

**4. Data Mining Solution**

To determine which model works best to predict whether or not a team will make the Sweet Sixteen, we decided to build three different types of models to ensure that we pick the highest performing model. We built a logistic regression, gradient boosting, and random forest model.

**4.1 Models**

When building the models, the target variable was set to the Sweet Sixteen feature that was binary coded during the preprocessing stage. We used an 80/20 split for training and testing data for all three models to ensure the same bounds were set and the data was equal for all models. The logistic regression model was a simple regression model built to analyze the binary feature, Sweet Sixteen. When building the gradient boosting model, the parameters were set 75 trees with a learning rate of 0.05 (Figure 7). For the random forest model, we used 200 trees with 5 attributes at each split (Figure 8). To determine which model best works for our problem we looked at the error metrics for all three models along with the confusion matrices.

**4.2 Performance Evaluation**

Looking at the error metrics for all three models we can see that the AUC is very high for all models in training and cross validation (Figures 9 and 10). However, for gradient boosting and random forest we can see that the training data AUC is 0.989 and 0.994 respectively which is almost one indicating that these two models could be possibly overfitting and therefore cannot be the best model to choose. Although the precision and recall are lower for logistic regression compared to both gradient boosting and random forest, the values are not too far apart. To avoid any problems with overfitting, we chose logistic regression to be the best model which can be backed up by looking at the confusion matrices for the models.

When examining the confusion matrices for the models (Figures 11, 12, and 13), we can see that the confusion matrix for logistic regression has the highest rate of predicting true negative occurrences and is tied with gradient boosting for predicting the highest rate of true positive occurrences. True negative is our problem refers to the model predicting a team will not make the sweet sixteen and the team not making the sweet sixteen, while true positive refers to the model predicting the team will make it to the sweet sixteen and the team did make it. The goal of this model is to determine with the highest success which teams will make the sweet sixteen and which teams will not so, having a high true negative and high true positive are good indicators for the model. Additionally, we can see in all three models there are instances of false positives and false negatives with the logistic regression model having the lowest number of false positives and false negatives. This is bound to happen with a large dataset as there will be outliers and as we see every year in March Madness there are unexpected upsets and “cinderella” teams that unexpectedly win.

**5. Conclusion**

**5.1 Recommendations**

This project has yielded a model with the potential to inform betting strategies and game predictions in March Madness. By analyzing core efficiency metrics like Adjusted Offensive and Defensive Efficiency, as well as Efficiency Margin for Strength of Schedule, the model highlights statistically significant factors that can influence a team's success in reaching the Sweet Sixteen. Betting strategies can be enhanced by prioritizing these high-influence metrics while placing less emphasis on factors with lower predictive power, such as Adjusted Tempo and Defensive Rebounds (as shown in Figures 3 & 5).

For continued refinement and broader applicability, we recommend incorporating additional data sources in future iterations of the model. Exploring the impact of factors like player injuries, coaching changes, and other external influences through further research could significantly improve the model's predictive accuracy.

**5.2 Limitations**

When running these models, you can still run into some limitations that are not factored into the model. The model’s capabilities are constrained by the scope of the historical data used for training, which thereby imposes limitations on the predictive accuracy and adaptability of the model. Also, while statistical analysis can provide valuable insights, it is also important to recognize that external factors such as injuries can have a major impact on the outcome and can make predictions difficult to gauge.

**5.3 Future Work**

Future work on this model can involve incorporating additional data columns. This could include injuries, coaching changes, and other currently omitted factors that might influence performance. Additionally, the data could be segmented to isolate teams that typically qualify for the tournament. This would allow for more practical application in setting betting lines before the tournament selection is finalized.

**Appendix**

**Figure 1: Data Dictionary for March Madness**

|  |  |  |  |
| --- | --- | --- | --- |
| Field Name | Data Type | Description | Units/Calculation |
| TeamNames | String | The name of the team | N/A |
| Year | Integer | The year of the NCAA tournament season | N/A |
| Seed | Integer | The team's seeding in the NCAA tournament | Tournament Seed (1-16) |
| Sweet.Sixteen | Integer | Indicates if the team reached the Sweet Sixteen round of the tournament (1 = Yes, 0 = No) | Binary Indicator |
| ADJOE | Float | Adjusted Offensive Efficiency (points scored per 100 possessions, adjusted for various factors) | Points per 100 possessions |
| ADJDE | Float | Adjusted Defensive Efficiency (points allowed per 100 possessions, adjusted for various factors) | Points per 100 possessions |
| efg. | Float | Effective Field Goal Percentage (shooting efficiency accounting for the value of 3-pointers) | Percentage |
| efgd. | Float | Effective Field Goal Percentage Defense (opponent's eFG%) | Percentage |
| tor | Float | Turnover Rate (turnovers per 100 possessions) | Turnovers per 100 possessions |
| tord | Float | Opponent Turnover Rate (turnovers forced per 100 possessions) | Turnovers per 100 possessions |
| orb | Float | Offensive Rebound Rate (percentage of available offensive rebounds secured) | Percentage |
| drb | Float | Defensive Rebound Rate (percentage of available defensive rebounds secured) | Percentage |
| FTR | Float | Free Throw Rate (free throws attempted per field goal attempt) | Ratio |
| FTRd | Float | Opponent Free Throw Rate | Ratio |
| X2p. | Float | 2-Point Field Goal Attempt Rate (percentage of field goal attempts that are 2-pointers) | Percentage |
| X2pd. | Float | Opponent's 2-Point Field Goal Attempt Rate | Percentage |
| X3p. | Float | 3-Point Field Goal Attempt Rate (percentage of field goal attempts that are 3-pointers) | Percentage |
| X3pd. | Float | Opponent's 3-Point Field Goal Attempt Rate | Percentage |
| adjt | Float | Adjusted Tempo (estimated possessions per 40 minutes, adjusted for pace) | Possessions per 40 minutes |
| WAB | Float | Wins Above Bubble (estimates wins contributed above a "bubble" tournament team) | Estimated Wins |
| [AdjEM.KP](http://AdjEM.KP) | Float | KenPom Adjusted Efficiency Margin (point differential per 100 possessions, adjusted) | Points per 100 possessions |
| [AdjO.KP](http://AdjO.KP) | Float | KenPom Adjusted Offensive Efficiency | Points per 100 possessions |
| [AdjD.KP](http://AdjD.KP) | Float | KenPom Adjusted Defensive Efficiency | Points per 100 possessions |
| [AdjT.KP](http://AdjT.KP) | Float | KenPom Adjusted Tempo | Possessions per 40 minutes |
| [Luck.KP](http://Luck.KP) | Float | KenPom Luck estimate | Estimate |
| [AdjEM.SOS.KP](http://AdjEM.SOS.KP) | Float | Efficiency Margin adjusted for Strength of Schedule | Points per 100 possessions |
| [AdjO.SOS.KP](http://AdjO.SOS.KP) | Float | Offensive Efficiency adjusted for Strength of Schedule | Points per 100 possessions |
| [AdjD.SOS.KP](http://AdjD.SOS.KP) | Float | Defensive Efficiency adjusted for Strength of Schedule | Points per 100 possessions |
| [AdjEM.NCSOS.KP](http://AdjEM.NCSOS.KP) | Float | Efficiency Margin adjusted for Non-Conference Strength of Schedule | Points per 100 possessions |
|  |  |  |  |

**Figure 2: Offensive Efficiency for strength of schedule and frequency of Sweet Adjusted Sixteen appearances.**

A diagram of a normal distribution

Description automatically generated

**Figure 3: Adjusted KenPom Tempo and the frequency of making the Sweet Sixteen.**

A diagram of a normal distribution

Description automatically generated

**Figure 4: Adjusted Defense vs. Adjusted Efficiency Margin.**

A blue and red pixelated object

Description automatically generated

**Figure 5: Boxplot of Defensive Rebounds per game for Teams that make and do not make the Sweet Sixteen.**

A diagram of a graph

Description automatically generated with medium confidence

**Figure 6: P values of all variables to determine statistical significance calculated in R**

A screenshot of a computer

Description automatically generated

**Figure 7: Gradient Boosting Parameters**

**A screenshot of a computer

Description automatically generated**

**Figure 8: Random Forest Parameters**

**A screenshot of a computer

Description automatically generated**

**Figure 9: CV with 10 Folds Error Metrics**

**A screenshot of a computer

Description automatically generated**

**Figure 10: Training Data Error Metrics**

**A screenshot of a computer

Description automatically generated**

**Figure 11: Logistic Regression Confusion Matrix**

**A screenshot of a graph

Description automatically generated**

**Figure 12: Gradient Boosting Confusion Matrix**

**A screenshot of a graph

Description automatically generated**

**Figure 13: Random Forest Confusion Matrix**

**A screenshot of a computer

Description automatically generated**