March Madness Report

Marketing Analytics Section 1

Team 6

Aditi, Matt, Paras, Pragya, Shreya

Introduction

March Madness is a college level Division I basketball single-elimination competition played in the month of March. March Madness started in 1939. Single-elimination means once a team plays and loses once they are out for the rest of the tournament. Teams play against each other based on their seed[1]. Teams are split according to region; the four regions are South, East, West, and Midwest. Then in each region they are ranked from one through sixteen. There are a total of sixty-four teams that participate in March Madness and every year the teams change depending on which teams thrived over the regular season and postseason. The winning university receives a rectangular, gold-plated trophy made of wood.

Every year people around the country participate in creating brackets on who they believe will win each game over the span of the month. This year alone there were over 22 million brackets created. The chances of getting a perfect bracket vary from one in 128 billion and one in nineteen quintillion. Both women's and men's division one participate in March Madness, however we only looked at the data of Men's March Madness. The goal of our research was to determine which team would most likely win March Madness. It is statistically unlikely to construct a flawless bracket for March Madness because of the unpredictable nature of college basketball, the single-elimination system, and the possibility of upsets. Although it's established that getting a perfect bracket is nearly impossible, we wanted to see what factors led to a better team whether it is with offense, defense, height or other variables.

Data Description

We used a Kaggle dataset that comes from an output dataset that evaluates the March Madness Data Analysis dashboard in Domo. This is the link to the dataset:

[1] A seed is determined by the ranking the team has according to the games played over the span of the season.

https://www.kaggle.com/datasets/jonathanpilafas/2024-march-madness-statistical-analysis. This dataset analyzes team performances during the March Madness seasons from 2002 to 2024. Every NCAA Division 1 men's basketball team is represented by a separate line item in each season. The dataset consists of 144 columns. A few of them are:TeamName**,** Effective Field Goal Percentage Rate (eFGPct) and its ranking (RankeFGPct); Turnover Percentage(TOPct) and its ranking (RankTOPct); Offensive Rebound Percentage (ORPct) and its ranking (RankORPct);Free Throw Rate (FTRate) and its ranking (RankFTRate). For each of the methodologies used, different variables were analyzed to evaluate team performance during the season.

Methodology #1 - Cross Tabulation Analysis

The first methodology we will be diving into is our cross tab analysis. Here, we have chosen to conduct this methodology via Excel instead of R as our data was not best synthesized from a coding standpoint.

To briefly walk through our analysis step by step, we will begin by downloading the appropriate dataset from Kaggle which can be accessed through this link[2]. This dataset will create a zip file consisting of approximately thirteen csv files. We were able to narrow down which specific CSV file would be appropriate based on our hypothesis. Just to reiterate, what our team is hoping to prove with these various methodologies is whether we can determine who is most likely to win March Madness. Therefore the CSV file in question that we ending up choosing for our Cross Tab Analysis is “INT _ KenPom _ Efficiency.csv”. This efficiency csv file was important in providing us with two vital categories - “Adjusted Offensive Efficiency Rank” and “Avg Possession Length (Offense)”. These two categories in conjunction allow us to

[2] https://www.kaggle.com/datasets/jonathanpilafas/2024-march-madness-statistical-analysis

gather insights into a team's offensive strategy, stability, efficiency, and potential areas for improvement.

After we have downloaded our csv file and identified the two respective columns deemed necessary for comparison, we created a new sheet with a consolidated table of just the two columns for the year 2024 to filter out any irrelevant data. As mentioned earlier the outputs of our data are not synthesized in the way that they can be grouped together. So as a result, we had to manually create our own intervals and group the data accordingly. In terms of "Adjusted Offensive Efficiency Rank," teams were grouped in intervals of ~100 and "Avg Possession Length (Offense)" were grouped in intervals of ~1. The table was color coded by these separated groups to utilize the count function and manually build our cross tab analysis. Please refer to Figure A in the appendix.

Now that we have created our manual cross tab analysis, it is important to use this information to dive deeper into our understanding of the data. We can use this table to interpret relationships and the two we are focusing on will be 1) Longer position lengths with high offensive efficiency and 2) Short position lengths with high offensive efficiency. The reason that we have identified these two relationships is because it shares which teams had a deliberate offensive strategy with efficient scoring and teams that had quick offensive strategies with efficient scoring respectively. In summation, we are looking for teams that, despite holding a position for a short or long amount of time, which teams were able to take advantage of the offensive opportunities that arose.

Methodology #2 - Classification

Neural network classification analysis can be very helpful in March Madness basketball analysis. Neural networks are strong models for machine learning that can identify various data patterns to help analyze player performance or game outcome prediction. Neural network classification analysis can estimate which team is likely to win a given contest by utilizing past data on players, teams, and game statistics. In order to train the neural network to predict the outcome of each game, data including team rankings, shooting percentages, turnovers, and defensive numbers must be fed into it. I used classification analysis and performed a neural network tes to find out the correlation between the target variable (`TOPct`) and predictors (`eFGPct`, `TOPct`, `ORPct`, `FTRate`). I opened the zip file and selected "INT _ KenPom _ Defense.csv" from Kaggle.Using the `neuralnet` function from the `neuralnet` package, a neural network model was constructed. Using a logistic activation function for classification, the model comprises two hidden layers with six and four neurons each. The input layer has 4 neurons corresponding to the 4 predictors in the dataset. The first hidden layer has 10 neurons, second hidden layer has 6 neurons, and the third hidden layer has 4 neurons. The output layer has a single neuron since it's a binary classification problem. The output of the neural network model resulted as 0.29% shows that the low accuracy indicates that there's significant room for improvement in the model's predictive performance for the "TOPct" target variable.

The classification analysis for defensive efficiency shows that the effective field goal percentage (Opponent_eFGPct) of the opponent: a lower Opponent_eFGPct indicates stronger defense in minimizing the opponent's scoring effectiveness. Turnover Percentage (Opponent_TOPct) of the Opponent: A higher Opponent_TOPct shows more potent defensive pressure and the capacity to force turnovers. Defensive Rebound Percentage (DRPct): A higher DRPct shows a stronger capacity to shut down opponents' second-chance opportunities and

secure defensive rebounds. In all, this test helps provide an insightful analysis of a team's performance by highlighting key factors which affect victory and can help predict a team's performance in future games.

Methodology #3 - Brand Conceptual Map

To become the winner within the tournament, teams want to figure out what is the most important stat to focus on to give them the best chance of winning. When looking at basketball, winning becomes a need for strategy, skill, and athleticism. We made a theory that we can figure out the defining stats for a basketball team that makes them the favorite to win their game. We found that field goal percentage (eFGPct) and offensive rebound percentage (ORPct) are two very important stats that most successful basketball teams dominated in. When looking at the recent NCAA tournament the two finalists, Purdue and UConn, both dominated in the two stats. Effective field goal percentage (eFGPct) is typically the main aspect when looking at a team's shooting efficiency. The eFGPct is a record for three-point shots, which gives a better understanding of how well a team scores. In our analysis of the top 16 teams in this year's NCAA tournament, it is clear that eFGPct is one of the most important stats to determine a potential good team as shown not just in the two finalists but also most teams that were in the top 16. Having high eFGPct shows how adept teams are in converting scoring opportunities and highlights the strategic decisions of each player in their shot. Another stat we found that was important in determining the offensive efficiency a team has is their offensive rebound percentage (ORPct). In a basketball match, getting offensive rebounds show important opportunities to have a chance to score again which can overall determine the outcome of the game. Both UConn and Purdue showed dominance in this statistic which proves the importance of ORPct. By taking advantage of a rebound, teams can increase their potential in scoring and

gaining the lead. The relationship between high eFGPct and ORPct that the two finalists show, highlights the importance of these stats within any match in basketball. When looking at these stats further than just numerical values, these metrics are clear examples of strategy and skill that show athletic success. The data shows that we can get a deeper understanding of the dynamics of basketball, revealing the relationship between certain stats and performance on-court. This study can help aid coaches and analysts understand where they should focus on building their teams and players. Analyzing the key statistics and the influence that effective field goal percentage (eFGPct) and offensive rebound percentage (ORPct) has on a game can help teams build around those key stats and improve to have a better chance at the championship. UConn has gone to the tournament two times in a row, and we can identify where they excel at so teams can work on either combating against eFGPCT and ORPct or even adapt their strategy to improve their own way of playing. Overall it is clear what stats both UConn and Purdue dominated in this year, and because of this, players and coaches alike can work to improve for their own success.

Methodology #4 - Factor Analysis

When attacking the hypothesis to decide what teams would win March Madness, I wanted to know which variables I should look at that would display the data the best. I realized the best way for me to do this would be to find the eigenvalues and perform a factor analysis. I wanted to search to see how many eigenvalues the summary.csv dataset had and the variance of those variables. Eigenvalues tell you the amount of variance in the observed values that is accounted for by each extra factor. There were three eigenvalues in the summary.csv in our dataset. Factor analysis helps to deal with data sets, where there are a large number of underlying dimensions that are thought to reflect a smaller number of observed variables. I

performed the Factor Analysis to explain the correlation between the observed values which in our case was defensive efficiency, offensive efficiency, and tempo. Offensive efficiency accounts for 32.7% of the total variance, defensive efficiency accounts for 28.5% of variance, and factor 3 accounts for 22.9% of the variance. All together, offensive efficiency, defensive efficiency, and temp account for 84% of the total variance, meaning that they capture a significant portion of the variability of the data, as you can see in **Figure B**[3]. Overall, I feel pretty confident about looking at these three variables when deciding what teams would make it to the final and would win since they explain a wide section of our data. These three factors explain a significant amount about teams that would thrive since the chi square statistic of 206,483 is large and the p-value is 0, which is less than 0.5. Offensive efficiency, defensive efficiency, and tempo explain a large significant part of our data in order for us to determine where to look when selecting a team to win over another in March Madness.

Methodology #5 - Regression

When studying basketball, regression analysis is crucial for finding correlations between players attributes and team performance metrics. Finding these correlations can allow coaches to better select what players to play for their team. The regression analysis that I used highlights the correlation between the height of the center and the average possession length of the offense. Our knowledge of player performance is further enhanced by the addition of average possession length (offense), which emphasizes the complex relationship between players' contributions to team success. My analysis in R which is in **Figure C** resulted in a computed p-value of 0.03 which indicates statistical significance between Center Height and Average Possession Length (Offense). This finding suggests a significant link between these two variables and offers strong

[3] In the appendix

evidence to reject the null hypothesis. As such, it forces one to reconsider received wisdom and provides a means of investigating the variables behind this correlation in more detail.

Strong evidence of the correlation between average possession length and center height can be found in the regression result. The analysis output which can be found in **Figure D** shows statistical significance with a coefficient of -0.10112 and a p-value of 0.0345, indicating that taller centers typically result with shorter holdings. The practical result of this is the idea that taller centers can lead to more points scored which can shorten possessions on offense. This research goes against the common belief that shorter players tend to have longer possessions because of their perceived quickness and ability to create plays. Regression analysis offers insightful information to coaches that is not seen by simply looking at the stats. Whether their centers are nimble playmakers or towering rim protectors, coaches can modify their offensive plans to take use of these qualities. It could be necessary to adjust player assessment measures to take into consideration the varied contributions that centers of different heights make to offensive efficiency. Insights from this correlation can also be used to guide strategic changes made during games, enabling teams to maximize offensive output.

Summary

In summation, we utilized a total of five methodologies to provide more insights and an accurate understanding of team performance and outcomes. With the data gathered, we are able to examine statistical factors and historical trends to make informed predictions. This predictive power in turn helps us to enable which teams can better allocate resources more effectively and optimize game plans based on opponent strengths and weaknesses, therefore having a competitive advantage.

Appendix

Cross Tab Analysis
Figure A:

Average Position Length (Offense)

Adjusted Offensive Efficiency Rank

	14.01-15	15.01-16	16.01-17	17.01-18	18.01-19	19.01-20+	Total
1-100	2	13	27	35	20	3	**100**
101-200	2	3	20	38	29	7	**99**
201-300	1	6	20	44	20	9	**100**
301+	0	2	11	23	19	7	**62**
Total	**5**	**24**	**78**	**140**	**88**	**26**	**361**

Factor Analysis Code:

```
factor.bball=read.csv("Desktop/INT _ KenPom _ Summary.csv",header=TRUE)
nScree(factor.bball[,3:16]) #estimate the number of factors from scree tests.
eigen(cor(factor.bball[,3:16])) #selection heuristics - eigenvalue 1
start_values <- matrix(0, nrow = ncol(factor.bball[,3:16]), ncol = 3)  # Assuming 3 factors
colnames(start_values) <- c("Factor1", "Factor2", "Factor3")
factanal(factor.bball[,3:16], factors = 3, start = start_values)
```

Figure B: Factor Analysis Output:

```
eigen() decomposition
$values
 [1] 6.942512e+00 3.495955e+00 1.955780e+00 8.613557e-01 2.667694e-01 1.962121e-01 1.255411e-01
 [8] 5.399948e-02 4.509336e-02 3.513267e-02 9.443543e-03 7.126122e-03 5.078565e-03 4.403322e-10

$vectors
```

```
Uniquenesses:
      Tempo    RankTempo     AdjTempo RankAdjTempo           OE       RankOE        AdjOE
      0.464        0.039        0.264        0.043        0.321        0.209        0.005
 RankAdjOE           DE       RankDE        AdjDE    RankAdjDE        AdjEM    RankAdjEM
      0.081        0.449        0.262        0.005        0.053        0.005        0.037

Loadings:
             Factor1 Factor2 Factor3
Tempo                        -0.731
RankTempo                     0.978
AdjTempo                     -0.856
RankAdjTempo                  0.976
OE           -0.820
RankOE        0.876   0.152
AdjOE        -0.974  -0.216
RankAdjOE     0.914   0.288
DE                    0.738
RankDE        0.131   0.846
AdjDE         0.298   0.949
RankAdjDE     0.339   0.908
AdjEM        -0.763  -0.644
RankAdjEM     0.733   0.651

               Factor1 Factor2 Factor3
SS loadings      4.573   3.992   3.202
Proportion Var   0.327   0.285   0.229
Cumulative Var   0.327   0.612   0.840

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 206483 on 52 degrees of freedom.
The p-value is 0
```

Figure C: Regression code

```
lm_model <- lm(`Offensive Rebounds` ~ `Center Height`, data = basketball_data)

summary(lm_model)

# Null hypothesis: There is no correlation between Center Height and Offensive Rebounds
# Alternate Hypothesis: There is correlation between Center Height and Offensive Rebounds
```

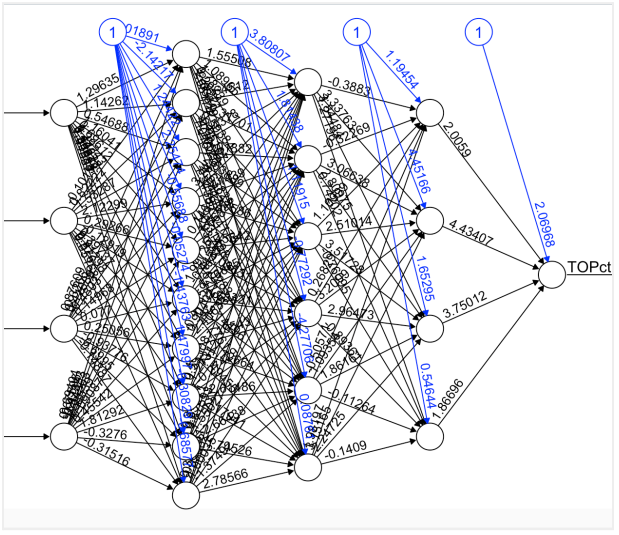
Figure D: Regression R Output

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.53231    0.05760 304.389   <2e-16 ***
CenterHeight -0.10112    0.04766  -2.122   0.0345 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.096 on 360 degrees of freedom
Multiple R-squared:  0.01235,           Adjusted R-squared:  0.009608
F-statistic: 4.502 on 1 and 360 DF,  p-value: 0.03453
```

Classification Neural Network Code:

```
library(readr)
```

```
INT_KenPom_Defense <- read_csv("Downloads/INT _ KenPom _ Defense.csv")
View(INT_KenPom_Defense)

library(neuralnet);
library(readr)
INT_KenPom_Defense <- as.data.frame(INT_KenPom_Defense)

# Specify the target variable and predictors
target <- "TOPct"
predictors <- c("eFGPct", "TOPct", "ORPct", "FTRate")

formula_str <- paste(target, "~", paste(predictors, collapse = " + "))
nn_model <- neuralnet(
  formula_str,
  data = INT_KenPom_Defense,
  hidden = c(6, 4),  # Two hidden layers with 5 and 3 neurons respectively
  linear.output = FALSE,  # For classification
  threshold = 0.5,  # Threshold for the threshold function
  act.fct = "logistic",  # Activation function
  lifesign = "minimal"
)

# Display the neural network model summary
print(nn_model)

# Predict on the entire dataset
predictions <- predict(nn_model, INT_KenPom_Defense[,
predictors])

# Convert predictions to class labels
predicted_classes <- ifelse(predictions > 0.5, 1, 0)

# Evaluate the model
confusion_matrix <- table(predicted_classes,
INT_KenPom_Defense$RankeFGPct)
print("Confusion Matrix:")
print(confusion_matrix)

# Calculate accuracy
accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
print(paste("Accuracy:", round(accuracy * 100, 2), "%"))
print (nm_model)
```

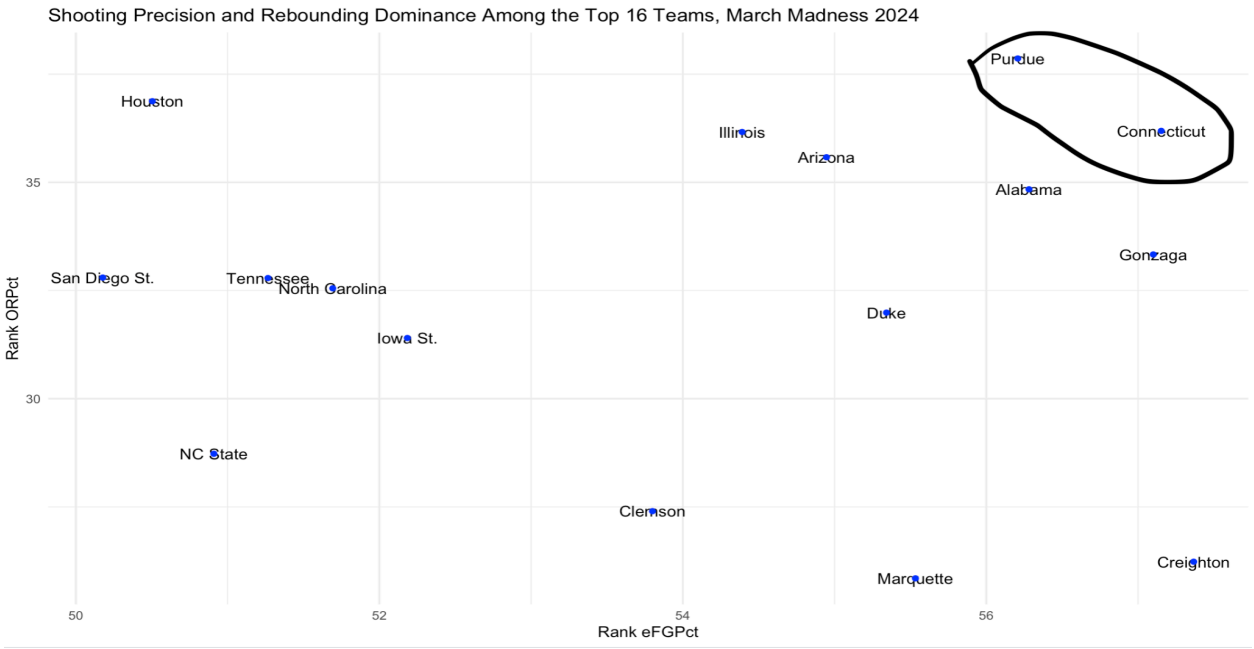


:

```
predicted_classes 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298
                1  28  26  26  27  26  26  25  25  26  25  26  26  25  27  26  27  25  26  27  26  26

predicted_classes 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319
                1  25  27  26  27  26  26  25  25  26  26  27  26  27  26  27  27  26  26  25  25  24

predicted_classes 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340
                1  23  24  23  23  23  23  23  22  20  20  20  19  19  19  19  18  18  17  17  17  17

predicted_classes 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361
                1  17  16  16  16  15  13  13  20  10  10  10   5   5   3   3   3   3   2   2   2

predicted_classes 362 363
                1   2   1
>
```

Brand Conceptual Map code:

```
basketball_data <- read.csv("Desktop/KenpomD.csv")
# Filter the dataset for the year 2024
basketball_data_2024 <- subset(basketball_data, Season == 2024)

library(ggplot2)

ggplot(basketball_data_2024, aes(x = eFGPct, y = ORPct, label = TeamName)) +
  geom_text() +
  geom_point(color = "blue") +
  labs(title = "Conceptual Map of Basketball Teams - 2024",
       x = "Rank eFGPct",
       y = "Rank ORPct") +
  theme_minimal()
```


Shooting Precision and Rebounding Dominance Among the Top 16 Teams, March Madness 2024
Purdue
Houston
Illinois
Connecticut
Arizona
Alabama
Gonzaga
San Diego St.
Tennessee
North Carolina
Duke
Iowa St.
NC State
Clemson
Marquette
Creighton
35
30
50
52
54
56
Rank ORPct
Rank eFGPct