NCAA Tournament Qualifications

Justin Bussey

University of North Texas

Toulouse Graduate School of Business

ADTA 5940 – Capstone

Dr. Denise Philpot

Spring 2023

Table of Contents

CHAPTER 1: INTRODUCTION

Background

At American higher education institutions, NCAA handles all things related to athletics. Since they are responsible for regulating different aspects of sports activities, their role is significant. Pomeroy created KenPom.com, which is an exceptional website that analyzes statistics from every game since 2002 to give each NCAA Men's Division I basketball team ratings and efficiency details.

Ken Pomeroy's site was most recognized for his proprietary statistics. Pomeroy invented statistics such as adjusted offensive efficiency. KenPom.com also provides non-numerical data for teams such as the conference they are in, strength of schedule, and offensive or defensive style of play.

For the NCAA Basketball tournaments, countless individuals are enthusiastic about predicting the victorious team during March Madness. The event is among the most awaited occurrences in American sports. This popularity makes predicting its outcome highly engaging. Wilco (2023) explains that a total of sixty-eight college teams compete in March Madness which draws millions of viewers nationwide every year. With some of the best college athletes competing for victory, there is no surprise that many are thrilled to engage in discussions about forecasts and conjectures. The excitement does not merely stem from the most potential contenders in the NCAA competition but also from unforeseen upsets and victories that bring fervor.

The uncertainty around which team will win adds to the excitement. for fans trying to guess who will come out on top. Another reason predicting outcomes is popular is how frequent office pools and online contests are where people can show off their bracket challenge skills. Interest in the upcoming NCAA Basketball tournament is extremely popular. Bracket challenges

offer multiple incentives that make forecasting game results even more enticing with cash prizes and bragging rights being two common motivators. Participating in these contests enables youth to highlight their abilities and engage in friendly competition with peers.

All 32-Division I Conferences receive an automatic bid that rewards one champion team from each postseason conference tournament, making each conference champion an automatic qualifier in March Madness. Teams that receive automatic bids in this way do not have to have a winning season but must be eligible for postseason play and win their respective conferences.

The remaining thirty-six teams go through a selection process known as Selection Sunday. The NCAA Division I Men's Basketball Selection Committee of ten members assembles to determine the pedigree of teams that did not receive an automatic bid as aforementioned and must receive an invitation to complete the bracket of sixty-eight teams. Each team then competes to make it to the Sweet 16, Elite 8, and Final Four and is named NCAA's Division I Men's Basketball Champion. The Selection Committee also assigns each of the sixty-eight teams a seed to better determine matchups in the first round of the tournament. Examining statistics and rankings assists in evaluating college basketball teams; however, there is no designated criteria presented by the NCAA website.

Research Questions

The NCAA Tournament Qualification will be the target variable, which means this variable will need to be added to the dataset. My research will investigate and answer the following questions:

- Do teams with super defense, lower turnover percentages, and higher scoring percentages have a better success rate of Tournament Qualification?

- Do teams with higher free throw rate percentages have a higher chance of making it to Tournament qualifications than teams with lower free throw rate percentages?

During the data analysis the four factors will be monitored, which are described by the KenPom website during the analysis to determine their significance in a team's qualification for the NCAA Tournament. These factors may play a role, and will assessed on their impact through analysis:

Shooting (eFG%) - Teams' shooting performances are quantified through Effective Field Goal Percentage (eFG%), a measuring tool that acknowledges the value of both two-point and three-point shots. (eFG% = (.5*3FGM + FGM) / FGA)

Turnovers (TOV%) - Measuring how many times teams turn over possession during play, TOV% provides insight into their ability to retain control of the ball. During a game, this statistic involves determining what percentage of a team's possessions end with them losing their grip on the ball. (TO% = TO / Possessions)

Rebounding (ORB% and DRB%) - Offensive and defensive rebounding is a measure of the possible rebounds that are gathered by the offense and defense. OR% = OR / (OR + DRopp)

Free Throws (FT Rate) - FT Rate refers to a measure used to estimate how frequently a particular team can accomplish successful free throws. (FTRate = FTA / FGA)

CHAPTER 2: LITERATURE REVIEW

Basketball History

Basketball originated in 1891 with Canadian physical education teacher James Naismith's invention of this new sport that assisted his students' activeness during winter months. The initial gameplay consisted of using a soccer ball while aiming at two peach baskets hanging from one of the gymnasium walls. To set specific boundaries for gameplay, Naismith wrote down thirteen rules disallowing running with the ball or making physical contact between competitors (Toole, 2021).

Following its introduction to numerous U.S institutions through local universities and colleges within an era ranging from late nineteenth-century till early twentieth-century, basketball quickly became an admired sport across America. Taking place in 1895, Hamline University competed against the University of Minnesota in a riveting basketball game marking meaningful change for collegiate sports. With gradual adaptation, teams eventually accepted the round ball and wooden backboards which modified gameplay guidelines forevermore (Hamline University, 2021).

From there on out, attending games for leisure increased with professionals favoring events hosted by both National Basketball League (NBL) and Basketball Association of America (BAA). During their prime years around mid-century, these two associations would merge forming what is currently recognized as one of the world's most respected sports organizations - National Basketball Organization (NBA) (A&E Television Networks, 2009).

The NCAA basketball tournament began with only eight teams back in 1939 but has since expanded, reaching a whopping sixty-four teams by 1985. Today, it is an eagerly awaited event that draws millions of viewers annually and brings thousands of dollars in revenue for

participating institutions. Since 1936, basketball has been a significant event at the Olympics, giving it enormous popularity and attention.

Statistics that feed into tournament qualifications

Basketball teams' potential success in tournament qualifications can be determined by evaluating various measurements, including adjusted efficiency margin (AdjEM). This metric considers both offensive and defensive capabilities while considering the quality of an opponent (Pomeroy, 2016). High AdjEM scores are crucial for achieving victory over lower-scoring competitors when it matters most in tournaments. In addition to this metric, several other variables such as rebounding skills, turnover rates, and shooting accuracy significantly impact winning games at the college level.

Overall success in college basketball involves bringing together multiple aspects effectively. To increase their chances of winning high-stakes games, teams must prioritize minimizing turnovers and maximizing rebounds while effectively sinking shots. Evaluating potential tournament contenders involves analyzing a range of factors, including win-loss ratios, strength of schedule rankings, efficiency ratings, turnover margins, as well as rebounding percentages and shooting accuracies (Pomeroy, 2004).

Improving weaknesses and reinforcing existing gameplay skills can help college basketball teams secure their eligibility for the tournament. Maintaining an active lifestyle through consistent exercise is integral for optimal functioning of the human body. Regular physical activity encourages efficient oxygenation within vital body systems such as the heart and lungs leading to better overall health outcomes. Additionally, keeping up with moderate intensities during workouts helps regulate metabolic processes resulting in improved energy levels throughout daily routines.

Challenges that affect predictions

Foreseeing the Final Four in March Madness poses an enormous challenge due to the unpredictable nature of the tournament. With underdogs consistently shattering brackets and defeating favored opponents, it becomes difficult to anticipate which four teams will advance. The following issues play a factor in predicting the outcomes of NCAA Basketball, upsets, injuries, the strength of schedules, momentum, and matchups:

Upsets - One of March Madness' most thrilling facets is its unpredictability. The tournament consistently sees low-seeded teams defeating their higher-seeded counterparts, rendering accurate game predictions incredibly tough. Upsets can also completely alter the trajectory of the tournament, making it challenging to determine which teams might prevail and reach the coveted Final Four

Injuries - Injuries sustained by critical players that could have grave repercussions for their respective teams. Injuries pose a serious threat to the success rate of any sports team, especially when involving influential players. An unfortunate injury sustained by one such critical component can result in immense vulnerability for their entire squad with increased chances for them to make an early exit from the competition. Singer (2015) authored an article explaining how injuries will play a factor when seeding a team for the NCAA Final Four. Statistics are affected by injuries in the NCAA. Injuries sustained by critical players could have grave repercussions for their respective teams. Injuries pose a serious threat to the success rate of any sports team, especially when involving influential players. An unfortunate injury sustained by one such critical component can result in immense vulnerability for their entire squad with increased of the team not making it to tournament qualifications. Singer (2015) authored an

article explaining how injuries will play a factor when seeding a team for the NCAA Final Four. It is predicted that injuries will knock teams out of NCAA Tournament Qualification

Strength of Schedule - The decision by some teams to tackle a tougher schedule during the regular season can come with its own set of obstacles, including encountering formidable opponents, and coping with tiredness. Projecting which of these teams will do well when matched against greater resistance is frequently challenging. Wilco (2019) explains that it is easy to look at a team's schedule and automatically assume a team will win or lose, this is how false predictions occur and upsets happen. For teams hoping to secure an invitation to the tournament, there is one crucial parameter beyond just their win-loss record and strength of schedule which is opponent difficulty.

Those who face off against top-ranked opponents and come out victorious have a leg up when it comes time for selection committees to make their decisions. Additionally, scheduling tough non-conference matchups can boost a team's chances even further. Depending on certain other variables involved, these components may make or break a squad's tournament prospects. In the "Calculating Strength of Schedule and Choosing Teams for March Madness" journal (Fearnhead, Taylor 2010), it explains how to accurately predict a college basketball team's likelihood of qualifying for tournaments, one must consider numerous factors beyond just their win-loss record and strength of schedule. Evaluating opponents' skill level is crucial in determining overall difficulty as well as advancing in postseason play. This article explains that strength of schedule would help inform decisions as to which teams are given bids to the national tournament (Fearnhead, Taylor 2010).

CHAPTER 3: METHODOLOGY: DATA PREPARATION

The data preparation took excel manipulation and a few formulas to clean the data, such as the TRIM, TEXTJOIN, and CONCATENATE functions. On January 29th, 2023, the KenPom dataset originally was collected and analyzed from KenPom.com and the data wrangling took place on February 11, 2023. The processes that took place in the data wrangling were discovery, structuring, cleaning, enriching, validating, and publishing of the dataset so it can be readily available when it was time to analyze information and build models. The first step of the data wrangling involved data discovery and navigating to KenPom.com while focusing on five different sections of the website, which were Pomeroy College Basketball Ratings, Efficiency and Tempo, Four Factors, Team Points Distribution, and Miscellaneous Team Statistics. Different datasets were downloaded for each year of data that was used, which were years 2019 through 2023. For the final dataset, different statistics were removed such as championship winnings. It's nice to know which teams have won in the past, but it wasn't going to beneficial when predicting the qualifying teams. Years 2019-2023 were also removed since the only data needed was data that went back until 2021. There was also a year column and a team column in the dataset. Instead of keeping them separate, the data was consolidated into the two columns so it would show both the team and the year of the statistic.

CHAPTER 4: METHODOLOGY: EXPLORATORY DATA ANALYSIS

Data Description

The data was collected from the following website, KenPom.com. These teams are part of the NCAA men's division, and their statistics will be used to determine the patterns that predict a certain win percentage for each 40-minute game and the patterns that predict the tournament qualifications for the NCAA. The data will be collected for seasons 2019 through 2023, which will be about four different NCAA basketball seasons.

This project aims to recognize the diverse types of data and the different traits in the data that will help predict win percentages for different teams. The dataset will have nominal, binary, and interval variables. Looking at the datasets, four different datasets represent each of the seasons starting from 2019 and ending in 2023. In each dataset, there are at least 350 rows and forty-six variables. All the years were consolidated together to create one dataset that represents 2019-2023 of the NCAA Men's Basketball Division I. This resulted in forty-six variables and 1,431 rows of data that represent all the college basketball teams that participated in the NCAA Men's Division I. All these data points will help build diverse types of models using exploratory data analysis.

Trends that feed into the Final Four - Having quality data is the most crucial factor in any analytical prediction. A common trend that is used throughout all NCAA analytical predictions is "game-level data" that has been consolidated to reflect the team characteristics. The following characteristics are labeled below and are crucial trends that help predict the outcome of March Madness. Each one of these sections can be found on kenpom.com, Pomeroy (2023), the website, and other NCAA statistical websites.

Pomeroy College Basketball Ratings - This section automatically populates the main page of KenPom.com. In this section, information is provided about several aspects pertaining to a team's performance. The data includes the team's current position or rank, number of games won and lost, its offensive and defensive efficiency, along with the total possible possessions in 40 minutes. Additionally, the luck rating and strength of schedule are also parameters that are given much attention.

Efficiency and Tempo - This section has variables that list the tempo (pace) of the team. Efficiency is the number of points scored or allowed per one hundred possessions and Tempo is described as the number of possessions per 40 minutes. KenPom.com explains that possessions are not an official NCAA statistic, so the stat is estimated.

Four Factors - Whether a basketball squad is excellent or terrible depends on how well they perform categories including the Effective Field Goal Percentage, Turnover Percentage, Offensive Rebounding rate and Free Throw Rate.

Team Points Distribution - This indicates where a team's offense is coming from. This indicates the percentage of points scored by each type of shot (Free throw, 2-points, or 3-points). KenPom.com explains that this data provides how a team plays offense or defense.

Miscellaneous Team Statistics Offense - This stat gives the overall scoring percentages for each team. The diverse types of variables are 3-point, 2-points, free throws, blocks, steals, non-steal turnovers, assists, and 3-point attempts.

Continuing with the trends that influence the outcome of the final four. From the Department of Statistics at North Dakota State University, it is said that free throw attempts, differences in defensive rebounds, and differences in turnovers play a significant role in determining the final four outcomes, followed by assists.

Missing Data

During data preparation web scraping techniques were employed to extract information from KenPom's website.

Outliers

Python was utilized to identify outliers in my dataset. The IQR method was utilized to identify outliers to set up a wall outside of Quartile 1 and Quartile 3. Any values that fell outside of the wall were considered outliers. All variables had outliers in the dataset except for Losses, Win %, AdjEM, ADjD, Strength of Schedule (SOS) AdjEM, SOS OppO, SOS, OppD, Def. Eff. Adj, and Stl%. This dataset will have outliers because in basketball statistics can vary depending on the player or team. None of the variables needed to be removed due to outliers, since basketball metrics fluctuate quite often. Below are just a few examples of the outliers that was found using a boxplot in python:

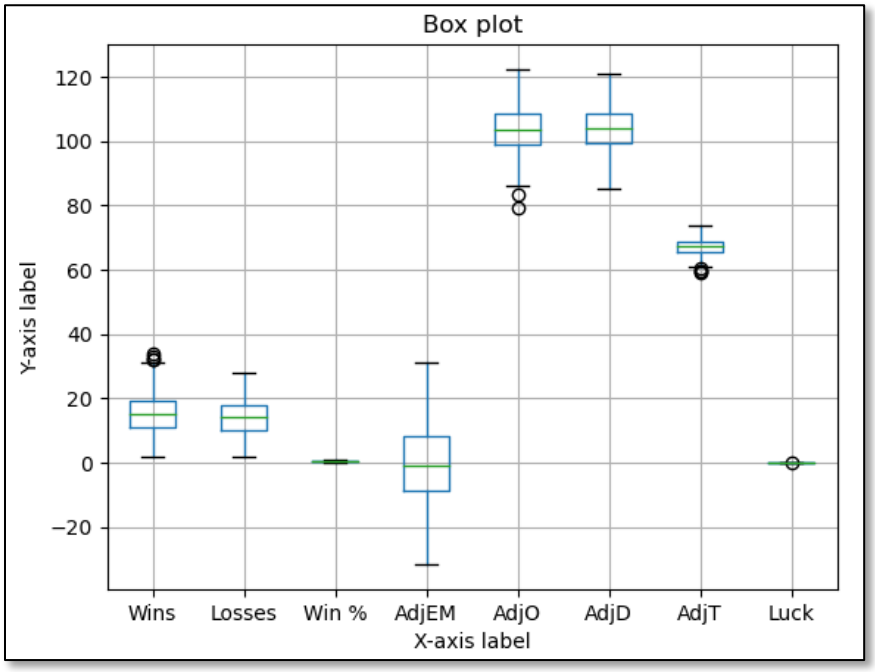

Fig. 4-1: Outlier Sample 1

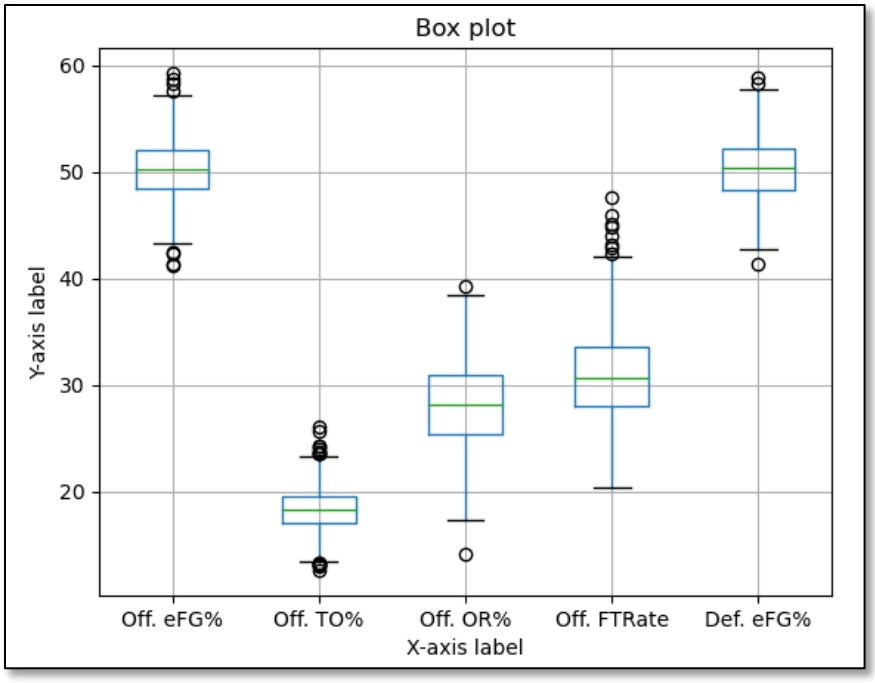

Fig. 4-2: Outlier Sample 2

CHAPTER 5: METHODOLOGY: MODELING

In the effort to answer the first question: Do teams with super defense, lower turnover percentages, and higher scoring percentages have a better success rate of Tournament Qualification? Models were performed to see which variables were more important when it comes to Tournament qualifications. The types of models used were Linear Regression, Decision Tree, and Clusters.

Linear Regression Model

To begin modeling, the dataset was imported into the file import node and after the file was imported the data was adjusted as shown in Fig. 5-1. NCAA Tournament Qualification was the Binary Target Variable. Conference was changed from interval to nominal, Team Year was changed to ID and nominal, and all other variables were left as input and interval.

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
_2P_	Input	Interval	No		No	.	.
_3P_	Input	Interval	No		No	.	.
_3PA_	Input	Interval	No		No	.	.
A_	Input	Interval	No		No	.	.
Adj__Off__Eff_	Input	Interval	No		No	.	.
AdjD	Input	Interval	No		No	.	.
AdjEM	Input	Interval	No		No	.	.
AdjO	Input	Interval	No		No	.	.
AdjOE	Input	Interval	No		No	.	.
AdjT	Input	Interval	No		No	.	.
Avg__Poss_Len(	Input	Interval	No		No	.	.
Avg__Poss_Len(	Input	Interval	No		No	.	.
Blk_	Input	Interval	No		No	.	.
Conf	Input	Nominal	No		No	.	.
Def__2_Pt_FG_	Input	Interval	No		No	.	.
Def__3_Pt_FG_	Input	Interval	No		No	.	.
Def__Eff__Adj_	Input	Interval	No		No	.	.
Def__Eff__Raw	Input	Interval	No		No	.	.
Def__FT_Attem(	Input	Interval	No		No	.	.
Def__FTRate	Input	Interval	No		No	.	.
Def__OR_	Input	Interval	No		No	.	.
Def__TO_	Input	Interval	No		No	.	.
Def__eFG_	Input	Interval	No		No	.	.
FT_	Input	Interval	No		No	.	.
Losses	Input	Interval	No		No	.	.
Luck	Input	Interval	No		No	.	.
NCAA_Tournam(	Target	Binary	No		No	.	.
NCSOS_AdjEM	Input	Interval	No		No	.	.
NST_	Input	Interval	No		No	.	.
Off_Eff__Raw	Input	Interval	No		No	.	.
Off__2_Pt_FG_	Input	Interval	No		No	.	.
Off__3_Pt_FG_	Input	Interval	No		No	.	.
Off__FT_Attem(	Input	Interval	No		No	.	.
Off__FTRate	Input	Interval	No		No	.	.
Off__OR_	Input	Interval	No		No	.	.
Off__TO_	Input	Interval	No		No	.	.
Off__eFG_	Input	Interval	No		No	.	.
Raw_Tempo	Input	Interval	No		No	.	.
Stl_	Input	Interval	No		No	.	.
Strength_of_Sch	Input	Interval	No		No	.	.
Strength_of_Sch	Input	Interval	No		No	.	.
Strength_of_Sch	Input	Interval	No		No	.	.
Team___Year	ID	Nominal	No		No	.	.
Win__	Input	Interval	No		No	.	.
Wins	Input	Interval	No		No	.	.

Fig. 5-1: Linear Regression Variables

Next, Stat Explore tool was used to get a better understanding of the data. Looking at figure 3-2 at the variable AdjEM, (A strength of schedule metric) is projected to be the most valuable variable, followed by Win Percentage, Adjusted Offense Efficiency, Losses, and Wins.

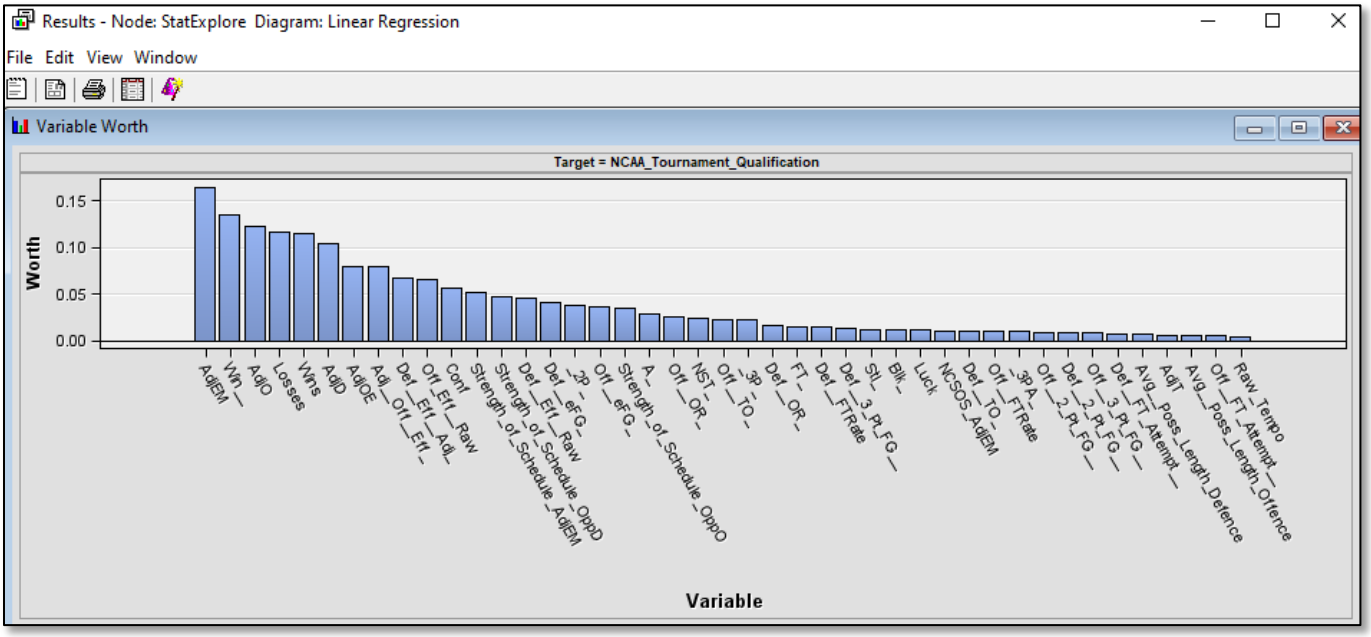

Fig. 5-2: Variable Worth

Looking at the output window in Fig. 5-3, we can see that some of the StatExplore data is skewed, and the kurtosis fluctuates. Based on the different models that were executed, the most important variable is strength of schedule AdjEM.

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
A_	INPUT	50.72334	5.234784	721	0	35.3	50.4	68.84	0.23501	0.124688
AdjD	INPUT	103.8683	6.423278	721	0	85	104.07	120.71	-0.20681	-0.34901
AdjEM	INPUT	-0.00006	11.48658	721	0	-31.57	-0.97	30.95	0.204298	-0.47502
AdjO	INPUT	103.8687	6.770562	721	0	79.4	103.67	122.3	0.06164	-0.08985
AdjOE	INPUT	103.8681	6.768237	721	0	79.4	103.7	122.3	0.061089	-0.08978
AdjT	INPUT	67.13718	2.448834	721	0	59.07	67.1	73.6	-0.00666	-0.02112
Adj__Off__Eff_	INPUT	103.868	6.768125	721	0	79.4	103.7	122.3	0.061113	-0.08963
Avg__Poss_Length_Defence	INPUT	17.59723	0.564376	721	0	15.5	17.6	19.4	-0.17457	0.209837
Avg__Poss_Length_Offence	INPUT	17.63245	1.10314	721	0	14.6	17.6	21.3	0.109849	0.102127
Blk_	INPUT	8.918724	2.074649	721	0	3.57	8.75	16.65	0.386562	0.198417
Def__2_Pt_FG__	INPUT	50.84588	3.777035	721	0	39.8	50.9	61.01	-0.04603	-0.14599
Def__3_Pt_FG__	INPUT	31.02053	3.768484	721	0	21.2	30.8	42.7	0.248877	-0.00297
Def__Eff__Adj_	INPUT	103.8691	6.424007	721	0	85	104.1	120.7	-0.2066	-0.34947
Def__Eff__Raw	INPUT	102.8047	5.379953	721	0	87.9	102.6	118.7	0.061377	-0.25975
Def__FTRate	INPUT	31.20674	5.317816	721	0	16.42	30.6	51.4	0.512564	0.37765
Def__FT_Attempt__	INPUT	18.132	2.574525	721	0	10.48	17.99	27.2	0.312058	0.089686
Def__OR_	INPUT	28.36816	2.994483	721	0	18.81	28.3	37.6	0.080337	-0.08717
Def__TO_	INPUT	18.28	2.327157	721	0	12.4	18.07	27.88	0.479883	0.524114
Def__eFG_	INPUT	50.31201	2.733288	721	0	41.3	50.4	58.85	-0.04578	-0.04017
FT_	INPUT	71.60017	3.678571	721	0	60.5	71.6	83	-0.1367	-0.0616
Losses	INPUT	14.04854	5.079821	721	0	2	14	28	0.195831	-0.49667
Luck	INPUT	0.001581	0.056293	721	0	-0.14	0	0.18	0.130125	-0.24863
NCSOS_AdjEM	INPUT	-0.56638	4.652387	721	0	-13.35	-0.77	16.98	0.283839	0.244949
NST_	INPUT	9.026852	1.332453	721	0	5.6	9	14	0.260966	0.324495
Off_Eff__Raw	INPUT	102.3447	5.887841	721	0	81.5	102.4	120.2	-0.08919	0.112672
Off__2_Pt_FG__	INPUT	50.91021	4.344887	721	0	37.35	50.9	63.17	0.032998	-0.15143
Off__3_Pt_FG__	INPUT	31.00907	4.764705	721	0	15.8	30.99	46.75	0.034012	0.091262
Off__FTRate	INPUT	30.96146	4.483387	721	0	20.4	30.7	47.6	0.356354	0.208633
Off__FT_Attempt__	INPUT	18.07993	2.337284	721	0	10.72	18	26.2	0.143213	0.072327
Off__OR_	INPUT	28.17741	3.98257	721	0	14.17	28.19	39.2	-0.08825	-0.15366
Off__TO_	INPUT	18.34194	2.036854	721	0	12.61	18.28	26.08	0.16818	0.434462
Off__eFG_	INPUT	50.11811	2.818932	721	0	41.25	50.2	59.25	-0.04491	0.107242
Raw_Tempo	INPUT	68.03135	2.54502	721	0	59.5	67.9	74.6	0.026195	-0.05099
Stl_	INPUT	4.689223	4.711019	721	0	0.06	0.14	13	0.112905	-1.86409
Strength_of_Schedule_AdjEM	INPUT	-0.28773	5.869347	721	0	-13.25	-1.48	13.38	0.445011	-0.77677
Strength_of_Schedule_OppD	INPUT	104.0372	3.206262	721	0	95.4	104.4	110.9	-0.36984	-0.64896
Strength_of_Schedule_OppO	INPUT	103.7494	3.242929	721	0	95.5	103.5	112.1	0.205475	-0.64638
Win__	INPUT	0.519709	0.174968	721	0	0.07	0.53	0.92	-0.1663	-0.62344
Wins	INPUT	15.38558	5.774803	721	0	2	15	34	0.250264	-0.22093
_2P_	INPUT	49.88305	3.250713	721	0	40.43	50	60.7	0.051904	0.094997
_3PA_	INPUT	37.51154	5.193069	721	0	22.2	37.51	54.8	-0.00099	-0.0083
_3P_	INPUT	33.68399	2.403807	721	0	25.2	33.7	43.7	-0.01236	0.227925

Fig. 5-3: Output Window

The first regression node, which is labeled as regression node one and determined which variables were statistically significant. The variables that remained can be seen in Appendix C.

After running a regression node and removing all the statistically insignificant variables, the Analysis of Variance (ANOVA) was checked, located in Fig. 5-4, which is a technique used to validate Tournament Qualification variables. The results were an F value of 13.25, a P-Value less than 0.001, and adjusted R-Sq value of 0.3665.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	34	43.741438	1.286513	13.25	<.0001
Error	686	66.605303	0.097092		
Corrected Total	720	110.346741			

Fig. 5-4: ANOVA Linear Regression 1 Model

The fit statistics shown below allows me to determine if the model is overfitting or underfitting the data. Since the misclassification rate is 0.14, there are fewer errors when predicting the outcome.

Model Fit Statistics

R-Square	0.3964	Adj R-Sq	0.3665
AIC	-1647.3174	BIC	-1641.7512
SBC	-1486.9950	C(p)	35.0000

Fig. 5-5:Model Fit Statistics for Linear Regression 1 Model

For my second regression model, the file was ran through a variable selection node. Which is shown in Fig. 5-6 below.

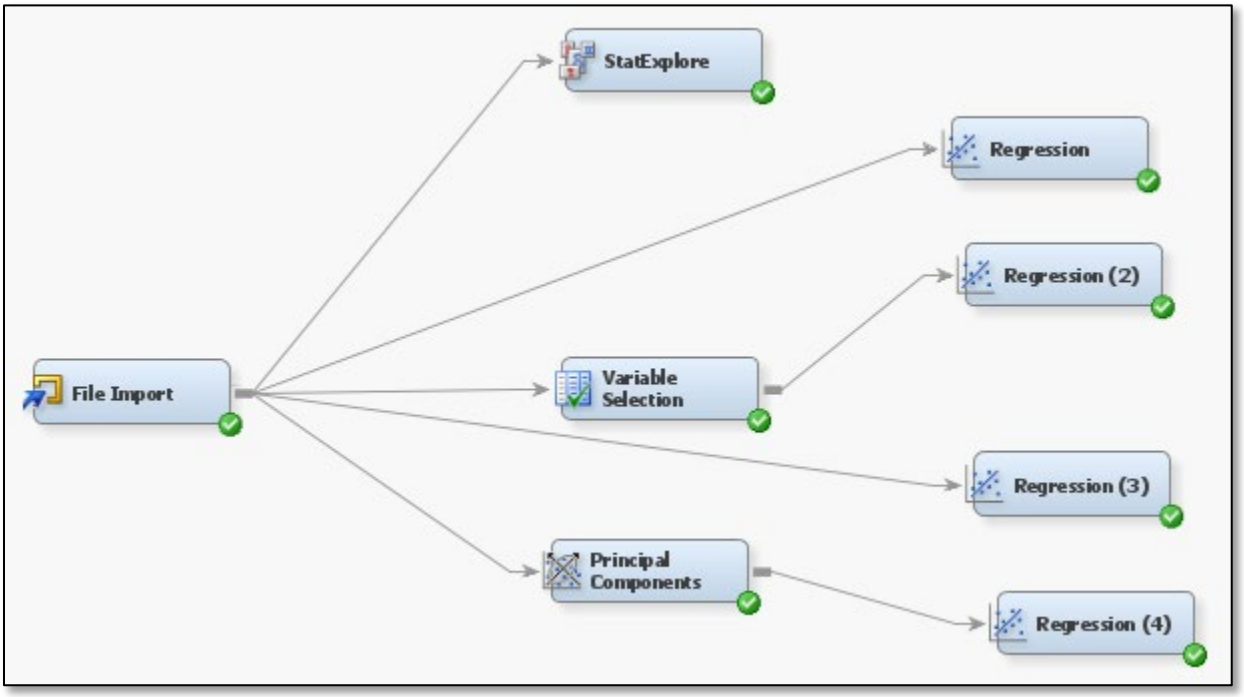

Fig. 5-6: Linear Regression Nodes

Shown in Fig 5-8, the results from the Variable Selection node show that the effects chosen for the target "NCAA_Tournament_Qualification" has chosen three different variables, which are number of losses, athletic conference affiliation, strength of schedule adjusted and efficiency metric.

The DMINE Procedure

Effects Chosen for Target: NCAA_Tournament_Qualification

Effect		DF	R-Square	F Value	p-Value	Sum of Squares	Error Mean Square
Var:	Losses	1	0.280698	280.579881	<.0001	30.974088	0.110393
Group:	Conf	4	0.103858	30.164487	<.0001	11.460360	0.094982
Var:	Strength_of_Schedule_AdjEM	1	0.001221	1.419768	0.2338	0.134774	0.094926

Fig. 5-7: Variable Selection Nodes Chosen Variables

Variables were only chosen by the variable selection node, and the ANOVA score that was received for the regression model two was an F-value of 74.74 and a P-Value of less than

.0001 which makes this model statistically significant. The adjusted R-Squared value for the second regression node is 0.3806.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	42.569222	7.094870	74.74	<.0001
Error	714	67.777518	0.094926		
Corrected Total	720	110.346741			

Model Fit Statistics

R-Square	0.3858	Adj R-Sq	0.3806
AIC	-1690.7386	BIC	-1688.6015
SBC	-1658.6741	C(p)	7.0000

Fig. 5-8: ANOVA & Model Fit Statistics for Regression 2 Model

Below are the ANOVA results for the third regression model, along with the Model Fit Statistics. Stepwise regression was used for this iteration. It resulted in a 13.67 F Value, P-value of <.0001, and an adjusted R-sq value of 0.3674 shown in fig. 5-10.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	33	43.738125	1.325398	13.67	<.0001
Error	687	66.608616	0.096956		
Corrected Total	720	110.346741			

Model Fit Statistics

R-Square	0.3964	Adj R-Sq	0.3674
AIC	-1649.2816	BIC	-1643.8209
SBC	-1493.5398	C(p)	33.0341

Fig. 5-9: ANOVA & Model Fit Statistics for Regression 3 Model

For the fourth regression model, it was ran through a Principal Component Analysis (PCA) node with a cumulative .90 Eigenvalue Cutoff. After removing the insignificant Principal

Component, the ANOVA ended up with an F-Value of 13.67, a P-value of less than .0001, and adjusted R-sq value of 0.3674 in the Model Fit Statistics.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	33	43.738125	1.325398	13.67	<.0001
Error	687	66.608616	0.096956		
Corrected Total	720	110.346741			

Model Fit Statistics

R-Square	0.3964	Adj R-Sq	0.3674
AIC	-1649.2816	BIC	-1643.8209
SBC	-1493.5398	C(p)	33.0341

Fig. 5-4: ANOVA & Model Fit Statistics for Regression 4 Model

Cluster Model

After the file was imported, the variable roles and levels stayed the same as the linear regression variables. The NCAA Tournament Qualification variable continued to be the binary target variable. Conference also stayed the same as nominal, along with Team Year being ID and nominal, and several variables were rejected due to inconsistencies. After creating a file import node, and then added a cluster node, linked it with the file import and ran the cluster node as shown in Fig 5-11.

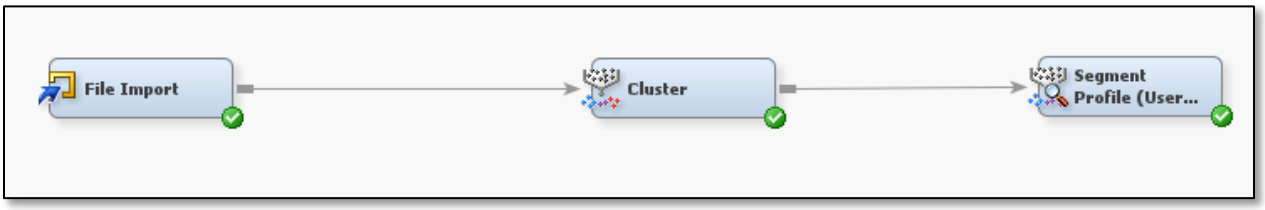

Fig. 5-5: Cluster Node Diagram

The cluster node is set to have a specification method that is set to automatic. After running the cluster node, Fig. 5-12 shows the results of a segment plot. The first cluster node had a specification method set to “automatic” and the second cluster node was set to “user specify,” it gave a resulting variable importance of only six variables.

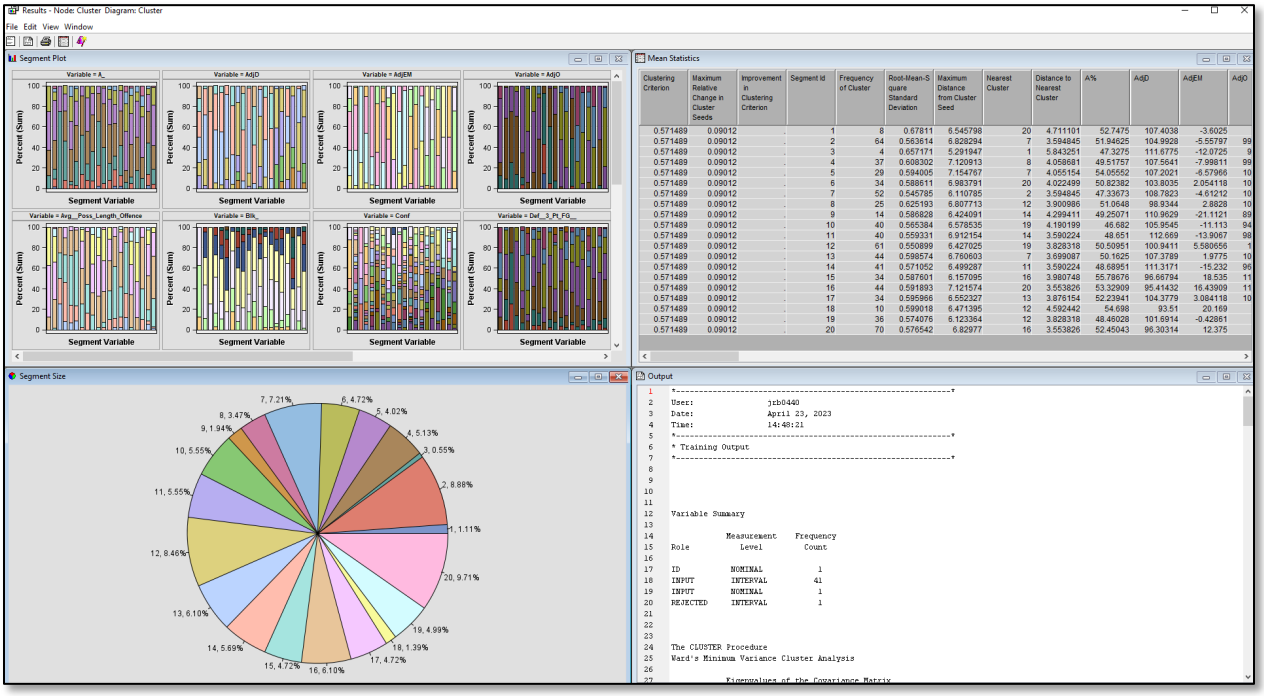

Fig. 5-6: 20 Clusters

The importance of each variable is shown under the output results. AdjO has the highest rate with an Importance of 1, and AdjEM is not far behind with 0.99733. Avg__Poss_Length_Defence gives the lowest rate of the cluster with 0.12640, second to last is Def_3_Pt_FG__ with 0.18015. The list of Variable Importance helps us see which variables of the data are most important for the target, NCAA Tournament Qualification. This type of data helps me answer my research questions on if defense, lower turnover percentage, and higher scoring percentages have better success rate of Tournament qualifications. Since most of the defensive variables are least important, and turnover percentage is also least important, the

conclusion was that free throw percentage, defensive 3 point field goals, and average possession length offense are not contributing to Tournament qualifications as much as Adj0, AdjEM, and Adj__Off__Eff_. The results can conclude that it is possible to get the same results about the second research question, since free throw rate percentage is not a top importance variable. The explanation for this data can be found on Fig. 5-13.

Variable Importance

Variable Name	Label	Number of Splitting Rules	Number of Surrogate Rules	Importance
Adj0		1	8	1.00000
AdjEM		2	3	0.99733
Adj__Off__Eff_	Adj. Off. Eff.	1	7	0.99211
Def__Eff__Adj_	Def. Eff. Adj.	1	6	0.97606
AdjD		0	4	0.82475
Win__	Win %	2	2	0.80189
Off_Eff__Raw	Off Eff. Raw	1	5	0.78228
Strength_of_Schedule_AdjEM	Strength of Schedule AdjEM	2	3	0.67777
Def__Eff__Raw	Def. Eff. Raw	1	3	0.66273
Losses		0	3	0.65684
Conf		0	2	0.63431
Strength_of_Schedule_OppD	Strength of Schedule OppD	1	3	0.63085
Strength_of_Schedule_Opp0	Strength of Schedule Opp0	0	5	0.58817
Off__3_Pt_FG__	Off. 3-Pt FG %	2	1	0.56538
Wins		0	2	0.55186
Off__eFG_	Off. eFG%	0	2	0.55186
_3PA_	3PA%	0	3	0.53971
Def__FTRate	Def. FTRate	1	3	0.53883
Off__2_Pt_FG__	Off. 2-Pt FG %	1	2	0.52703
_3P_	3P%	0	3	0.50479
Off__TO_	Off. TO%	1	3	0.47777
Def__FT_Attempt__	Def. FT Attempt %	0	2	0.46054
NST_	NST%	0	3	0.45985
Def__eFG_	Def. eFG%	0	2	0.43779
Def__TO_	Def. TO%	2	2	0.42775
Off__FTRate	Off. FTRate	0	2	0.36007
Off__OR_	Off. OR%	0	1	0.34492
AdjT		0	2	0.33767
Off__FT_Attempt__	Off. FT Attempt %	0	1	0.30637
A_	A%	0	1	0.29708
Blk_	Blk%	0	1	0.27627
Avg__Poss_Length_Defence	Avg. Poss Length Defence	0	1	0.27461
FT_	FT%	0	1	0.18015
Def__3_Pt_FG__	Def. 3-Pt FG %	0	1	0.18015
Avg__Poss_Length_Offence	Avg. Poss Length Offence	0	1	0.12640

Fig. 5-7: Importance Variables for Cluster Node with User Specific as Specification Method

A segment profile node was added to the cluster node. A node was executed and reviewed the results. This will give me some more detailed information about the segments. For the segment profile there are twenty results. Pie Properties was added to show a percentage and value of each cluster shown in Fig. 5-14.

Fig. 5-8: Segment Profile

Fig 5-15 has an output of twenty different segments, most variables that relate to the segments are defensive variables. The conclusion is that the worth for each variable is the same which raises some concern for this cluster model segment. AdjO is worth more in all the segments. We can see in segment 20 and sgment 2 the adjO variable is a top performer.

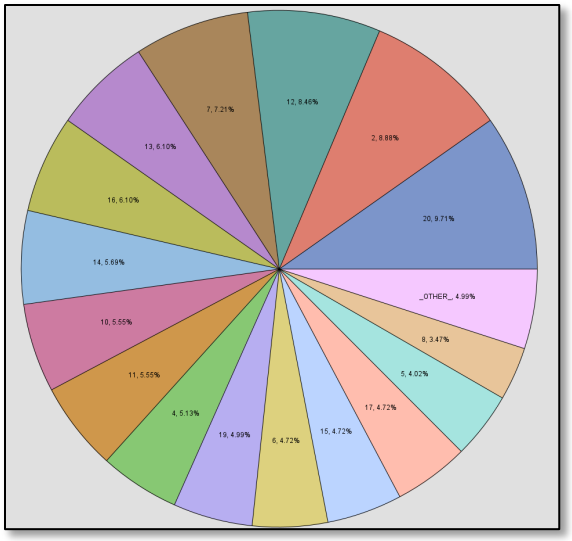

Fig. 5-9: Segments

Looking at the diagrams, there are different segments and their different counts and percentages. The red lines help show the average of each segment and the importance of them.

Decision Tree Model

Variables were adjusted for the decision tree model which can be found in Appendix D. All win and loss statistics have been rejected due to the target of identifying tournament qualifiers before the season has begun. Conference was rejected because SAS was incorrectly predicting this as an important variable. However, it is clear to understand that simply because you are in the Big 12 does not mean you will make it into the tournament. With the conference variable included in the modeling, SAS Enterprise Miner was saying that was the case. The "VAR" variables are not identifiable, so those were rejected as well (Fig. 5-19).

The data has been partitioned to train, validate, and test the decision tree. It has split of 40% for train, 40 % for validate and 20% for test. The decision tree shows that Kenpom's proprietary Adjusted Efficiency Margin metric is the most important when it comes to Tournament qualifications (Fig. 5-16).

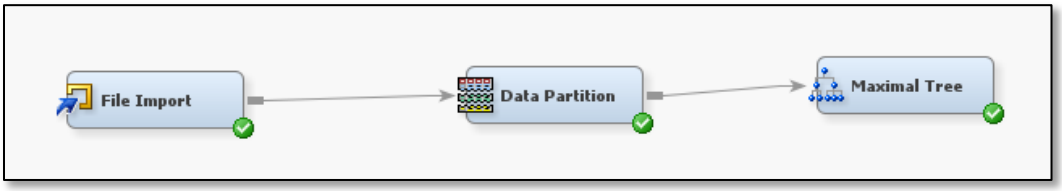

Fig. 5-16: Maximal Tree Diagram

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
AdjEM		1	1.0000	1.0000	1.0000

Tree Leaf Report

Node Id	Depth	Training Observations	Training Percent 1	Validation Observations	Validation Percent 1
2	1	236	0.06	241	0.08
3	1	52	0.79	46	0.76

Fit Statistics

Target=NCAA_Tournament_Qualification Target Label=NCAA Tournament Qualification

Fit Statistics	Statistics Label	Train	Validation	Test
NOBS	Sum of Frequencies	288.000	287.000	146.000
MISC	Misclassification Rate	0.083	0.105	0.103
MAX	Maximum Absolute Error	0.945	0.945	0.945
SSE	Sum of Squared Errors	41.914	52.085	26.371
ASE	Average Squared Error	0.073	0.091	0.090
RASE	Root Average Squared Error	0.270	0.301	0.301
DIV	Divisor for ASE	576.000	574.000	292.000
DFT	Total Degrees of Freedom	288.000	.	.

Fig. 5-17: Variable Importance

A run of StatExplore further strengthens AdjEM's standing as the most important variable (Fig. 5-18). Some reflection on this metric has brought us to the following conclusion. It is the most important indicator because it considers the fact that not every team plays the same opponents. As a strength of schedule metric, it helps compare teams with the same record but different sets of opponents.

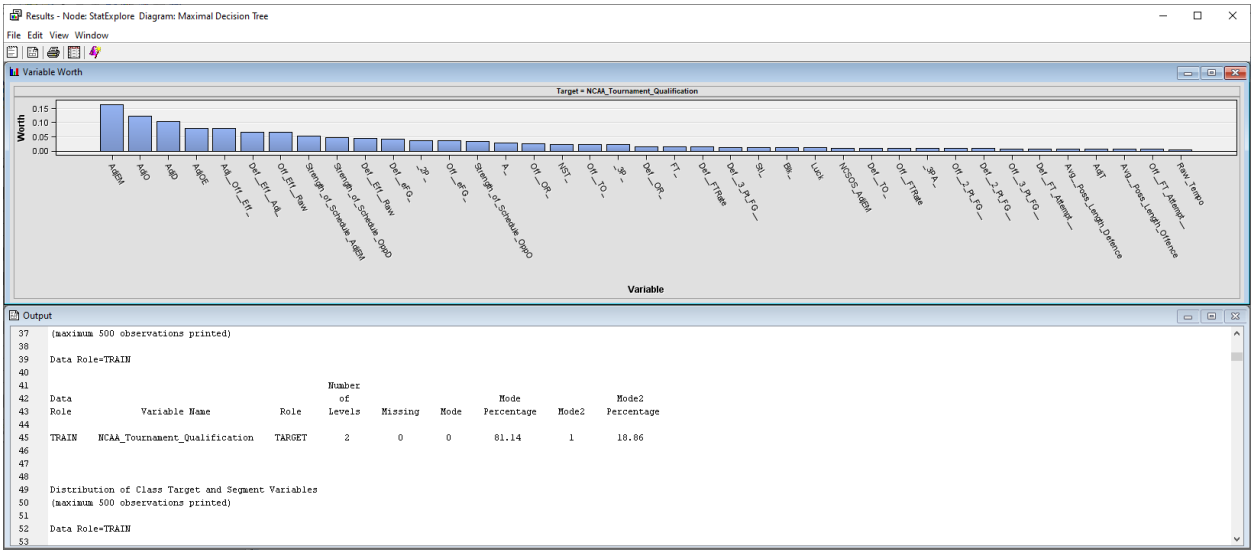

Fig. 5-18: StatExplore Results

For a second decision tree to compare the original against, win, wins, and losses variables were kept. In the second decision tree, AdjEM remains the most important variable and Win__ is now the second most important variable.

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
AdjEM		1	1.0000	1.0000	1.0000
Wins		1	0.3836	0.2619	0.6828

Tree Leaf Report

Node Id	Depth	Training Observations	Training Percent 1	Validation Observations	Validation Percent 1
2	1	236	0.06	179	0.07
7	2	45	0.89	33	0.82
6	2	7	0.14	4	0.25

Fit Statistics

Target=NCAA_Tournament_Qualification Target Label=NCAA Tournament Qualification

Fit Statistics	Statistics Label	Train	Validation	Test
NOBS	Sum of Frequencies	288.000	216.000	217.000
MISC	Misclassification Rate	0.066	0.093	0.097
MAX	Maximum Absolute Error	0.945	0.945	0.945
SSE	Sum of Squared Errors	35.171	35.962	38.103
ASE	Average Squared Error	0.061	0.083	0.088
RASE	Root Average Squared Error	0.247	0.289	0.296
DIV	Divisor for ASE	576.000	432.000	434.000
DFT	Total Degrees of Freedom	288.000	.	.

Fig. 5-19 Decision Tree 2 Variable Importance

A model comparison shown in figure 5-20, reveals that Decision Tree (2) had a smaller average squared error. The misclassification rate is important because it shows how many times the model is wrong; it is a measure of error. The further down the classification rate goes, the better. Since the decision tree is not too complex, it is less error. Looking at Fig. 5-21, anything less than 12.685 or has missing values is split to the left, and anything greater than 12.685 or equal to is split to the right. This also shows that the whole tree has been trained from the root and down.

Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Tree2	Decision Tree (2)	0.09259	0.061061	0.065972	0.083245
	Tree	Maximal Tree	0.10453	0.072767	0.083333	0.090741

Fig. 5-20: Model Comparison Fit Statistics

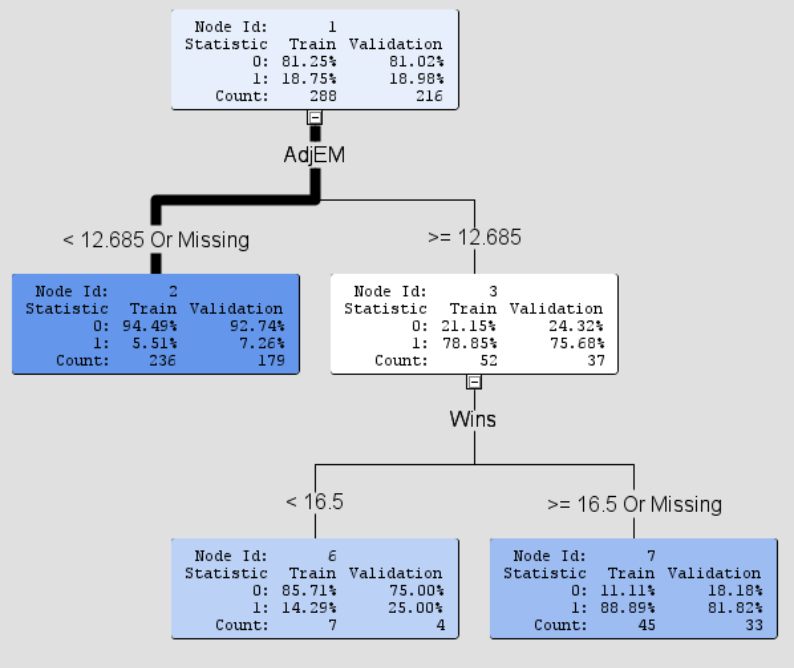

Fig. 5-21: Decision Tree (2)

CHAPTER 6: MODEL EVALUATION

Findings

Linear Regression - To determine the factors that impact Tournament Qualification. The data was compiled from the linear regression models in the table below for comparison. To report the methods, p-value, F-value, r^2 value and the variables selected as relevant for each respective linear regression method. All linear regression models have p-values of <.0001, but varying F-values. The Manual Data Cleaning and Stepwise methods have F-values of 13.25 and 13.67, respectively. The Variable Selection had an F-value of 74.74, while the Principal Component method had a F-value of 13.67. The R^2 values were similar across the methods with the lowest R^2 value of 0.3665 reported for the Manual Data Cleaning and the highest R^2 value of 0.3806 for the Variable Selection method. In terms of the variables selected for each method, the Stepwise method resulted in the fewest with only two of the possible forty-five variables. The Manual Data Cleaning and Variable Selection methods resulted in three variables selected, and the Principal Component method resulted in over 10 Principal Components selected, the most of any regression method.

Linear Regression				
	Linear Regression 1	**Linear Regression 2**	**Linear Regression 3**	**Linear Regression 4**
Method	Manual Data Cleaning	Variable Selection	Stepwise	Principal Component
P-value	<0.0001	<0.0001	<0.0001	<0.0001
F-value	13.25	74.74	13.67	13.67
R^2 value	0.3665	0.3806	0.3674	0.3674
Variables Selected	• Conference • Losses • Strength of Schedule AdjEM	• Strength of Schedule AdjEM • G_conf • Losses	• Conference • Losses	• PC 1 • PC 10 thru PC 20 • PC 2 thru PC 9

Fig. 6-1: Linear Regression Models

Cluster Analysis - Cluster is a collection of data objects that divides the data into multiple groups, which then helps determine what each of the groups are afterward. The clustering analysis helps us identify the multiple segments in the data. It also helps find similarities and differences between the data and helps predict the outcomes for the analysis. This data consisted of twenty clusters. This gave a segment size range of 0.55%-9.71%. The Variable Importance data did not come as much of a shock, given the other results from previous models. The most important variable has an importance rate of 1.0 which is the opponents adjusted offensive efficiency (AdjO) Right under adjusted offensive efficiency is adjusted efficiency metric for each team (AdjEM) with 0.99733. The variables that are least important in this cluster, are the average possession length for offense (Avg__Poss_Length_Offense) with 0.18015 and defensive three point field goal percentage (Def__3_Pt_FG__) with 0.12640

Implications

Linear Regression - The p-value is a measure of model significance. P-value of less than 0.05 is statistically significant. All the linear regression models resulted in a p-vale of <0.0001, suggesting that all models are significant representations of the data. The F-value is a hypothesis test for the significance of the data. The null hypothesis suggests that linear regression does not have any significance. If the F-value is greater than 1, then it means we can discard the null hypothesis and infer that the regression model carries significance. Similar to the p-value, for all the models, F-value exceeds 1; which is making all regression models statistically significant. The ability of a linear regression model to accurately fit into data is measured by r^2 value with a score of 1 indicating maximum fit whereas a value of 0 representing no explanatory capability. Of the four methods, the Stepwise method produced an r^2 value closest to 1.

All methods are statistically significant, so each subset of selected variables needs to be considered. The variable selection is the category where the models diverge. The Principal Component method groups together several variables to create each specific principal component. Therefore, this comparison will focus on the first three linear regression methods. The variables selected are a subset of the initial forty-five variables which are considered important predictors based on the linear regression method. For this study, these are variables which are important predictors of whether a team makes it to the NCAA tournament. There are only two variables which appear in each of the first three linear regression models: Conference and Losses. Strength of schedule AdjEM appears in two regression models. This suggests that Conference, Losses, and Strength of Schedule AdjEM are the most consistent predictors of NCAA Tournament Qualification.

Cluster Analysis - Based on the findings from the Cluster Analysis, the AdjO is the most important variable in the data for the NCAA Tournament Qualification. To become victorious in any competition, one must challenge themselves against top-tier competitors. Achieving championship status requires a team's capability to perform exceptionally well under pressure and against opponents. A team's longevity in a tournament hinge on its schedule composition leading up to the big event. AdjO is one of the three components of strength of schedule. The least important variable being Avg__Poss_Length_Offense.

Opportunities for Improvement

The opportunities for improvement include using seeds to predict the outcomes of March Madness. Stekler & Klein (2012) explain that one forecasting procedure is to use rankings based on the seeds of the teams as determined by the NCAA committee and to always select the team that has the better seeding. Seeds determine the respective positions of the participating teams.

The committee in charge of selecting and arranging the tournament teams assigns these seeds, which shows their performance during both regular season games and conference tournaments. Dean Oliver (2004) and Cecchin (2022) introduced a structure to assess basketball performance called the "Four Factors" during his time with the Denver Nuggets from 2002 to 2004. These four factors are shooting, turnovers, rebounding, and free throws. The four factors are used to determine a team's efficiency in offense and defense. Analyzing a team's metrics of performance allows coaches, analysts, and sports betters to understand a team's weaknesses and strengths. To improve one's predictions utilizing the four factors, coaches, analysts, and sports betters would need to analyze the data carefully, understand matchups between two players, and the two teams, and utilize statistical models so trends and patterns can be identified.

CHAPTER 7: CONCLUSION

Discussion

Using various data mining models including Linear Regression, Cluster Analysis and Decision Trees, they produced interesting findings and implications related to the NCAA Basketball Tournament Qualification. The dataset was sourced from Kenpom.com and included statistics from the 2021-2022 season through the 2022-2023 season.

The research questions were answered by the analysis:

1. Do teams with super defense, lower turnover percentages, and higher scoring percentages have a better success rate of Tournament Qualification?

After reviewing the variable importance output for the cluster analysis, the defensive variables never were a top variable by importance. Defense was important, but not more important than the strength of schedule variables and win percentage. Offensive turnovers were important for the cluster analysis, as it was ranked higher than defensive variables. When the decision tree was used, there were two decision trees that were created, the first was a maximal decision tree and the other was a normal decision tree. The maximal decision tree can create more splits. During the analysis, the normal decision tree was the better model. This was concluded by running a model comparison for both for trees. The normal decision tree had a misclassification rate of .083 and an average squared error of 0.072767. After reviewing this decision tree, the results stated that AdjEM (adjusted Efficiency Margin) is the most important variable. This helps decide on tournament qualifications.

2. Do teams with higher free throw rate percentages have a higher chance of making it to Tournament qualifications than teams with lower free throw rate percentages?

After reviewing the outputs from all models, free throws are not a top priority. In a basketball game, free throws matter, but if a person never goes to the free throw line, then it

would be hard to judge a team based off their free throw percentages. In the data, free throws were not a top metric, it was ranked extremely low. So, do teams with higher free throw rate percentages have a higher chance of making it to the tournament qualifications? My answer to this is no.

By using StatExplore in SAS Enterprise Miner and select models, all showed “AdjEM” to be the most important variable in determining Tournament Qualification. The Adjusted Efficiency Metric identifies the overall strength of schedule of a team (Appendix A). “AdjEM” considers the fact that not every team plays the same opponents, so competing and winning against the best often produces the best chances of qualifying. According to results in SAS Enterprise Miner using the Model Comparison node, the Maximal Decision Tree is the best model to use, however, high errors in multiple categories across all Decision Tree models ran led me to determine that the Linear Regression model is the strongest for the project.

APPENDIX

Appendix A: Variable Names, Labels, Levels, and Descriptions (Data Dictionary)

NAME	LABEL	DESCRIPTION
_2P_	2P%	Offensive 2-Point Field Goal Point Distribution, percentage of team points resulting from 2-point shots.
_3P_	3P%	Offensive 3-Point Field Goal Point Distribution, Percentage of team points resulting from 3-point shots.
_3PA_	3PA%	3-Point Attempt %: Three-point attempts divided by all attempts.
A_	A%	Assists percentage: Assists divided by field goals made.
Adj__Off__Eff_	Adj. Off. Eff.	Adjusted offensive effiency
AdjD	AdjD	Adjusted defensive efficiency – An estimate of the defensive efficiency (points allowed per 100 possessions) a team would have against the average D-I offense.
AdjEM	AdjEM	The difference between a team's offensive and defensive efficiency.
AdjO	AdjO	Adjusted offensive efficiency – An estimate of the offensive efficiency (points scored per 100 possessions) a team would have against the average D-I defense.
AdjOE	AdjOE	Adjusted offensive efficiency – An estimate of the offensive efficiency (points scored per 100 possessions) a team would have against the average D-I defense.
AdjT	AdjT	The number of possessions a team has per 40 minutes.
Avg__Poss_Length_Defence	Avg. Poss Length Defence	Average possession length for defense.
Avg__Poss_Length_Offence	Avg. Poss Length Offence	Average possession length for offense.
Blk_	Blk%	Block %: This is the percentage of opponents' two-point shots that are blocked by the player while he is on the court. It is computed by Blocks/(%Min * Opponents' two-point attempts). Anything greater than 8% is very good.
Conf	Conf	Team's athletic conference affiliation.
Def__2_Pt_FG__	Def. 2-Pt FG %	Defensive 2-point field goal percentage
Def__3_Pt_FG__	Def. 3-Pt FG %	Defensive 3-point field goal percentage
Def__Eff__Adj_	Def. Eff. Adj.	Defensive efficiency adjusted.
Def__Eff__Raw	Def. Eff. Raw	Defensive effciency raw.
Def__FT_Attempt__	Def. FT Attempt %	Defensive free throw attempts.
Def__FTRate	Def. FTRate	Defensive free throw rate.
Def__OR_	Def. OR%	Defensive offensive rebound percentage.
Def__TO_	Def. TO%	Defensive turnover percentage.
Def__eFG_	Def. eFG%	Defensive effective field goal percentage (eFG%).
FT_	FT%	Free Throw % – Free throws made divided by free throws attempted.
Losses	Losses	Number of losses.
Luck	Luck	A measure of the deviation between a team's actual winning percentage and what one would expect from its game-by-game efficiencies. It's a Dean Oliver invention. Essentially, a team involved in a lot of close games should not win (or lose) all of them. Those that do will be viewed as lucky (or unlucky).
NCAA_Tournament_Qualification	NCAA Tournament Qualification	Process by which teams are selected to participate in the NCAA Division I Men's Basketball Tournament.
NCSOS_AdjEM	NCSOS AdjEM	Non-Conference Strength of Schedule Adjusted Efficiency Margin.
NST_	NST%	Non-Steal Turnover %: Percentage of turnovers not accredited as a "steal".
Off_Eff__Raw	Off Eff. Raw	Offensive efficiency raw.
Off__2_Pt_FG__	Off. 2-Pt FG %	Offensive 2-point field goal percentage
Off__3_Pt_FG__	Off. 3-Pt FG %	Offensive 3-point field goal percentage
Off__FT_Attempt__	Off. FT Attempt %	Offensive free throw attempts percentage.
Off__FTRate	Off. FTRate	Offensive free throw rate.
Off__OR_	Off. OR%	Offensive rebound percentage.
Off__TO_	Off. TO%	Offensive turnover percentage.
Off__eFG_	Off. eFG%	Effective field goal percentage (eFG%).
Raw_Tempo	Raw Tempo	Raw tempo.
Stl_	Stl%	This is the percentage of possessions that a player records a steal while he is on the court. It is computed by Steals/(%Min * Team Possessions). Anything greater than 5% is very good.
Strength_of_Schedule_AdjEM	Strength of Schedule AdjEM	A team's strength of schedule: Overall strength of schedule of a team.
Strength_of_Schedule_OppD	Strength of Schedule OppD	A team's strength of schedule: Opponent's average adjusted defensive efficiency .
Strength_of_Schedule_OppO	Strength of Schedule OppO	A team's strength of schedule: Opponent's average adjusted offensive efficiency.
Team___Year	Team & Year	Team name and year.
Win__	Win %	Win percentage.
Wins	Wins	Number of wins.

Appendix B: Descriptive Statistics of Quantitative Variables

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
A_	INPUT	50.72334	5.234784	721	0	35.3	50.4	68.84	0.23501	0.124688
AdjD	INPUT	103.8683	6.423278	721	0	85	104.07	120.71	-0.20681	-0.34901
AdjEM	INPUT	-0.00006	11.48658	721	0	-31.57	-0.97	30.95	0.204298	-0.47502
AdjO	INPUT	103.8687	6.770562	721	0	79.4	103.67	122.3	0.06164	-0.08985
AdjOE	INPUT	103.8681	6.768237	721	0	79.4	103.7	122.3	0.061089	-0.08978
AdjT	INPUT	67.13718	2.448834	721	0	59.07	67.1	73.6	-0.00666	-0.02112
Adj__Off__Eff_	INPUT	103.868	6.768125	721	0	79.4	103.7	122.3	0.061113	-0.08963
Avg__Poss_Length_Defence	INPUT	17.59723	0.564376	721	0	15.5	17.6	19.4	-0.17457	0.209837
Avg__Poss_Length_Offence	INPUT	17.63245	1.10314	721	0	14.6	17.6	21.3	0.109849	0.102127
Blk_	INPUT	8.918724	2.074649	721	0	3.57	8.75	16.65	0.386562	0.198417
Def__2_Pt_FG__	INPUT	50.84588	3.777035	721	0	39.8	50.9	61.01	-0.04603	-0.14599
Def__3_Pt_FG__	INPUT	31.02053	3.768484	721	0	21.2	30.8	42.7	0.248877	-0.00297
Def__Eff__Adj_	INPUT	103.8691	6.424007	721	0	85	104.1	120.7	-0.2066	-0.34947
Def__Eff__Raw	INPUT	102.8047	5.379953	721	0	87.9	102.6	118.7	0.061377	-0.25975
Def__FTRate	INPUT	31.20674	5.317816	721	0	16.42	30.6	51.4	0.512564	0.37765
Def__FT_Attempt__	INPUT	18.132	2.574525	721	0	10.48	17.99	27.2	0.312058	0.089686
Def__OR_	INPUT	28.36816	2.994483	721	0	18.81	28.3	37.6	0.080337	-0.08717
Def__TO_	INPUT	18.28	2.327157	721	0	12.4	18.07	27.88	0.479883	0.524114
Def__eFG_	INPUT	50.31201	2.733288	721	0	41.3	50.4	58.85	-0.04578	-0.04017
FT_	INPUT	71.60017	3.678571	721	0	60.5	71.6	83	-0.1367	-0.0616
Losses	INPUT	14.04854	5.079821	721	0	2	14	28	0.195831	-0.49667
Luck	INPUT	0.001581	0.056293	721	0	-0.14	0	0.18	0.130125	-0.24863
NCSOS_AdjEM	INPUT	-0.56638	4.652387	721	0	-13.35	-0.77	16.98	0.283839	0.244949
NST_	INPUT	9.026852	1.332453	721	0	5.6	9	14	0.260966	0.324495
Off_Eff__Raw	INPUT	102.3447	5.887841	721	0	81.5	102.4	120.2	-0.08919	0.112672
Off__2_Pt_FG__	INPUT	50.91021	4.344887	721	0	37.35	50.9	63.17	0.032998	-0.15143
Off__3_Pt_FG__	INPUT	31.00907	4.764705	721	0	15.8	30.99	46.75	0.034012	0.091262
Off__FTRate	INPUT	30.96146	4.483387	721	0	20.4	30.7	47.6	0.356354	0.208633
Off__FT_Attempt__	INPUT	18.07993	2.337284	721	0	10.72	18	26.2	0.143213	0.072327
Off__OR_	INPUT	28.17741	3.98257	721	0	14.17	28.19	39.2	-0.08825	-0.15366
Off__TO_	INPUT	18.34194	2.036854	721	0	12.61	18.28	26.08	0.16818	0.434462
Off__eFG_	INPUT	50.11811	2.818932	721	0	41.25	50.2	59.25	-0.04491	0.107242
Raw_Tempo	INPUT	68.03135	2.54502	721	0	59.5	67.9	74.6	0.026195	-0.05099
Stl_	INPUT	4.689223	4.711019	721	0	0.06	0.14	13	0.112905	-1.86409
Strength_of_Schedule_AdjEM	INPUT	-0.28773	5.869347	721	0	-13.25	-1.48	13.38	0.445011	-0.77677
Strength_of_Schedule_OppD	INPUT	104.0372	3.206262	721	0	95.4	104.4	110.9	-0.36984	-0.64896
Strength_of_Schedule_OppO	INPUT	103.7494	3.242929	721	0	95.5	103.5	112.1	0.205475	-0.64638
Win__	INPUT	0.519709	0.174968	721	0	0.07	0.53	0.92	-0.1663	-0.62344
Wins	INPUT	15.38558	5.774803	721	0	2	15	34	0.250264	-0.22093
_2P_	INPUT	49.88305	3.250713	721	0	40.43	50	60.7	0.051904	0.094997
_3PA_	INPUT	37.51154	5.193069	721	0	22.2	37.51	54.8	-0.00099	-0.0083
_3P_	INPUT	33.68399	2.403807	721	0	25.2	33.7	43.7	-0.01236	0.227925

Appendix C: Regression Node 1 Variables

Name	Use	Report	Role	Level
A_	No	No	Input	Interval
AdjD	No	No	Input	Interval
AdjEM	No	No	Input	Interval
AdjO	No	No	Input	Interval
AdjOE	No	No	Input	Interval
AdjT	No	No	Input	Interval
Adj__Off__Eff_	No	No	Input	Interval
Avg__Poss_Leng	No	No	Input	Interval
Avg__Poss_Leng	No	No	Input	Interval
Blk_	No	No	Input	Interval
Conf	Yes	No	Input	Nominal
Def__2_Pt_FG_	No	No	Input	Interval
Def__3_Pt_FG_	No	No	Input	Interval
Def__Eff__Adj_	No	No	Input	Interval
Def__Eff__Raw	No	No	Input	Interval
Def__FTRate	No	No	Input	Interval
Def__FT_Attemp	No	No	Input	Interval
Def__OR_	No	No	Input	Interval
Def__TO_	No	No	Input	Interval
Def__eFG_	No	No	Input	Interval
FT_	No	No	Input	Interval
Losses	Yes	No	Input	Interval
Luck	No	No	Input	Interval
NCAA_Tourname	Yes	No	Target	Binary
NCSOS_AdjEM	No	No	Input	Interval
NST_	No	No	Input	Interval
Off_Eff__Raw	No	No	Input	Interval
Off__2_Pt_FG_	No	No	Input	Interval
Off__3_Pt_FG_	No	No	Input	Interval
Off__FTRate	No	No	Input	Interval
Off__FT_Attemp	No	No	Input	Interval
Off__OR_	No	No	Input	Interval
Off__TO_	No	No	Input	Interval
Off__eFG_	No	No	Input	Interval
Raw_Tempo	No	No	Input	Interval
Stl_	No	No	Input	Interval
Strength_of_Sch	Yes	No	Input	Interval
Strength_of_Sch	No	No	Input	Interval
Strength_of_Sch	No	No	Input	Interval
Win__	No	No	Input	Interval
Wins	No	No	Input	Interval
_2P_	No	No	Input	Interval
_3PA_	No	No	Input	Interval
_3P_	No	No	Input	Interval

Appendix D: Decision Tree Variables

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
_2P_	Input	Interval	No		No	.	.
_3P_	Input	Interval	No		No	.	.
_3PA_	Input	Interval	No		No	.	.
A_	Input	Interval	No		No	.	.
Adj__Off__Eff_	Input	Interval	No		No	.	.
AdjD	Input	Interval	No		No	.	.
AdjEM	Input	Interval	No		No	.	.
AdjO	Input	Interval	No		No	.	.
AdjOE	Input	Interval	No		No	.	.
AdjT	Input	Interval	No		No	.	.
Avg__Poss_Leng	Input	Interval	No		No	.	.
Avg__Poss_Leng	Input	Interval	No		No	.	.
Blk_	Input	Interval	No		No	.	.
Conf	Rejected	Nominal	No		No	.	.
Def__2_Pt_FG_	Input	Interval	No		No	.	.
Def__3_Pt_FG_	Input	Interval	No		No	.	.
Def__Eff__Adj_	Input	Interval	No		No	.	.
Def__Eff__Raw	Input	Interval	No		No	.	.
Def__FT_Attemp	Input	Interval	No		No	.	.
Def__FTRate	Input	Interval	No		No	.	.
Def__OR_	Input	Interval	No		No	.	.
Def__TO_	Input	Interval	No		No	.	.
Def__eFG_	Input	Interval	No		No	.	.
FT_	Input	Interval	No		No	.	.
Losses	Rejected	Interval	No		No	.	.
Luck	Input	Interval	No		No	.	.
NCAA_Tourname	Target	Binary	No		No	.	.
NCSOS_AdjEM	Input	Interval	No		No	.	.
NST_	Input	Interval	No		No	.	.
Off_Eff__Raw	Input	Interval	No		No	.	.
Off__2_Pt_FG_	Input	Interval	No		No	.	.
Off__3_Pt_FG_	Input	Interval	No		No	.	.
Off__FT_Attemp	Input	Interval	No		No	.	.
Off__FTRate	Input	Interval	No		No	.	.
Off__OR_	Input	Interval	No		No	.	.
Off__TO_	Input	Interval	No		No	.	.
Off__eFG_	Input	Interval	No		No	.	.
Raw_Tempo	Input	Interval	No		No	.	.
Stl_	Input	Interval	No		No	.	.
Strength_of_Sch	Input	Interval	No		No	.	.
Strength_of_Sch	Input	Interval	No		No	.	.
Strength_of_Sch	Input	Interval	No		No	.	.
Team___Year	ID	Nominal	No		No	.	.
Win__	Rejected	Interval	No		No	.	.
Wins	Rejected	Interval	No		No	.	.

REFERENCES

A&E Television Networks. (n.d.). *NBA is born*. History.com. Retrieved April 23, 2023, from https://www.history.com/this-day-in-history/nba-is-born

Beaudoin, D., & Duchesne, T. (2018). *Prediction of the margin of victory only from Team Rankings for regular season games in NCAA men's basketball*. Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology, 232(4), 315–322. https://doi.org/10.1177/1754337117754181

Birthplace of intercollegiate basketball. First Intercollegiate Basketball Game | Hamline University - MN. (2021). Retrieved April 23, 2023, from https://www.hamline.edu/about/first-intercollegiate-basketball-game#:~:text=Did%20you%20know%20Hamline%20is,the%20Blue%20Garden%20is%20today).

Boozell, J. (2017, March 28). *Is momentum overrated in March madness?* NCAA.com. Retrieved March 20, 2023, from https://www.ncaa.com/news/basketball-men/article/2017-03-27/momentum-overrated-why-march-madness-has-gone-script

Can the NCAA basketball tournament seeding be used to predict the margin of victory? Taylor & Francis. (n.d.). Retrieved March 19, 2023, from https://cogentoa.tandfonline.com/doi/abs/10.1080/00031305.1999.10474438

Cecchin, A. (2022). Oliver's four-factor model: Validation through causality. *International Journal of Sports Science & Coaching*, *17*(4), 838–847. https://doi.org/10.1177/17479541211049287

Conte, D., Tessitore, A., Gjullin, A., Mackinnon, D., Lupo, C., & Favero, T. (2018). Investigating the game-related statistics and tactical profile in NCAA Division I men's basketball games. *Biology of Sport*, *35*(2), 137–143. https://doi.org/10.5114/biolsport.2018.71602

Education, M. I. T. P. (n.d.). *The Science of Strength: How Data Analytics is transforming college basketball*. MIT News | Massachusetts Institute of Technology. Retrieved March 19, 2023, from https://news.mit.edu/2022/data-analytics-transforming-college-basketball-1017

Fearnhead, P., & Taylor, B. M. (2010). Calculating strength of schedule, and choosing teams for March madness. *The American Statistician*, *64*(2), 108–115. https://doi.org/10.1198/tast.2010.09161

Katz, J., & Fang, A. (2023, March 15). *N.C.A.A. bracket picks: Where fans and experts diverge*. The New York Times. Retrieved March 19, 2023, from https://www.nytimes.com/interactive/2023/03/15/upshot/ncaa-bracket-picks-table.html#commentsContainer

Magel, R., & Unruh, S. (2013). Determining factors influencing the outcome of college basketball games. *Open Journal of Statistics*, *03*(04), 225–230. https://doi.org/10.4236/ojs.2013.34026

Musial, N. (2023, March 15). *How Kenpom's Advanced Stats & the "four factors" can help with 2023 NCAA Tournament bracket picks & betting*. Sporting News. Retrieved March 19,

2023, from https://www.sportingnews.com/us/ncaa-basketball/news/kenpom-advanced-stats-four-factors-2023-ncaa-tournament-bracket-picks-betting/wqkdfjur1odxo1crovj6btga

Mutimer, B. T. (1997). Basketball: Its origin and development. *Sport History Review*, *28*(1), 73–74. https://doi.org/10.1123/shr.28.1.73

Oliver, D. (2004). *Basketball on paper: Rules and tools for performance analysis*. Brassey's, Inc.

Pomeroy, K. (2023) Pomeroy College Basketball Ratings. (n.d.). Retrieved March 19, 2023, from https://kenpom.com/

Singer Mar 12, M. (2015, June 2). *NCAA committee chair: Injuries will play a factor in seeding*. CBSSports.com. Retrieved March 20, 2023, from https://www.cbssports.com/college-basketball/news/ncaa-committee-chair-injuries-will-play-a-factor-in-seeding/

Springfield college. Springfield College. (n.d.). Retrieved April 23, 2023, from https://springfield.edu/where-basketball-was-invented-the-birthplace-of-basketball

Stekler, H. O., & Klein, A. (2012). Predicting the outcomes of NCAA Basketball Championship Games. *Journal of Quantitative Analysis in Sports*, *8*(1). https://doi.org/10.1515/1559-0410.1373

Stone, D. F., & Arkes, J. (2018). March madness? underreaction to hot and cold hands in NCAA basketball. *Economic Inquiry*, *56*(3), 1724–1747. https://doi.org/10.1111/ecin.12558

Toole, T. C. (2021, May 4). *Here is the history of basketball-from peach baskets in Springfield to global phenomenon*. History. Retrieved April 23, 2023, from

https://www.nationalgeographic.com/history/article/basketball-only-major-sport-invented-united-states-how-it-was-created?loggedin=true&rnd=1682298461411

Wilco, D. (2019). *Explaining college basketball's strength of Schedule*. NCAA.com. Retrieved March 20, 2023, from https://www.ncaa.com/news/basketball-men/article/2019-01-16/explaining-college-basketballs-strength-schedule

Wilco, D. |. (2023, March 15). *What is March Madness: The NCAA tournament explained*. NCAA.com. Retrieved March 20, 2023, from https://www.ncaa.com/news/basketball-men/bracketiq/2023-03-15/what-march-madness-ncaa-tournament-explained

Zimmermann, A. (2016). Basketball predictions in the NCAAB and NBA: Similarities and differences. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *9*(5), 350–364. https://doi.org/10.1002/sam.11319