Final Paper

STOR 320 Group 15

December 06, 2023

INTRODUCTION

Being students at UNC, we love UNC basketball and college basketball as a whole. Every year we watch and hope for a successful season from our Tar Heels. From our passion for this team we discovered some questions that we'd like to answer to get a better understanding of the numbers behind college basketball and how they can be used for predictions.

The first question that we wanted to find an answer to is can we predict the win percentage of a college basketball team using key basketball metrics? As well as, can we identify which specific factors have the greatest impact on win percentage? This question came to our minds first as a good winning percentage is the first step to making it to The Big Dance (March Madness Tournament). We believe that the answer to this question is most helpful to college coaches, analysts, and players, helping them best prepare for success. Fans may also like to see these answers to help them predict whether or not their favorite team will have a high winning percentage and therefore make the tournament.

Next we thought that if we can predict a step to making The Big Dance, could we develop an accurate predictive model using key basketball metrics to predict a team's chances of making the tournament, and if so, becoming the champion? Finding out the answer to this question would help out players and coaches find what to spend more time practicing on, as well as bracketologists to build accurate postseason brackets. Who knows, maybe by using our model a fan will be able to predict a perfect NCAA March Madness Bracket!

DATA

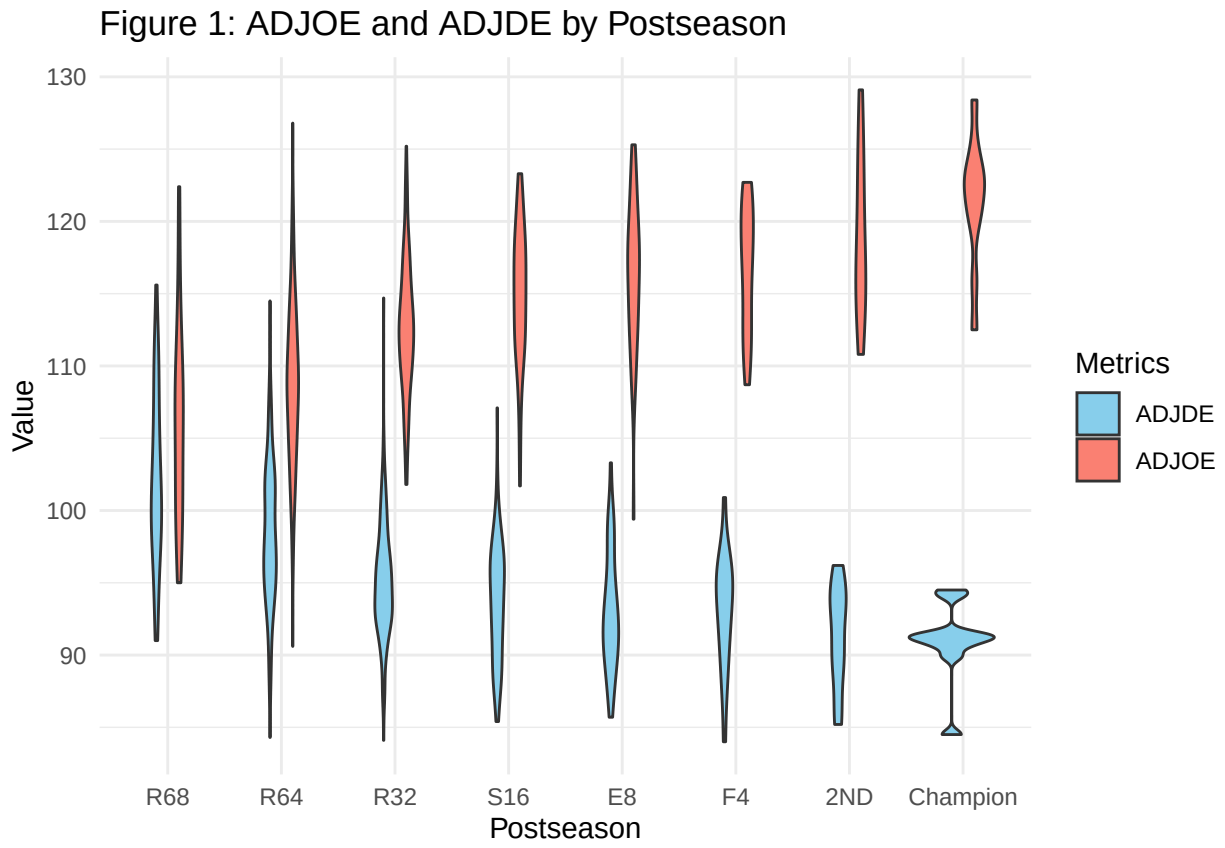
The dataset utilized in this study encapsulates Division I college basketball seasons spanning from 2013 to 2023. This dataset consolidates information primarily obtained from multiple sources, particularly (http://barttorvik.com/trank.php). Initially collected and cleaned from this website, it was subsequently enriched with supplementary variables like *POSTSEASON*, *SEED*, and *YEAR*. It encompasses a comprehensive array of team-centric statistics and performance metrics, incorporating offensive and defensive efficiency ratings, shooting percentages, turnover rates, rebounding metrics, tempo estimates, postseason achievements, and more.

Highlighting key variables central to our research, our group introduced essential elements such as "*Postseason_Success*" and "*WP*". *WP*, a derived variable representing "win percentage," offers a tangible metric derived from Games played and Win variables, enhancing its utility in our models. Postseason_Success categorizes teams based on their tournament success levels, distinguishing between "**Did_Not_Make_Tournament**," "**Early_Rounds**," "**Final_Four**," "**Runner_Up**," and "**Champions.**" This detailed variable proves instrumental in our models and conclusions regarding the variables' impact on predicting tournament outcomes. Postseason_Success is critical in gauging a team's success, given the limited number of teams that advance to later tournament rounds, as shown in Table 1 below.

Table 1: Percentage of Teams at Each Stage of the Postseason over the past 10 Years

Postseason_Success	Percentage
Champions	0.28
Did_Not_Make_Tournament	80.70
Early_Rounds	18.17
Final_Four	0.57
Runner_Up	0.28

Additionally, focusing on crucial statistics, "*EFG_O*" and "*EFG_D*" detail a team's effective field goal percentage on offense and their opponent's percentage during defensive plays. These metrics encompass both two-pointers and three-pointers, providing more weight to three-pointers due to the larger reward from the shot, summarizing shooting efficiency in a single metric. Furthermore, "*BARTHAG*," a power rating, represents a team's chance of beating an average Division 1 team. Lastly, "*ADJOE*" and "*ADJDE*," denoting adjusted offensive and defensive efficiency respectively, estimate points scored or allowed per 100 possessions, adjusted for opponents' defenses or offenses. Figure 1 below shows the *ADJOE* and *ADJDE* distribution for each round of the tournament. As you can see, a team with a low *ADJDE* has a better defensive rating and a team with a high *ADJOE* has a better offensive rating. These metrics collectively offer insights into a team's success, ability, and both offensive and defensive capabilities.



Comprising **3523** observations spanning from **2013 to 2023**, each observation corresponds to a specific Division I college basketball team within a particular season. This comprehensive volume of data enables an intricate analysis of team performances, trends, and patterns across multiple seasons.

Acknowledging limitations within the dataset, a significant anomaly is the absence of data for the 2020

college season due to the unprecedented COVID-19 pandemic disruption. This absence of postseason play in 2020 poses challenges in establishing continuous trends or evaluating the impact of regular-season performance on postseason outcomes for that year. Moreover, while this dataset encompasses extensive team-level statistics and performance metrics, external factors such as player injuries, coaching changes, or unforeseen circumstances not accounted for in the dataset may influence team performances. Consideration of these external factors is vital when interpreting the results or drawing conclusions from the analysis conducted using this dataset.

RESULTS

Question 1

To answer the first question we chose to predict the **win percentage** of college basketball teams using a combination of key basketball metrics including adjusted offense, adjusted defense, turnover rates, free throw percentage and more. In this case both the response variable and the explanatory variables are quantitative and thus we chose to use multiple regression modeling techniques.

We began by building several types of models: empty, full, full interaction model, and two step-wise models. The full model was built using all of the quantitative variables in the dataset. The full interaction model was built using all of the quantitative variables and all of the interactions between them. The stepwise models (**step.out.full**, **step.out.fullinteract**) were using the stepwise model selection algorithm. They were built from the scope of empty to the scope of the full model and the full interaction model respectively. Stepwise model building is a model selection algorithm that implements both forwards selection to add new variables and backwards elimination to remove redundant variables. The algorithm aims to minimize the Akaike's Information Criterion (AIC) at each step. AIC is a criterion used to evaluate models while penalizing for unnecessary complexity of models.

After building these models, we compiled the metrics for each of these models. For comparison, we decided to use 4 metrics: Cross-Validation Root Mean Squared Error (RMSE), Cross-Validation Mean Absolute Error (MAE), AIC, and Adjusted R-Squared (AdjRSQ). Adjusted R-Squared is a metric that assesses the amount of total variability in the response variable explained by the set of explanatory variables. The Cross-Validation RMSE and MAE were calculated using the K-Fold Cross-Validation technique with 10 folds. K-Fold Cross Validation randomly splits the data into K folds. **K-1** folds are used to train the model and then predictions are calculated for the remaining fold not used to train the data. This is repeated K times such that each data point has been in the testing data set exactly once.

Using the calculated metrics as seen in Table 2 below, we see that the **Stepwise2** model seems to be the most optimal model. It minimizes RMSE, MAE, and AIC and maximizes AdjRSQ. The **Stepwise2** model is the model that utilizes stepwise model building up to the scope of the full interaction model. The **Stepwise2** model explains 84.71% of the total variability in Win Percentage providing us with a specific metric to answer how well we can predict Win Percentage using a set of basketball metrics.

Table 2: Model Comparison

Model	RMSE	MAE	AIC	AdjRSQ
Stepwise2	7.1537	5.6530	23848.77	0.8471
Stepwise1	7.2538	5.7112	23954.75	0.8415
Full	7.2660	5.7189	23962.00	0.8414
Full Interact	7.3104	5.7610	23950.15	0.8471
Empty	18.1685	14.8945	30432.08	0.0000

The second part of our question aimed to understand the specific factors which had the greatest impact in predicting win percentage. When a model is built, we can conduct individual t-tests for the coefficients of each variable to assess its usefulness in predicting the response variable. We conducted individual t-tests using the optimal **Stepwise2** model to identify the variables which had significance in predicting Win Percentage. Table 3 contains the list of significant variables. Each of these variables have a p-value of less than 0.05 . This represents significance and provides us with evidence that the variable is useful in predicting Win Percentage. The p-value specifically represents the probability that this happened by chance and thus we want to minimize the probability of this being due to a one-off random occurrence. The set of variables and variable interactions in Table 3 provide the most significant variables in predicting Win Percentage.

Table 3: Significant Variables in Stepwise2

term	estimate	std.error	statistic	p.value
2P_D	0.611917	0.087507	6.992748	0.000000
TOR	-9.866706	1.450131	-6.804010	0.000000
EFG_O	5.078681	0.765062	6.638260	0.000000
FTR	2.363255	0.399623	5.913718	0.000000
ADJOE	-3.412031	0.586382	-5.818782	0.000000
ADJOE:ADJDE	0.026493	0.005401	4.905340	0.000001
EFG_O:DRB	-0.113797	0.023407	-4.861682	0.000001
DRB:TOR	0.139672	0.028805	4.848941	0.000001
TORD:FTRD	0.037869	0.009034	4.192020	0.000028
(Intercept)	383.177000	91.738666	4.176832	0.000030
ADJOE:DRB	0.061100	0.014961	4.083918	0.000045
FTR:ADJDE	-0.015951	0.003920	-4.069191	0.000048
ORB	2.731067	0.695550	3.926486	0.000088
DRB:ORB	-0.051121	0.013184	-3.877375	0.000108
DRB:ADJDE	0.028157	0.008330	3.380335	0.000732
TORD	-5.226433	1.647435	-3.172466	0.001525
DRB	-5.632247	1.832260	-3.073934	0.002129
EFG_D:TORD	0.063578	0.020905	3.041308	0.002373
ADJDE	-2.398555	0.789971	-3.036256	0.002413
FTRD	-1.376845	0.489693	-2.811653	0.004956
EFG_O:TORD	0.052781	0.020890	2.526607	0.011561
DRB:FTRD	-0.019091	0.007649	-2.495843	0.012612
ADJOE:EFG_D	-0.021308	0.010651	-2.000518	0.045522
TOR:ORB	0.030290	0.015161	1.997859	0.045809
TOR:ADJDE	0.024751	0.012601	1.964188	0.049588

Finally, we use diagnostic plots to assess the conditions and assumptions made when building a multiple regression model. Figure 2 shows the Residual vs. Fit plot. The scatterplot is generally centered around 0 and depicts homoscedasticity for the most part satisfying the zero mean and constant variance conditions of regression. There is a small worry that the model consistently underpredicts at low fitted values as seen by the cluster of points in the top left of the graph. Figure 3 illustrates a Normal Quantile Plot of the Residuals. The plot contains minimal skew or deviation from the linear line and it seems in line to assume that the errors are normally distributed.

Figure 2: Residual vs Fit

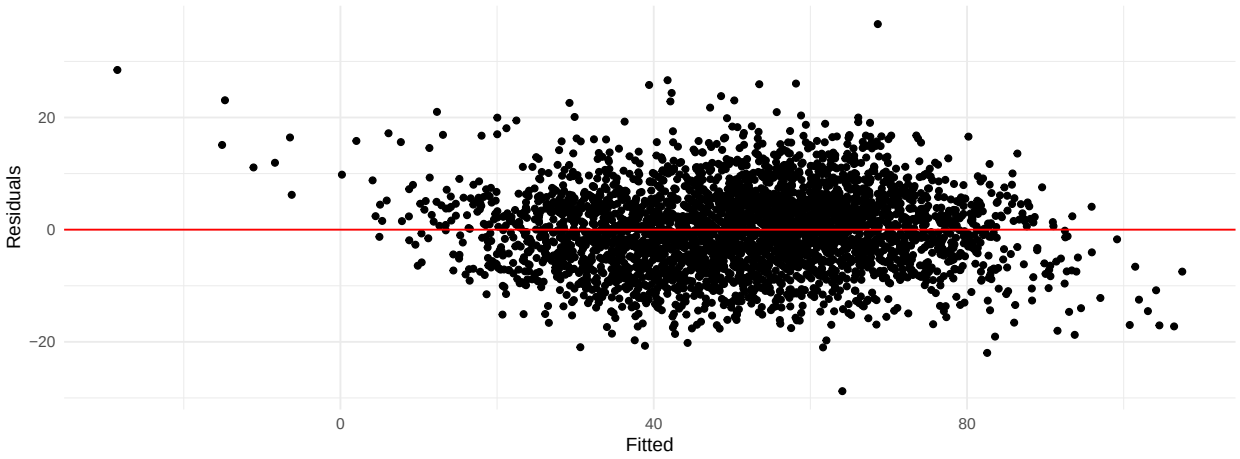
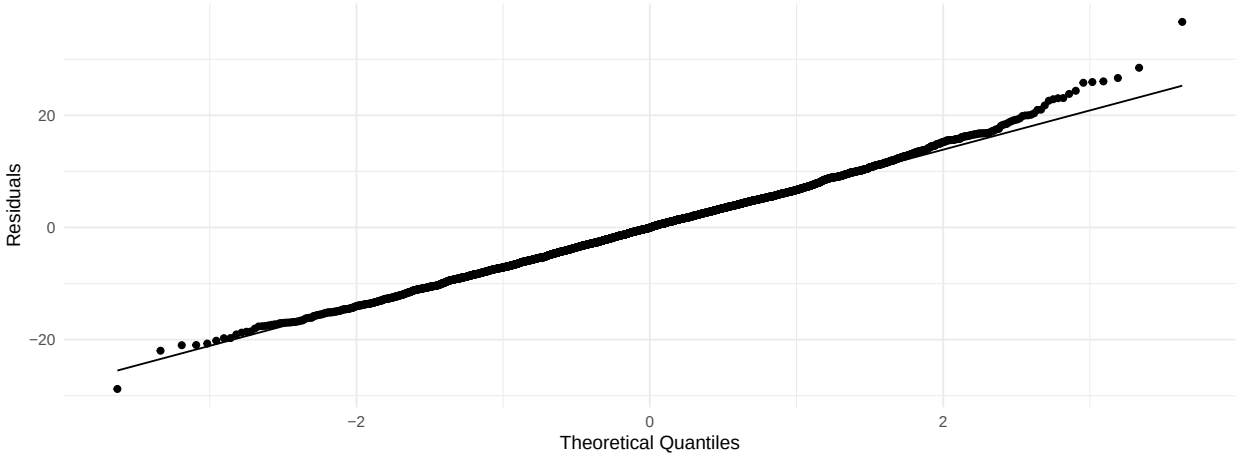


Figure 3: Normal Quantile Plot of Residuals



Question 2

For our projects second question, we tackled developing predictive models to predict where college basketball teams might end up during the postseason. We categorized them into five groups: **Champions**, **Runner-Up**, **Final Four**, **Early Rounds**, and **Not Making the Tournament**. To do this, we used the '*Postseason*' variable to create a new one named '*Postseason Success*', treating any missing (NA) values as not qualifying for the tournament. Our approach included two types of models: **Multinomial Logistic Regression** and **K-Nearest Neighbors** (KNN). Both models relied on important basketball stats like adjusted offensive and defensive efficiencies (*ADJOE* and *ADJDE*), effective field goal percentages (*EFG_O* and *EFG_D*), and win percentage (*WP*). To improve our models, we used cross-validation (CV). With the Logistic Regression models, we used our entire dataset in the CV phase, which automatically divides the data into training and testing sets. Following this, we assessed their performance on a separate test set that we had initially partitioned from our data. This approach allowed us to thoroughly evaluate the models

after we completed cross-validation. For the KNN model, we standardized the dataset to maintain data consistency and comparability as well as cross-validating the model. This process of training and testing was very important in refining the models' parameters, significantly improving their accuracy and usefulness in predicting the different possible postseason outcomes. This process ensured that we were accurately and correctly assessing the true predictive strength of our models.

In developing our **Multinomial Logistic Regression** models, we used varying predictor variables to see how they affected accuracy and predictions. We had two main types: the '**Multinom Model Based**', which used key predictors like adjusted offensive efficiency ($ADJOE$), adjusted defensive efficiency ($ADJDE$), Barthag rating ($BARTHAG$), effective field goal percentages for offense (EFG_O) and defense (EFG_D), and win percentage (WP). This model kept a balance between having a too many variables and being able to predict well. Then we have the '**Multinom Model All**', which used every single numeric variable in the dataset pertaining to basketball stats (not including seed and year). This gave us a really detailed view, but we had to be careful about overfitting so our model isn't too tailored to the training data and doesn't work well in real situations. We also made two specialized models: '**Multinom Model Offense**' and '**Multinom Model Defense**'. The offense one focused only on stats like EFG_O, two-point shot percentage ($TwoP\_O$), and three-point shot percentage ($ThreeP\_O$), while the defense one looked at $ADJDE$, EFG_D, and defensive rebounding (DRB). These models let us look into how a team's offensive or defensive skills affect their chances in the postseason. On top of these, we created a **K-Nearest Neighbors** (KNN) model. We decided to use KNN because it doesn't work like the logistic regression models. It groups teams based on how similar they are in key metrics. By adding KNN into the mix, we got to look at team performance in college basketball from a different angle, which made our overall analysis more well-rounded and gave us a better understanding of what drives a team's success in the postseason.

In evaluating our models' performance, which we've detailed in Table 4, we used a collection of metrics including Sensitivity, Specificity, Precision, Negative Predictive Value (NPV), and F1 Score. Sensitivity is about how well the model picks up actual winners (correctly calling a team as 'Champions'), while Specificity looks at its ability to rightly spot teams that won't go far (not making the 'Final Four'). Precision is how accurate the model is with its positive calls, NPV deals with the accuracy of negative predictions, and the F1 Score is a mix of Precision and Sensitivity. We chose these metrics to get a complete picture of how well each model is able to predict the five different postseason outcomes. Our '**Multinom Model Based**' was great at identifying 'Champions' with 100% Sensitivity, but it dropped to 0% for predicting 'Final Four' and 'Runner-Up', showing some clear gaps. Its Specificity and NPV was solid across the board, meaning it was good at recognizing when teams wouldn't hit certain levels. But, its Precision, like Sensitivity, was mixed - great for 'Champions', 'Early Rounds', and 'Did Not Make Tournament', but not so much for the mid-level categories. The F1 Score, which balances out Precision and Sensitivity, was perfect for 'Champions' but zero for 'Final Four' and 'Runner-Up', pointing out the model's uneven performance. The '**Multinom Model All**' did a bit better at predicting 'Final Four' but still had limited success. Its Specificity stayed high across all categories, and Precision and NPV were generally good, especially for 'Champions' and teams that didn't make the tournament. The F1 Scores followed this trend - high for some categories but lower for others. As for the **KNN Model**, it was great at predicting teams that wouldn't make the tournament but struggled with the 'Final Four' and 'Runner-Up'. Its Specificity was consistent, but Precision and NPV varied, suggesting that there's room for improvement. The F1 Scores mirrored these findings. Looking at the specialized models, '**Multinom Model Offense**' and '**Mutlinom Model Defense**' showed different levels of Sensitivity and Precision depending on the category. For instance, the 'Offense' model was okay for 'Early Rounds' but not great for 'Final Four' and 'Runner-Up'. These models, focusing on either offense or defense, had their strengths and weaknesses in predicting certain outcomes, as seen in their NPV and F1 Scores. Overall, these stats show that our models have complex predictive abilities. They're really good at the extremes like predicting 'Champions' or teams that won't make the tournament, but they find it tricky to get the predictions for the middle stages of the tournament correct. The varied performance across models in terms of Specificity, Precision, NPV, and F1 Scores really highlights the intricate nature of predicting postseason success in college basketball.

Table 4: Model Performances in Each Class

	Model	Sensitivity	Specificity	Precision	NPV	F1
Class: Did_Not_Make_Tournament	Multinom Model Based	0.9859155	0.6911765	0.9302326	0.9215686	0.9572650
Class: Early_Rounds	Multinom Model Based	0.6718750	0.9756944	0.8600000	0.9304636	0.7543860
Class: Final_Four	Multinom Model Based	0.0000000	1.0000000	0.0000000	0.9943182	0.0000000
Class: Runner_Up	Multinom Model Based	0.0000000	1.0000000	0.0000000	0.9971591	0.0000000
Class: Champions	Multinom Model Based	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000
Class: Did_Not_Make_Tournament1	Multinom Model All	0.9771127	0.7426471	0.9406780	0.8859649	0.9585492
Class: Early_Rounds1	Multinom Model All	0.7187500	0.9687500	0.8363636	0.9393939	0.7731092
Class: Final_Four1	Multinom Model All	0.2500000	0.9985714	0.5000000	0.9957265	0.3333333
Class: Runner_Up1	Multinom Model All	0.0000000	1.0000000	0.0000000	0.9971591	0.0000000
Class: Champions1	Multinom Model All	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000
Class: Did_Not_Make_Tournament2	Multinom Model Offense	0.9683099	0.5220588	0.8943089	0.7977528	0.9298394
Class: Early_Rounds2	Multinom Model Offense	0.4921875	0.9548611	0.7078652	0.8943089	0.5806452
Class: Final_Four2	Multinom Model Offense	0.0000000	1.0000000	0.0000000	0.9943182	0.0000000
Class: Runner_Up2	Multinom Model Offense	0.0000000	1.0000000	0.0000000	0.9971591	0.0000000
Class: Champions2	Multinom Model Offense	0.0000000	1.0000000	0.0000000	0.9971591	0.0000000
Class: Did_Not_Make_Tournament3	Multinom Model Defense	0.9577465	0.3970588	0.8690096	0.6923077	0.9112228
Class: Early_Rounds3	Multinom Model Defense	0.3593750	0.9444444	0.5897436	0.8690096	0.4466019
Class: Final_Four3	Multinom Model Defense	0.0000000	1.0000000	0.0000000	0.9943182	0.0000000
Class: Runner_Up3	Multinom Model Defense	0.0000000	1.0000000	0.0000000	0.9971591	0.0000000
Class: Champions3	Multinom Model Defense	0.0000000	1.0000000	0.0000000	0.9971591	0.0000000
Class: Champions4	KNN Model All	0.0000000	1.0000000	0.0000000	0.9971923	0.0000000
Class: Did_Not_Make_Tournament4	KNN Model All	0.9688599	0.5579420	0.9019029	0.8102886	0.9341831
Class: Early_Rounds4	KNN Model All	0.5364317	0.9568527	0.7334730	0.9031478	0.6196657
Class: Final_Four4	KNN Model All	0.0000000	1.0000000	0.0000000	0.9943846	0.0000000
Class: Runner_Up4	KNN Model All	0.0000000	1.0000000	0.0000000	0.9971615	0.0000000

Table 5 in our paper breaks down a side-by-side comparison of how each model did across various statistics. The standout is the '**Multinom Model All**', which compared to the other models has the most accurate predictive capabilities, with an Accuracy of 92.33% and a Kappa statistic of 0.740. This shows it's really good

at correctly classifying teams in different postseason scenarios. It's pretty balanced too, with an Average Sensitivity of 58.92% and Specificity of 94.20%, proving it's effective in spotting both true winners and teams that aren't going to go far. Its Precision and NPV are also solid, confirming it's reliable for accurate predictions. In comparison, the '**Multinom Model Based**' is a bit behind with an Accuracy of 92.04% and Kappa of 0.720. It's still strong, with decent Sensitivity and Specificity, but there's some room to get better at picking out all the true positives. Its Precision and NPV are pretty good, showing it's quite accurate in its predictions, whether positive or negative. The '**KNN Model All**' is alright overall, with an Accuracy of 87.96% and Kappa of 0.562. It's not quite as consistent or accurate as the logistic regression models, especially in Sensitivity and Precision. It's better at predicting what won't happen than what will, based on its Specificity and NPV. Then, there's the '**Multinom Model Offense**' and '**Defense**', each with its own performance pattern. The 'Offense' model has moderate effectiveness, but it could be better at identifying true positives. The 'Defense' model shows similar trends, with a need for improvement in correctly spotting positive cases. These different results across the models really highlight their unique strengths and weaknesses, for example, although **Multinom Model All** is the most accurate model according to this table(5), according to Table 4, **Multinom Model Based** has significantly greater predictive capabilities when it comes to specifically forecasting the Champion class. This further highlights why it's crucial to pick the right model for the specific analysis you're doing as well as the intricacies of predicting postseason success in college basketball.

Table 5: Side-by-side Model Comparison Across Various Statistics

Model	Accuracy	Kappa	Avg_Sensitivity	Avg_Specificity	Avg_Precision	Avg_NPV
Multinom Model Based	0.9204545	0.7201511	0.5315581	0.9333742	0.5580465	0.9687019
Multinom Model All	0.9232955	0.7403350	0.5891725	0.9419937	0.6554083	0.9636489
Multinom Model Offense	0.8707386	0.5251142	0.2920995	0.8953840	0.3204348	0.9361396
Multinom Model Defense	0.8380682	0.3829499	0.2634243	0.8683007	0.2917506	0.9099907
KNN Model All	0.8795779	0.5629071	0.3010583	0.9029589	0.3270752	0.9404350

Our models excelled in picking out the '**Champions**' and **those who wouldn't make the tournament**. This accuracy shows they're great at telling which teams have a strong or slim chance of postseason success. This kind of insight is super useful for making strategic calls and forecasts in basketball analytics. But, we hit some bumps when trying to predict the middle stages of the tournament, like the '**Final Four**' and '**Runner-Up**'. This part was tricky and it points to how complex and unpredictable these stages can be. It seems like there are factors affecting these outcomes that our models aren't fully catching. This might be because of the unpredictable elements in sports competitions - things like how the game flows, player performance, team spirit, coaching strategies, and other hard-to-measure metrics that really affects how games turn out. The mixed results we saw across different tournament stages mean there's room to make our models even better. It also makes us think about how adding more up-to-date, dynamic data could help. Like, if we include stats on how teams and players are doing during the games, we could get a clearer picture of those harder-to-predict stages. Adding these kinds of details could boost our prediction accuracy, especially for the trickier parts of the tournament. This way, we could offer a more rounded tool for predicting tournament outcomes and help out with strategic planning in basketball analytics.

Our research really puts a spotlight on how powerful statistical models like Multinomial Logistic Regression and K-Nearest Neighbors can be in sports analytics, especially for predicting college basketball postseason outcomes. These models were particularly **good** at calling the **extreme cases** - teams that would become 'Champions' or those not making the tournament at all. This shows a solid link between the models and key performance indicators in these scenarios. But, we also noticed some areas where the models weren't as sharp, like predicting who'd make it to the 'Final Four' or be the 'Runner-Up'. This part of the study really highlights how complex and uncertain sports can be, and it suggests that our models may need more advanced

and dynamic ways to predict these tournaments. We think some key factors that could really benefit the models predictive ability could be player-specific analytics, live game data, or even psychological or situational factors like match ups. These could give us a deeper insight into what influences team performance. By exploring these more detailed aspects, we could boost the accuracy of our predictions, especially for those crucial, game-changing stages of the tournament. In short, this study not only deepens our understanding of what current statistical models can and can't do in sports analytics but also opens the door to more detailed and explicit ways of predicting sports tournament outcomes. The findings could be super useful for analysts, sports strategists, and bracketologists, and they really show the varied and dynamic nature of sports competitions.

CONCLUSION

When looking at the results for question 1, "can we predict the win percentage of a college basketball team using key basketball metrics? As well as, can we identify which specific factors have the greatest impact on win percentage?", as expected we can conclude that we can predict the win percentage of a team using key basketball metric by using the **Stepwise2** model as it is the most optimal model we created. We can also conclude that there are many variables that are useful in predicting win percentage with *2P_D, TOR, EFG_O, FTR, and ADJOE* having the greatest impact for prediction. The answers to this question has many real world applications. The most import ones being for coaches and players. By using this model coaches and players are able to see what will help them to win more games, as well as see which type of statistic they have to work on to see great improvement in their win percentage. This will also change the way coaches will strategize for each game. An alternate application would be for sports gamblers(in legal states). If they can predict the winning percentage of teams, it can inform them on who to pick for their next bet. If someone was to continue to work on this problem, we would suggest to dig deeper into the most impactful variables and to figure out what goes into calculating these variables. This would most likely take more web scrapping to find the factors of the variables, but in the end it would give the coaches and players more information on what they'd have to improve on to get better records.

For question 2, "can we develop an accurate predictive model using key basketball metrics to predict a team's chances of making the tournament, and if so, becoming the champion?", we got the conclusion we expected. We can conclude that after test multiple different models that we have an accurate predictive model to predict a team's chances of making the tournament and the final four. The most accurate model we developed was the **Multinom Model All** model, however, we learned that this model is very good at predicting the extremes. It is a good way to determine if a team will make the tournament at all or if the team has a high chance to become champion. However, predicting what round a team will be eliminated within the tournament would be a challenge for this model. We believe that the models we developed for this question would be most useful for bracketologists trying to predict their March Madness champion. To continue this work, we would recommend expanding on our model to try to make a model that can predict what round a team will be eliminated in the tournament. By doing this the bracketologists will be able to have an accurate prediction for their whole March Madness bracket. This will most like need a lot more data that we didn't have, as there are a lot of factors that go into things once in the tournament.

Overall, we were able to find conclusions to our questions and they were what we expected them to be. We expected to build a model to to explain a significant amount of variability of the response variable ***win percentage*** and we were able to accomplish this by building an optimal model which explained 84.71% of the variability. We were also able to build models to predict a categorical variable, ***postseason success*** with over 90% accuracy. To further continue our work, we would recommend researchers to dive deeper into more complex models. Especially to answer the second question with multi-classification, untested methods of Random Forests, Neural Networks, and more could potentially be used to achieve greater accuracy specifically within the intermediate categories. Our models had high accuracy to predict champions or teams that did not make the tournament but more complex methods may be useful in further determining these differences. To further develop our ideas in the first question there is a potential that additional data such as data regarding player dynamics within teams, team injury lists, etc could help explain a greater percentage of the variation in ***win percentage***.