Basketball analysis

Tianrui Ye

January 8, 2023

1 Introduction

The purpose of this paper is to explore the performance of UNC basketball team in the NCAA, what factors affect the performance of the basketball team, and to analyze them quantitatively. This report mainly uses correlation analysis, regression analysis, machine learning and other methods to analyze the data set. The data set I used came from the cbb file in Kaggle's College Basketball Dataset. All the models and graphs in the text are from R and can be found under the ”Source Code” Branch.

2 Variables

$ADJOE$: Offensive efficiency, the higher the better
$ADJDE$: Defensive efficiency, the lower the better
$BARTHAG$: Chance of beating an average Division I team
EFG_O: Effective Field Goal Percentage Shot
EFG_D: Effective Field Goal Percentage Allowed
TOR: Turnover Percentage Allowed (Turnover Rate)
$TORD$: Turnover Percentage Committed (Steal Rate)
ORB: Offensive Rebound Rate
DRB: Offensive Rebound Rate Allowed
FTR: Free Throw Rate (How often the given team shoots Free Throws)
$FTRD$: Free Throw Rate Allowed
$2P_O$: Two-Point Shooting Percentage
$2P_D$: Two-Point Shooting Percentage Allowed
$3P_O$: Three-Point Shooting Percentage
$3P_D$: Three-Point Shooting Percentage Allowed
ADJ_T: Adjusted Tempo (An estimate of the tempo (possessions per 40 minutes) a team would have against the team that wants to play at an average Division I tempo)
WAB: Wins Above Bubble (The bubble refers to the cut-off between making the NCAA March Madness Tournament and not making it)

3 What factors will affect the performance of the UNC basketball team?

First of all, I explored which variables would affect the performance of the UNC basketball team and how to make UNC get better results. I compared the offensive and defensive efficiency of all the teams, which shows on 1, UNC is one of the most efficient offensive teams in the NCAA, as well as a highly efficient defense. But at the same time, it can be seen that the offensive efficiency of UNC in two years was not that high compared with other years. Through comparison of data, it can be confirmed that the two years were 2013 and 2014 respectively.

As the chart shows 1, both offensive and defensive efficiency was much worse than North Carolina's average in 2013 and 2014. The main difference was in the WAB, which was more losses. $EFGO$ means the UNC basketball team's shooting percentage is lower, while $EDGD$ is higher compared to all North Carolina stats, which means the opponent's shooting percentage is higher. At the same time, the team's turnover rate is higher, and the offensive rebounding efficiency has decreased. These are

	$ADJOE$	$ADJDE$	WAB	EFG_O	EFG_D	TOR	$TORD$	DRB	ORB	FTR
All team	103.3	103.3	-7.802	49.81	50.0	18.76	18.69	30.08	29.88	35.99
2013 all team	100.8	100.8	-7.994	48.54	48.74	20.03	19.95	31.82	31.60	35.97
North Carolina	118.5	93.59	6.757	51.43	47.94	16.83	18.56	28.20	38.20	33.01
2013 North Carolina	100.8	100.8	-7.994	48.54	48.74	20.03	19.95	31.82	31.60	35.97
2014 North Carolina	104.58	104.6	-7.44	49.48	49.69	18.37	18.28	31.4	31.17	40.47

Table 1: Average stats for all team and North Carolina.

the main reasons for the team's offensive and defensive efficiency. And, to avoid the possibility that all teams will become more efficient offensively and defensively because of The Times, I've compiled an average for all teams in 2013. As you can see, there are no significant differences in each column from 2013 compared to all teams. Therefore, it can be shown that in 2013 and 2014, the offensive efficiency was significantly reduced due to the poor performance of UNC.

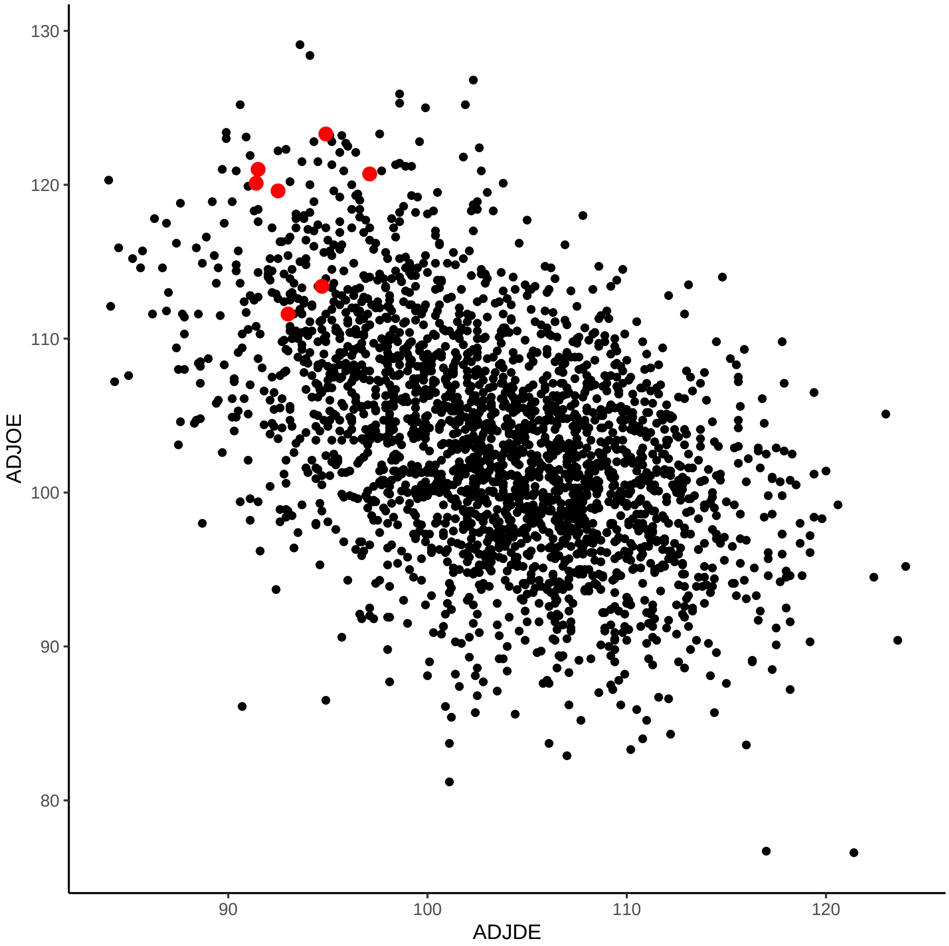

Figure 1: Offensive and defensive efficiency charts for all teams, the red dots are UNC basketball stats

Next, I wanted to quantify the impact of each factor on team performance. In this regard, I selected some variables with strong correlations, established a multiple regression model, and obtained the coefficient of each variable. And the model has been verified and is significant, and reasonable. All the coefficients of the variable are listed in figure 2. For example, the 1% increase in EFG_O will lead to a 0.852 increase in offensive efficiency.

So the team needs to increase EFG_O as much as possible and decrease TOR as much as possible to be more efficient offensively. In terms of defensive efficiency, the team needs to reduce $2P_D$ as much as possible and increase $TORD$. It's all very common sense.

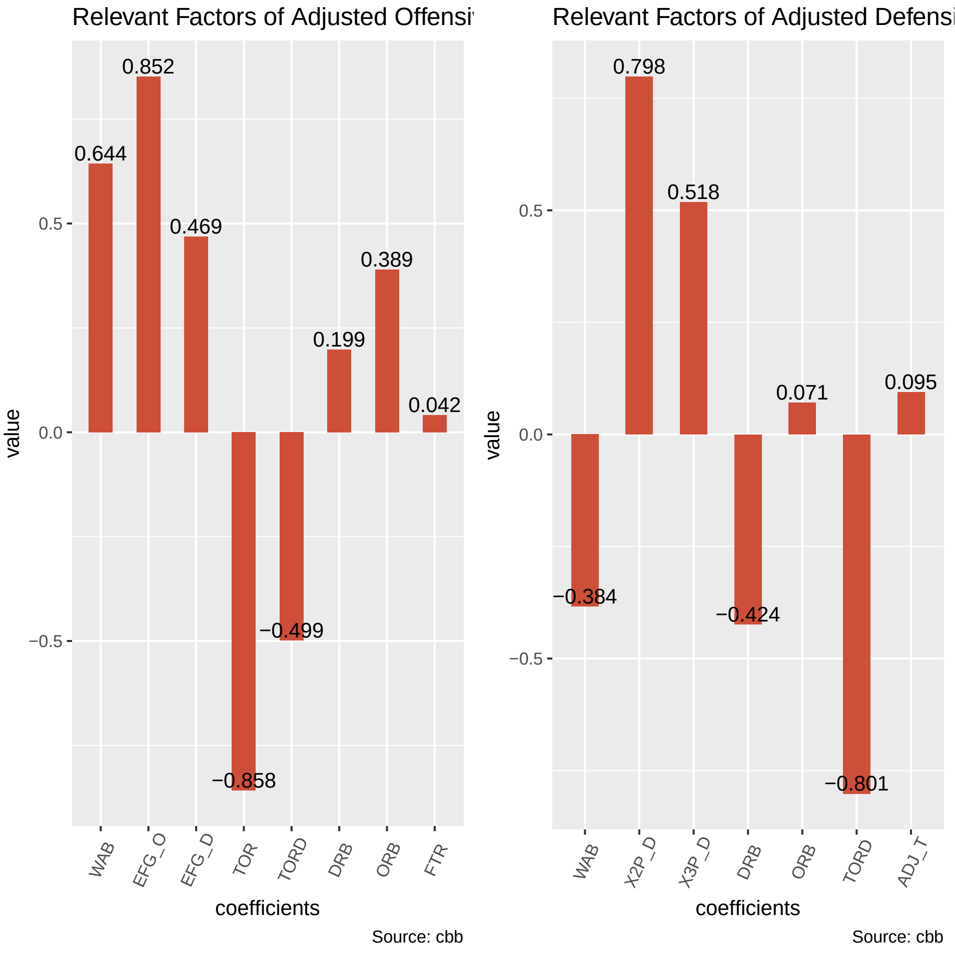

Figure 2: Relative Coefficients for each variable

	Multiple Regression Model	Random Forest Model	XGBoost Model
RMSE	0.0677	0.0857	0.0824

Table 2: Average stats for all team and North Carolina.

So far, I have explored and understood this problem. First, I compared the offensive and defensive efficiency of the UNC basketball team and made clear what aspects were not in place to cause the team's offensive and defensive efficiency to decrease. Then I used the data of all the teams to build a model, as a whole to illustrate a team to achieve better offensive and defensive efficiency needs to pay attention to what factors and quantified.

4 Predict the winning rate of the UNC basketball team

I created a new variable $Winrate$ using the ratio of team wins to the total number of team games in the data set. Also, use this variable as a response variable for the model. For this, I use three methods to build three different prediction models, which are the multiple regression model, random forest model, and xgboost model. Here, I use RMSE as the evaluation standard of the model. RMSE is a common index to measure the error of the model in the prediction, that is, the smaller the RMSE is, the better the model will be in the prediction. The RMSE for the three models are shown in table 2

It can be seen that the RMSE of the three prediction models is similar, but the multiple regression model has a relatively better RMSE. Therefore, we will use a multiple regression model in the later

	Actual Win Rate	Predicted Win Rate	Prediction Bias
2013	0.6857143	0.6937745	1.1754%
2014	0.7058824	0.7060511	0.0239%
2015	0.6842105	0.7044272	2.9547%
2016	0.825	0.7620720	7.6276%
2017	0.8461538	0.8140548	3.7935%
2018	0.7027027	0.6641283	5.4894%
2019	0.8055556	0.7838097	2.6995%

Table 3: The Prediction and Actual Win Rate for North Carolina each year

predictive analysis. At the same time

Figure 3 depicts the importance of each variable in the model, that is, the degree to which it affects the win ratio. The more important the variable, the stronger the influence on the winning rate. As you can see from the figure, $TORD$ is the most important variable for winning percentage, along with EFG_D, DRB, TOR, EFG_O and ORB, which are more than 15% important. These are relatively more important variables, so teams should put more effort into improving them if they want to improve their winning percentage.

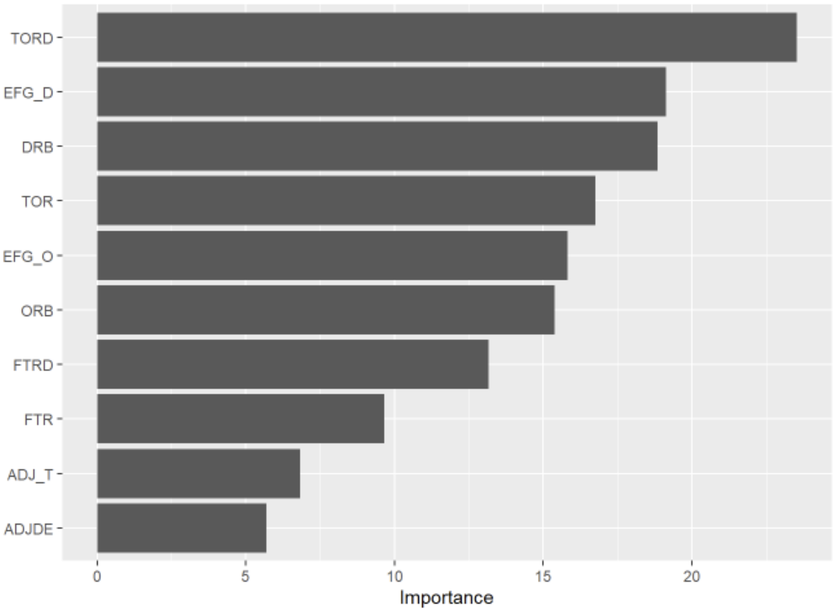

Figure 3: Relative Importance for each variable in prediction model

In Table 3, I present the predicted results, and I use data from North Carolina specifically to present the results. Prediction Bias refers to a ratio of the difference between the true value and the predicted value. As can be seen from the table, the predicted value and the real value are relatively close, and the error is not more than 5% except in 2016 and 2018. It indicates that I have established a relatively accurate model and can predict the data of UNC basketball team in the future.

To sum up, I established a new variable $Winrate$ to measure the team's performance, and used it as a response variable to establish three prediction models using three different methods. The optimal multiple regression model was selected by comparing RMSE. Then I tested the importance of the variables and picked out some variables that were relatively more important for the winning percentage. That's where teams could consider spending more time. I used North Carolina data to test my model, and the results were good.

5 Classify Teams

In my knowledge, UNC basketball team has always been regarded as one of the top teams in the NCAA. However, we have discussed in the first part that UNC basketball team did not perform well in 2013 and 2014. I wonder if the UNC basketball team can still be called a top team these two years. And what are the top teams statistically speaking? To explore this problem, I use the K-mean clustering algorithm. This is a machine-learning algorithm for categorization. As I see it, I'm going to divide all the teams into three groups: "top teams," "average teams," and "bad teams." The classification results are shown in Figure 4.

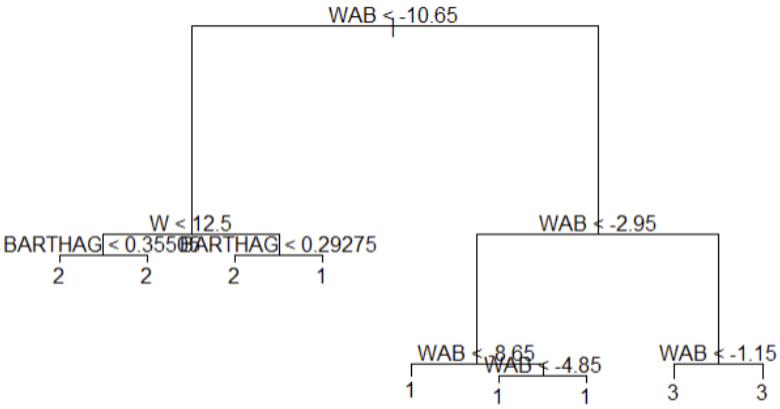

Figure 4: Classfication Results: 1,2,3 means three groups

It can be seen clearly from the picture. Teams with $WAB > -2.95$ can be considered as group 3, or the top teams; $-10.65 < WAB < -2.95$ can be considered as an average team; $WAB < -10.65$ could be considered as a bad team. WAB means the difference between the number of actual wins and the number of wins needed for playoff games, positive if above, negative if below. So all teams in the playoffs and those 2.95 wins away from the playoffs can be considered top teams. Any team that falls more than 10.65 games short of the number of wins needed to make the playoffs can be considered a bad team. But there are some exceptions. Teams that fall more than 10.65 games short of the number of wins needed to make the playoffs, but exceed 12.5 wins and have a $BARTHAG > 0.2975$ are considered average teams. Finally, teams that are more than 2.95 wins short of the number needed to make the playoffs, but less than 10.65, can be considered average teams.

The whole category is based on WAB, which I think makes sense because the team that ends up with more wins should be considered the better team.

It's time to get back to our main question, which is whether the 2013 and 2014 UNC basketball teams can still be considered top teams. The data set gives the UNC basketball team with WAB 2.5 and 4.2 in 2013 and 2014, respectively. That's a poor number compared to UNC's performance in other years, but it's above my model's standard for top teams. So even though UNC's performance in these two years is not very good, it is still one of the top teams.

In this part of the analysis, I used the K-mean clustering algorithm to classify the team's performance. I've given a clear breakdown of the top teams, the average teams, and the bad teams. And this part of the analysis also explains that the performance of the UNC basketball team has always been among the top teams.