NCAA Men's March Madness Analysis

Julia Mengxuan Yu

Data

In this analysis, I have used the Andrew Sundberg's College Basketball Dataset via Kaggle, which contains data from 2013 to 2021 season. This data contains 2455 observations and 24 variables. Among these variables, TEAM, CONF, and POSTSEASON are character variables and others are numerical. There are 1979 missing values in bother POSTSEASON and SEED, since POSTSEASON stands for round where the given team was eliminated or where their season ended and SEED stands for seed in the NCAA March Madness Tournament. Hence, I have replaced the null value with "Eliminated" and "0" respectively.

Explanatory Plots

First, I explore ADJOE (adjusted offensive efficiency) and ADJDE (adjusted defensive efficiency) since they are considerable factors related to the success of a basketball game. ADJOE and ADJDE are important measures of a team's offensive and defensive performance. And intuitively, a team with a higher ADJOE will be more likely to score more points and vice versa. Figure I is aligned with my expectation. Champions' offensive performances are quite outstanding among these team. And even though ADJDE is not as significant as ADJOE, champion teams are still superior to most of the teams.

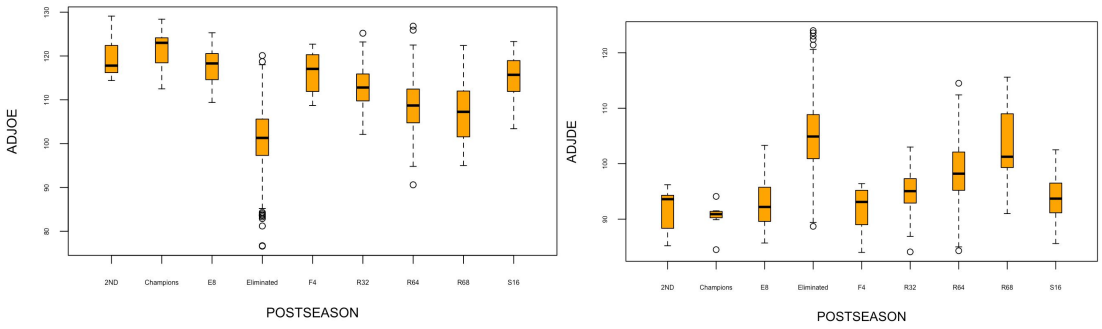

Figure I

To explore the relationship between ADJOE (Adjusted Offensive Efficiency), ADJDE (Adjusted Defensive Efficiency) and other factors, I extracted two dataset consisting the offensive statistics and defensive statistics. With these two correlation heat maps in Figure II, I found that the correlation between ADJOE and EFG_O

(Effective Field Goal Percentage Shot) is 0.73, the one between ADJDE and EFG_D (Effective Field Goal Percentage Allowed) is 0.8. Besides, the negative correlation between ADJOE and TOR (Turnover Percentage Allowed) is quite significant. Other correlation such as between ADJOE and ORB (Offensive Rebound Rate), between ADJDE rate and DRB (Defensive Rebound Rate), etc are also reasonable.

Moreover, from Figure II, I find that the correlation between two point shooting percentage and three point shooting percentage is 0.42, which fits my intuition that even though attacking strategies may differ, the reasons behind success of attacks are related, which might include players' abilities, team strategy, etc.

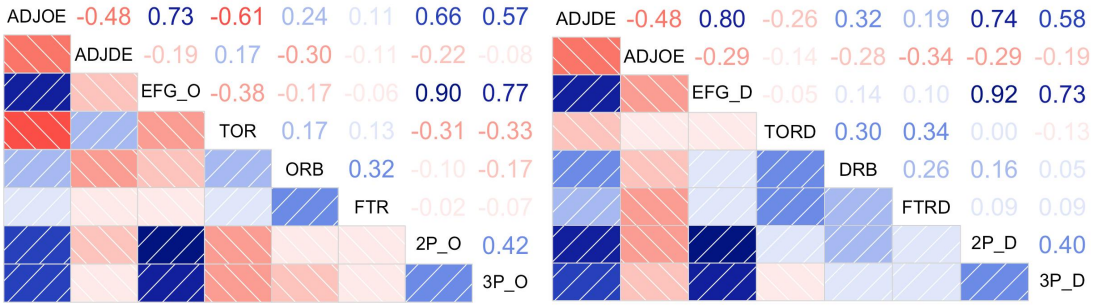

Figure II

Then, I try to find the relationship between turnover rate and adjusted offensive efficiency. A lower turnover rate means more offensive possessions and hence more opportunities to score, which can potentially lead to a high adjusted offensive efficiency. Since adjusted offensive efficiency could be a significant factor influencing the performance of a team in the race, I also plot TOR (turnover rate) versus POSTSEASON. From Figure III, it is quite obvious that champion teams or those entering into Runner-up seem to have quite low turnover rate.

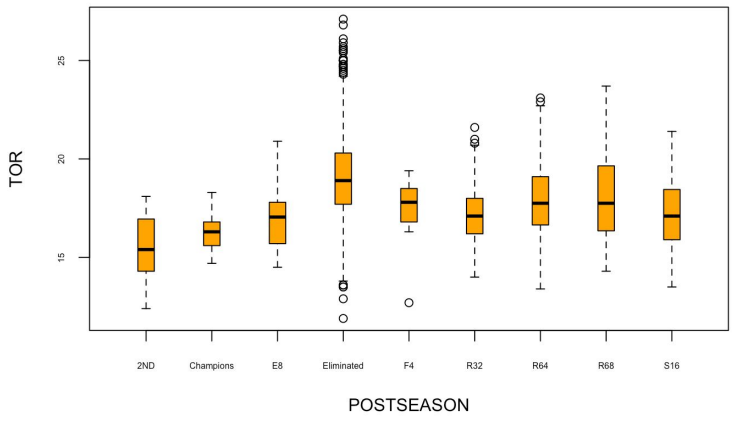

Figure III

Model

To predict the performance of a team (POSTSEASON) with given features, I use Naive Bayes and K-Nearest Neighbors (KNN) to build models.

First, I divide the dataset into training set (80%) and test set (20%). Then I run Naive Bayes algorithm on the training set, and use test set for the misclassification rate, which turned out to be about 0.014 (Figure IV). The extreme overfitting of this model is not surprising here. Even though Naive Bayes model is an efficient model that can be used when there are a large number of features, it makes the assumption of independence between features, which is often not the case in real-world datasets.

```
             y.test
nb.class      2ND Champions  E8 Eliminated  F4 R32 R64 R68 S16
  2ND           0         0   0          0   0   0   0   0   0
  Champions     1         1   0          0   0   0   0   0   0
  E8            0         0   6          1   0   0   0   0   0
  Eliminated    0         0   0        391   0   0   0   0   0
  F4            1         0   0          0   0   0   0   0   0
  R32           0         0   1          0   1  23   0   1   0
  R64           0         0   0          0   0   1  48   0   0
  R68           0         0   0          0   0   0   0   4   0
  S16           0         0   0          0   0   0   0   0  11
[1] 0.01425662
```

Figure IV

In this dataset, the number of observations is small, and the features are mostly related to our target variable. Hence, I decide to use KNN algorithm. KNN classification is a type of supervised machine learning algorithm. The basic idea behind KNN is to find the k-number of data points in the training set that are closest (or "nearest") to a given test point, and then use the majority class among these k-nearest neighbors to classify the test point. Therefore, I use cross-validation to find the optimal k value, which is 49. (Figure V) With the optimal k value, I get the test error of KNN model, 0.1079.

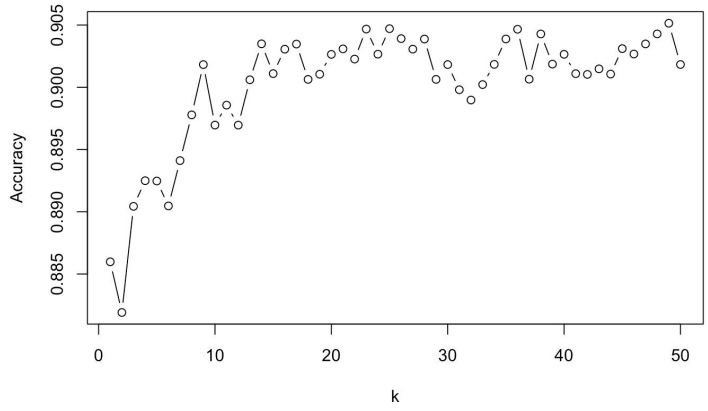

Figure V

Although KNN model provides a quite low test error, there are also other factors should be taken into consideration in the future research such as the number of observations, the distribution of data, etc. And we also need to try other possible algorithms.