MIS 776 - NCAA Basketball Executive Summary

Group Members: Dhiraj Chavva, Jordan Eisinger, Alireza Javid, Tsion Melaku, Chih-Hui (Kevin) Weng, Eric Yoon

Dataset Overview

The dataset we used had data on every NCAA Division 1 Basketball game from 1894 to 2016. In order to perform our analysis, we had to use the BigQuery API as the dataset was not a CSV file. There were 10 tables in total, each ranging in size from 4 to 132 columns. The largest table had over 500,000 rows. In total, there are 351 teams in the dataset.

We chose the NCAA Basketball dataset for three reasons: it is a unique dataset and an interesting topic for analysis, we wanted a more involved dataset compared to the examples in class, and we were all fans of the sport.

Our analysis focused on the following three business questions:

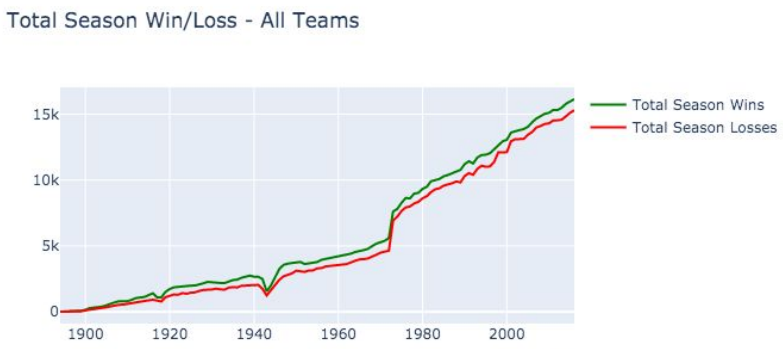
1) General trends in college basketball over time
2) Predicting season venue attendance
3) Predicting game winning factors for home and away teams

The following is a summary of our analysis, methods, and conclusions for each of these questions.

Business Question 1 - Trends in College Basketball Over Time

Since the NCAA Basketball dataset has data on every Division 1 game played from 1894 to 2016, the first logical area to focus on for an analysis is general trends over time. This gives us a general understanding of the data. The methods used in this section were primarily descriptive statistic and plots of the data over every season.

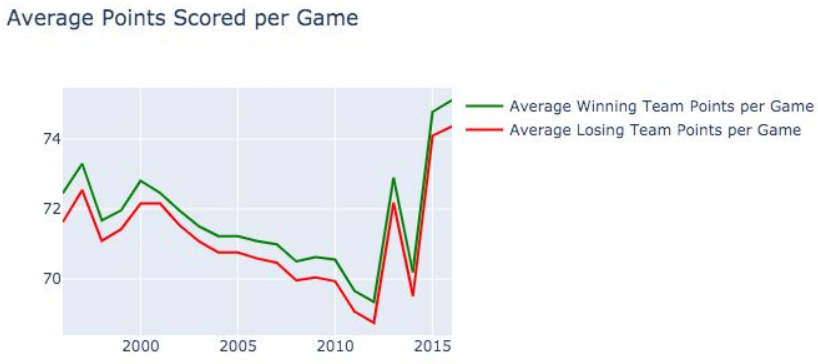
First, we looked at answering the question “Do modern teams win more on average than teams in the past?”. By aggregating and summarizing every win and loss for all 351 teams in every season and plotting this data, we are able to see if there are any trends:



From the plot above we see that there is a significant increase in the number of games won and lost over time. This is primarily explained by more teams/schools being in the dataset as the years progress. For example, the first tournament in 1939 was only an 8 team tournament compared to today where there are 64 teams. Nonetheless, the number of games won over any particular season has increased over time.

This trend is the same when looking at a team's average record over time as well. For example, at the start of the dataset the average number of wins in a season was about 5 where in 2016 the average was 14.

Next, we looked at the average score of a NCAA Basketball game over time to see if there were any major differences between seasons:



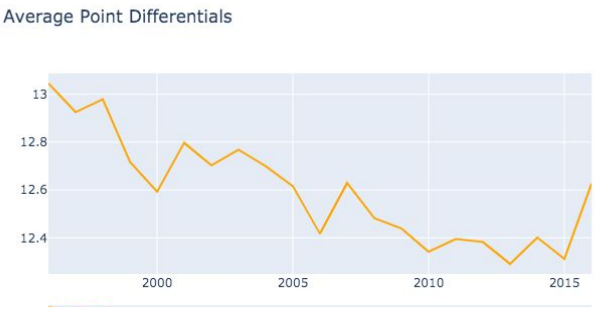
From the plot we see that the average score of a NCAA Basketball game decreased since 2000, hitting a low of 69. Only in the last four years of the data did the average spike up to a high of 75 points per game.

Lastly, we looked at point differentials over time to see if NCAA Basketball games today are more competitive. An example of a point differential is the following: If a game ends 110-100, then the point differential would be 10.

```
In [16]: # Descriptive statistics for every variable
         point_diff.describe().round(0)
```

Out[16]:

	season	points_game	opp_points_game	point_diff
count	506653.0	506653.0	506653.0	506653.0
mean	2007.0	72.0	71.0	13.0
std	6.0	14.0	14.0	10.0
min	1996.0	0.0	0.0	0.0
25%	2002.0	62.0	62.0	5.0
50%	2007.0	71.0	71.0	11.0
75%	2012.0	81.0	80.0	18.0
max	2016.0	179.0	167.0	117.0



From the table we see that the average point differential from 2007 to 2016 was 13 points, where the median was 11. From the plot of the average point differential over time we can see that it has actually decreased slightly. This implies that NCAA Basketball games have become more competitive over time.

In summary, from the analysis above we see that NCAA Basketball has changed significantly since the beginning of the dataset.

Business Question 2 - Predicting Season Attendance

Although the NCAA is considered ameteur athletics, everyone knows that it is a business. Athletic departments, athletic conferences, and the NCAA spend billions on marketing and fan experience to make games entertaining and fun for fans. However, the bottom line is that teams must win for fans to come. With the NCAA Basketball dataset, we will explore what factors affect attendance and try to predict season and game attendance. This will help the college athletic departments have an idea for demand for attendance to optimize ticket sales and cost of hosting the games.

Venues differ in size across the country. The Syracuse Orange have the biggest venue with 35,446 seats and the USC Upstate Spartans had a venue capacity of 818. There are 351 teams in the dataset.

A correlation between wins and average percent of venue filled throughout the season was 0.5 which is an okay correlation. Let's see if we can build a model with other predictors to determine how full arenas will be.

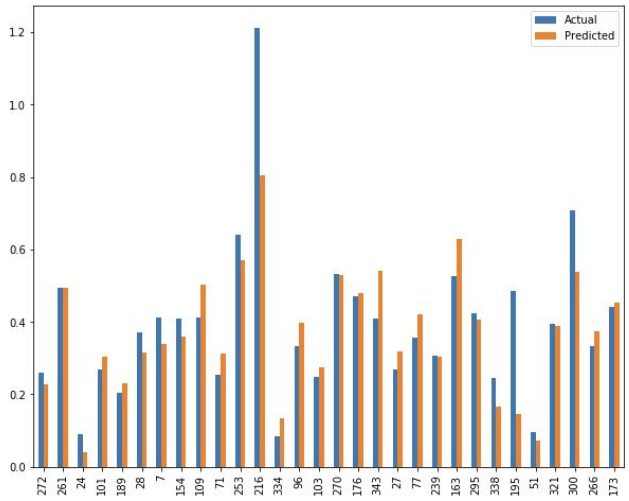
	wins	Average_Attendance	Average_VenueCapacity	avg_percent_full
wins	1.000000	0.463256	0.318925	0.500876
Average_Attendance	0.463256	1.000000	0.830279	0.732662
Average_VenueCapacity	0.318925	0.830279	1.000000	0.320189
avg_percent_full	0.500876	0.732662	0.320189	1.000000

Unlike the NBA, college basketball has a lot of disparity in the level of talent on teams and their fanbases. Colleges like Kentucky and North Carolina always seem to have talented teams as well as passionate fanbases. Other schools, like SMU and Princeton, may have successful seasons here and there. Fans may come and go with the wins for these teams. Lastly there are low major teams that may not get a lot of attention despite some winning seasons. For example, New Mexico State won 28 games in 2016 but only saw an average attendance of 39.1%. To account for this we created 3 clusters using K-Means. The variables used included what conference they belong to (one hot encoded), wins over the last several years, and venue capacity. Here is a summary of the centers of the clusters:

```
In [78]: # summary of cluster 0
         df_team[df_team['cluster']==0][['market','venue_capacity','wins_2013','wins_2014','wins_2015','wins_2016']].mean()

Out[78]: venue_capacity    9378.881481
         wins_2013           16.807407
         wins_2014           16.703704
         wins_2015           17.777778
         wins_2016           17.622222
         dtype: float64

In [79]: # summary of cluster 1
         df_team[df_team['cluster']==1][['market','venue_capacity','wins_2013','wins_2014','wins_2015','wins_2016']].mean()

Out[79]: venue_capacity    17244.480769
         wins_2013            21.903846
         wins_2014            20.980769
         wins_2015            20.711538
         wins_2016            21.057692
         dtype: float64

In [81]: # summary of cluster 2
         df_team[df_team['cluster']==2][['market','venue_capacity','wins_2013','wins_2014','wins_2015','wins_2016']].mean()

Out[81]: venue_capacity    4139.006098
         wins_2013           14.756098
         wins_2014           15.579268
         wins_2015           15.006098
         wins_2016           15.195122
         dtype: float64
```

Cluster 0 could be classified as mid-tier programs that have an average size fan bases and mid-tier talent and therefore average number of wins. Cluster 1 could be classified as top tier programs with large fan bases and lots of talent and therefore a large amount of wins. Cluster 2 could be considered teams with small fan bases and not a lot of talent and therefore less wins.

We'll use these clusters to help us predict attendance. The predictive model on top of the variables used in clustering included the percent venues were filled on average in previous seasons. Our predictive model only benefited from clustering in one cluster: cluster 0 - the mid tier teams. It had an adjusted R squared of 0.87 and a mean squared error of 0.006. The mean squared is quite low which means our model does a fairly good job of predicting attendance close to the actual value. Meanwhile the adjusted R squared explained 87% of the variability of the percentage that the arena was full (response data) around the mean. Without clustering the model had an adjusted R squared of 0.836 and a mean squared error of 0.0097.



Actual and Predictive (Percent of Venue Capacity) for the non-clustered model.

In conclusion, we believe athletic departments can use our research to help determine what market they fit into (clustering) as well as use our predictive model to help determine how many people will show up to games and how full their arenas are going to be. We observed one of the

variables that had the most predictive power was, not surprisingly, attendance in the previous seasons, meaning that athletic departments can know if teams have great attendance in previous season they can expect to see great attendance in the upcoming season.

Business Question 3 - Game Winning Factors

Finding the strongest determinants of winning a NCAA Basketball game has strong implications for athletic departments as well sports bettors. Using post-game statistics, we used a multiple linear regression model to find the best game winning factors. We used two models: one for home-team and one for away-team. The dependent variable in both cases was the total points scored in a game.

h_three_points_made	2.67
h_two_points_made	1.84
h_rebounds	0.18
a_turnovers	0.10
h_steals	0.09
h_blocks	0.05
h_turnovers	0.04
h_assists	-0.02

Home-Team Model Results

a_three_points_made	2.12
a_two_points_made	1.45
h_turnovers	0.72
a_assists	0.05
a_blocks	0.05
a_rebounds	-0.08
a_turnovers	-0.57
a_steals	-0.80

Away-Team Model Results

In both cases, the number of three and two points made were the strongest determinants of a team winning. This is not surprising at all, as a team that scores the most will obviously win. More interestingly, in the Home-Team model the number of home-team assists made actually hurts their chances of winning. Similarly, in the Away-Team model the number of rebounds and steals the away team gets also hurts their chance of winning. This is an interesting result and suggests that the offensive and defensive strategies in the sample are not consistent.

Conclusion

The results and conclusions to the three business questions our analysis was focused on can be summarized in the following points:

- College basketball has changed significantly over time, with modern teams scoring and winning more on average
- The model used to predict venue attendance can give athletic departments a good indicator of the number of fans that will buy tickets to home games, especially for mid-tier teams
- Predicting game winning factors based solely on prior game statistics is difficult, as there are un-measurable variables that help to explain the final outcome of most basketball games