Ordinal Bucketing of AP Poll Rankings Concerning NCAA Division 1 Basketball Attendance

Daniel Moeller
Marquette University
Milwaukee, WI 53233, USA
daniel.moeller@marquette.edu

Abstract

Many studies regarding sports attendance attempt to include rankings into their model. These outside third party rankings tend to be the primary method for which newscasters and the average fan compare teams. As it stands today there is not a standard way of dealing with this intrinsically ordinal data. We propose a new way of bucketing these rankings into statistically significant separate buckets. This allows for great understand of how fans make decisions and how schedule makers should factor these in order to maximize ticket sales revenue.

Author Keywords

NCAA; Attendance; Basketball; Rankings.

ACM Classification Keywords

H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous

Introduction

The Associated Press (AP Poll) has provided a weekly top 25 ranking for NCAA Division 1 basketball since 1949. These rankings have served as the primary measurement for comparing university basketball teams from across the country. With the explosion of

Coefficients	Estimate
Capacity	-0.0301
Home Win %	0.3649
Away Win %	-0.003
Home AP Rank	-0.003
Away AP Rank	0.0037
Weekend	0.0978
Prev Season Attnd	0.0318

Table 1: linear regression model coefficients

sports analytic popularity in recent years, researchers have enhanced their accuracy around predicting wins and losses, the score of the games, and attendance figures. An important factor that seems to be in every report is the ranking of a team in the AP Poll. General findings tend to point towards a positive correlation between a higher ranked opponent and more tickets sold supporting a very common belief that people tend to want to see "better" teams. Where our interest lies is in how these top 25 rankings can be interpreted. While the measurement is surely ordinal, we set out to explore and suggest statistically significant buckets for use in future attendance models.

Data Collection

There were no readily available datasets published for public access that have game-by-game attendance details. For this reason we turned towards an open source python framework for website scraping called Scrapy. A series of web crawlers were written to navigate www.espn.com and extract the information we needed [1]. This process was set up to be repeatable for any university and any year available on the website. This allows for simplicity for others to replicate this study.

The main dataset used for this study was collected through the process designed above and includes game data from The Big East Conference and the Big 12 Conference from the 2013-14 season to the 2015-16 season. Data collected includes but is not limited to:

- Attendance (tickets sold)
- Stadium capacity (1000)
- Home/Away team win percentage
- Scoring line
- Betting line
- Date and time
- Day of the week
- TV coverage
- Previous seasons average attendance (1000)
- Home/Away AP ranking
- Television network

After collection, the data was cleaned up following Hadley Wickham's tidy data standards [8]. This included removal of games with insufficient data (~2%) and bolstering games incomplete data due to abnormal format of website (~1%).

Exploratory Analysis

With dataset in hand, we construct a linear regression model in order to explain attendance with the features introduced previously to validate our data with expected results [6]. We standardize attendance based on capacity of the stadium by setting our independent variable to be the attendance divided by the capacity in order to determine the percentage full the stadium is. Table 1 shows a subset of these coefficients, which result in a R^2 value of 0.598. This model does not intend to all encompass the explanation of how ticket sales number are determine, but rather it paints a picture of our features.

Most models included in our research use both the home teams AP Poll rank as well as the away team AP Poll rank as they find them both to be statistically significant. Through our own analysis we tend to see a much clearer trend in away AP Poll rankings for our dataset. This distinction guides our further research to focus only on the visiting basketball team. Our exploratory analysis tells us there does appear to be

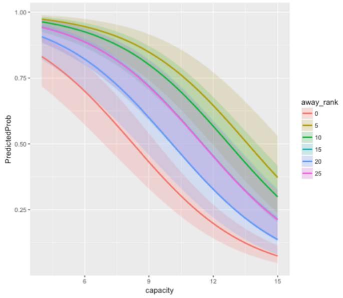


Figure 2: initial predicted probabilities and 95% confidence intervals for away AP ranking against capacity

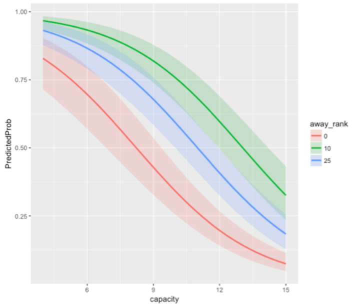
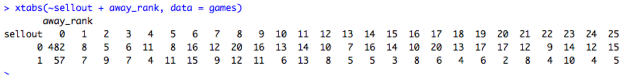


Figure 3: final predicted probabilities and 95% confidence intervals for away AP ranking against capacity

similar behavior in the home team, but that is not discussed here and is an opportunity for future work.

Initial Bucketing

With general theory in mind our attention turns towards natural buckets of the ranking system. Due to the ranking system only spanning the number one team in the country to the 25th, we make the generally supported assumption that any team that is not ranked change our outcome variable from how close to capacity ticket sales were, to considering games as sellouts or not. We define a sellout as any game that has ticket sales equal to or exceeding the capacity of the stadium. While this sacrifices specificity in our model, it allows us to use supportive tools associated with logistical regression [4].



```
> xtabs(~sellout + away_rank, data = games)
       away_rank
sellout   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25
      0 482   8   5   6  11   8  16  12  20  16  13  14  10   7  16  14  10  20  13  17  17  12   9  14  12  15
      1  57   7   9   7   4  11  15   9  12  11   6  13   8   5   5   3   8   6   4   6   2   8   4  10   4   5
>
```

Figure 1: Two-way contingency table of categorical outcome and predictors. Away rank represents the AP Poll ranking of the visiting team where 1 is the best team and 0 represents a team that is not in the top 25.

With figure 1 in mind we choose an initial 6 buckets of not ranked (0), 1-5, 6-10, 11-15, 16-20, and 21-25. We can then use a logistical regression model using the generalized linear regression model before to map behavior of our ranking system. This model can then be used to generate predicted probabilities for the average value of each feature and different rankings. Figure 2 shows us how we can visualize these behaviors while holding all other things constant. We begin to see that while there is a significant difference between not being ranked and being ranked, we can see that teams ranked between 21 and 25 seem to be more likely to have a sellout than a team ranked between 20 and 16. This curious finding may tell us there is some general bias in our model do to data size, but we move past that by re-bucketing and re-running our model in order to determine the best set.

Results

Through multiple iterations of the process described previously, we suggest further research uses three distinct buckets when considering how to handle visiting teams AP ranking:

- An unranked opponent outside the top 25
- An opponent ranked in the top 25 but outside of the top 10
- An opponent ranked within the top 10

These results are supported by general knowledge and basketball theory along with statistical evidence. Our final predicted probability chart in figure 3 gives a clear image of how different our three ordinal categories are. We see very little to almost no overlap between these three buckets and their 95% confidence interval. We also can see through the Chi-squared test that they each are statistically significantly different than each other.

With our bucketing in order we move forward with our logistical regression as to interpret the results. We see that the difference between our buckets in terms of log odds is a whole unit. This tells us that for every bucket we move up in rankings, the log odds of a sellout increases by ~1. We all see odds rations of 6.04 and 2.81 for top 10 teams and top 25 teams respectfully. More specifically, a visiting opponent who is ranked in the top 25 has an increase by a factor of 2.81 in the chance of a sellout over a team that is not ranked.

Impact

While our findings are nowhere near groundbreaking, they do shine light on an area that is relatively uncharted. The advancement of techniques in data science have still not found there way to all branches of corporate and freelance work and this simplified approach and a relatively naive study should help those who are interested in understanding how categorical, nominal, or ordinal data should be considered and interpreted. Our findings also serve to help though scheduling games for their university. While it most likely true that a higher ranked opponent is more difficult to schedule they can see weigh these costs with the significant advantage of drawing in more revenue from tickets sales. A team may choose to pay the small extra cost to get a top 10 team rather than a team that is just outside the top 10.

Future Work

This exploratory analysis serves only as an initial plunge into how the AP Poll and sports ranking systems in general should by interpreted and handled. More work will be needed with extensive datasets and more experienced minds to come to any foregone conclusion. There is a cornucopia of available approaches to this issue with this proposal being just one. We hope to push this research further to better understand how fans make purchasing decisions.

It is also important to note here that I am in no way an expert on these topics. This is a preliminary report at the most and should only be utilized as such. I would highly recommend gathering more data from a more diverse team set in order to verify these results.

References

1. Due, J. (2016) *How to Crawl A Web Page with Scrapy and Python 3.* Retrieved from https://www.digitalocean.com/community/tutorials/how-to-crawl-a-web-page-with-scrapy-and-python3
2. Howell, P. 2015. NCAA_MBB. [Software]. Retrieved from https://github.com/pathow/NCAA_MBB
3. Jupyter (1.0) [Software]. 2017. Project Jupyter. Retrieved from https://github.com/jupyterhub/jupyterhub
4. Logit Regression. UCLA: Statistical Consulting Group. From http://stats.idre.ucla.edu/r/dae/logit-regression/ (accessed April 22, 2016).
5. Moeller, D. 2017. *Ncaa-bball-attendance.* Retrieved from https://github.com/danmoeller/ncaa-bball-attendanceg
6. Mueller, K. 2013. *Marquette Basketball Attendance Analysis.* Retrieved from http://epublications.marquette.edu/cgi/viewcontent.cgi?article=1052&context=cps_professional
7. Paul, R., Weinbach, A., Plaut, J. *Information, Prediction Markets, and NCAA Basketball Attendance: The Big East Conference*. Retrieved from http://buildingthepride.com/jobie/uploads/JOBIEV19B31.pdf
8. Wickham, H. *Tidy Data*. The Journal of Statistical Software, vol. 59. 2014. Retrieved from http://vita.had.co.nz/papers/tidy-data.html