Daniel Moeller
April 10, 2017

COSC/MATH 4931: Easter Milestone Report

All the documentation and code referenced in this report can be found in the following github repository:

https://github.com/danmoeller/ncaa-bball-attendance

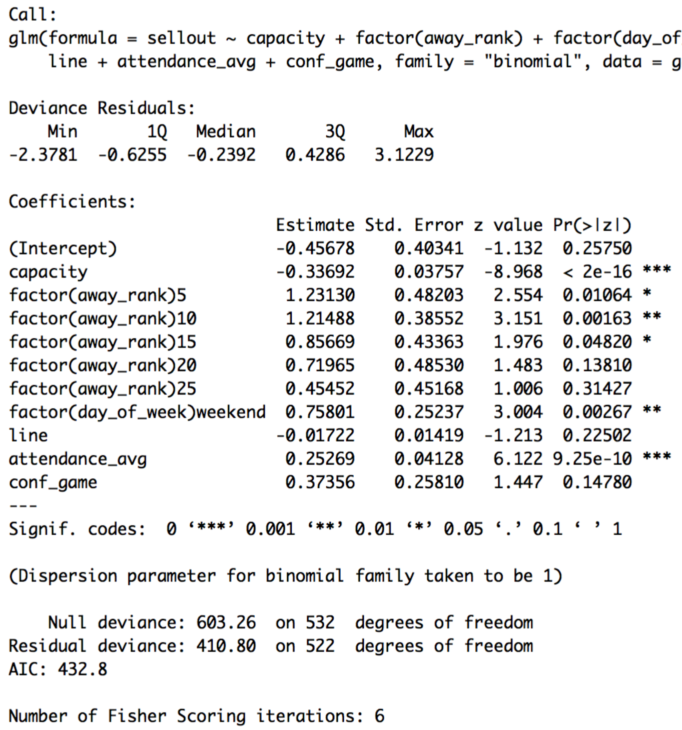
This report assumes the reader understands what this project is at a high level and has a casual knowledge of college basketball. If there is any doubt to either of these please visit the README.md or the introduction PowerPoint in the doc directory.

We left off the last check-in with our base dataset extraction method completed as well as an exploration into how our designated features appear to impact stadium attendance in NCAA Division I basketball. More specifically our focus has been on the Big East Conference from the 2012-13 season to the 2015-16 season. These years represent the complete existence of what has become the "new" Big East Conference. Additional data sources have been identified and explored, but have not been included as to focus on the methodology of the analysis. There is a hope and a desire to work in this information to standardize the model.

To begin our analysis I constructed a very simple OLS linear regression model in R to get a more mathematical basis behind our visual observations. This code can be found in /analysis/linear_regression_big_east.R which gives us valuable information around which variables are statistically significant to how full a stadium is. Here we learned from the results of our regression that most of our inclinations about significant features where encouraged.

Feature	Coefficient	P-Value
Capacity (1000)	-0.0338783	< 2e-16 ***
Home Win Percentage	0.0668923	0.074197
Away Win Percentage	0.0512768	0.140985
Home Rank	-0.0030212	0.157653
Away Rank	0.0052690	8.40e-07 ***
Weekend	0.0926387	1.66e-09 ***
Line	0.0025790	0.226305
Scoring Line	-0.007961	0.708464
Previous Season Attendance Average (1000)	0.0318243	< 2e-16 ***
Conference Game	0.0609151	0.000155 ***

The biggest take away from these results is not what is statistically significant, but what is not. The most shocking of which is the betting line and the scoring line of the game both appear to have minuscule effect on the attendance numbers. I hypothesis this is due to the large number of games that have missing lines defaulting to zero. The scoring line even has a negative coefficient, which is counter to our research. Our model would suggest that a game with more predicted points would actually result in less tickets sold. This information was then used to construct a logistic regression model in order to predict whether a game is a sellout or not. I used UCLA idre's article around logistic regression as a guide to do this [1]. While following this guide we could construct useful information to better understand how each individual feature impacts a teams chance at reaching a sellout.



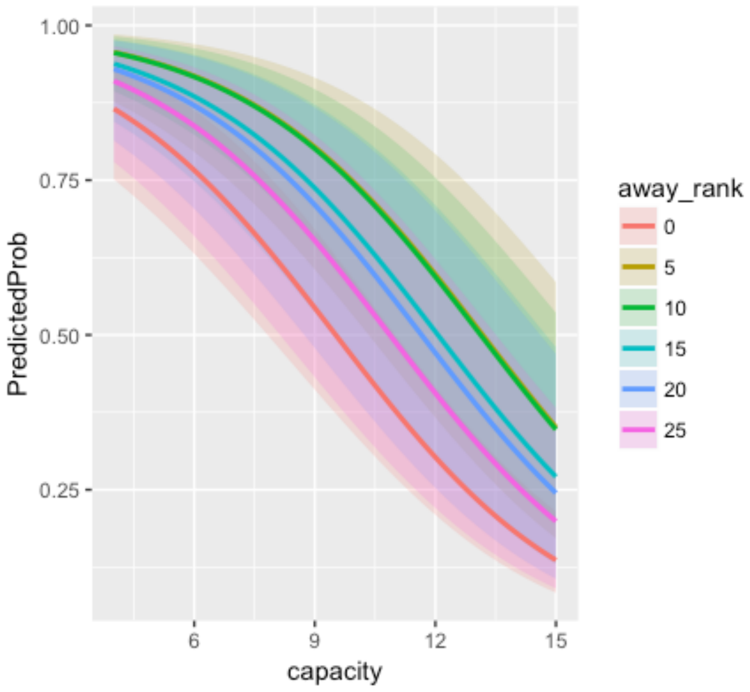
```
Call:
glm(formula = sellout ~ capacity + factor(away_rank) + factor(day_of_week) +
    line + attendance_avg + conf_game, family = "binomial", data = games)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3781  -0.6255  -0.2392   0.4286   3.1229

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                -0.45678    0.40341  -1.132  0.25750
capacity                   -0.33692    0.03757  -8.968  < 2e-16 ***
factor(away_rank)5          1.23130    0.48203   2.554  0.01064 *
factor(away_rank)10         1.21488    0.38552   3.151  0.00163 **
factor(away_rank)15         0.85669    0.43363   1.976  0.04820 *
factor(away_rank)20         0.71965    0.48530   1.483  0.13810
factor(away_rank)25         0.45452    0.45168   1.006  0.31427
factor(day_of_week)weekend  0.75801    0.25237   3.004  0.00267 **
line                       -0.01722    0.01419  -1.213  0.22502
attendance_avg              0.25269    0.04128   6.122 9.25e-10 ***
conf_game                   0.37356    0.25810   1.447  0.14780
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 603.26  on 532  degrees of freedom
Residual deviance: 410.80  on 522  degrees of freedom
AIC: 432.8

Number of Fisher Scoring iterations: 6
```

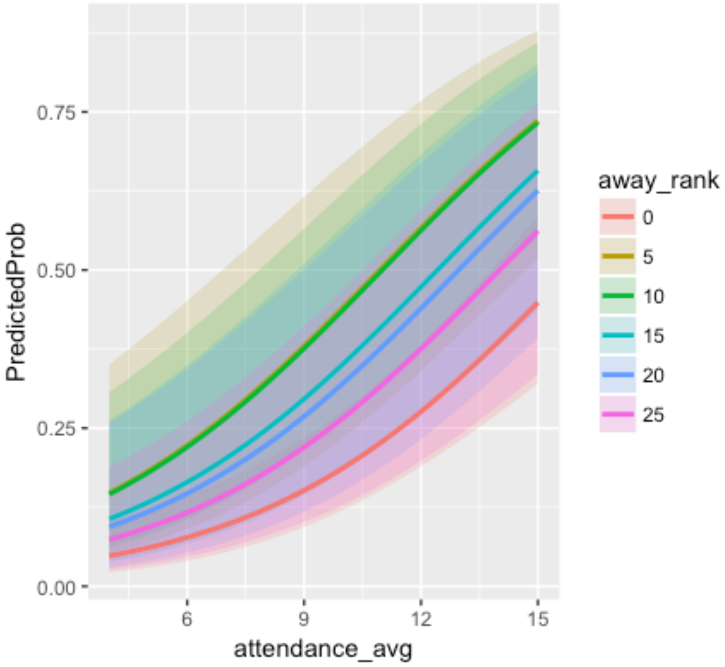
This tells us that a majority of our features are statistically significant or are at least very close to being statistically significant. We see a switch in the sign of the coefficient in front on the betting line of the game, which fits more on line with the idea that a team that is favored should see more tickets sold and a greater chance at a sellout. A few important coefficients to notice is that a game on the weekend increases the log odds by 0.75801 as well as a conference game has a 0.37356 more log odds than a nonconference game. We also see a very opposite and counter acting effect between the capacity of the stadium and the average attendance last year, as they seem to very closely balance each other out. A Wald's test was also conducted to prove that each team ranking bucket as well as the day of the week were significantly statistically different respectfully.

We then could compute our odds ratios and their confidence intervals shown below and see the odds of a sellout verses not a sellout increases by a factor of 1.287 for every 1000 more average attendance in the previous season. We can also see that teams ranked between 1-10 have a very similar odds ratio. Our Wald's test tells us that these are statistically different, but wee see very similar behavior. As well as for teams ranked between 11-20. More work will be needed to flush these out as separate buckets, but through our analysis we have determined that these nominal values are very influential to the likelihood of a sellout.

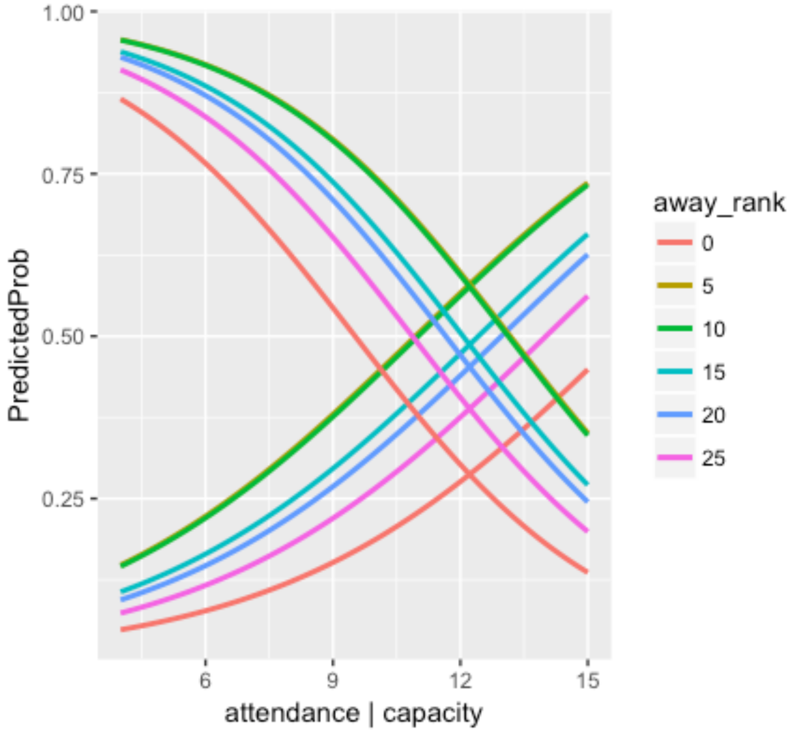
```
                                OR     2.5 %    97.5 %
(Intercept)              0.6333167 0.2839600 1.3859834
capacity                 0.7139658 0.6602566 0.7654195
factor(away_rank)5       3.4256847 1.3251111 8.8530376
factor(away_rank)10      3.3699037 1.5896622 7.2464385
factor(away_rank)15      2.3553437 1.0050710 5.5479327
factor(away_rank)20      2.0537133 0.7829616 5.3197972
factor(away_rank)25      1.5754137 0.6316992 3.7665895
factor(day_of_week)weekend 2.1340359 1.3084778 3.5261024
line                     0.9829259 0.9558085 1.0106440
attendance_avg           1.2874793 1.1899827 1.3997370
conf_game                1.4528978 0.8776231 2.4189335
```

For a better explanation and visual of how capacity and previous season average attendance work in opposite directions I have produced the following charts.





This tells us while there is a benefit from having a larger previous years attendance, this comes at a cost due to having a larger stadium, which makes it harder to sellout because more tickets must be sold. We can then visualize the idea capacity with respect to tickets sold in hopes of having more sellouts.



From our chart we can see that it is most likely that there will be a sellout when the average attendance is as close as possible to the capacity. At this intersection point we will have nearly a 30% higher chance of selling out the available tickets for an opponent ranked in the top ten verses an appoint that is not ranked at all. While this is insightful we need to take a step back and think about what types of decisions are made around this information. For one if we are seeking to help a university schedule games we must only consider factors for which they have control over. Thus as we move forward with this project in the coming months we must focus on these more controllable features in order to better help people make these decisions.

Reference:

1. Logit Regression. UCLA: Statistical Consulting Group. from http://stats.idre.ucla.edu/r/dae/logit-regression/ (accessed April 22, 2016).