MA678 Midterm Project: Analyzing NCAA Division I FBS Game Attendance

Jack Carbaugh

12 December 2021

Abstract

REWRITE THIS AFTER REST OF REPORT IS DONE. For right now, this abstract will some notes for the rest of the report. Introduction: I'll talk about the reasons I find this topic interesting, taking a lot of information from the project proposal. Method: Talk about the model selection, why I made the choices I did, mention the transformations that were done, add in some short EDA to help with explanation. Result: Talk about the coefficients that were found, as well as showing fixef and ranef plots. Discussion: Take deeper dive into what this may mean, highlight any interesting random effects, talk about what I would like to do next. As the conclusions may not be incredibly mind-blowing, it may make sense to treat this project as a stepping stone that could include more school-specific information (student body as well?) as opposed to just game information. DONT FORGET CITATIONS.

Introduction

College football within the NCAA's Division I Football Bowl Subdivision (FBS) is one of the most popular sports in the United States, and consists of teams constructed by some of the country's largest and most competitive universities. Due to the popularity of the sport, these programs can be a major source of revenue for the universities. In particular, ticket sales and other sponsorships are major part of this revenue, all being linked to attendance to the games. Thus, this project is mainly meant to better understand when fans attend college football games, in order to assist universities in increasing their attendance numbers, and thus increasing revenue.

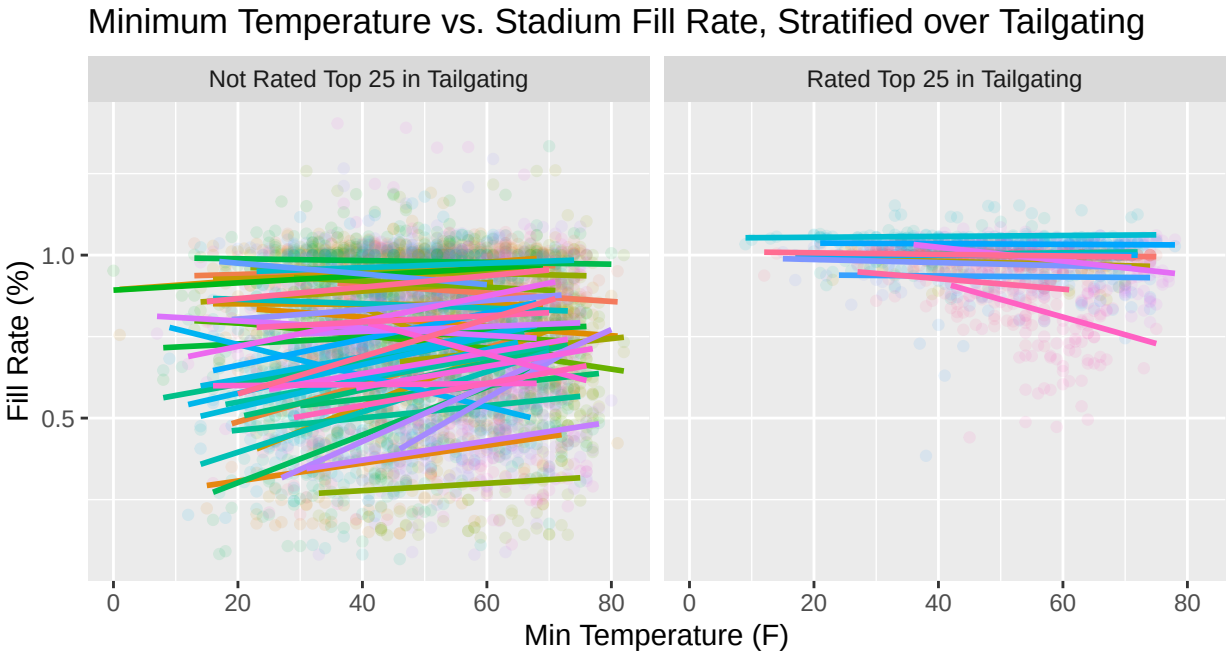
Method

Planning

This analysis utilizes the College Football Games (2000 to 2018) posted by Jeff Gallini on Kaggle. This dataset has a majority of the factors I expected to have an effect on attendance, such as weather data, game data, and team performance data (as well as attendance numbers themselves). The dataset logs over 6000 games for 63 home teams. Even among NCAA Division I FBS, the highest tier of college football, there is a great difference in the size of pedigree of school football programs. Because of this, attendance numbers are expected to vastly differ based on the school in question; this will require the use of a multilevel linear model with grouping based on school.

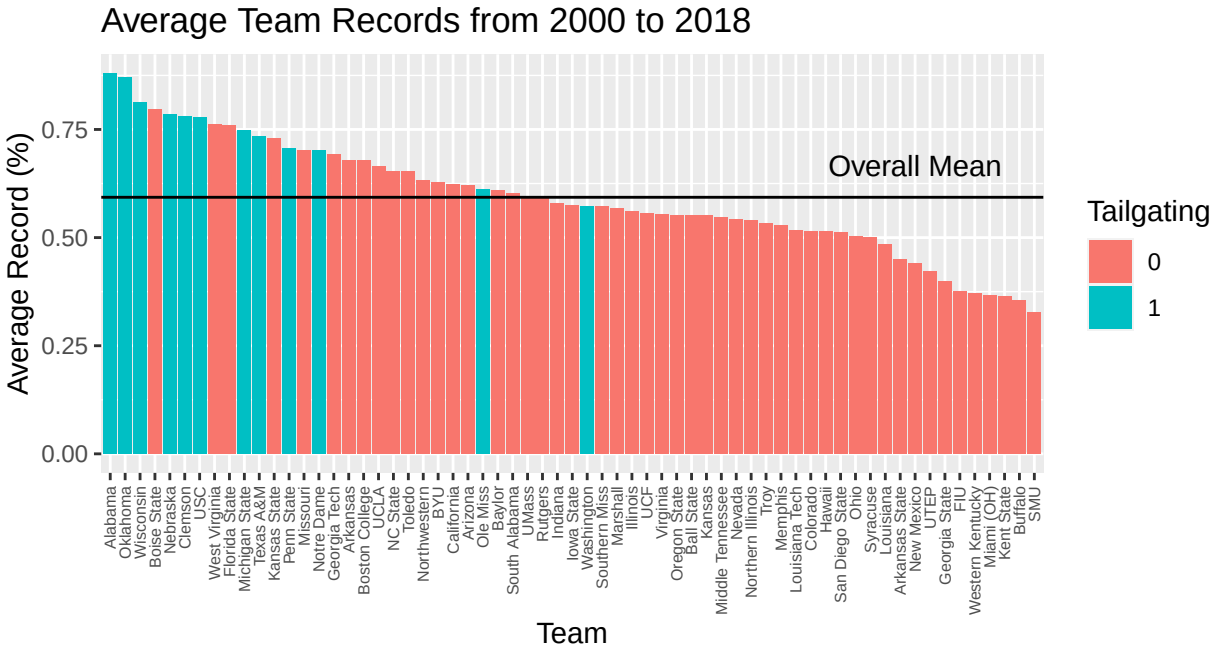
Exploratory Data Analysis

At the start of the analysis, it will be helpful to visualize some of the variables to get a sense of potential effects. In particular, temperature during the game and the quality of tailgating have the potential to be predictors.



This scatter plot shows how the minimum temperature during the day of the game affected the fill rate of the stadium. Each line summarizes the relationship for each individual school. As one would expect, it appears that as the temperature increases, the fill rate of the stadium will increase for most schools. The plot also separates out the schools that were rated to have a top 25 tailgating scene prior to the game. Notably, the majority of these schools seem to contain some of the highest average fill rates, while the fill rates also do not depend as much on temperature. The lack of dependence is likely due to these schools continuously having high attendance, and almost alwasy selling out regardless of weather.

We would also expect these schools have highly rated tailgating to be some of the stronger programs in the country. The following visualization can help see this.



We can now better compare the average records of each team against one another. More importantly, when we highlight the teams with highly rated tailgating, we see that a majority also consistently have strong records. In fact, we see that of the 15 teams with the highest records over the data's time frame, 10 of them are rated to have a top 25 tailgating scene. Due to this correlation, it will be necessary to interact these two variables during modeling.

Model Fitting

Due to the sizable variance in size, location, and recruiting prowess of Division I FBS schools, a multilevel linear model grouping by school will be used to predict the stadium fill rate of college football games. Fixed effects will include the amount of precipitation during the game, whether or not the game is Non-conference or not, and whether or not it is a "Big Game," such as a rivalry or championship game. The amount of games into the season will interact with the minimum temperature during the game, as all locations will generally get colder as the season progresses from late summer to winter. Record will interact with tailgating, as there was likely correlation seen in the EDA. Record will also be a random effect, as fan bases can range from expecting success every year, expecting failure every year, or being surprised based on the team's usual quality.

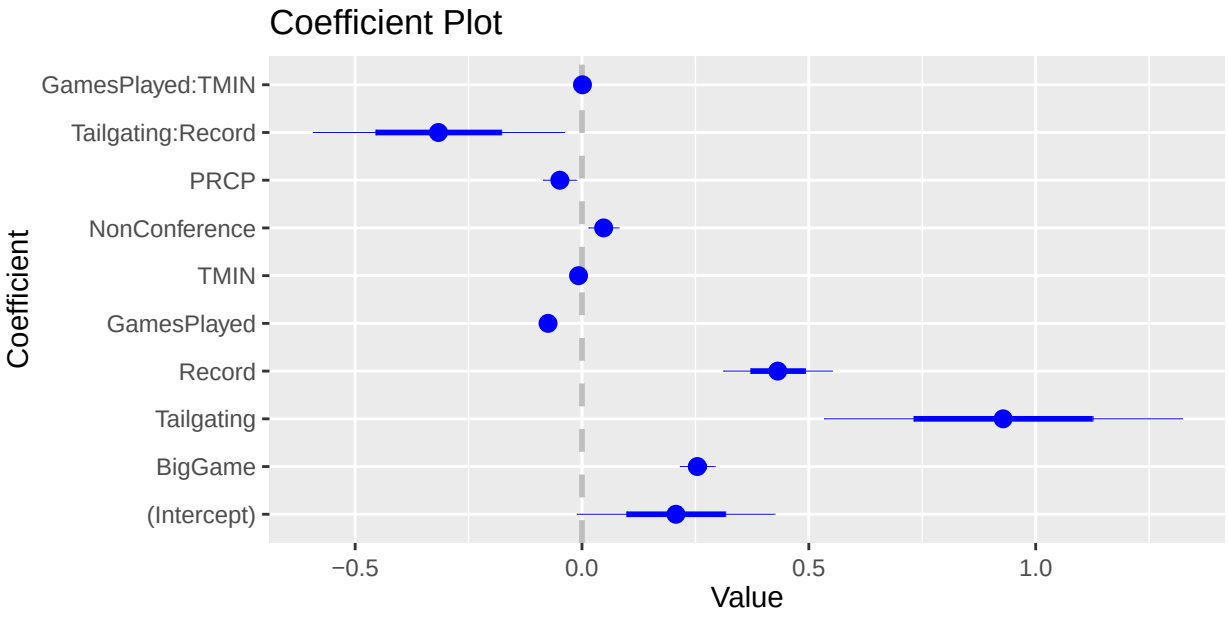
The variable used to represent attendance is fill rate, which is calculate by dividing the attendance number by the stadium capacity. Using fill rate instead of raw attendance numbers will provide a better standardization for smaller schools. Due to some stadiums allowing for overflow capacity, the fill rate will occasionally be greater than 1. To fix this, the fill rates were scaled to fit within the range of 0 to 1, and then a logit transformation was used to get closer to normality. The model used is below:

```
model <- lmer(qlogis(Fill_Fixed)~ BigGame + Tailgating * Record +
                GamesPlayed * TMIN + NonConference + PRCP + (1 + Record|Team),data_t)
```

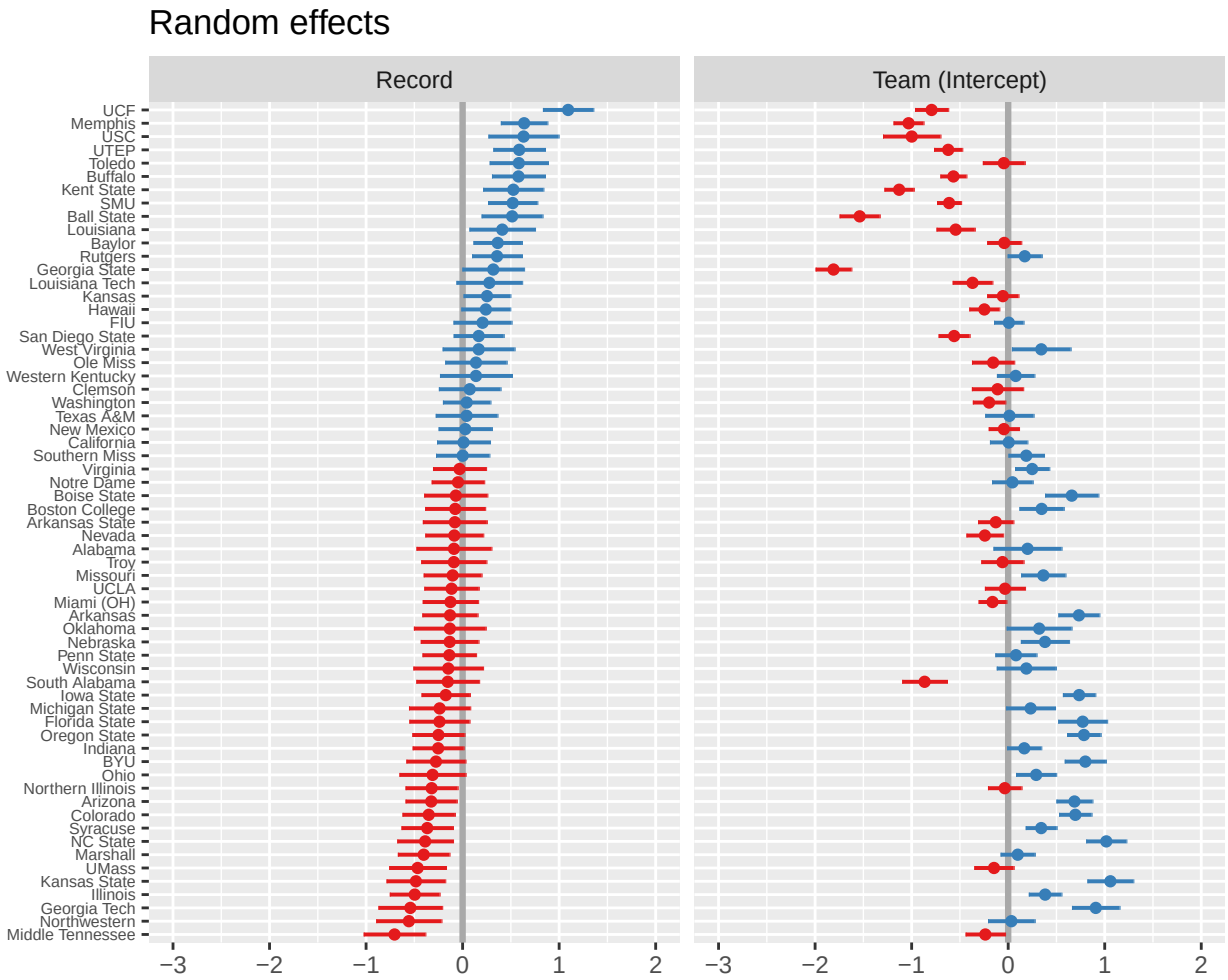
Results

The fixed effects have the following estimates and errors:

	Estimate	Std. Error	t value
(Intercept)	0.207	0.109	1.904
BigGame	0.254	0.019	13.266
Tailgating	0.928	0.197	4.711
Record	0.431	0.060	7.208
GamesPlayed	-0.075	0.008	-9.892
TMIN	-0.008	0.001	-6.583
NonConference	0.048	0.017	2.894
PRCP	-0.049	0.018	-2.688
Tailgating:Record	-0.316	0.138	-2.288
GamesPlayed:TMIN	0.001	0.0001	7.704



The random effects can be visualized like so:



Discussion

While many of the fixed effects seem to be significant, not as many are impactful in determining the fill rate of stadiums. As expected, precipitation deters people from attending games, but temperature strangely has a negative effect as well, meaning people are more likely to attend colder games. The small impact of these environmental effects may be due to students and younger folks, who are hardy than average people, consisting of much of the attendees. The unexpected sign switch on temperature may be due to the large number of southern schools that are hot early in the year. The number of games played being negative tells us that more people attend games early in the year, which is expected as this is when schools usually increase their marketing push, and fans are hungry for football after the near 7 month break. As neither games played nor temperature were very impactful, neither was their interaction. Fans also are slightly more likely to attend games featuring nonconference opponents, potentially due to these games providing rarely seen marquee matchups, or sometimes reigniting old rivalries that were broken up due to past conference realignment.

The fixed effects with the greatest impact on the fill rate of stadiums are the record of the team, the tailgating measure, and whether or not the game is of high importance. These results are all fairly expected, as one would wager that a good team is more likely in demand than a bad one, and rivalry and championship games usually invoke strong emotions in fans and are targets of higher marketing from the schools. The tailgating is much higher than the rest, but seems a bit unstable due to its high error and the penalization from the tailgating-record interaction term. This may be due to the much smaller sample size of schools with highly rated tailgating. Regardless, the impact of this variable signals that incentivizing fans to attend tailgating before the game and outside the stadium usually leads to higher attendance at the game itself. One must wonder if schools could set up additional events or games beforehand to entice fans.

The random effects allow us to understand which team's fan bases are most affected by their record. Interestingly, the school's with the greatest random coefficients for record are mainly mid-sized schools such as UCF, Memphis, UTEP, Toledo, and Buffalo. These types of schools are not in a power 5 conference, and thus do not have access to the same level of recruiting as larger schools. However, as the largest of the smaller schools, they are occasionally able to have "flash in the pan" seasons due to being somewhat comparable to the smaller power 5 schools. Thus, their large random effect in record likely aligns with these schools occasionally having strong seasons, exciting the fanbase. This is most clear in UCF, which is an outlier above all the others. UCF had been a below-average to bottom tier team for years up until 2017, when they went undefeated and started a national following. Their success continued into 2018, with the fans being well known for their voracity during this period. This wide variance in record and following likely causes the large positive random effect. The volatility of these mid-sized schools can be seen in their largely negative intercepts as well.

The schools around the 0 mark are those teams not as affected by record. This group consists of traditionally strong programs like Clemson, Texas A&M, Notre Dame, and Alabama which usually always have strong records, and even when they don't have the historical fanbases to attend regardless. The section also contains many smaller schools such as New Mexico, Southern Miss, and Arkansas State, which almost never have good records, and likely never cultivate a strong fanbase. The schools with negative record random effects consist of an odd group of mostly power 5 schools with traditionally weak football programs, such as Syracuse, NC State, Illinois, and Northwestern. While these teams have had occasional strong seasons, it seems odd for the random effect to be so low. One potential reason could be that while these teams have weak football teams, they traditionally have much stronger basketball teams, the second most popular college sport. It could be that these schools with stronger basketball teams do not care so much about football teams, and would even skip strong season games late in the year for basketball games, when the two sports' seasons begin to overlap.

Finally, I hold that this project is a good stepping stone for further analysis in football game attendance. Notably, only half of the Division I FBS teams were present in the dataset. Having a greater sample size, particularly among the schools with highly rated tailgating, could be quite helpful. It would also be interesting to add more school-related variables to this dataset, such as funding numbers, student population data, etc. Many of the variables studied in this dataset are either out of the schools control, such as weather, or are correlated with simply fielding a stronger team. Regardless, from this study, we can see that increasing

tailgating quality, or perhaps simply a stronger pre-game experience, may lead to higher game attendance. Along with this, it seems that a winning culture attracts fans, along with stronger rivalries.

Citations