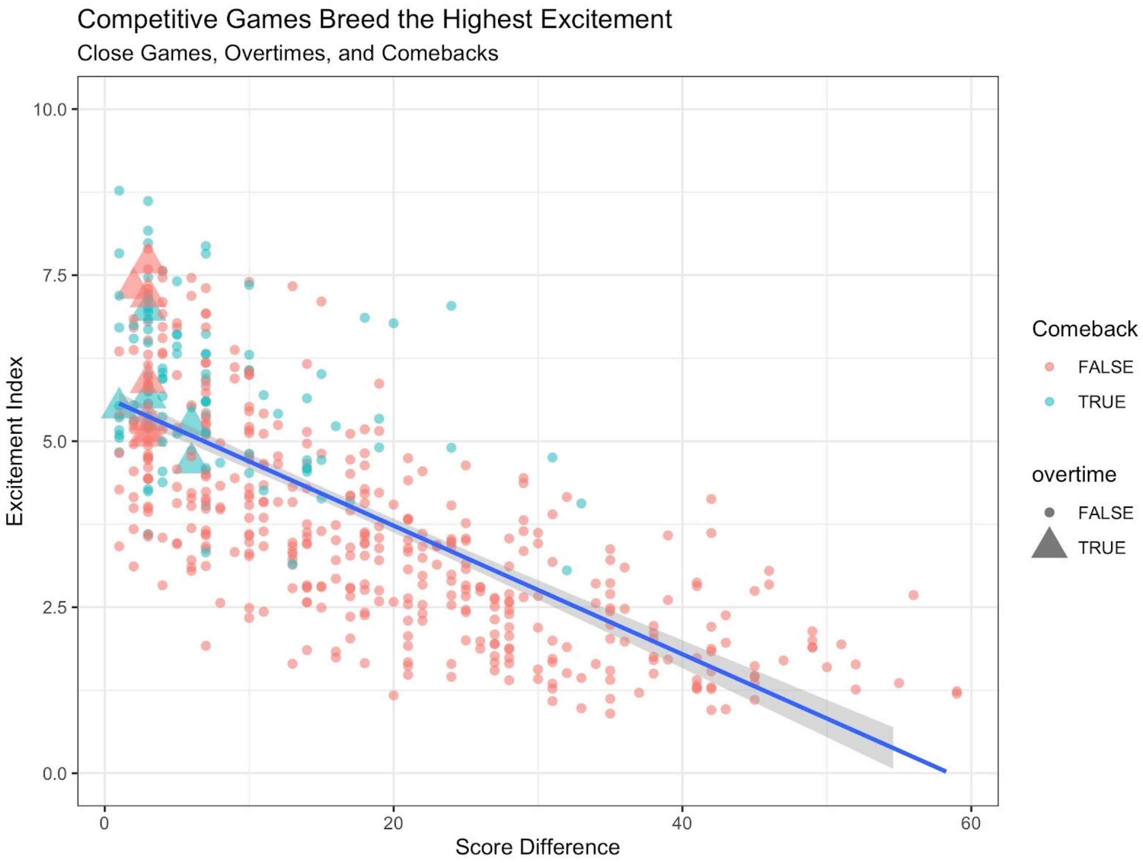
College Football Excitement Index

Jill Reiner reiner_j1@denison.edu
Raj Dasani rajdasani@berkeley.edu



<u>Executive Summary:</u>

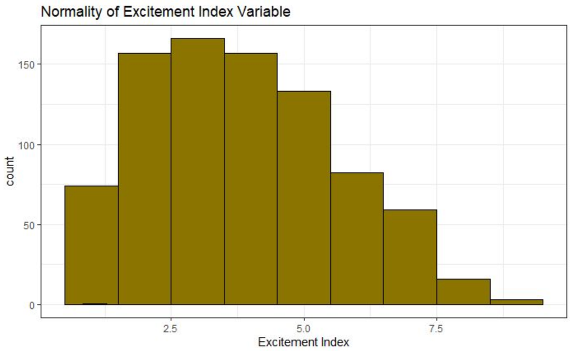
The excitement in college football makes the sport one of the most popular in the United States. In the 2019 season, games averaged over 1.8 million viewers and reached over 145 million unique fans. But with hundreds of games happening each weekend and a wide variety of results, we wanted to find out: what makes for an exciting game in college football? In order to answer our question, we set out to find the main variables that had a significant effect on our response variable, a game's Excitement Index. Across Division I games with a calculated excitement index, we tested common assumptions of excitement via exploratory analysis and cross-validation to build a linear model of the best combination of variables that maximized correlation and minimized error rates. The four variables that we found to have a statistically meaningful effect on a game's Excitement Index are whether the game was considered to be a comeback, whether a game went to overtime, the overall score difference between the two teams in a game, and the score differential at the end of the first half.

Introduction:

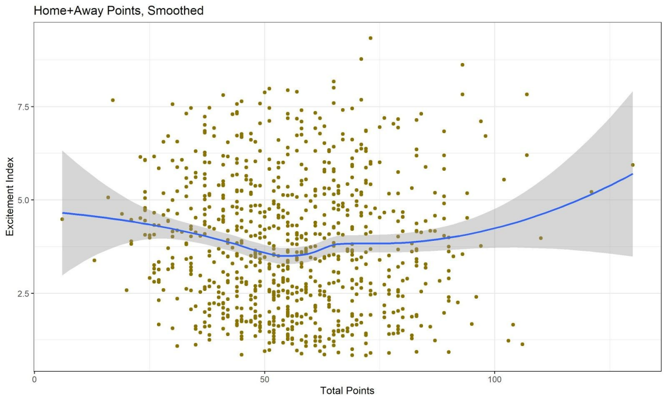
NCAA college football consists of three divisions: Division I, Division II, and Division III, with Division I being the most competitive. There are 130 schools around the United States in the Division I Football Bowl Subdivision (FBS). Within the FBS, teams are split into several different conferences, and as of 2018, there are 10 conferences. Most FBS teams play 12 regular season games per season, with eight or nine of those games being inter-conference matchups. Our dataset consists of 849 regular season FBS games, with one row representing one game in the 2019 FBS season. We were originally given 29 variables, with some of these being "game identifier" variables, such as the week the game was played, start time, as well as information about where the game was played. The dataset also includes quantitative variables such as our response variable, Excitement Index, as well as variables relating to the score of the game, including points by quarter and total points scored by each team. From these, we created four additional variables that we thought may have an effect on a game's Excitement Index, including score differential at the end of the first half. These helped us quantify a comeback variable and an overtime variable, indicating whether the game fell in either of these categories.

Response Variable: Excitement Index

Our response variable of choice was the excitement index, because a) as seen below, it has a relatively normal distribution, which is great for our linear modeling and b) it was a variable neither of us had heard of or dealt with before, and we wanted to test our common assumptions of excitement with this index. There is a slight tail that makes it a bit left-skewed but this comes natural with a stat like this based on the few CFB games that break the boundary in terms of excitement for that year. The Excitement Index is defined by ESPN's win probability model. The index is calculated by the sum of swings in win probability over the course of the game. For example, the most exciting game this season, according to the EI, was Utah State @ Wake Forest, with an index of 9.34, indicating the win probability changed by a total of 934% over the course of the game.

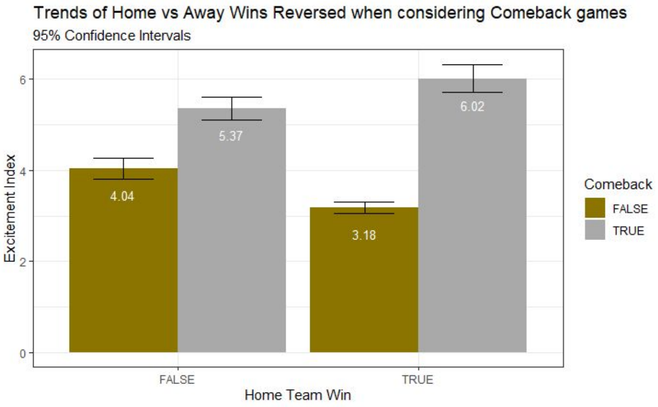


Distinguishing between Excitement Index and the Assumptions of Excitement (EDA):



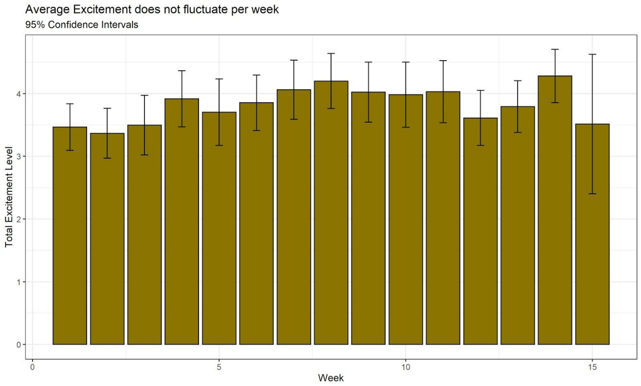
The first step we took in analyzing our data was to test some common assumptions a general fan would have about excitement in sports in general.

Our first assumption revolved around the notion that more total points scored in a game makes the game more exciting. However, in terms of Excitement Index, this is not the case. As displayed in this plot, there is no clear trend between a game's point total and excitement index. The only slight trend with this graph and the graph of just using 4th quarter points, was seen in the highest scoring games (90+ total, 30+ in the 4th quarter), where more points had a higher index, but nothing conclusive came of it because the amount of games in each group were minimal.

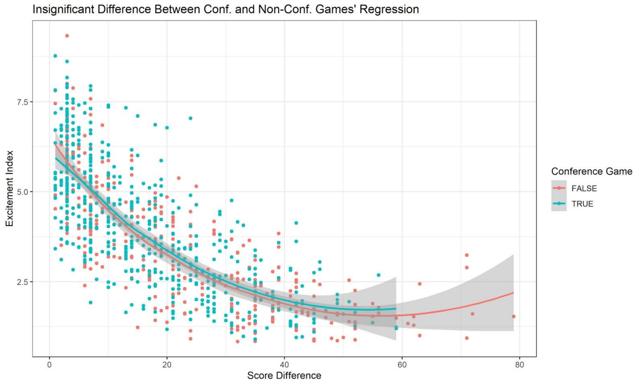


Our second assumption came in the consideration of whether the home team won the game or not. Typically, in terms of stadium excitement levels, a home team winning would appear more exciting. However, this brings the main distinction between excitement and excitement index. We found that on average, a home win had an average excitement index of 3.54 while an away win had an average excitement index of 4.32, which ended up becoming our best indication of an "underdog win," which by definition has high fluctuation in win probability. We in fact did look deeper, and found that while an away win had an overall higher average, when we are looking at purely comeback games, the average home win had a higher excitement index than away win, 6.02 to 5.37 respectively, as shown above.

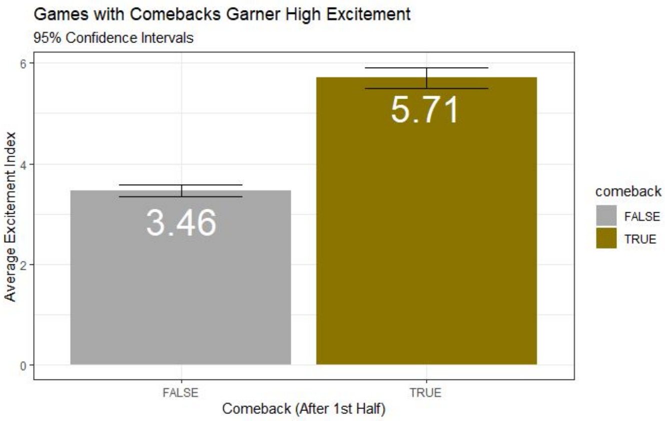
Our third assumption focused whether playing a game later in the season made a game more exciting. When we plotted the average Excitement Index per week, we found that there is not a clear trend as imagined as most of the 95% confidence intervals overlap. The average Excitement Index peaked at Week 14, but in general, it did not vary much throughout the season. With the total excitement index per week, the first and last weeks had the highest amount, but that was based on the high amount of games and lack of "byes," when teams are given a week break in the middle of the season, during those weeks. Visually, it looked like there may be a difference between the first couple weeks and the rest of the season, so we decided to keep conference games, the latter 8-9 games of the season, in mind as a potential variable for our linear regression model.



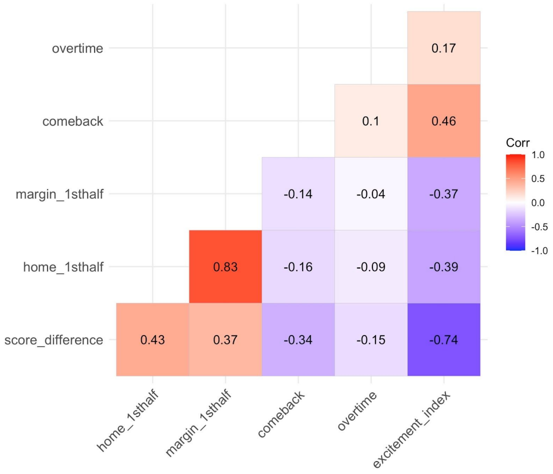
Regression Model Variable Selection



To create the best linear regression for excitement index, we went through various different forms of variable selections to decide which variables could add to our model. We started with "score difference", as from the get go, it was clear that it was a very important regresser to excitement index. On its own, it had a correlation of -0.74 and an adjusted R^2 term of about 0.55. We then looked into conference games, where teams in the same conference (aka Pac-12, SEC, Big 12, etc) play against each other, to see if the rivalry/similar level of competition would breed excitement. While there seemed to be a somewhat significant difference in number (4.05 average for conference games, 3.42 for non-conference games), in terms of our regression model, as shown above, it did not have a significant impact with score difference already in the model. In terms of excitement index, it seems the score difference already encapsulates the difference within conference vs non-conferences games.

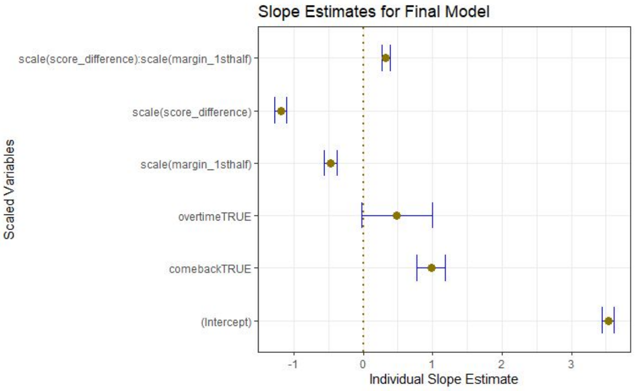


Next, we looked into a variable we created ourselves: comeback. The comeback boolean indicates if the team winning in the 2nd half was different than the team that won the game (created using the non-absolute difference in score after each half). We found a big difference here with comeback games having an excitement index of 2.25 higher than non-comeback games, as indicated on the right. But not all comebacks come equally, and we knew we wanted to quantify these comebacks somehow. We looked at two variables primarily: the home team's 1st half points (home_1sthalf) and the margin of the score at half (margin_1sthalf). While the home team's 1st half points variable had a slightly higher individual correlation to excitement index, we found within the model, the choice that minimized RMSE and maximized our R^2 value was the interaction between the 1st half margin and score difference. This made the most sense qualitatively as well, as we were looking for the best variable to help quantify these games' intensities. Therefore, this interaction became the core of our final model.

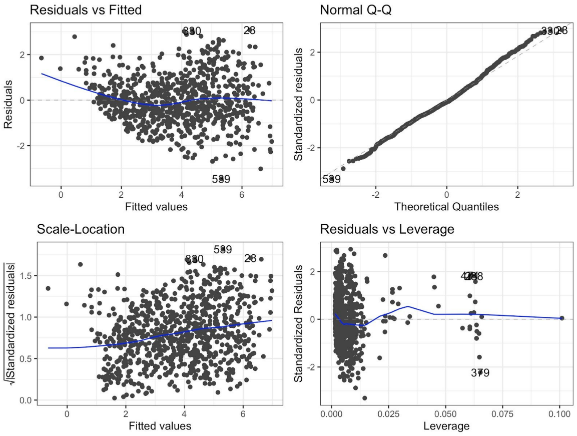


The Model

So with our chosen variables: *score_difference, margin_1sthalf, overtime (boolean), and comeback (boolean)*, we decided to conduct 5-fold cross validation to determine which combination of variables would minimize RMSE. We started by generating holdout predictions and RMSE values for the different combinations of the home team's 1st half points, the margin at half, and the absolute score difference, and as stated before, we chose the interaction of score difference and margin 1st half because it minimized RMSE. We then used that to test it with the combination of our booleans and found that the best model was the interaction between the margin at half and score difference with both the overtime and comeback booleans, which gave us an RMSE of about 1.06. Finally, when creating the final best fit model, we implemented a key manipulation: scaling the quantitative variables (margin and score difference) so the slopes could indicate their true overall impact to the final model as indicated below.



As stated before, the variable that had the greatest impact on the Excitement Index was *score_difference*. For our final model, Excitement Index decreases by 1.18 for every one (scaled) point increase in score difference. Following score difference, Excitement Index decreases by 0.47 for every one (scaled) point index in *margin_1sthalf*. The theme of closer, competitive games showed impact in our boolean variables as well such as with our overtime variable, Excitement Index increased by 0.98 if the team winning at the end of each half was different. In addition to this, Excitement Index increased by 0.49 if the game went to overtime. Additionally, the slope of the interacted variable came to be 0.33, and the starting excitement index intercept was 3.53.

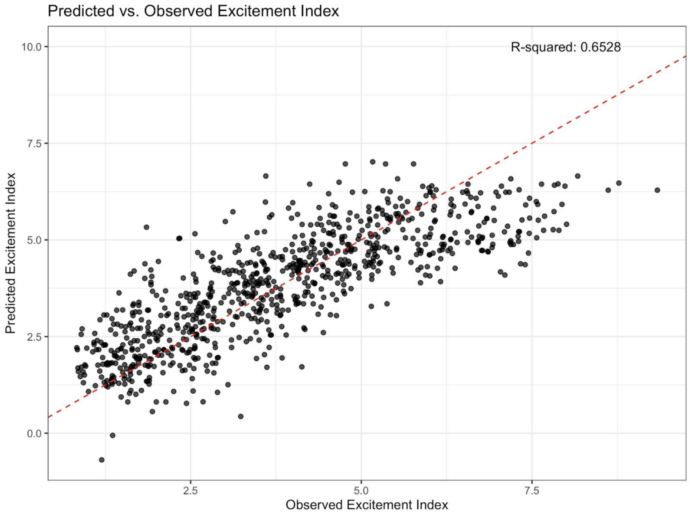


After choosing our variables, we needed to check how well our model actually fits the data. When looking at the Residuals vs. Fitted plots, it appears that there are no clear patterns in the data or no obvious outliers. However, a couple points with very low and high excitement indices proved to be overestimated and underestimated respectively, indicating to us that the best model to regress the excitement index is probably not perfectly linear.

The Normal Q-Q plot shows that our model's observations are generally very close to the dashed reference line. However, looking towards the bottom left and top right of the plot, the points tend to diverge from the dashed line, showing again that our model predicted games with mid-range Excitement levels generally well, but did not do as well in predicting games with very low and high Excitement levels. The Scale-Location plot demonstrates homoscedasticity of errors (equal, constant variance). In the Residuals vs. Leverage plots, the points with the highest leverage (in the right half of the plot) represent obscure games such as blowouts (dot 288: Penn State - 59 @ Maryland - 0, EI = 1.2) or competitive games without huge fluctuations (dot 379: Tulsa- 43 @ SMU - 37, Overtime Game, EI = 4.67) where one team, in this case Tusla, made a huge comeback that counted as one big fluctuation and therefore not a huge addition to excitement index.

Modeling Analysis Results

Our final step in our model analysis was to look at our model's predicted Excitement Index versus the observed Excitement Index. With an R^2 value of 0.6528, our model generally did a good job in predicting games with mid-range Excitement Index values, but as seen in the plot, it does not predict the extremes very well. Our model tends to underestimate games with very high Excitement Index values. In addition to this, our model underestimates some games with very low Excitement Indices but also overestimates a few games with low observed Excitement Indices, as seen in the bottom left quarter of the plot.



Conclusion & Discussion

The variables that had the greatest effect on Excitement Index were *comeback*, *overtime*, and the interaction of *score_difference* and *margin_1sthalf*. Additionally, the total point scoring in a game, the home team winning, and whether a game occurs at a later point in the season does not correlate with a higher Excitement Index, indicating that a fan's notion of a game's excitement level does not correspond with the true Excitement Index given in this dataset.

If we were to do additional analysis and modeling using the same dataset, we would look into a logarithmic or quadratic regression to attempt to predict the games with very high Excitement Index values as well as games with very low Excitement Index. In addition to this, we would also want to look into where both teams ranked on the day of the game. This would give us more context in seeing whether a team was considered to be an "underdog" or whether teams were ranked very closely, which may be considered a better matchup. It would be interesting to see whether rank had any effect on a game's Excitement Index.