Lacey Griffin
MA 346
Prof Carter
10 December 2020

Big Brother: Biased or Not?

For my final project in MA 346, I analyzed data regarding the long running TV Reality show, Big Brother. For years, Big Brother has faced backlash regarding their lack of racial diversity and the fact that there has never been an African American winner on a standard season. For this reason, I wanted to create a model which could predict what place a player would come in and so I could evaluate what factors affect the likelihood of players coming in a certain placement. I used Logistic Regression to model what place a player is expected to come in given their demographics, game type, and competition wins. What I learned is that for the first half of the game (the latter half of placements), demographics such as race and sexual orientation are very significant. If the player makes it past the first half of the game, competition wins are far more significant in determining where the player falls in the placement.

Since 2000, each summer, 12-17 people are moved into the Big Brother house where hundreds of cameras and microphones track their every move and word for up to 100 days. Each week, there are competitions for the powers of Head of Household (HoH) and Power of Veto (PoV). The winner of the HoH competition has the obligation to nominate two other players for eviction. Then, the HoH, the nominees, and three other players selected by random draw play in the PoV competition. The winner of the PoV has the ability to remove one of the nominees from the eviction block, leaving the HoH to select a replacement. Finally, one of these houseguests who are nominated will be evicted, then the process restarts with a new HoH competition.

To begin, I found a dataset created by Vince Dixon, whose report can be seen here. I read in and cleaned the dataset. I removed season one from the dataset because the season is not comparable to all subsequent seasons. Next, I worked through the dataset, replacing “na” values with a 0 for all columns which indicate frequency. For example, in season 2, there was no PoV competition yet, so the dataset used “na,” but for the purpose of this project, there needed to be numeric values in these cells.

Next, I analyzed the subset of just the non-white houseguests. Based on my analysis of just the non-white players in comparison to each other, it seems there are many races which are severely

underrepresented. For example, there is only one player which belongs to each of the Portuguese, Armenian, Pacific Island, and Black/Asian races. It is also notable that the majority of the Black players seem to be evicted earlier in the season, where the other races seem more fairly distributed.

Next, I moved into an analysis of the players who won the most competitions total. Some players play in multiple seasons, so I considered total wins per player, rather than wins per season. I then move into an analysis of the winners, to show that competition wins are important but certainly not everything.

To begin my model building, first, I took the dataset I had already cleaned and added a new column to it, outside of Python, which analyzes the game type of the player. This column categorizes players into athletic type, intellectual type, social type, manipulative type, and floater type. Most of these game types are very self-explanatory, but floater requires some detail. Floater type players do not have a particular strategy and generally continue in the game because the other players do not consider them a threat, so they are permitted to stay and float along while more threatening players are evicted.

Once I had all my columns, I had to ready the data for modeling so I made all of the categorical variables into Boolean columns such that the models will be able to use them. After this, I split the data into training and testing data, so I can test for overfitting once the model is built.

Next, I built a Logistic model using my new Boolean columns as well as some numeric columns such as competition wins and age. This model determines the probability of a player coming in each possible place and assigns the player to the place with the highest probability. I believe this to be the better model, not only because it is accurate on the training data about 30% of the time and it is within two places about 65% of the time, but because in Big Brother, the evictions are moderately independent. The evictions are not totally linear, where coming in seventh is closer to 6 than 9. The evictions tend to be in clusters, based on where the power in the house is. It is often the case that the intended week one evictee does not get evicted, then makes it very far in the game. For example, in Season 10, Memphis Garrett is supposed to be evicted in week one. At the last minute, the house shifted and evicted Brian Hart instead. Memphis Garrett went on to come in second in that season. When I test this model using the testing dataset, it performs moderately comparably, with about 10% accurate and 60% within two places. For the Logistic model, both the testing and training Root Mean Squared Error indicates that the model, on average, comes within 3 places of accuracy.

Next, I make a Linear model using the same predictors and test/train split. For the Linear model, the testing and training RMSE come within about 4 places of accuracy and predict correctly about 10% of the time and within 2 places about 40% of the time. For this reason, the Logistic model is the one I will reference for the remainder of this report.

The most interesting conclusion I drew from this model is that different placements have very different coefficients which help to predict if a player will come in that place. I ran the model many times and it was very consistent that the top half of places ~(1-8) are most heavily affected by the PoV and HoH wins. The latter half of places ~(9-16) are more heavily controlled by race, gender, sexual orientation, and game type. This indicates that for the first half of the game, people tend to be evicted for being unlike other players. This somewhat reinforces the complaints of the media regarding the implicit bias in the Big Brother Game. I don't know if there is anything CBS can do to mitigate this.

On the other hand, it is noteworthy that in the times I ran the model, Black was not often a significant coefficient for any of the placements, where Asian and Athletic came up quite frequently, which tends to oppose the claim that the game is directly biased against the African American Community. The fact that the top half of the placements are more heavily controlled by competition wins indicates that in the latter half of the game, one's abilities really come to the forefront of determining placement, which is sensible. Once a person has been in the house for 8 weeks, their abilities are far more significant to their continuation of the game than their sexual orientation or race. In the first half of the season, there is not much for the players to base their nominations and evictions on and many would argue that it is just human nature to oppose the people least like yourself. Whether this is acceptable or not is certainly not for me to decide, that is a question for a philosopher or sociologist, and I am neither.

For a deeper description of the game, necessary definitions, and to see all of my code which were discussed in this paper, click here. To see my dashboard which will allow you to see how certain types of players fared, click here. And to see my git repository which includes all of those links, data, and more, click here.