Movie Earnings Methods Page

As we investigated which factors had the biggest impact on movie earnings for our dataset, we tried many different methods in order to solidify our results. We started by bringing in data about inflation in order to create a data adjusted gross value for each movie in order to more accurately compare movies over time. We then created a heatmap of correlations for our quantitative explanatory variables. After, we turned these results into R^2 values to get a more uniform comparison of each of the variables. We also created histograms for each of these explanatory variables in order to see how our data was distributed for each of them. We then realized that after performing some preliminary calculations about the directors, the results were extremely right skewed due to outliers like James Cameron, Christopher Nolan, and others. So we created two separate directors categories and performed calculations on each of them, and our results were more correlated after splitting them apart. We then used t-tests to figure out even more how these director variables related to movie earnings. As many of the other variables hadn't given us great results, we turned to the genre variables provided by our dataset in order to see if they could help us find stronger results. We calculated the average inflation adjusted gross of each genre just to see if we could visually tell if certain genres made more than others. We also calculated the RMSE for genres as predictors. At this point though, even after calculating genre, we still hadn't revealed any strongly correlated variables, so we decided to calculate the R^2 and adjusted R^2 for all of our variables combined and other combinations of some of the variables to see if what combinations would give us the highest R^2 and adjusted R^2 values. As many of our preliminary results weren't very strong, we decided to turn to machine learning to possibly get better results.

Our hope was that machine learning would be able to perform calculations and find correlations between variables that our other models and calculations hadn't been able to thus far. We started with linear regressions of our various explanatory variables. After we received less than impressive results, we turned to decision trees and random forest classifications to try and predict blockbusters, a variable we created that contained only the top 90th percentile of movies. We ended by creating confusion matrices for the above models to get a better visual as to how accurately our models were predicting blockbusters. Through most of our models, whether they be exploratory or machine learning, we managed to only find weak relationships between our variables and movie earnings, however, our final classification models were somewhat accurate in predicting blockbuster films.