As with the logistic regression competition, I started with the same code from the linear regression assignment. I started with exploration where I looked at the dataset and saw if there was anything I could see from visualizations or if there were a lot of missing values. I found that one column had about 7% zeros, but the rest were good and there were no null values. Other than that, I could not find that much from exploration so I had to go into testing and selection somewhat blind. I played around with some different steps for feature engineering which slightly improved the accuracy, but the biggest payoff here in accuracy was using gridsearchcv. Thankfully it didn’t take that long to run and I could run it for every different set of outputs with different features which helped me find the best model. In the end I used a function to replace the zeros in the one column with the mean of the column, took the log of all the columns (which is why the zeros had to be replaced) and used the standard scaler. I was pretty happy with the improvement in accuracy I got for this competition over the base model. If I had more time I would probably try sqrt columns, feature selection, and just overall more time exploring to find better features to try.