Question 3

In order to determine the best approach, several models were built. First a linear model of shares on explanatory variables was run, and then the results were classified as viral or not viral. Two models were built using this method using stepwise selection. Stepwise selection was used for both a linear model with interactions, and one without interactions, and the model without interactions was chosen in both cases because the interactions did not significantly improve the performance of the model. The first was a linear model of shares, and the second was a linear model of log of shares. The log transformation was used to reduce the impact of outliers on the regression. The linear model of log of shares was more successful. Over multiple train test splits, it achieved a lower overall error rate, a lower false positive rate, and a higher true positive rate. This makes sense because the nature of viral articles (number of shares) is likely not linear. This model outperformed the best baseline model which assumes all articles are not viral. The baseline model has an overall error rate and false positive rate of approximately 0.49 and a true positive rate of zero over 100 train/test splits. The log linear model had, averaging over 100 train/test splits, an approximate overall error rate = 0.41, a false positive rate = 0.67, and a true positive rate = 0.86.

The second approach was to create an indicator variable for whether or not each article was viral based on its shares. To determine the best model to use for a logit regression, two methods were used. The first was a stepwise model, and the second was a lasso model. There were no major differences between the performance of these models. The logit model using stepwise variable selection outperformed the linear models and the best baseline model because it has more of an increase in true positive rate than false positive rate compared to the log linear model. The logit model had, averaging over 100 train/test splits, an approximate overall error rate = 0.37, a false positive rate = 0.37, and a true positive rate = 0.63.

The first method is effectively using a linear regression to predict a binary outcome. Logit performs better for classification problems like this because it is designed for binary classification using maximum likelihood estimation, while linear models use ordinary least squares, which may be less efficient in this case in terms of the true positive rate is lower. In the case of the linear model, the model is more likely to guess viral, because of the outliers (this was even more apparent in the model for shares). The overall error rate for the logit model is lower, and the false positive rate outperforms the baseline and the linear model. However, if the true positive rate were very important in determining a model's performance, the linear model would be better.