Question 4 - Building a Predictive Model with a Subset of Predictors

Optimal model(RFmodel_selected_4var):

Random Forest model on kw_avg_avg/ self_reference_avg_sharess /kw_max_avg / n_unique_tokens

OOB: 0.8372848

Out of sample mse: 0.7049856

Demonstration of our method for question 4 optimal model

There are mainly two parts in the work for question4: variable selection and model selection. In variable selection, we will use LASSO and random forest to have some potential optimal-subset of variables. Then, we will use splines plot from the GAM function to have a general picture of the relations between each variable and the log(share). Once we narrow down on 1/2 potential subset of variable, we will move on to the model selection part. We some analysis on what we got in the variable selection and previous work, we decide to apply non-linear model KNN and random forest to the subset of variables. With MSE from each potential model, we will finalize the best model and make prediction on test data.

Question 6 – Predictive Model with One Variable

Optimal model: Spline (GAM) using kw_avg_avg.

In-sample EPE: 0.8057390

Out-of-sample EPE: 0.8149446

Method: Our approach for this problem is to run all models on each of the 59 variables, calculate in-sample and out-of-sample EPE, and record the best variable for each model.