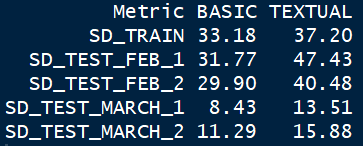
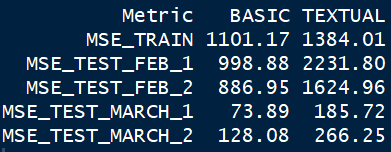
This document gives brief overview of the R program code sybmitted alongside which aims to predict the number of feedbacks that are likely to be received within next 24-hours after it’s posting.

The approach follows the standard sequential steps right from datasets preprocessing to model evaluation and final selection to achive the best acceptable solution in the context.Below is the sequence of the steps undetaken during the model builiding.

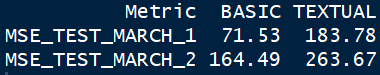
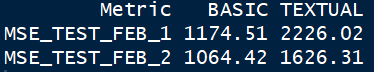
Step 1 : Importing the Train and Test datasets (the 3rd and the 21st day of Feb & Mar each)

Step 2 : As a pre-processing step, we retained only the reavant columns as needed based on “Basic” and “Textual” experimentation.

Step 3: I applied Linear Model using lm() function and produced the results as mentioned in the snippet below:



Step 4: To supplement our class learnings and contemplating a better fitting model, I also applied the Generalized Linear Model using glmnet (). I chose alpha = 0 (Lasso function) and randomly picked a fairly low lambda = 1.69 which was the 100th record in the model run.



Step 5: We observe the MSE (Mean Squared Error) to be lesser in Mar Test datasets in case of both Basic and Textual experiments and so is a better fit as compared to Feb Test dataset. We see lesser MSEs across the Train/Feb/Mar datasets for our experiment 1 as compared to experiment 2. This infers that resorting to the set of these 10 basic features yields better predictive capability for the likely no. of feedbacks to blogs as compared to set of 200 textual features.

In terms of strength of the model, we see a not-so-encouraging R-squared values (around 23 %) for both Basic and Textual experiments (even with glmnet). However, the lesser MSEs and SDs on Test datasets gives a feel-good factor. The less than 5% p-values for about 8 of these 10 features also gives a sense of good influence on the count of feedbacks received. The further data points in the past (earlier than year 2010), any additional features or a non-linear model might potentially improve upon train fit and yield a better explaining power (R-squared value).