Problem- prediction of the number of comments in the upcoming 24 hours on those blogs, The train data was generated from different base times that may temporally overlap. Therefore, if you simply split the train into disjoint partitions, the underlying time intervals may overlap. Therefore, the you should use the provided, temporally disjoint train and test splits to ensure that the evaluation is fair.

a. Read the dataset and identify the right features

### Data Preprocessing ###

library(data.table)

library(foreach)

# retrieve filenames of test sets

test\_filenames = list.files(pattern = "blogData\_test")

# load and combine dataset

train = fread("blogData\_train.csv")

test = foreach(i = 1:length(test\_filenames), .combine = rbind) %do% {

temp = fread(test\_filenames[i], header = F)

}

# log-transform

train[, V281 := log(1 + V281)]

test[, V281 := log(1 + V281)]

# drop continous variables without variation

drop = c(8, 13, 28, 33, 38, 40, 43, 50, 278)

train[, (drop) := NULL]

test[, (drop) := NULL]

# write to files

write.csv(train, "BlogFeedback-Train.csv", row.names = F)

write.csv(test, "BlogFeedback-Test.csv", row.names = F)

b. Clean dataset, impute missing values and perform exploratory data analysis.

### Basic Models ###

library(data.table)

library(MatrixModels)

library(e1071)

library(FNN)

library(glmnet)

library(ranger)

library(xgboost)

# load and combine dataset

train = fread("BlogFeedback-Train.csv")

test = fread("BlogFeedback-Test.csv")

# error measure

mse = function(y\_hat, y) {

mse = mean((y - y\_hat)^2)

return(mse)

}

c. Visualize the dataset and make inferences from that

# try kNN

pred\_knn = knn.reg(train\_x, test\_x, train\_y, k = 19)$pred

mse(pred\_knn, test\_y)

# try LASSO

mdl\_lasso = cv.glmnet(train\_x\_sparse, train\_y, family = "gaussian", alpha = 1)

pred\_lasso = predict(mdl\_lasso, newx = test\_x)

mse(pred\_lasso, test\_y)

# try random forest

mdl\_rf = ranger(V281 ~ ., data = train, num.trees = 1000, mtry = 120, write.forest = T)

pred\_rf = predict(mdl\_rf, test)

mse(pred\_rf$predictions, test\_y)

d. Perform any 3 hypothesis tests using columns of your choice, make conclusions

e. Create a linear regression model to predict the number of comments in the next 24 hours (relative to basetime)