kaggle report

Yiqun Xia

March 17, 2016

Preparation

We firstly load the data and remove non-predictors.

```
library(caret)

#read data and remove non-predictors
training <- read.csv("news_popularity_training.csv")[,-c(1,2,3)]
test.real <- read.csv("news_popularity_test.csv")[,-c(1,2,3)]
training <- training[-22686,]
```

We remove row 22686 as it has strange scale: ratios over 1000. Then we do some preprocess. We assume data are class balanced.

```
#remove n_non_stop_words
#since its low variance and 0 has already been included in other #variables

table(round(training$n_non_stop_words,4))
```

```
##
##     0     1
##   882 29117
```

```
table(round(test.real$n_non_stop_words,4))
```

```
##
##    0    1
##  299 9345
```

```
training$n_non_stop_words <- NULL
test.real$n_non_stop_words <- NULL
```

We cancel repetative variables(linear combination of other variables)

```
#cancel repetative variables
rep.dummy <-findLinearCombos(training)$remove
training[,rep.dummy] <-data.frame(NULL)
test.real[,rep.dummy] <- data.frame(NULL)
```

We then transfer population and num_keywords from integer to factor. We also make the two series of dummy variables to two factor. We omit these verbose code.

Finally we normalize the non-categorical predictors.

```
category.name <- c("weekday", "data_channel", "num_keywords")

training.pp <- preProcess(training[, !names(training) %in% c(category.name,"popularity")],
                          method = c("center", "scale"))

training <- predict(training.pp, newdata = training)
test.real <- predict(training.pp, newdata = test.real)

dim(training)
```

```
## [1] 29999    45
```

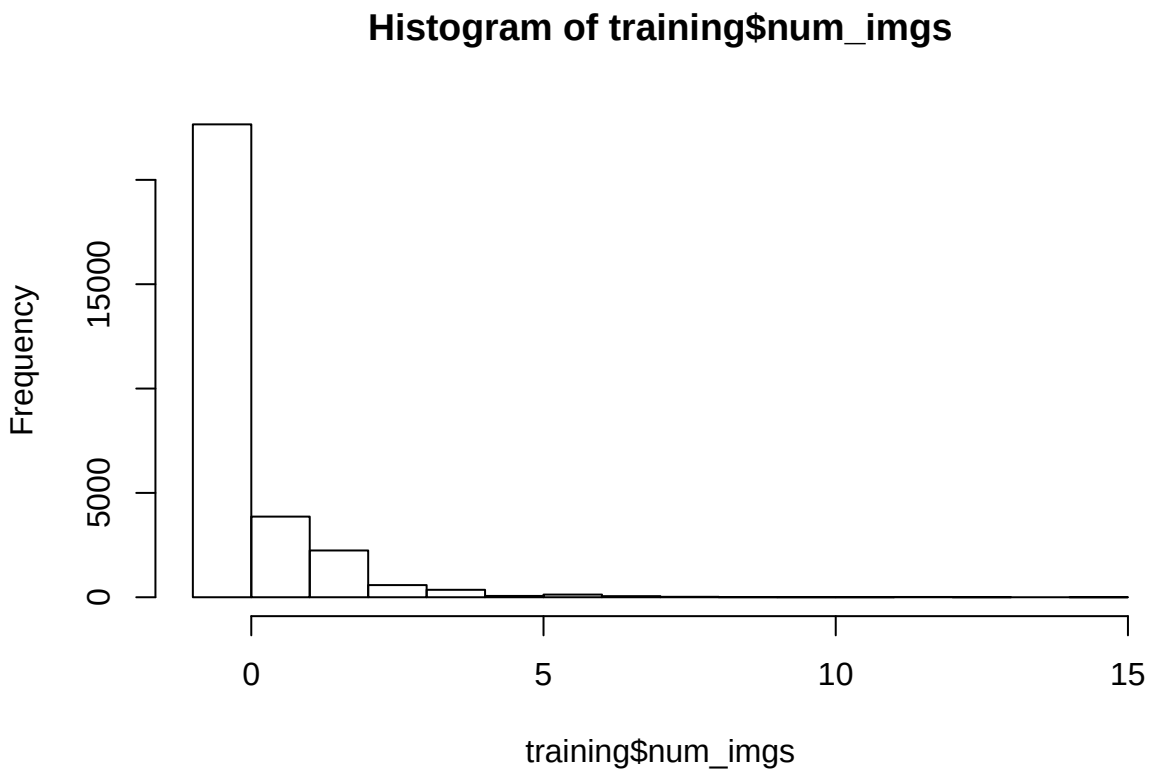
```
dim(test.real)
```

```
## [1] 9644   44
```

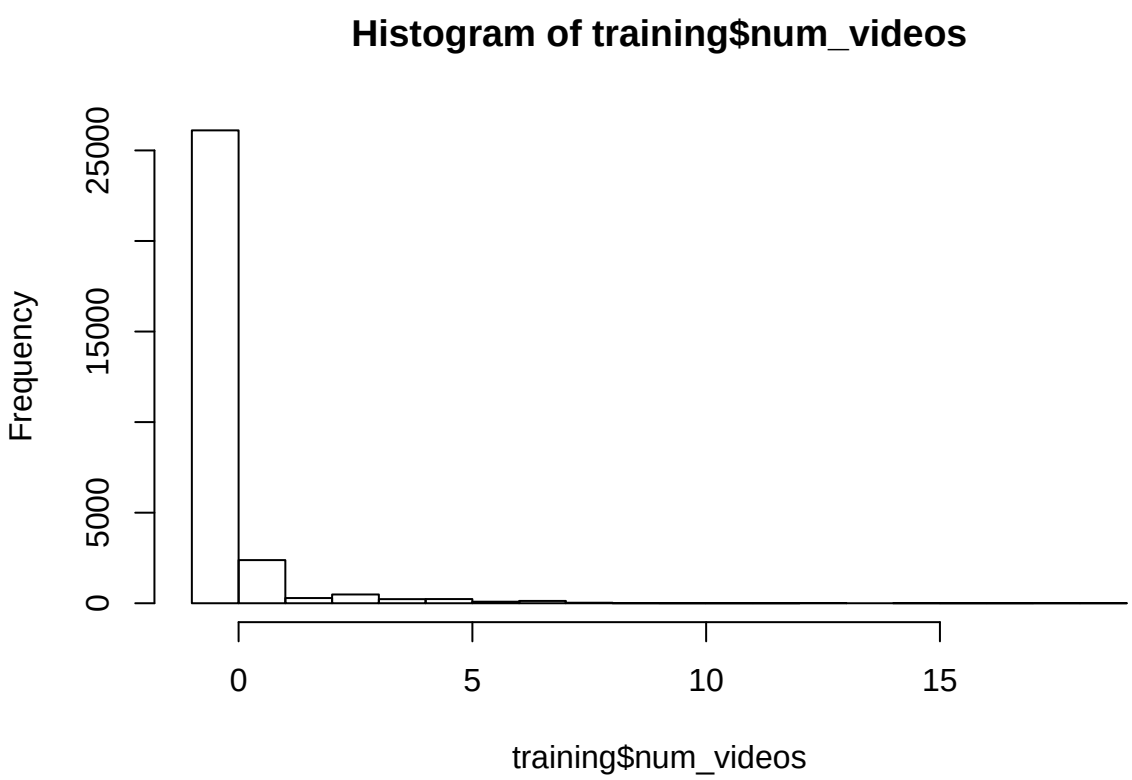
We found that many predictors are seriously right tail and we have both continuous and catigorical predictors, so we didn't think SVM might do a good job especially for multiclass.

```
hist(training$num_hrefs)
```


```
hist(training$num_imgs)
```




```
hist(training$num_videos)
```



We then do multinomial logistic regression and some ordinal logistical regression like proportional odds, it seems diffcult to improve the fits. And becasue there bunch of outliers, these parametric models might not be very suitable.

Finally, we focused on tree based models with ensemble methods. We chose the most powerful two tools: randomForest and gbm. Refer to caret package, we use 3-repeat 5-fold CV to tune the paremeters. We select 3 models with the best estimated accuracy from each type. Our strategy is just to do simple model average taking majority rule.