833933

```
library(tictoc)
tic()
PATH <- "https://raw.githubusercontent.com/aldosolari/DM/master/docs/DATA/"
train <- read.csv(paste0(PATH,"train.csv"), sep="")
test <- read.csv(paste0(PATH,"test.csv"), sep="")
library(dplyr)
library(ranger)
library(xgboost)
y <- train$y
n <- nrow(train)
m <- nrow(test)
combi <- bind_rows(train[,-ncol(train)], test)
combi <- combi %>% as_tibble() %>%
  mutate_at(.vars = c("vas1", "vas2", "payment.method",
       "gender", "status", "tariff.plan", "activ.area", "activ.chan"), factor)
combi$activ.area[which(combi$activ.area==0)] <- 1
combi$activ.area <-  factor(combi$activ.area, levels = levels(combi$activ.area)[-1])
subs <- combi %>% filter(q03.in.dur.tot < 0) %>%
  select(ends_with("in.dur.tot"), -q03.in.dur.tot) %>%
  rowMeans()
combi[which(combi$q03.in.dur.tot < 0), "q03.in.dur.tot"] <- as.integer(subs)
subs <- combi %>% filter(q09.out.dur.peak < 0) %>%
  select(ends_with("out.dur.peak"), -q09.out.dur.peak) %>%
  rowMeans()
combi[which(combi$q09.out.dur.peak < 0), "q09.out.dur.peak"] <- as.integer(subs)
train <- combi[1:n,]
train <- bind_cols(train, y)
colnames(train) <- c(colnames(train)[1:ncol(train)-1], "y")
test <- combi[(n+1):(n+m),]

best_subset <- c("q09.out.ch.peak", "q09.out.val.peak", "q09.out.dur.peak",
                 "q09.in.dur.tot", "q09.out.val.offpeak","q09.out.dur.offpeak",
                 "q09.in.ch.tot", "q09.out.ch.offpeak", "age", "q07.out.val.peak",
                 "q08.out.dur.peak","q07.out.ch.peak","tariff.plan",
                 "q07.out.dur.peak", "q08.out.val.peak","q08.out.ch.peak",
                 "q08.out.dur.offpeak", "q08.out.val.offpeak", "q08.in.dur.tot",
                 "q08.in.ch.tot", "q08.out.ch.offpeak", "q07.in.dur.tot",
                 "q06.out.dur.peak", "q07.in.ch.tot", "q06.out.ch.peak",
                 "q06.out.val.peak", "q06.in.dur.tot", "q05.out.dur.peak",
                 "q05.out.val.peak", "q06.in.ch.tot", "q05.out.ch.peak",
                 "q04.out.dur.peak", "activ.area", "activ.chan",
                 "q04.out.val.peak", "q05.in.dur.tot", "q05.in.ch.tot",
                 "q04.in.dur.tot", "q04.out.ch.peak", "q07.out.dur.offpeak",
                 "q03.out.dur.peak", "q07.out.val.offpeak", "q03.out.val.peak",
                 "q04.in.ch.tot", "q09.ch.sms", "q03.out.ch.peak",
                 "q03.in.dur.tot", "gender", "q03.in.ch.tot",
```

```
                "q02.out.val.peak", "payment.method", "q01.out.val.offpeak",
                "q01.out.val.peak", "q09.ch.cc", "q02.out.dur.peak",
                "q01.out.dur.peak", "q02.in.dur.tot", "q01.in.dur.tot", "status")
prev <- function(train, test, best_subset){
  train <- train %>% mutate(across(starts_with("q"), .fns = log1p)) %>%
    mutate(y = log1p(y))
  test <- test %>% mutate(across(starts_with("q"), .fns = log1p))
  y_train <- train$y

  beta_rf <- 0.5854
  beta_boost <- 0.4243

  fit1 <- ranger(y ~., data = train[,c(best_subset, "y")],
                 mtry = 15, num.trees = 2000, verbose = F)
  yhat1 <- predict(fit1, data = test[,best_subset])$predictions

  fit2 <- xgboost(data = model.matrix(y ~., train)[,-1], label = y_train,
                  params = list(eta = 0.005, max_depth = 6, gamma = 0,
                                colsample_bytree = 0.5,
                                min_child_weight = 1, subsample = 1),
                  nrounds = 4000, verbose = 0)
  yhat2 <- predict(fit2, newdata = model.matrix(~., test)[,-1])

  yhat <- beta_rf * yhat1 + beta_boost * yhat2
  yhat[yhat<0] <- 0
  yhat <- expm1(yhat)


  return(yhat)

}

yhat <- prev(train = train, test = test, best_subset = best_subset)
toc()
```

```
## 622.723 sec elapsed
```