arima

Nikhil Muthukrishnan

20 March 2019

```
library(data.table)
library(caret)
library(tidyverse)
library(data.table)
library(readxl)
library(zoo)
library(forecast)
library(lubridate)
options(max.print=10,digits = 10)
setwd("C:/Users/in0166/Desktop/ycorrelation/assembled")
df0<-fread("2017,5min,.pv.csv",header = T)
df<-data.frame(df0)
t1<-df$t1
df<-df[,c(F,T)]
df$t1<-NULL
phasemaps<-read.csv("phasesmaps.csv")
sensornames<-names(df)
sensornames<-data.frame(sensornames)
names(sensornames)<-"Name"
phasenames<-left_join(sensornames,phasemaps,"Name",all.x=T)
phasenames$Name<-NULL
names(df)<-phasenames$PhaseNum
df<-apply(df,2,as.numeric)
df<-data.frame(t1,df)
df$t1<-dmy_hm(df$t1)
mind<-as.POSIXct("2017-02-07 00:00:00 UTC")
maxd<-as.POSIXct("2017-12-30 00:00:00 UTC")
gts<-seq.POSIXt(mind,maxd,"mins")
gts<-data.frame(gts)
names(gts)<-"t1"
df<-merge(gts,df,sort = T,all.x = T)
Batchtimemapping <- data.frame(read_excel("Batchtimemapping.xls"))
Batchtimemapping$t1<-as.POSIXct(round(Batchtimemapping$t1,units = "mins"))
df<-merge(df,Batchtimemapping,"t1",all.x = T)
df$BatchNumber<-na.locf(df$BatchNumber,fromLast = T,na.rm = F)
batchduration<-data.frame(table(df$BatchNumber))
names(batchduration)<-c("BatchNumber","hours")
batchduration$hours<-round(batchduration$hours/60)
df<-merge(df,batchduration,"BatchNumber",all.x = T)
qty<-read_excel("qty.xls")
df<-merge(df,qty,"BatchNumber",all.x = T)
df<-na.locf(df,na.rm = F)
specs<-read_excel("allspecs.xls")
df<-merge(df,specs,"BatchNumber")
df<-na.locf(df,na.rm = F)
grabs<-read.csv("grabsamples.csv")
grabs$t1<-dmy_hm(grabs$t1)
```

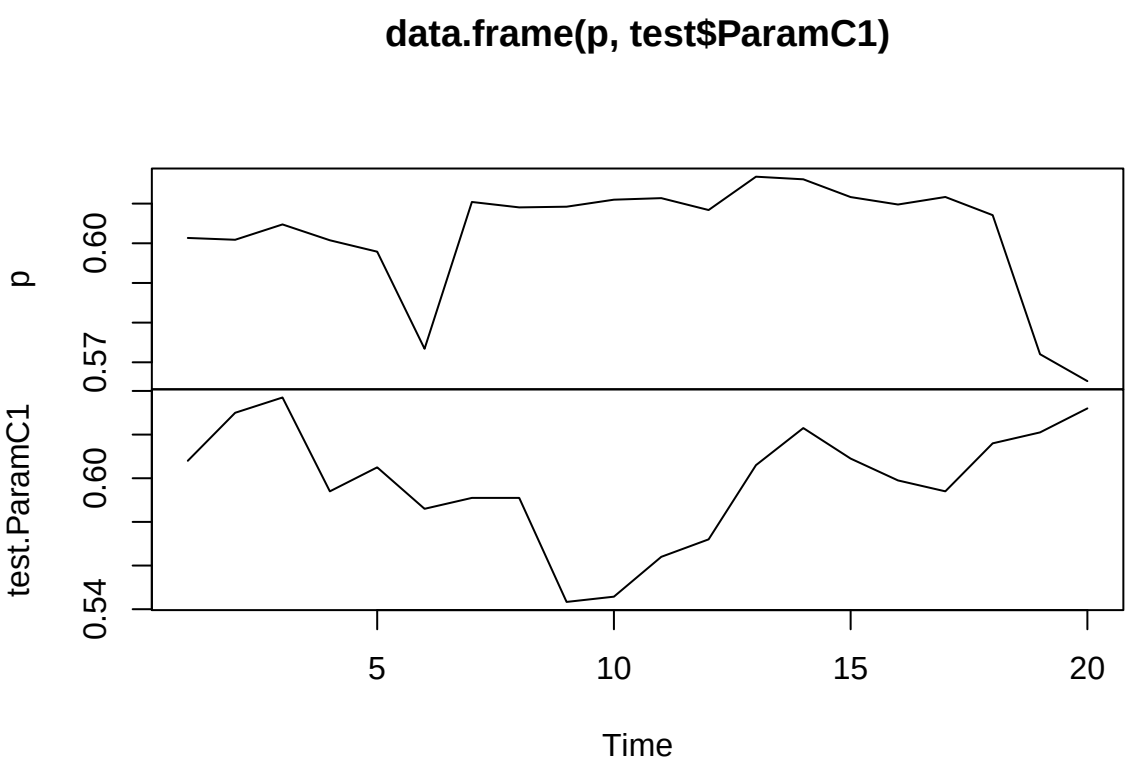
```
df<-merge(df,grabs,"t1",all.x = T)
df<-na.locf(df,na.rm = F)
weather<-read.csv("weather.csv")
weather$t1<-dmy_hm(weather$t1)
weather<-filter_all(weather,all_vars(.>0))
df<-merge(df,weather,"t1",all.x = T)
PO<-read.csv("PO.csv")
PO$t1<-dmy_hm(PO$t1)
df<-merge(df,PO,"t1",all.x = T)
df<-na.locf(df,na.rm = F,fromLast = T)
df$index<-ave(df$hours,df$BatchNumber,FUN = seq_along)

df$SC2_LIC55001.pv_Ph_4<-NULL
df$SC2_LIC23501.pv_Ph_4<-NULL
df$SC2_FIC20461.pv_Ph_2<-NULL
df$SC2_FIC20462.pv_Ph_2<-NULL
df$SC2_FIC20464.pv_Ph_2<-NULL
df$SC2_LIC55022.pv_Ph_8<-NULL
df$SC2_TIC20760.pv_Ph_2<-NULL
# df$ParamC10<-NULL
# df$ParamC11<-NULL
# df$ParamC14<-NULL
# df$ParamC13<-NULL
# df$ParamC15<-NULL

t1<-df$t1
df$t1<-NULL
df<-aggregate.data.frame(x = df,by = list(df$BatchNumber),FUN = mean)
df<-na.aggregate(df)
df$BatchNumber<-df$Group.1
df$Group.1<-NULL
df1<-df[,grepl("SC2|t1|ParamC1",names(df))]
yclass<-cut(df1$ParamC1,
            breaks = c(-Inf,0.605,0.626,Inf),
            labels=c("low","middle","high"))
train<-df1[1:200,c(1:39,40)]
test<-df1[201:220,c(1:39,40)]

#lm
model<-lm(ParamC1~.,train)
p<-predict(model)
postResample(p,train$ParamC1)
```

```
##          RMSE      Rsquared           MAE
## 0.01597499313 0.44657051913 0.01209700295
```

```
p<-predict(model,test)
postResample(p,test$ParamC1)
```

```
##          RMSE      Rsquared           MAE
## 0.03360470216 0.10171942055 0.02601005437
```

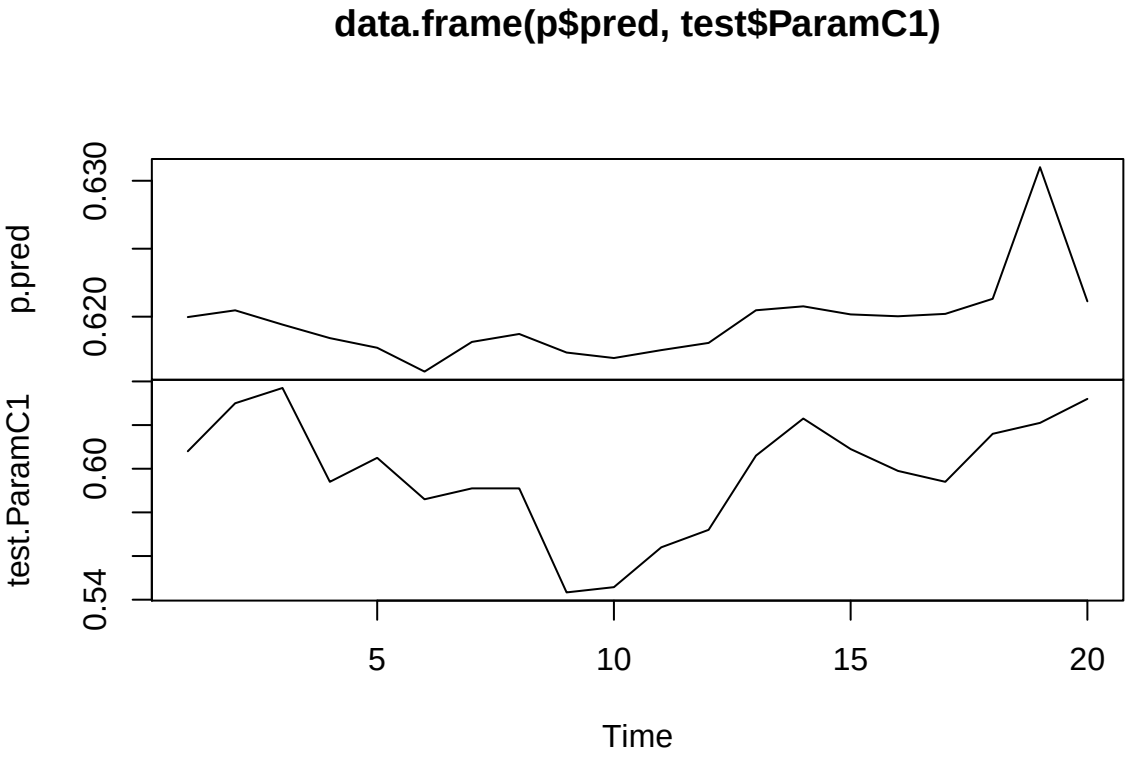
```
plot.ts(data.frame(p,test$ParamC1))
```




```
#ts
train<-df1[1:200,c(1:39,40)]
test<-df1[201:220,c(1:39,40)]

trregs<-as.matrix(train[,c(1,2,13,18,4,6,7,log(21))])
tsregs<-as.matrix(test[,c(1,2,13,18,4,6,7,log(21))])
model<-auto.arima(train$ParamC1,xreg = trregs)

p<-predict(model,n.ahead = 20,newxreg = tsregs)
postResample(p$pred,test$ParamC1)
```

```
##          RMSE      Rsquared           MAE
## 0.03237081006 0.28975808769 0.02535872549
```

```
plot.ts(data.frame(p$pred,test$ParamC1))
```




```
df1$yclass<-cut(df1$ParamC1,
                breaks = c(0,0.616,0.7),
                labels=c("1","2"))
table(df1$yclass)
```

```
##
##   1   2
## 117 103
```

```
yclass<-df1$yclass
model<-lm(yclass~.,df[,2:39])
p<-predict(model)
p<-ifelse(p<1.5,1,2)
postResample(p,yclass)
```

```
##     Accuracy        Kappa
## 0.7727272727 0.5430755172
```

```
confusionMatrix(as.factor(p),as.factor(yclass))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  1  2
##          1 93 26
##          2 24 77
##
##                Accuracy : 0.7727273
```

```
##                  95% CI : (0.7115897, 0.8263579)
##     No Information Rate : 0.5318182
##     P-Value [Acc > NIR] : 1.224491e-13
##
##  [ reached getOption("max.print") -- omitted 14 rows ]
```

```
p<-predict(model,test[,-40])
p<-ifelse(p<1.5,1,2)
postResample(p,yclass[201:220])
```

```
## Accuracy    Kappa
##     0.85     0.50
```

```
confusionMatrix(as.factor(p),as.factor(yclass[201:220]))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  1  2
##          1 15  3
##          2  0  2
##
##                Accuracy : 0.85
##                  95% CI : (0.6210732, 0.9679291)
##     No Information Rate : 0.75
##     P-Value [Acc > NIR] : 0.2251560
##
##  [ reached getOption("max.print") -- omitted 14 rows ]
```