# Forecasting customer demand for packaging in SIG Indonesia market for better customer sales promotion


## Business Analytics Using Forecasting

## Team 2

Shang-Chi Tu
Beverly Lin
Ken Wu
Jimmy Wang

# Executive Summary


**Introduction of Our Client: SIG**

In this project, our client is SIG Combibloc. SIG is a leading systems & solutions provider of carton packaging and flexible filling machines for beverages and food, helping bring food products to consumers in a safe, sustainable and affordable way.

**Business Problem**

The main business goal for this project is to offer next-12-month forecasts for future monthly sales volume of packages on a customer and product type level at the beginning of each month by us. With the forecasts, the salesperson of SIG Indonesia can use it as a reference during their monthly meeting to revise their month marketing strategy and design tailor-made promotions for customer sales in advanced.

**Data Description**

We obtained data from SIG which included fields such as Customer, Product hierarchy, Month and Plan qty, etc. The time period of the series are from January, 2009 to December 2018 and it recorded every sale for 45 different customer and product type level.

**Forecasting Solution**

Before forecasting, we did data preprocessing and customer segmentation. Then, we applied different models to our 3 types customer. For the inactive customers, we forecast zero in the next 12 months. For the new customers, we used Naive to forecast their sales demand for packaging. For the repeat customer, because of some customers' extreme behaviors, we used linear regression and modeled these months separately. For more information, please refer to the detailed report below.

**Recommendations**

According to the results, we observe orders from some customers are steadily growing. Hence, the salesperson of SIG Indonesia market can pay more attention on those customers with potential to purchase more in the future while developing a sales promotion strategy. Besides, SIG can try more external data such as customer satisfaction scale can help discover customers' ordering behavior and improve the forecast accuracy.

# Detailed Report

## Problem Description

In this project, our client is SIG Combibloc. SIG is a leading systems & solutions provider of carton packaging and flexible filling machines for beverages and food, helping bring food products to consumers in a safe, sustainable and affordable way.

- **Business Goal**
  We aim to offer next-12-month forecasts for future monthly sales volume of packages at the beginning of each month. We think that with our forecast data, the salesperson of SIG Indonesia market can compare the forecasts conducted by us with their forecasts data to arrange and purpose new sales promotions which are tailor-made for each customer during their monthly meeting. The forecasts data is presented with our R code and spreadsheet.
- **Forecasting Goal**
  In order to match our business goal, we attempted to forecast the customer demand for packaging in the next 12 months on a customer and product type level. This is a forward-looking goal, and the forecast horizon is set to be from 1 to 12 (a-year-ahead forecasts per month). It can help SIG to develop the marketing strategy for the next year in advance. The forecasting result will provide SIG a better way for customer sales promotion.

## Data Description

We obtained data from SIG, which included fields such as Customer, Product hierarchy, Month and Plan qty (thousands of package sales volume). This data had entries from January, 2009 to December and it recorded every sale for 45 different customer and product type level.

| Customer | Name4 | Product hierarchy | Plan qty. | Month |
|---|---|---|---|---|
| 951433 | Indonesia | PC010150A | 1,830.400 | 201806 |
| 951433 | Indonesia | PC010150A | 1,830.400 | 201808 |
| 951433 | Indonesia | PC010150A | 1,830.400 | 201809 |
| 951433 | Indonesia | PC010150A | 1,830.400 | 201810 |
| 951433 | Indonesia | PC010150A | 1,830.000 | 201811 |
| 951433 | Indonesia | PC010250J | 15,416.960 | 201701 |
| 951433 | Indonesia | PC010250J | 10,381.440 | 201702 |
| 951433 | Indonesia | PC010250J | 14,080.000 | 201703 |
| 951433 | Indonesia | PC010250J | 11,116.800 | 201704 |
| 951433 | Indonesia | PC010250J | 7,040.000 | 201705 |

Figure 1. Sample of a 10 rows per raw data series

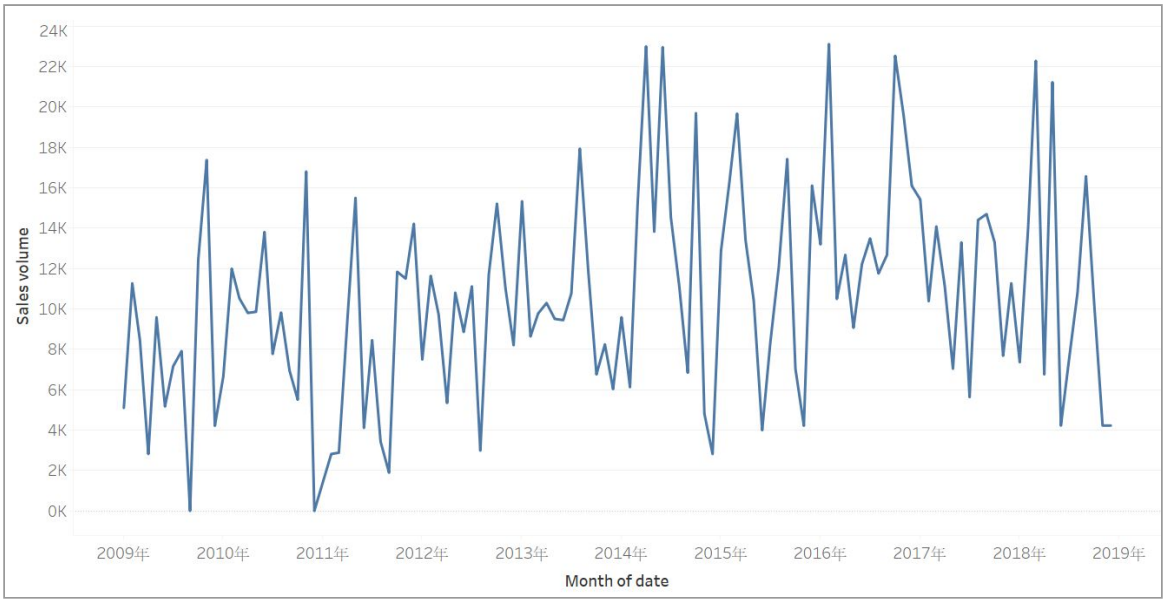


Figure 2. Time plot for a customer and product type level (Customer: 951433 product hierarchy: PC010250J)

## Brief Data Preparation Details

- **Data Preprocessing**

  Before forecasting, we did data preprocessing to aggregate the raw data into a usable format.

| | 200901 | 200902 | 200903 | 200904 | 200905 | 200906 | 200907 | 200908 | 200909 | 200910 | 200911 | 200912 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 951433 PC070300J | 13464 | 17222.76 | 12117.6 | 10771.2 | 13464 | 12111.12 | 13839.48 | 16122.24 | 3916.8 | 16156.8 | 16401.6 | 16156.8 |
| 951433 PT020500J | 0 | 0 | 0 | 0 | 0 | 0 | 1900.8 | 0 | 0 | 1800.9 | 2438.1 | 0 |
| 951433 PT021000J | 475.2 | 946.8 | 0 | 157.2 | 3517.2 | 0 | 1421.7 | 934.8 | 0 | 0 | 3314.1 | 1880.4 |
| 951870 PC010125A | 6336 | 12638.08 | 8448 | 12672 | 4147.84 | 12663.68 | 12672 | 15755.52 | 6510.72 | 9651.84 | 18742.4 | 4190.72 |

Figure 3. Raw data after data preprocessing

- **Customer Segmentation**

  We plotted time series and found out that some customers didn't place orders for a while. Hence, we divided customers into three groups based on their ordering period.

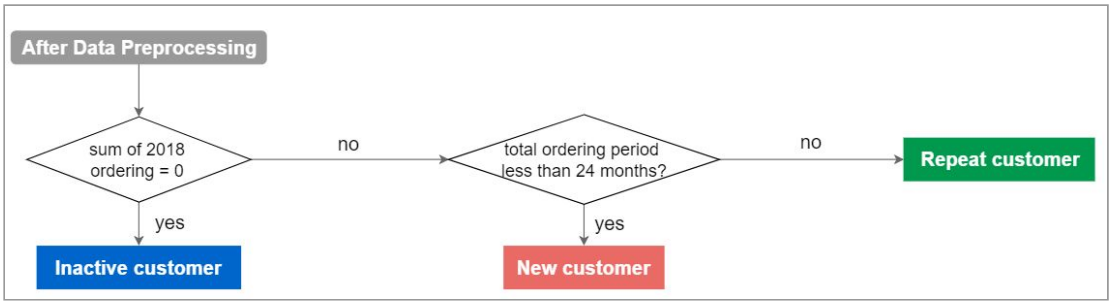


Figure 4. Customer Segmentation

## Forecasting solution

- **Methods Applied**

  We used seasonal naïve forecast model as our benchmark because it didn't consider external variables and it had a relatively high level of predictive accuracy. Then, we applied different models to our 3 types customer.

| *Customer* | *Methods & Detailed* |
|---|---|
| **Inactive** | These customers who didn't purchase any products in 2018, we **forecast zero** in the next 12 months. |
| **New** | The ordering period was less than 24 months. Therefore, we used **Naive** to forecast their sales demand for packaging. |
| **Repeat** | Many forecasting models were considered and applied on this data, such as naive, seasonal naïve, exponential smoothing, neural networks and linear regression in order to compare their predictive performance. We chose **Linear Regression with trend and seasonality** which had the best performance. In addition, we found Dec and Jan are not properly forecasted because of some customers' extreme behaviors (peaks in Dec and dips in Jan). We tried to model these months separately. Feb-Nov as first group, Jan as a second group, and Dec as a third group. Finally, we combined the results of these groups together and used it as our forecasting model. |

Table 1. Method and detailed of different types of customer

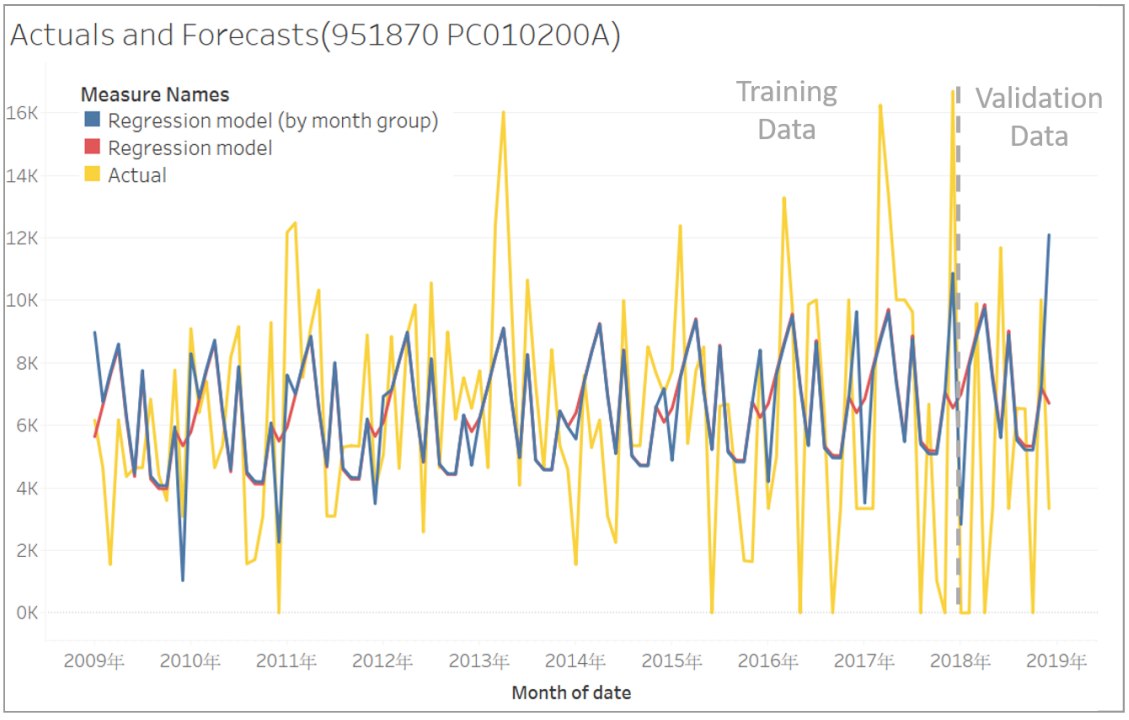


Figure 5. Time plot of actuals and the forecasts by regression models

- **Performance Evaluation**
  To evaluate the performance of the methods, we used RMSE and error chart to compare performance of different types of customer.

| | Benchmark: Snaive | Inactive customer: Forecast zero | New customer: Naive |
|---|---|---|---|
| **RMSE** | 2699.6 | 0 | 2334.8 |

Table 2. Performance of Inactive customer and new customer

| | Benchmark: Snaive | Repeat customer: Linear Regression | Repeat customer: Linear Regression (by month group) |
|---|---|---|---|
| **RMSE** | 2699.6 | 4418.6 | 4240.2 |

Table 3. Performance of repeat cusomer

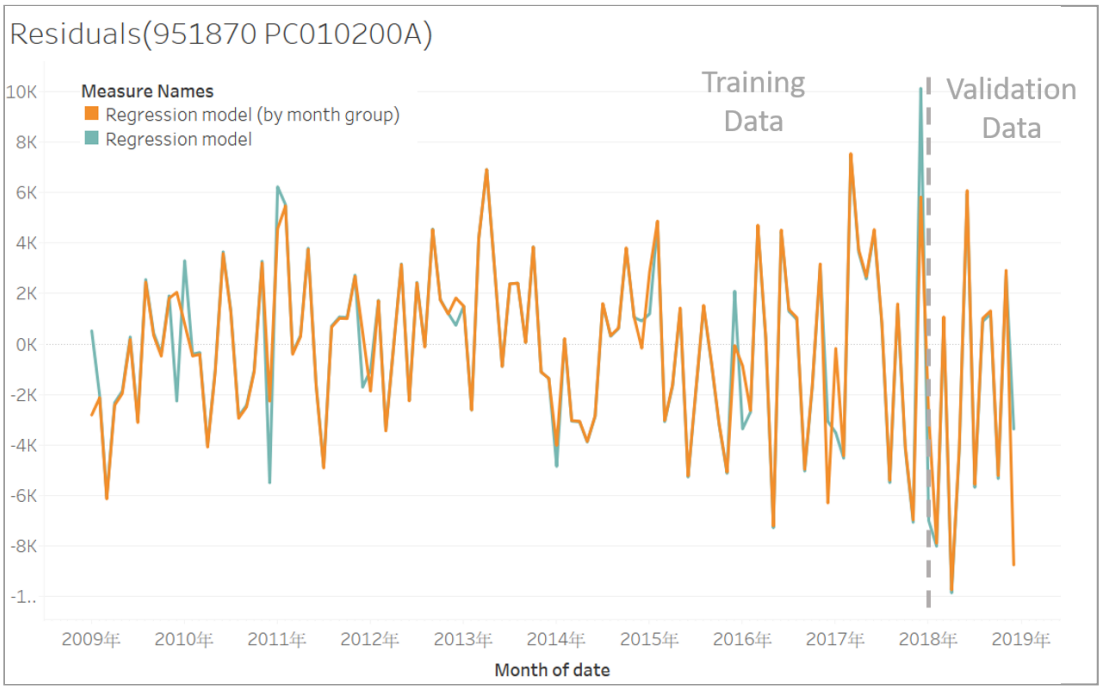


Figure 6. Time plot of residuals by regression models

From the error chart, the performance of regression model which modeled months separately is better than model all months together especially in January and December.

**Note.** The time plots of 5 series forecasting results of are shown in Appendix 2.

## Conclusions

Through the whole forecasting project, several limitations and recommendations show as follows:

- **Limitations**

  1. For the new customers: Due to the short period of ordering, if we try to forecast their next 12 months demand for packs, the results might be inaccurate. The more data we have, the forecast model becomes more accurate.

- **Recommendations**

  1. From the time plots, we observe orders from some customers are steadily growing. Hence, to pay more attention on those customers with potential to purchase more in the future while developing a sales promotion strategy.
  2. Try more external data such as customer satisfaction scale can help discover customers' ordering behavior and improve the forecast accuracy.

# Appendix 1: R code

```
library(readxl)
library(reshape2)
library(dplyr)

## Data Preprocessing
# read all sheets
y2009_2010 <- read_excel("APS_Actual_Sales.xlsx",sheet = "2009-2010")
y2011_2012 <- read_excel("APS_Actual_Sales.xlsx",sheet = "2011-2012")
y2013_2014 <- read_excel("APS_Actual_Sales.xlsx",sheet = "2013-2014")
y2015_2016 <- read_excel("APS_Actual_Sales.xlsx",sheet = "2015-2016")
y2017_2018 <- read_excel("APS_Actual_Sales.xlsx",sheet = "2017-201811")

#change all negative value into zero
y2009_2010$`Plan qty.`[which(y2009_2010$`Plan qty.`< 0)] <- 0
y2011_2012$`Plan qty.`[which(y2011_2012$`Plan qty.`< 0)] <- 0
y2013_2014$`Plan qty.`[which(y2013_2014$`Plan qty.`< 0)] <- 0
y2015_2016$`Plan qty.`[which(y2015_2016$`Plan qty.`< 0)] <- 0
y2017_2018$`Plan qty.`[which(y2017_2018$`Plan qty.`< 0)] <- 0

# creare new aggregate dataframe
filter_y2009_2010 <- aggregate(`Plan qty.`~ Name4 + Customer + `Product
hierarchy`+ Month, sum, data = y2009_2010)
filter_y2011_2012 <- aggregate(`Plan qty.`~ Name4 + Customer + `Product
hierarchy`+ Month, sum, data = y2011_2012)
filter_y2013_2014 <- aggregate(`Plan qty.`~ Name4 + Customer + `Product
hierarchy`+ Month, sum, data = y2013_2014)
filter_y2015_2016 <- aggregate(`Plan qty.`~ Name4 + Customer + `Product
hierarchy`+ Month, sum, data = y2015_2016)
filter_y2017_2018 <- aggregate(`Plan qty.`~ Name4 + Customer + `Product
hierarchy`+ Month, sum, data = y2017_2018)

# Combine 2009 - 2018 dataframe
sig <-
rbind(filter_y2009_2010,filter_y2011_2012,filter_y2013_2014,filter_y2015
_2016,filter_y2017_2018)

# Concate customer and product hierarchy
sig$`Link(cust+pro)` <- paste(sig$Customer,sig$`Product hierarchy`)
sig <- sig[-c(2,3)]

# create pivot table of month, package, and demand numbers
sig_pivot <- as.data.frame(tapply(sig$`Plan
```

```
qty.`,list(sig$`Link(cust+pro)`,sig$Month),sum))
sig_pivot[is.na(sig_pivot)] <- 0

# merge country back with package
sig_pivot$Country <-
sig[match(rownames(sig_pivot),sig$`Link(cust+pro)`),]$Name4

# write to csv
write.csv(sig_pivot,"APS_clean_version.csv")

## Customer Segmentation
#load time series
SIG.data<- read.csv("APS_clean_version.csv")
#subset time series from none zero period
first_nonzero <- function(ts){
  for(i in 1:length(ts)){
    if(ts[i] > 0){
      break}
  }
  return(subset(ts, start = i))
}

#put SIG.data in function
customer_segement <- function (timeseries_dataframe){

  #filter
  data.indonesia <- SIG.data %>%
    filter(Country =='Indonesia')%>%
    select(-Country) %>%
    t(.)

  #rename, dataframe, data type
  colnames(data.indonesia) <- data.indonesia[1,]
  df <- data.indonesia[2:dim(data.indonesia)[1],] %>%
    as.data.frame()
  for(i in 1:dim(data.indonesia)[2])df[, i] <- as.character(df[,i]) %>%
    as.numeric()

  #add columns
  tag_df <- cbind(t(df), Inactive=apply(df[(dim(df)[1]-11):dim(df)[1],],
2, sum),
                  New=rep("",dim(data.indonesia)[2]),
                  Repeat=rep("",dim(data.indonesia)[2])) %>%
    as.data.frame()
```

```
#set factor level
Inactive.levels <- levels(tag_df$Inactive)
levels(tag_df$Inactive) <- c(Inactive.levels, "Inactive","")
New.levels <- levels(tag_df$New)
levels(tag_df$New) <- c(New.levels, "New","")
Repeat.levels <- levels(tag_df$Repeat)
levels(tag_df$Repeat) <- c(Repeat.levels, "Repeat","")

#Inactive customer
for (i in 1:dim(tag_df)[1]){
  if(tag_df$Inactive[i]==0){
    tag_df$Inactive[i] <- "Inactive"
  }else{
    tag_df$Inactive[i] <- ""
  }
}

#New & Repeat
for (i in 1:dim(df)[2]){
  data.indonesia.ts <- ts(df[,i],
                          start = c(2009,1),
                          end = c(2018, 12),
                          #monthly data
                          freq = 12)
  if(length(first_nonzero(data.indonesia.ts))<24 ){
    tag_df$New[i] <- "New"
  }else{
    tag_df$Repeat[i] <- "Repeat"
  }
}

#remove already "Inactive"
for (i in 1:dim(tag_df)[1]){
  if(tag_df$Inactive[i]=="Inactive"){
    tag_df$New[i] <- ""
    tag_df$Repeat[i] <- ""
  }
}

#combine category
df_segement <- cbind(tag_df[,1:(dim(tag_df)[2]-3)],
                     Segment=str_trim(paste(tag_df$Inactive,
                                            tag_df$New,
                                            tag_df$Repeat)))
write.csv(df_segement, "df_segement.csv")
```

```
  return(df_segement)
}

customer_segement(SIG.data)

## Forecasting model
#load time series
df_segement<- read.csv("df_segement.csv")
# load inactive customer
data.indonesia.inactive <- df_segement %>%
  filter(Segment=='Inactive')%>%
  select(-Segment) %>%
  t(.)

# load new customer
data.indonesia.new <- df_segement %>%
  filter(Segment=='New')%>%
  select(-Segment) %>%
  t(.)

# load repeat customer
data.indonesia.repeat <- df_segement %>%
  filter(Segment=='Repeat')%>%
  select(-Segment) %>%
  t(.)

first_nonzero <- function(ts){
  for(i in 1:length(ts)){
    if(as.numeric(ts[i]) > 0){
      break}
  }
  return(subset(ts, start = i))
}

forecast_result <- function(matrix_){

  forecast.result <- data.frame("date"=c(201901:201912))

  for (i in 1:ncol(matrix_)){

    data.indonesia.ts <- ts(matrix_[2:121,i], start = c(2009,1), end =
c(2018, 12), freq = 12)
    train.ts <- window(data.indonesia.ts, start=c(2009,1), end = c(2018,
12))
    nValid<-12
```

```
    if(forecast_zero(data.indonesia.ts) == FALSE){
      nonzero.ts <- first_nonzero(data.indonesia.ts)
      # forecast repeat customer
      if (length(nonzero.ts)>=24){

        # seperate months (Jan as a first group, Dec as a second group,
Feb-Nov as a third  group)
        Jan. <- ts(subset(data.indonesia.ts, cycle(data.indonesia.ts) ==
1),frequency = 1)
        Dec.<- ts(subset(data.indonesia.ts, cycle(data.indonesia.ts) ==
12),frequency = 1)
        # Feb to Dec
        Rest.temp <- ts(subset(data.indonesia.ts,
cycle(data.indonesia.ts) != 1),frequency = 11)
        Rest. <- ts(subset(Rest.temp, cycle(Rest.temp) != 11),frequency
= 10)

        # data partition
        Jan.train.ts <- subset(Jan., end = length(Jan.))
        Dec.train.ts <- subset(Dec., end = length(Dec.))
        Rest.train.ts <- subset(Rest., end = length(Rest.))

        Jan.train.linear <- tslm(Jan.train.ts ~ trend)
        Jan.train.linear.pred <- forecast(Jan.train.linear, h= 1)
        Dec.train.linear <- tslm(Dec.train.ts ~ trend)
        Dec.train.linear.pred <- forecast(Dec.train.linear, h= 1)
        Rest.train.linear <- tslm(Rest.train.ts ~ trend+season)
        Rest.train.linear.pred <- forecast(Rest.train.linear, h= 10)

        tem.Jan <- c()
        tem.Dec <- c()
        tem.Rest <- c()
        if ((Jan.train.linear.pred$mean[1]) < 0){
          tem.Jan <- 0
        }else{
          tem.Jan <- Jan.train.linear.pred$mean[1]
        }

        if ((Dec.train.linear.pred$mean[1]) < 0){
          tem.Dec <- 0
        }else{
          tem.Dec <- Dec.train.linear.pred$mean[1]
        }

        for (j in (1:10)){
```

```
          if ((Rest.train.linear.pred$mean[j]) < 0){
            tem.Rest[j] <- 0
          }
          else
          {
            tem.Rest[j] <- Rest.train.linear.pred$mean[j]
          }
        }

        forecast. <- c(tem.Jan, tem.Rest, tem.Dec)
        forecast.result[i+1] <- forecast.

      }else{
        # forecast new customer
        nfit <- naive(as.numeric(train.ts))
        forecast.result[i+1] <- rep(nfit$mean[1], nValid)
      }

    }
    else{
      # forecast inactive customer
      forecast.result[i+1] <- numeric(12)
    }
    colnames(forecast.result)[i+1] <- matrix_[1,i]
  }
  return(forecast.result)
}

# export
inactive.result <- forecast_result(data.indonesia.inactive)
inactive.forecast<-
data.frame(cbind(colnames(inactive.result),t(inactive.result)))
write.xlsx(inactive.forecast, file = "inactive_customer.xlsx",
sheetName="inactive.forecast", row.names=FALSE, append=TRUE)

new.result <- forecast_result(data.indonesia.new)
new.forecast<- data.frame(cbind(colnames(new.result),t(new.result)))
write.xlsx(new.forecast, file = "new_customer.xlsx",
sheetName="new.forecast", row.names=FALSE, append=TRUE)

repeat.result <- forecast_result(data.indonesia.repeat)
repeat.forecast<-
data.frame(cbind(colnames(repeat.result),t(repeat.result)))
write.xlsx(repeat.forecast, file = "repeat_customer.xlsx",
sheetName="repeat.forecast", row.names=FALSE, append=TRUE)
```

# Appendix 2: Time plot of series with forecasts