Australian Retail Sales

**Executive Summary:** Goal is to Explore the process of selecting the right model to express the trajectory of this retail sales data. We will try to use base model such as snaive to see how it will perform and make modifications only to the data feed side of things and then use an auto arima model to see if we could do better. As the primary goal is to have an efficient model which will help us understand any irregularities or variations in data, which helps in troubleshooting all the while saving memory usage.

**Data set:** We have Australian retail sales data in an excel file as our dataset.

**Process Flow in RStudio:**

#Loading necessary libraries

library(forecast)

##a. Reading the excel and loading the dataset

retaildata <- readxl::read\_excel("retail.xlsx", skip=1)

##b. Selection (I have chosen South wales food Industry sales data)

Southfood <- ts(retaildata[,"A3349398A"], frequency=12, start=c(1982,4))

##c. Exploration (Using Five different plots)

autoplot(Southfood)

A graph showing a line

Description automatically generated

ggseasonplot(Southfood)

A graph of different colored lines

Description automatically generated

ggsubseriesplot(Southfood)

A graph showing the time of the day

Description automatically generated with medium confidence

gglagplot(Southfood)

A graph of different colored lines

Description automatically generated

ggAcf(Southfood)

**A graph with lines in the middle

Description automatically generated**

**Comments on the exploration:** Autoplot makes the positive trend of the data series apparent. Although seasonality is visible, It gets much clearer from ggsubseries plot and ggseason plot, where we see that data is following similar pattern. We can also see cyclicity in the data when we observe the gglagplot. The plot indicates how at every lag the data is conforming to the dataset and they showcase heavy correlation. The final ggacf plot solidifies our observations, as acf values at every lag are high.

###2. For the chosen retail time series:

##a. Splitting the data

Southfood.train <- window(Southfood, end=c(2010,12))

Southfood.test <- window(Southfood, start=2011)

##b. Verifying the split is carried out effectively.

autoplot(Southfood) +

autolayer(Southfood.train, series="Training") + autolayer(Southfood.test, series="Test")

A graph showing a line

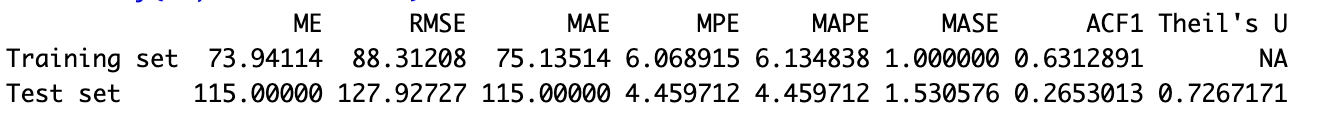
Description automatically generated with medium confidence

##c. Forecasting using snaive

fc <- snaive(Southfood.train)

##d. Comparison of accuracy for our forecast wrt test set data

accuracy(fc,Southfood.test)



**Comment on the results from snaive model:** Values of ME, RMSE, MAE are very high for training and test sets. Whereas MAPE value of 6.13% seems average, MASE value of 1 and 1.5 for training and test express that the model is worse than Naïve forecast. ACF1 is the only value which has gone down favourably.

##e. Checking the residuals.

checkresiduals(fc)

A graph of a graph

Description automatically generated with medium confidence

**Comments on the residuals:** Residuals are where snaive seems to have done a decent job but not a great one. The ACF values have come down at most of the higher lags and the distribution seems almost normal. ACF1 values from accuracy have also expressed the improvement in that aspect.

##f. How sensitive are the accuracy measures to the training/test split?

###Scenario 1: Decreased training set length

Sf.train <- window(Southfood, end=c(2009,12))

Sf.test <- window(Southfood, start=2010)

fc1 <- snaive(Sf.train)

##d. Accuracy results

accuracy(fc1,Sf.test)



###Scenario 2: Increased training set length

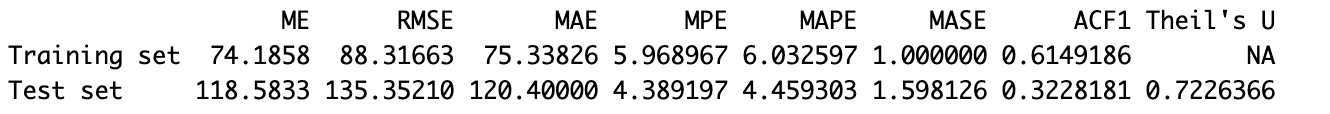
Sf.train2 <- window(Southfood, end=c(2011,12))

Sf.test2 <- window(Southfood, start=2012)

fc2 <- snaive(Sf.train2)

##d. Accuracy results

accuracy(fc2,Sf.test2)

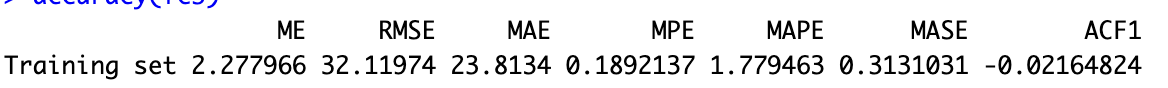


##g. Applying auto ARIMA

**Comments on sensitivity:** All the metrics are relatively similar irrespective of training set length changes, except for ACF1. So, the data doesn’t seem sensitive to the split changes.

fc3 <- auto.arima (Southfood, seasonal = TRUE)

accuracy(fc3)



checkresiduals(fc3)

A math equation with black text

Description automatically generated

A graph of a graph

Description automatically generated with medium confidence

##h. According to auto.arima() what is the model fit? Explain the model that is generated by the function.

ARIMA (3,1,3)(0,1,2)[12] is the generated model. Based on the notation ARIMA(p,d,q)[P,D,Q][m], which can also be called as SARIMA, due to seasonal autoregressive components. In the generated model for our data,

p(3) - Model has 3 autoregressive terms,

d(1) - Data was differenced once to make it Stationary (One of the main criteria),

q(3) - Three moving average terms,

P(0) - No seasonal autoregressive terms,

D(1) - Differenced once seasonally,

Q(2) - Two seasonal moving average terms involved,

m(12) - Seasonality period is 12

##i. Interpret the accuracy of your auto.arima model.

Mean error (ME) is very little and positive which indicates slight positive bias in forecasts. RMSE although smaller than snaive, it is still moderate. MPE is 0.19% which shows slight overestimation in forecasts, MASE of 0.31 showcases the superior performance when compared to naïve forecast. Finally, the ACF1 value is -0.02, which is almost zero and indicative of the no autocorrelation remains in the residuals, and it is also visible in the residuals plot we have generated.

The residuals after auto arima have reduced into white noise, as they follow a normal distribution and their acf values dropped significantly. Also there seem to be no pattern remaining in the residuals, so auto Arima seem to have done a good job in capturing all the meaning from the residuals.

Finally, I would say that by applying those components sequentially could be a better approach than Auto Arima modeling. As the insights will be much clearer, rather than reverse Interpretation of given results.

***Formatting model:***

Comments added within Code are in Green color.

Answers and explanation-based comments for report are in Boxes

Plots and console Output in Blue Boxes