Case Study #1: Sales Forecasting

In this case, consider the data on worldwide monthly sales from a large chain of grocery stores (*673\_case1.csv*). This dataset was extracted from [www.kaggle.com](http://www.kaggle.com/) and then adopted for this course. The monthly sales are given for a period of 2015-2022 and measured in millions of dollars. The goal is to identify the best forecasting model to predict monthly sales in 12 months of 2023.

**Questions**

# Identify time series components and plot the data.

* 1. Create time series data set *sales.ts* in R using the *ts()* function.

Text

Description automatically generated with medium confidence

* 1. Employ the *plot()* function to create a data plot of the historical data, provide it in your report, and explain what data patterns can be visualized in this plot.

Text

Description automatically generated with medium confidence

Chart, histogram

Description automatically generated

DATA PATTERNS-

* We can clear observe additive seasonality.
* The highest drop in sales was experienced in the first quarter of 2018.
* We can clearly see that the sales in last six months of each were drastically increased and when the year finished it experienced a drastic drop.
  1. Apply the *Acf()* function to identify possible time series components. Provide in the report the autocorrelation chart and explain time series components existing in the historical data.

Logo, company name

Description automatically generated

Chart, box and whisker chart

Description automatically generated Text, table

Description automatically generated with medium confidence

* We can observe that there is a positive ACF in the first 3 lags and the last 4 lags. ●
* We see that lags 4 to 8 are negative and rest are positive
* For lags and lag 8, we can see that the ACF is very close to 0, which means there might not be any correlation for lag 4 and lag 8.
* Lags 4, 5,6, 7, 8 and 9 are not significant as they are below the significance level

# Use trailing MA for forecasting time series.

* 1. Develop data partition with the validation partition of 24 monthly periods (2 years) and training partition of 72 monthly periods (6 years). Provide the data partition’s R code in your report.

Text

Description automatically generated

* 1. Use the *rollmean()* function to develop 3 trailing MAs with the window width of 3, 8, and 12 for the training partition. Present the R code for these MAs in your report.

A screenshot of a computer

Description automatically generated with medium confidence

* 1. Use the *forecast()* function to create a trailing MA forecast for each window width from question *2b* in the validation period, and present one of them, e.g., with window width of 3, in your report.

A screenshot of a computer

Description automatically generated with medium confidence

Chart, line chart, histogram

Description automatically generated

* 1. Apply the *accuracy()* function to compare accuracy of the three trailing MA forecasts in the validation period. Present the accuracy measures in your report, compare MAPE and RMSE of these forecasts, and identify the best trailing MA forecast.

Text, letter

Description automatically generated

The first model (i.e., window width 3) is the best model as it has the lowest MAPE(11.697) and RMSE(47.264)

# Apply the two-level forecast with regression and trailing MA for residuals.

* 1. Develop using the *tslm()* function a regression model with linear trend and seasonality. Present the model summary in your report. Present and briefly explain the model equation in your report. Using this model, forecast monthly sales in the validation period with the *forecast()* function. Present the forecast in your report.

trend.seas <- tslm(train.ts ~ trend + season)

summary(trend.seas)

Text

Description automatically generated

trend.seas.pred <- forecast(trend.seas, h = nValid, level = 0)

trend.seas.pred

Table

Description automatically generated

* 1. Identify regression residuals in the training period, apply a trailing MA (window width of 3) for these residuals using the *rollmean()* function, and identify trailing MA forecast of these residuals in the validation period (use the *forecast()* function). Provide the trailing MA forecast for residuals in the validation period in your report.

Table

Description automatically generated with low confidence

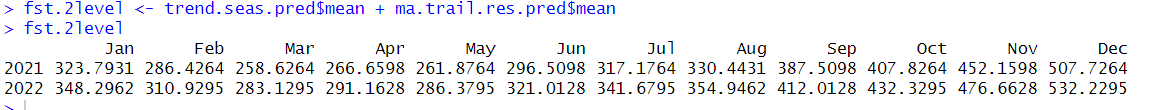
A picture containing text

Description automatically generated

Table

Description automatically generated

* 1. Develop two-level forecast for the validation period by combining the regression forecast and trailing MA forecast for residuals. Present in your report a table that contains validation data, regression forecast, trailing MA forecast for residuals, and two-level (combined) forecast in the validation period. Apply the *accuracy()* function to compare accuracy of the regression model with linear trend and seasonality and the two-level (combined) model with the regression and trailing MA for residuals. Present the accuracy measures in your report, compare MAPE and RMSE of these forecasts, and identify the best forecasting model for the validation period



Table

Description automatically generated

Text

Description automatically generated

The two-level model with the regression and trailing MA for residuals is a better model as the MAPE(10.55) and RMSE(31.624) is the lowest

* 1. For the entire data set, identify the regression model with linear trend and seasonality and trailing MA with the window width of 3 for the regression residuals. Use these models to forecast the 12 months of 2023 and develop a two-level forecast for the 12 future months as a combination of the specified forecasts. Present in your report a table that contains the regression forecast, trailing MA forecast for residuals, and two-level (combined) forecast in the 12 months of 2023.

Text

Description automatically generated

* 1. Develop a seasonal naïve forecast for the entire historical data set and apply the *accuracy()* function to compare accuracy of the three forecasting models: seasonal naïve forecast, regression model with linear trend and seasonality, and two-level (combined) model with the regression and trailing MA for residuals. Present the accuracy measures in your report, compare MAPE and RMSE of these forecasts, and identify the best forecasting model for forecasting monthly sales in 2023.

Text, letter

Description automatically generated

The two level model has the is better as it has the lowest MAPE(5.3554) and RMSE(17.3811)

# Use advanced exponential smoothing methods.

* 1. For the training partition (from question *2a*), use the *ets()* function to develop a Holt- Winter’s (HW) model with automated selection of error, trend, and seasonality options, and automated selection of smoothing parameters for the training partition. Presentthe model summary (output) and explain the model in your report. Use the model to forecast monthly sales for the validation period using the *forecast()* function, and present this forecast in your report.

Graphical user interface, text, application, email

Description automatically generated

Text

Description automatically generated

* 1. To make a forecast in the 12 months of 2023, use the entire data set (no partitioning) to develop the HW model using the *ets() function* for the model with the automated selection of error, trend, and seasonality options, and automated selection of smoothing parameters. Present the model summary (output) and explain this model in your report. Use the model to forecast monthly sales in the 12 months of 2023 using the *forecast()* function, and present the forecast in your report.

Graphical user interface, text

Description automatically generated

We used “ZZZ” in R to select an automatic model and hence an “AAA” model was generated from it. The below model is called Exponential Smoothing model with Additive errors, Additive trend, and Additive seasonality. The optimal value of the exponential smoothing constant is alpha = 0.2052. Also, the Smoothing constant for seasonality is gamma = 0.0001

Text

Description automatically generated

We used “ANA” in R to select an automatic model and hence an “ANA” model was generated from it. The below model is called an Additive Error, No Trend and Additive Seasonality model. The optimal value of the exponential smoothing constant is alpha = 0.3072. Also, the Smoothing constant for seasonality is gamma = 0.0001

Text

Description automatically generated with medium confidence

Text, table

Description automatically generated

* 1. Apply the *accuracy()* function to compare the two models: seasonal naïve forecast (applied in question *3e*) and the HW model developed in question *4b*. Present the accuracy measures in your report, compare MAPE and RMSE of these forecasts, and identify the best forecasting model.

Text, letter

Description automatically generated

The AAA model is better as it has the lowest RMSE(23.526) and MAPE(7.577)

* 1. Compare the best forecasts identified in questions *3e* and *4c*. Explain what your final choice of the forecasting model in this case will be.

(3E)

Text, letter

Description automatically generated

(4C)

Text, letter

Description automatically generated

I would use the 2 level model out of all of them because it performed the best as it has the lowest RMSE(17.3811) and MAPE(5.3554)

