Time Series and Product Analysis for a Demand Forecasting Tool

Faye Beatriz C. Tumaliuan

Central European University
July 2018

Table of Contents

Time Series and Product Analysis for a Demand Forecasting Tool

"We live in a world of data" - this is a fact that not just businessmen, analysts, and scholars know and understand, but every person who recognizes how much data they produce and need in order to interact and function in a rapid-moving, data-driven society. The data we produce and collect now was scarcely imaginable in the past decades, and companies have been using data in the past years in all types of industries for top-level management decisions.

All manufacturing organizations, no matter the size, face certain levels of risk and uncertainty. Business decisions are made under these conditions, and one aspect in lessening risks is by determining future product and sales prospects. The company that this study focuses on faces the same challenge - that is, to predict the optimal production quantities for more than 300 products to lessen inventory costs and optimize sales quantities. They produce various hygiene paper products like kitchen towels, tissues, and toilet paper. Since the company started from humble beginnings, with steady growth, current company decisions regarding production manufacturing quantities are still based on experienced intuition. Demand forecasting helps in this aspect, as it helps in predicting product demand for future periods. Quantitative techniques will be discussed as we try to determine product demand, and as the goal is to aid the company in decision making by exploring data-driven decisions, a predictive analytics tool – based on the results focusing on demand forecasting, was created. This tool should be able to process new data and new products and update the forecasting models. Thus, the models used are simple models that could be updated automatically without manual intervention.

Demand forecasting can be done using sales history data and other demand inputs like promotions, product pricing, and inventory data. These demand inputs are what differentiates between sales forecasting and demand forecasting, since both uses sales history data. These demand inputs eliminate biases when predicting demand from sales history – like when products are put on promotion and demand goes up, or stock outs happen and no sales are produced due to unavailability of products. In this study, actual sales data for five (5) years from January 2013 to April 2018 was used. This includes invoice and purchase numbers, product quantities, prices, total sales, customer and shipping information, currency, country, sales channel, and sales type (discounted or base price). The goal is to forecast demand with a forecast horizon of four (4) months for each of the products. Initially, 19 products were included in this data, and the company states that possibility of stock-outs is very low in this sales history. Thus, without inventory data, demand forecasting was performed by assuming there are no product stock-outs.

Since our sequence of measurements of sales data are with the time component, product analysis and forecasting were done using time series methods and models. The data was aggregated on a monthly basis, wherein monthly product quantities are the data points of the time series for each product. Price corrections, credit notes, and goods returned were not included in the data. By data exploration of summaries and totals using simple charts like histograms, we were able to see products with the most sales, and that the company produced new products over the years - with some of these products only sold for specific events or months like soccer or summer season.

Also a result of the data exploration, it was identified that the products are mostly at discounted prices more than they are at their base prices. For this reason, history of average prices were analyzed per product time series using X11 decomposition to see underlying patterns of trend, season, and remainder components. Decomposition was performed on every product, and seasonality of prices for each product whether additive (the same seasonality over time) or multiplicative (changes seasonality over time) was identified. Of course, whether average prices also have an increasing or decreasing trends were determined. Upon removing trend and seasonality, the remainder series is checked for extreme values and outliers. Average price for some products had outliers from level shifts that were because of the products being new to the market. Easter holiday had some effect on the average prices of some products. Since the X11 method also uses ARIMA (Autoregressive integrated moving average) components to describe autocorrelations in the data, it was seen that the average price of some products have a relationship with its own previous average prices (autoregressive component). The average price of some products also has a relationship with the errors from the average price of the previous values (moving average component).

Focusing on the goal, sales history by quantity sold was analyzed next, and the demand forecasting is based on this time series data. By looking further at X11 decompositions, sub series plots and season plots, each product demand time series was tested if it is seasonal or non-seasonal. Having seasonality means that there are variations that occur at regular intervals - in this case, we are testing for monthly intervals. Like with the average price analysis, seasonal and trend components, outliers, holiday effects, and ARIMA components were identified for the demand patterns of each product. Once seasonality was determined for each, they were tested if they were stationary and non-stationary. A stationary time series is one whose properties do not depend on the time at which the series is observed. Thus, time series with trends, or with seasonality, are not stationary - the trend and seasonality will affect the value of the time series at different times. On the other hand, a stationary series is like white noise - it does not matter when you observe it, it should look much the same at any point in time. (Hyndman, R., Athanasopoulos, G. Forecasting: principles and practice. 2018). Determining these features helped in building time series models later on the next phases of the analysis. For example, when testing for stationarity, Box-Ljung test was used in a way wherein seasonal products had different lag input than non-seasonal ones. Autocorrelation coefficients plot (ACF plot) was also used to determine stationarity between the monthly demands. In addition to these two tests, the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test was also used. These tests are automatic in R and this software was used in doing the analysis and the forecasting tool dashboard.

Since we want to forecast how much demand for a specific product is for the next four (4) months, various time series models for every product on a 64-month long time series data were created. Simple benchmark methods like seasonal naïve and drift methods were used. In this case, we set each forecast to be equal to the last observed value from the same season of the year (e.g., the same month of the previous year). This is the same with the drift method, but it allows the forecasts to increase and decrease over time.

Seasonal and Trend decomposition using Loess was also used, like the X11 method earlier. Exponential smoothing methods like Simple Exponential Smoothing, Holt’s Linear Smoothing, and Holt-Winter’s Seasonal Method were also used. It is a combination of the naive method (where the forecast are equal to the latest values) and the average method (where the forecast is equal to the average of all historical values). Forecasts are calculated using weighted averages, where the weights decrease exponentially as observations come from further in the past - the smallest weights are associated with the oldest observations. ARIMA (Autoregressive integrated moving average), as discussed earlier, was also used. We can set the parameters for ARIMA, but we use the auto ARIMA functionality with R on this analysis since our main goal is a dashboard which updates best models from new data.

The analysis will fit all the time series methods, using 60 months of training data (Jan 2013 to Dec 2017). The training data changes, as the tool chooses the four latest months as test data and the rest as training data. The best model was chosen using Mean Absolute Percentage Error (MAPE) on the remaining data as test set, and forecast results for this 4-month period from Jan 2018 to April 2018. Chosen models for each product is stored in an output file which is accessed by the forecast tool. The tool trains the new data using the chosen models, and can be updated by rerunning the script which analyzes the best models. Forecasted values include the Forecast Points and the 95% Prediction Intervals.

Since some forecast results are noisy, an analysis on the effect of price is done by linear regression on time series data. For the linear model, the average price is computed for every month including base and discounted prices. This average price is transformed in log form, and is used as the predictor for the total demand for the month. The average price as a predictor has a significant effect on some products, and on some not. In order to forecast demand for each product after training the linear regression model, the average price would have to be also forecasted using its trend and season. The forecasted average prices will be the new data fed into the forecast model for the linear regression. We compare the MAPE of the linear regression model with the MAPE we got from the best models from the previous result. If the MAPE is lower, then we use linear regression with price as predictor for the specific product. Currently, MAPE ranges from 5% to 30% for 17 products, and 80% to 90% for the remaining two.

After having time series results using the different models we used, we want to set up benchmarks to help the decision making process. Since we have Forecast Point and Intervals for our results, we want to know the level of inventory we are required to produce to achieve a certain service level by using the mean and standard deviation of the product demand data fit to a normal distribution. Service level is the probability that all demand is satisfied without any stock-out as defined by Salvendy. (Salvendy, G.Handbook of Industrial Engineering Technology and Operations Management. 3rd Edition. 2001.) Knowing the level of inventory we need, we can minimize costs and can help decision makers to dynamically investigate the forecast results.

The final implementation of the forecasting tool includes totals and summaries which aim to help the company visualize their data. It shows sales by volume sold, sales by sales channel (comparing export and domestic channels), volume sold by price types, and price histories. Finally, it includes the demand forecasts together with inventory level benchmarks as defined by chosen service levels. The company can choose out of more than 300 products they have, update the models per product, and visualize their forecast results.