Data driven decision making

Authors: Lúcia Moreira, Nirbhaya Shaji

Newsvendor problem

Summary

The benefits from an improved forecast dominates the succeeding optimization phase in the newsvendor problem [1]. We have addressed forecasting by classical times series approach using SARIMA and ETS and by a non-classical forecasting method using neural networks (NN). In this pedagogical exercise, we have observed that neural networks showed to be more powerful for forecasting than the classical time series methods herein considered. Probably owing to the fact that the single classic model chosen was not the best model for all the time series at same time,, as herein considered and as discussed in this report. But for simplification purposes, we have previously defined our scope in the classical approach to use a single classic model for all the 96 time series herein considered.

Another learning outcome from this exercise, is that the optimization step assuming a normal distribution indeed increases profit in comparison to simply using the forecast solution (a 42 % increase in profit was observed for forecasting 3840 instances ahead in a test dataset when using a classical time series forecasting approach). Discovering the structural pattern present in the data set was relevant for the classical methods forecasting while not strongly relevant for the non-classical NN approach, but if the structural information discovered is inserted in the model, NN forecasting is substantially improved.

Introduction

Retailers that offer perishable items are required to make ordering decisions for hundreds of products on a daily basis. This task is non-trivial because the risk of ordering too much or too little is associated with overstocking costs and unsatisfied customers. The well-known newsvendor model captures the essence of this trade-off [1,2,3]. Traditionally, this newsvendor problem is solved based on a demand distribution assumption. In order to set appropriate stock levels, many inventory models assume a specific demand distribution. The problem is then solved in a two-step procedure. First, the parameters of a given demand distribution are estimated, and second, an optimization problem based on this distribution is solved. However, the growing availability of large datasets and historical data may help overcome this issue and improve the performance of inventory models in real-world situations. Data that are indicative of future demand provide an opportunity to make better-informed decisions. There are usually two levels for data abstraction.

On the first level, data is exploited for demand estimation. The available data may contain information about future demand that can be extracted by suitable forecasting methods by using the historical demand data.

On the second level, several approaches can be considered [1,2]: 1- The estimate-as-solution (EAS) approach, involves first forecasting the demand, and then simply treating the point forecast as a deterministic demand value, i.e., setting the newsvendor solution equal to the forecast. 2- The previous approach, however, ignores the key insight from the newsvendor problem, namely, that we should not simply order up to the mean demand, but rather choose a level that strikes a balance between underage and overage costs using the distribution of the demand. In this case, the inventory decision is optimized based on the demand forecast by assuming a demand distribution (model-based approach). This approach is referred to as separated estimation and optimization (SEO). The main disadvantage of this approach is that it requires us to assume a particular form of the demand distribution (e.g., normal), whereas empirical demand distributions are often unknown or do not follow a regular form. 3- The empirical quantile approach is a non-parametric approach that is used for obtaining the demand distribution, so it does not assume any particular form for that. 4- Another approach uses several machine learning methods on a general optimization problem where several different explanatory features (eg. related to calendar, weather or location of stores) of the demand data are available for a data driven-decision. 5- A final approach can yet be considered that integrates the demand estimation (first level) and the optimization step (second level) into a single model by directly predicting the optimal decision from the historical data and feature data.

Approaches 3 to 5 to the newsvendor problem are usually classified as data-driven approaches as they do not rely on a distribution assumption (approach 1 is not a newsvendor problem!) [1,2]. However, it is not yet clear whether and under which circumstances data-driven approaches are preferred to model-based approaches [1]. Furthermore, the question of the conditions under which separate or integrated estimation and optimization is superior remains open [1].

Regarding the first level for data abstraction, classical parametric approaches for demand forecasting in time series include SARIMA, GARCH and Exponential smoothing (ETS) [1,2]. While non-classical time series approaches for forecasting are e.g. neural networks and decision trees [1,2].

Exploratory Data Analysis

Mapping of inventory vs sales (Fig. 1) shows that inventory meets demand needs really close (quite nice linear 45 degrees trend between inventory and sales) indicating that most inventory numbers are slightly above the main linear trend ratio what indicates that demand was historically quite well predicted and also that stock-outs are not that strongly dominant in this dataset (indeed, most of the points are in the upper part of the 45 degrees trend line).

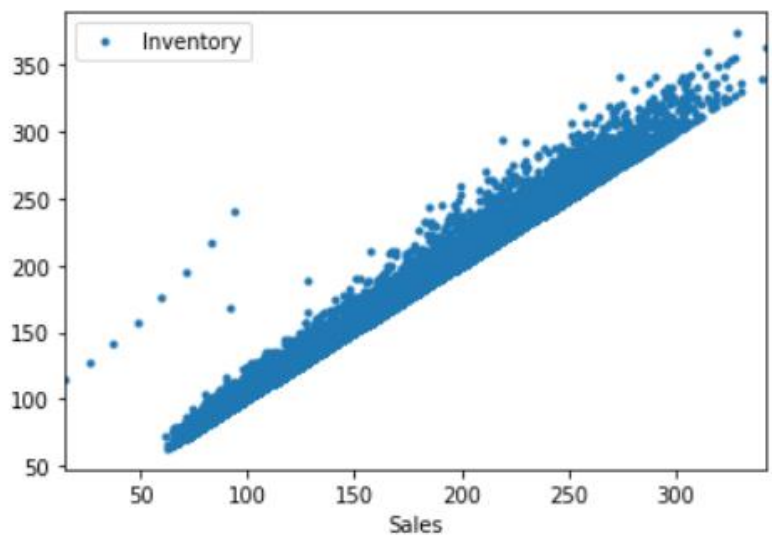

Fig.1 - Scattering plot of Inventory against Sales.

Actually, for 62 % of the instances there was no stock-out, while in the other 38 % of the cases there might have been a shortage of inventory. Moreover, in the cases where no stock-outs were observed, the inventory values are only 6 % above demand on average. Historical data shows that indeed a quite good prediction of the demand has been performed. This assumption is supported by the quite small difference (6 %) between inventory and demand (sales) when stock-out does not occur. Actual demand /Forecast (A/F) ratio near 1 (as shown in Fig. 2). Regarding the demand distribution, it will be assumed in the optimization step as a normal distribution. In Fig.2, if the Actual demand /Forecast (A/F) ratio corresponding to 1 could be split in the true A/F ratios, a distribution similar to normal distribution would be achieved in the limit.

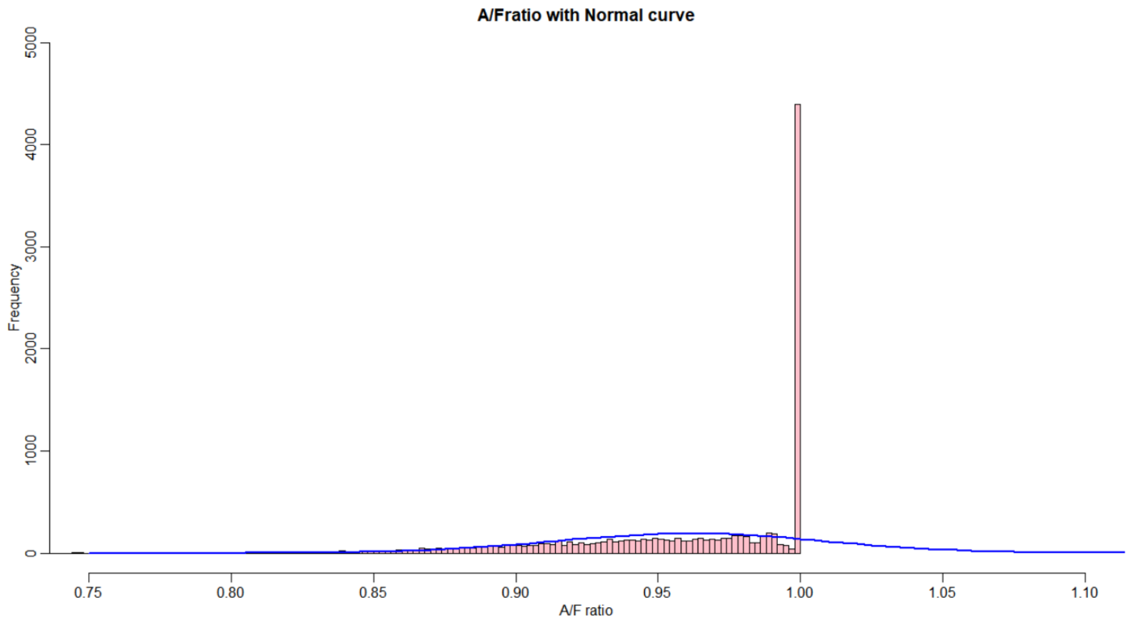

Fig. 2 - Actual demand/Forecasting ratio (A/F) distribution.

Figure 3 shows the evidence for a weak seasonality on sales/demand over the first 1344 instances. Assuming that, Fig. 4 shows the boxplot indicating that weekend days present lower sales than on weekdays and the median demand on Sunday is lower than it is for all other days.

Fig. 5 shows the time series for one representative product through the 120 days period. With a red color are intercalated the 5 first weeks, showing the evidence of a lower sales on Sundays. Fig. 6 shows the overlay of the 96 products for the first 4 days of the dataset. One can see that some products have sales that are more day-of-the-week dependent while others (eg. the first 20 products) are less seasonal dependent.

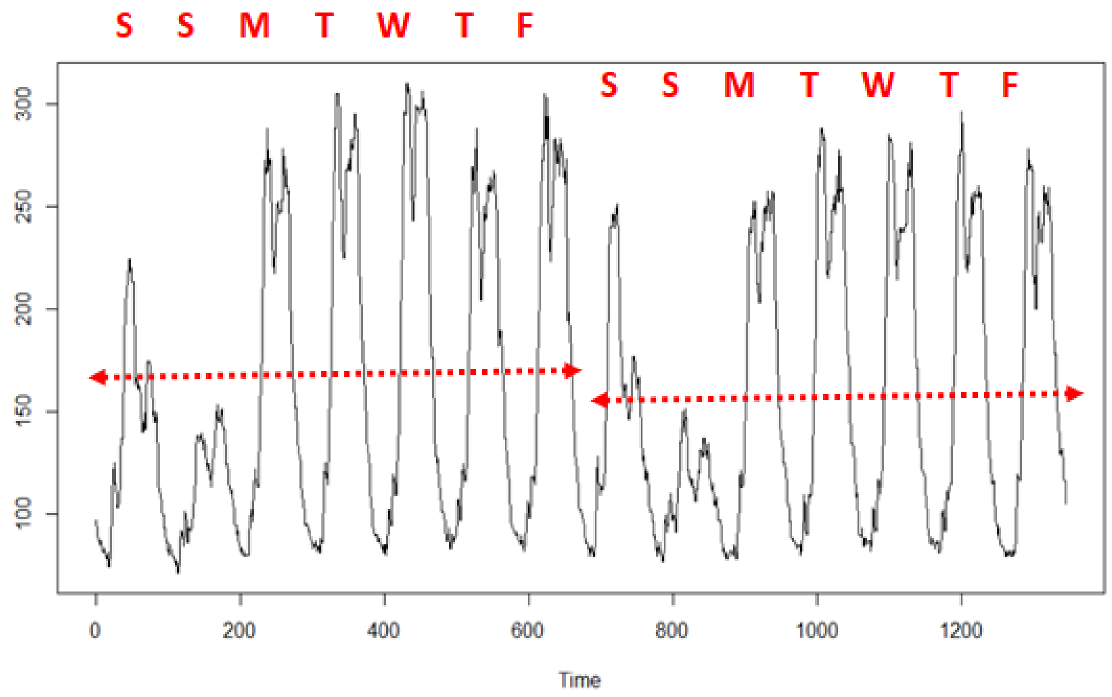

Fig. 3 - Seasonal pattern observed in the first 1344 instances of the data set.

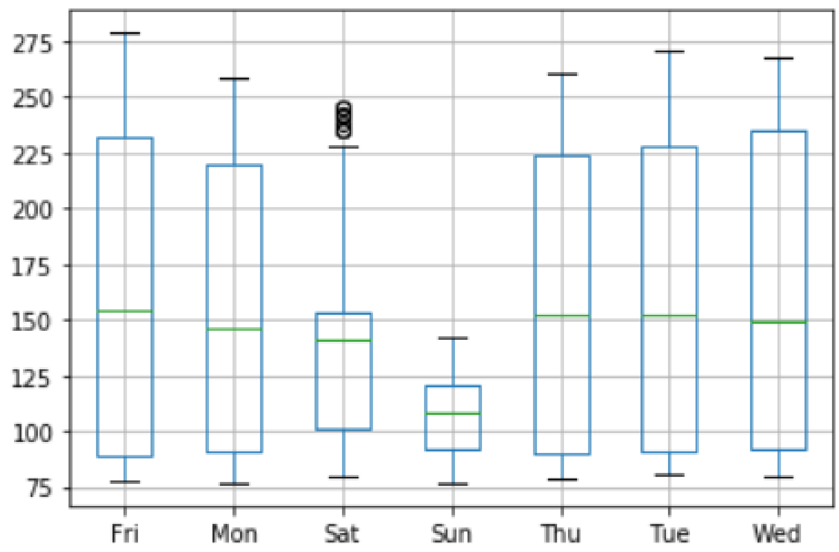

Fig. 4- Boxplot representation evidencing the averaged number of sales per product in each day of the week.

The search for structure in the available dataset was required to allow better forecastings with the classical approaches (see below). This is shown to be not highly relevant for the non-classical approach, but with the input of such structure (96 products and a weekly seasonal pattern) in the model, profit is indeed improved further, as will be shown below.

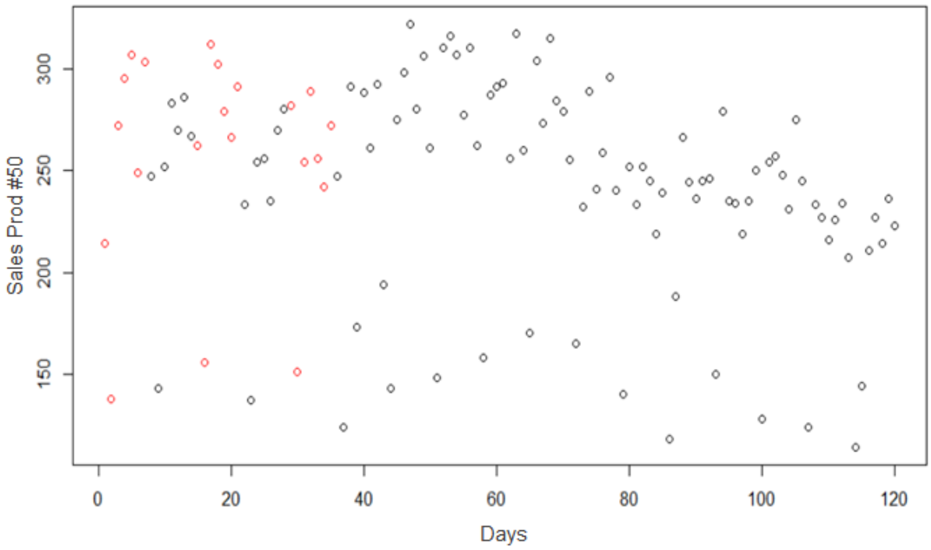

Fig. 5- Time series for one representative product evidencing a 7 days period.

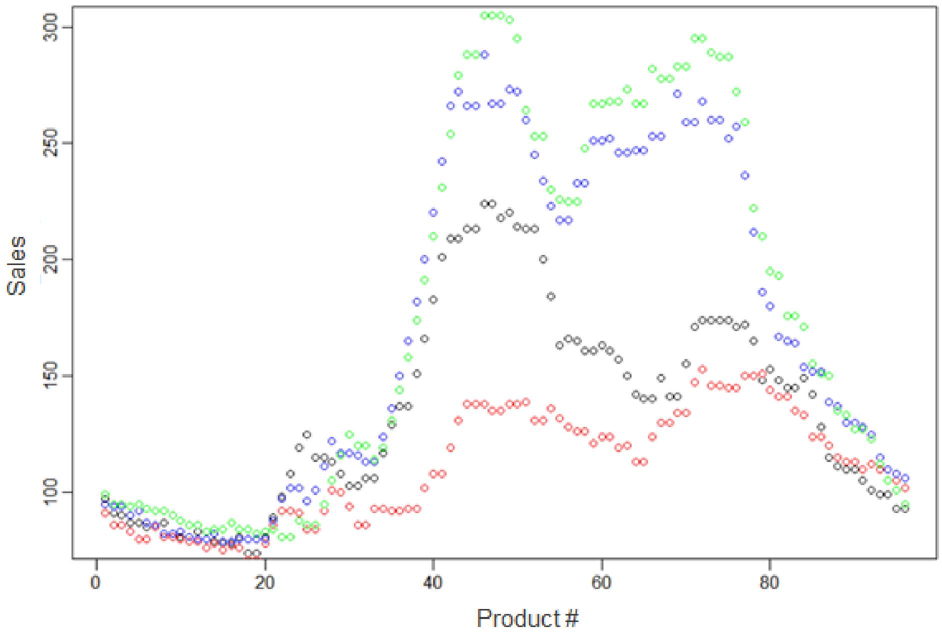

Fig. 6- Sales for the first 4 days (Saturday – Black, Sunday – Red, Monday- Blue, Tuesday – Green) of the dataset for the 96 products.

Strategy

We have used two methods for evaluating the best model for the demand forecasting task: a) two classical time series methods (SARIMA and exponential smoothing (ETS)) that deal well with seasonality, and b) a non-classical approach, Neural Networks. This task was time consuming because the benefits from an improved forecast usually dominates the succeeding optimization phase [1].

Classical Forecasting approaches

Based on the structure discovered in our EDA, a set of 96 time series was created by attributing one time series to each of the discovered products. This overview analysis was performed in R once we have stronger skills using this language for inferring the best models, based on our previous time series course. In the evidence that a classical forecast is the best forecasting method, then the most promising model directly used in the order() function with python code. The time series were divided in train and validation sets by saving the last 7 (or 40) days of the time series for each product for evaluation (in-sample forecasting).

Regarding SARIMA function, analysis of the autocorrelation function for all the 96 products showed evidence of some series with a low seasonal component pointing to a AR(1) or a AR(2) statistically significant model while some time series show a more evident seasonal component pointing to a $SAR(1)_7$ or $SAR(2)_7$ model, moreover the analysis indicated that the time series requires one differentiation (*ndiffs*() function). According to that, the choice of the best SARIMA model was performed for a short range of AR orders as well as AR seasonal orders using AIC and BIC criteria in order to confirm if the data set points to a SARIMA (2,1,0)(2,0,0)7 process. For that, modelling included auto-regressive p orders from 0 to 3 simultaneously with auto regressive seasonal p_7 orders also from 0 to 3, using sarima() function from forecast library.

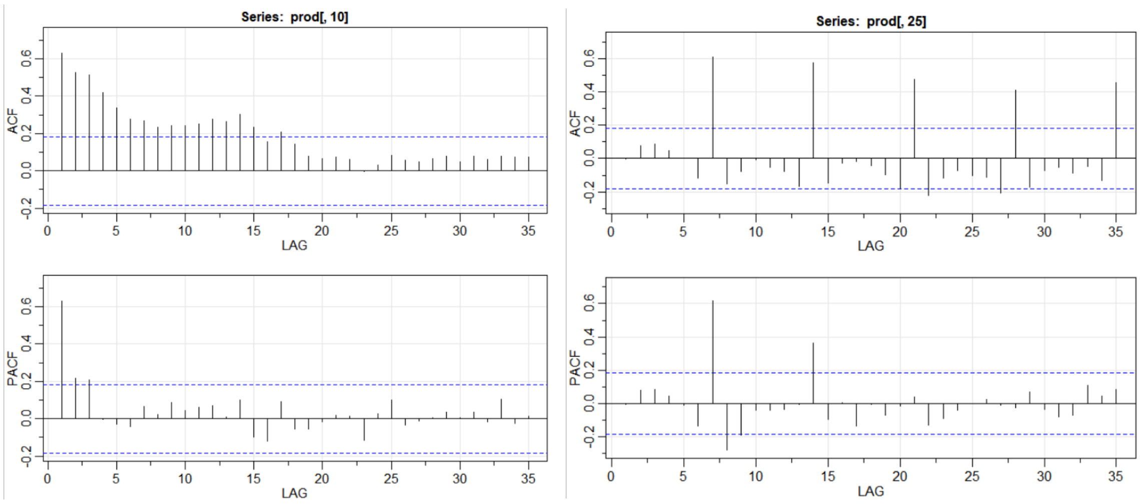

Fig 7- Autocorrelation function of two representative products: left- less seasonal time series, right-strongly seasonal time series.

Table 1- SARIMA models and averaged AIC values.

Base Model 1	(p,1,0)(p7,2,0)7			
	AIC=7.98			
Order p\p7	*0*	*1*	*2*	*3*
0	0	0	0	0
1	0	0	8	0
2	0	0	**88**	0
Base Model 2	(p,1,0)(p7,1,0)7			
	AIC=7.72			
Order p\p7	*0*	*1*	*2*	*3*
0	0	0	0	0
1	0	0	0	7
2	0	1	16	**72**
Base Model 3	(p,1,0)(p7,0,0)7			
	AIC=8.18			
Order p\p7	*0*	*1*	*2*	*3*
0	0	0	0	0
1	0	1	0	0
2	0	17	17	**61**
Base Model 4	(p,0,0)(p7,1,0)7			
	AIC=7.67			
Order p\p7	*0*	*1*	*2*	*3*
0	0	6	0	0
1	0	8	18	0
2	0	3	**61**	0

Table 1 shows the different confusion matrices obtained for the different SARIMA models and the 96 time series for the AIC criteria. AIC values correspond for the averaged AIC after fitting the highest voted (votes on bold) model on the all 96 product time series. One can see that base model 1 with the following orders (2,1,0)(2,2,0)7 was selected for 88% of the products and producing a product averaged AIC value of 7.98 (red) while base model 2 based on the orders (2,1,0)(3,1,0)7 was selected for 72% out of the 96 product time series and a averaged AIC value of 7.72. Finally, base model 4 with the orders (2,0,0)(2,1,0)7 was selected for 61 % out of the 96 product time series and an averaged AIC value of 7.67.

Based on the two lowest averaged AIC values, we have further inspected the models with the highest votes based on 2 and 4 models into more detail.

Fig. 8 shows the output of the SARIMA function for base model 2 with orders $(2,1,0)(3,1,0)_7$ and for one representative product (with a seasonal pattern). One can see that model residuals present a normal distribution and that they are not statistically correlated with each other. Moreover, all the model coefficients showed to be statistically significant. This study was performed for several products and the 2 different base models and conclusions can be considered quite similar.

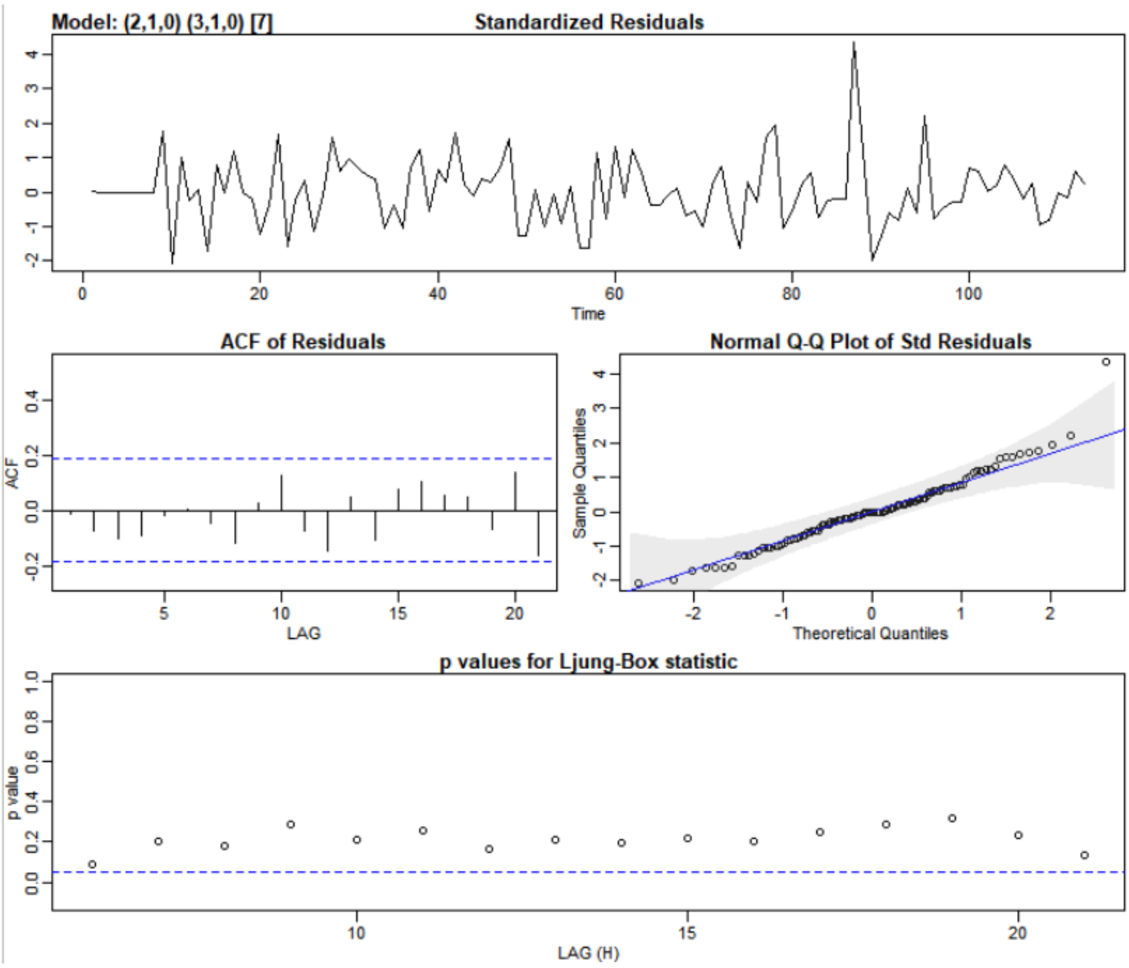

Fig. 8 - Residual analysis of one representative product from the output of model SARIMA $(2,1,0)(3,1,0)_7$

The choice of the classic model between the SARIMA $(2,1,0)(3,1,0)_7$ and $(2,0,0)(2,1,0)_7$ was based on in-sample forecasting and evaluation was performed either 1- by directly comparing with the sales of that specific day in terms of root mean squared error for all the points considered in the test sample and 2- profit obtained considering a normal distribution for the actual demand:forecasting ratio, as explained above (see Fig. 2). Fig. 9 shows as an example the forty days ahead prediction for one representative product with best base model 2. Table 2 summarizes the conclusions obtained from the classical forecasting tasks. One can see that based on RMSE and the profit, model SARIMA $(2,1,0)(3,1,0)_7$ is the best one. But one should consider that this model is indeed the best one only for ca. 72% out of the 96 time series (see Table 1). This information is relevant when comparing profit with the non-classic forecasting approach. Table 2 also shows forecasting using exponential smoothing , ets() function, but RMSE is the highest so details of the approach are not specified here in the report.

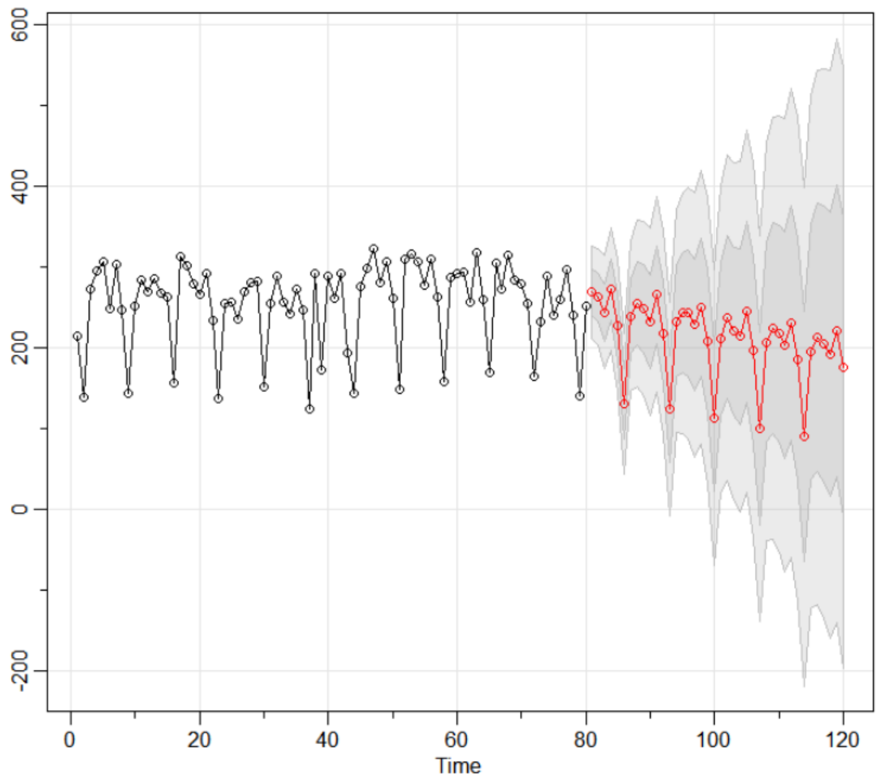

Fig. 9 - Forecasting (40 days ahead) for one representative product with SARIMA (2,1,0)(3,1,0)$_7$

Profit was calculated for 40 days or 7 days ahead using approach #1 in the optimization step and assuming a normal distribution, with the respective parameters retrieved from the A/F ratio distribution. It is considered that if no stock outs occur A/F distribution would tend to a normal distribution. In this case the optimized forecast is the forecast prediction penalized by 0.94. This penalization happens due to the quite low critical fractile ratio: overage /(underage+overage costs) = 0.1 and the lower than 1 mean (0.96) of the A/F ratio (sdved=0.048). This could be largely improved if the true A/F ratio could be observed (i.e. without stock outs). One can see that effectively model SARIMA (2,1,0)(3,1,0)$_7$ produces the highest profit for 7 days ahead. Profit obtained for 40 days ahead is 42 % higher when using the optimization procedure in comparison to only setting the newsvendor solution equal to the forecast (between parenthesis) - approach 1-, showing the advantage of the newsvendor solution to increase profit.

Table 2 - Comparison of the forecasting classical methods for 7 days (672 instances) and 40 days (3840 instances) ahead for each product.

Model	**Averaged RMSE (7 days ahead for all the products - 672 instances)**	**Profit (after 7 days)**	**Averaged RMSE (40 days ahead for all the products)**	**Profit (after 40 days - 3840 instances)**
SARIMA (2,1,0)(3,1,0)$_7$	15.9	54496	19.0	386256 (272234)*
SARIMA (2,0,0)(2,1,0)$_7$	20.0	67	-	-
ETS	27.0	-	-	-

*approach #1 – no optimization: newsvendor solution is equal to the forecast.

Non-classical approach

For the non-classical approach a neural network model for the time series prediction was considered using Keras deep learning library. The advantage is that no previous knowledge of the structure of the data is required as input, contrary to the classical approach above. It is the model itself that learns the structure from data but using the previous knowledge of the structure will be able to improve performance of the model.

First the problem is framed as a regression problem. Then the neural network model is framed using the window approach so that multiple recent time steps can be used to make the prediction for the next time step, the size of the window being the tuning parameter.

Given the current time (t), if we want to predict the value at the next time in the sequence (t + 1), we can use the current time (t) as well as the two prior times (t-1 and t-2). When phrased as a regression problem the input variables are t-2, t-1, t and the output variable is t+1 for a window size of 3.

We started with a simple network with 1 input, 1 hidden layer with 8 neurons and an output layer. The model is fit using mean squared error, which if we take the square root gives us an error score in the units of the dataset. Later after trying a few rough parameters, settled on a configuration with two hidden layers, with 12 and 8 neurons each. This resulted in a profit of 442845 and RMSE value of 7.38. A train test sample of 7718 and 3802 each was chosen for the model.

Through brute force it was seen that the optimal window size to be 3. Also an increase in the number of epochs never guareted any increase in profit. It even gave much lower profit most of the time.

Since it was seen that the profit varies heavily with the choice of the number of neurons in the hidden layers values we considered, the seasonality and the presence of multiple product findings from the classical approach, to vary the hidden layer neuron count.

The highest profit that Neural Network could bring was 458584 on 7.58 RMSE, with three hidden layers of 120, 96 and 7 neurons each in them, with an epoc value of 1. A very similar profit of 452286 on 8.18 RMSE was obtained with two hidden layers of 96 and 7 neurons each. A profit of 431894 was obtained with 120 and 7 neurons each on the two hidden layers. This could be due to the fact that deep learning offers insights into time series forecasting, such as the automatic learning of temporal dependence and the automatic handling of temporal structures like trends and seasonality. The RMSE value for all the predictions from the neural network stayed quite lower than in the classical approaches.

The profit is calculated against the initial test split of the past sales which is passed as future_demand to the order function.

Even though each prediction of neural network varies due to the different paths the deep learning approach takes for learning, as the final model we chose the one with 3 hidden layers with 120,96 and 7 neurons each considering that configuration gave the highest number of profit for majority of the tries.

Conclusions

Neural networks showed to be more powerful for forecasting than the classical time series methods herein considered. Probably owing to the fact that the single classic model chosen was not the best model for all the time series at same time, as herein considered and as discussed in this report. But for simplification purposes, we have previously defined our scope in the classical approach to use a single classic model for all the 96 time series herein considered.

Another learning outcome from this exercise, is that the optimization step assuming a normal distribution indeed increases profit in comparison to simply using the forecast solution (a 42 % increase in profit was observed for forecasting 3840 instances ahead in a test dataset when using a classical time series forecasting approach). Discovering the structural pattern present in the data set was relevant for the classical methods forecasting while not strongly relevant for the non-classical NN approach, but if the structural information discovered is inserted in the model, NN forecasting is substantially improved.

References

1- European Journal of Operational Research, Volume 278, Issue 3, 2019, Pages 904-915, https://doi.org/10.1016/j.ejor.2019.04.043

2-Applying Deep Learning to the Newsvendor Problem, https://doi.org/10.1080/24725854.2019.1632502

3- Evan L. Porteus, The Newsvendor Problem *in* Building Intuition: Insights From Basic Operations Management Models and Principles (2008). doi:10.1007/978-0-387-73699-0