Forecasting demand of medical equipment at Médecins Sans Frontières (MSF) Supply

Carlos García Meixide, Frithiof Ekström, Jonathan Koch, Kathrin Durizzo
In collaboration with: Eugénie Cochin (MSF) and Yevgeniy Ilyin (AWS)
ETH Zurich, Hack4Good, 4th Edition, 23. November 2021

Abstract

Accurate demand forecasting is crucial for humanitarian equipment suppliers, such as Médecins Sans Frontières (MSF) Supply, to optimally allocate their resources to save lives. In this project, we show that with our two models, based on ARIMA and the Gaussian Process Regression, forecasting of demand data can be improved compared to the model currently used by MSF. Additionally, we provide important insights to forecastability of items as well as patterns in the ordering behavior.

1. Introduction

Médecins Sans Frontières (MSF) Supply coordinates the logistics of over 14,000 medical and non-medical items for MSF or other international humanitarian medical NGOs all around the globe. Like all humanitarian equipment suppliers, they face two major challenges: avoiding stock shortages to ensure that missions can be provided with the equipment they need to save lives and at the same time to minimize not-needed items to allocate resources optimally and avoid expired items. Therefore, for MSF it is crucial to improve forecasting of demand for items to better allocate life-saving resources.

This project aims to contribute to this issue threefold. First, we provide more analytical insights about which items can be forecasted and on what time scale. Second, we identify patterns in ordering behavior by exploring time, country, and item trends. Third, we improve forecast demand compared to the currently used rolling average model by MSF.

2. Data

We use two datasets provided by MSF of demand data and item version management data. The demand data lists 340,785 orders of a mission on a specific day for the period January 2016 - November 2021. Information about 21 features was provided; however, for this analysis we only use 10 of them: *date, country, mission, mission details, article version, item description, article status, order quantity, order price total, and urgency*. The version management dataset consists of a list of old and new item versions since codes changed over time.

In the first step, we merged the two datasets by harmonizing the item versions in the demand data. In a second step, we defined - together with the advice of MSF - the scope of the analysis. In particular, we excluded (i) all urgent items, (ii) orders from other NGOs, (iii) items that were not ordered at all within the last 12 months, (iv) items for which the total ordered value is less than 10,000$ over the five years.

In total, we ended up with a sample of 124,661 orders for 1,039 items.

3. Methods

Various models were compared using different error metrics to find an improvement in comparison to MSF's current solution. Currently, the rolling average of the prior 6 months is used by MSF to forecast the next month. We took this as a baseline to compare other models too. In particular, we looked at models based on linear regression, random forest, XGBoost, ARIMA,

and Gaussian processes. Only the latter two provided results that notably improved upon the baseline.

We split our dataset into training and validation sets by holding out the most recent year during training. The trained models were then used to make forecasts for the next year, which we were able to then compare to our validation set. To measure the error we used several different metrics, such as the mean absolute error and the normalized root mean squared error. We believe the latter to be the most significant, as it normalizes the different order quantities for different items.

4. Results

4.1 Forecastability

Demand classification results

Demand classification is a common approach to any forecasting problem to help get a better understanding of the data. Items are assigned either of 4 categories based on their variability in demand quantity (the square of the coefficient of variability, also CV^2) and the average demand interval (ADI). The categories smooth, intermittent, erratic, and lumpy are assigned based on certain predefined thresholds. Naturally, items with smooth demand patterns should be easier to forecast than items of the other categories, especially lumpy ones. And indeed, when comparing the results of rolling average using the normalized root mean squared error, we see that the error is smallest for smooth items and largest for lumpy ones. We also see the same behavior when using other models.

Before applying the classification to our dataset, we aggregated the demand of each item into monthly and quarterly buckets. We then applied the classification to both the monthly and quarterly datasets and also looked at classifications for each individual year as well as the whole time period of our data.

Naturally, we find that with larger demand buckets (eg. quarterly rather than monthly data) we get more smooth items. This is useful information, as it tells us which items we should be able to forecast successfully on a monthly basis and which ones are more sensitive to forecast quarterly.

Further, by looking at the whole dataset rather than individual years, we get more smooth items as well. Interesting insights here are that the smooth items for individual years change over time, so the demand behavior for items is inconsistent over multiple years. Our forecasting models performed best on items that were categorized as smooth in the year prior to the forecast.

4.2 Patterns in ordering behavior

Time trends

Over the five years, MSF coordinated the supply of over 1.2 billion items included in 124,661 orders for a total price of over 114 million $. Based on a linear regression with years and months, we provide evidence that in December the number of orders and the value of the order was significantly smaller compared to January. This confirms the assumption of MSF that in December there is a lower demand potentially due to final year budget behavior as well as holidays. On the contrary, we find for January and November an increased demand.

Country trends

Our sample consists of orders from 73 different countries globally. Since MSF assumes that some countries only have certain supply windows to import equipment, we explored country-specific trends in the demand data. We could detect several countries where we expect import windows for certain months. However, these results should be confirmed with country managers of MSF to fully understand the trends.

Item trends

Trends in item orders were explored by analyzing correlations. The correlation between items was studied by first aggregating the demand for each item on a monthly basis, removing outliers, and then finding the

correlation coefficient for the demand of each combination of items. Outliers were removed to prevent single large orders from blowing up the correlation coefficients.

We found 157 item combinations with a correlation greater than 0.7 and 37 combinations with a correlation greater than 0.8.

This information can be used to infer the anticipated demand of one item given predictions of another, strongly correlated item, which can potentially increase the accuracy of predictions for items with limited historical data.

No item combinations with strong negative correlations were found. Thus, no cannibalization patterns could be identified from the correlation study.

4.3 Forecasting of demand

ARIMA

The autoregressive integrated moving average (ARIMA) is a useful model to analyze and forecast time series data. The AR part of ARIMA indicates that the evolving variable of interest is regressed on its own lagged (i.e., prior) values. The MA part indicates that the regression error is actually a linear combination of error terms whose values occurred contemporaneously and at various times in the past. The I (for "integrated") indicates that the data values have been replaced with the difference between their values and the previous values (and this differencing process may have been performed more than once). The purpose of each of these features is to make the model fit the data as well as possible.

There are a few things that need to be considered in order to understand the results of using this model:

- We applied a basic ARIMA model without much fine-tuning to our dataset.
- As mentioned before, we passed 4 years of training data to directly forecast the whole next year. However, since this is a model based on the rolling average, we believe that its results would improve notably if the actual data of the prior months were always available. In that case, we would only always forecast the next month. This is actually what we did in our baseline model with the rolling average.
- We applied the model to each item on a monthly basis and did not take any information gathered from our classification into account. Forecasting certain items on a quarterly rather than a monthly basis might make a lot more sense in terms of accuracy.

Despite these caveats and potential improvements, we were still able to get some interesting results.

ARIMA performed best on 60% of all items whereas the rolling average performed best on only 27% of items. For the items that were better forecasted using ARIMA, the model decreased the normalized root mean squared error by an average of 33% per item compared to the rolling average.

An obvious next question would be to look at what differentiates the items that are better forecasted using ARIMA from the ones that work better with the rolling average or other models. Another point to consider is that the model was evaluated only on the cleaned-up dataset and we do not yet know about its performance on the whole dataset including sparse items. Still, especially considering these results were achieved with very limited time and effort, it is apparent that using predictive models can notably increase the accuracy and efficiency of forecasts.

Gaussian Process framework

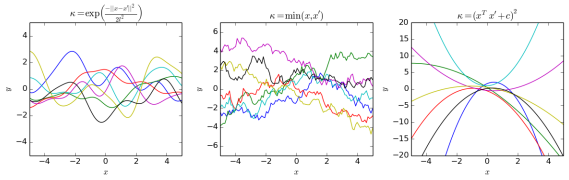
Gaussian Process Regression is a Bayesian regression approach formally based on assigning prior probabilities directly to functional shapes that the model can take. This provides us with the first improvement with respect to parametric modeling, in which it is parameters themselves that are assigned a prior distribution.

In this sense, the big step is to assume that the prior of the conditional expectation of the response variable given the covariates (regression function) is a stochastic process with a fancy property: every finite collection of

random variables within the process follows a multivariate gaussian distribution.

This makes Gaussian Processes a powerful construction that allows one to keep an eye only on a finite number of evaluation points. Inference will tell us the same as if we looked at the corresponding infinite-dimensional objects. Other strong theoretical properties include being the posterior distribution a Gaussian Process again, among many others.

While the clever step of Gaussian Processes is this finite-wise modeling, setting up how such sets of function values are correlated completely determines the shape of a drawn sample function. In this sense, the shape of a Gaussian Process is completely determined by a mathematical operator called kernel that stands for correlation between functional evaluations.



Choosing a kernel rigorously in a general context remains to be an open problem today although it can be handmade for each specific data science setup.

In our case, time series forecasting is a very particular regression setting in which predictions are desired to be launched far away from the temporal region in which the observations in the dataset lie. Therefore, traditional forecasting settings tend to lose expressivity due to epistemic uncertainty (lack of observations) when predictions about the future are launched.

With this in mind, it makes sense to set functions that reflect periodicity as the kernel of our Gaussian Process. Fourier analysis is the key to automatically choose the relevant “frequencies” in our data in order to build a robust kernel choice methodology for our setup.

As preliminary results, for the particular case of a randomly chosen item (more concretely, glucometer strips) a more efficient allocation of 72 euros per order placed to MSF would have been possible by using our strategy. Taking into account all the items present in the dataset, around 21 million euros could have been kept track of along a 2 year period, approximately. Consequently, by using our method, suboptimal management of a considerable amount of stock would be avoided. Very interesting future extensions may involve Bayesian deep learning to accurately model time series peaks.

5. Conclusion & Recommendation

Recommendations & Conclusion

This project clearly showed that using predictive models beyond the rolling average can have a notable positive impact on the accuracy of demand forecasts. The models used are still very basic and can be improved by fine-tuning parameters, refining loss functions, and selectively applying them to certain items.

In this regard, the classification part of the project should also be of value to help understand better which items to forecast on what timescale.

Difficulties, Limitations & Risks

There are some obvious challenges with any forecasting problem: Much of the dataset is rather sparse with too few orders to make any meaningful forecasts. We did clean our dataset at the beginning of the project so as not to take these into account. This improves our ability to make sensible forecasts but it also means that a large number of items were not considered in this project.

Furthermore, this sort of forecasting only takes into account historical demand data. This means that any events that lead to a change in demand that are not naturally part of the historical dataset cannot be considered and may render forecasts useless. This could be disease outbreaks or any other unforeseen events.