Streamlining Inventory Management for Storable Healthcare Products in a Unified Demand Environment

Nathaniel Mitrani, Alex Serrano, Jan Tarrats, Joel Solé
Polytechnic University of Catalonia

November 12, 2023

Abstract

In this project, we propose a mathematical model and implementation based on a collaborative scheme designed to optimize the storage and distribution of medical products to hospitals given historical data.

1. Introduction

Supply chain robustness is of paramount importance for businesses, as it not only ensures operational continuity but also provides significant financial benefits. Especially in medical settings, robustness is imperial as the effects of a shortage can cost human lives [3]. Additionally, it is of paramount importance for a company to minimize the ecological footprint stemming from its activities, especially related to transportation which is a major source of pollution, 28% as of 2021 [1].

2. Problem Statement

The goal is therefore to find cost-optimality whilst satisfying robustness and environmental constraints. We suggest a unified demand approach to tackle the problem, that is an agreement reached by an array of medical institutions to combine the demand for medical products in one single order to reduce cost [2], followed by a mathematical model and further implementation to solve the problem in this setting. For simplicity, we consider a time granularity of months.

3. Proposed solution

In the unified demand scenario, we assume that all the hospitals group their orders for a given product in one order, with the referenced economical benefits that this supposes [2]. We consider therefore each product separately and build a model to optimize the processes for each product.

Given a specific product, we consider a unique provider and distribution center (as we are in a given region we can assume a certain locality). Our challenge is to optimize the costs given a certain environmental footprint and resilience score.

We quantify the environmental impact as proportional to the number of orders, as referenced in [1] and in the problem specification. This amounts to choosing when and how much to order given robust satisfaction of demand, fixed number of orders, and storage costs.

The demand we obtain comes from a prediction for the purchase plan, from which we yield the amount of units needed for the coming year.

4. Modelling product demand

Our goal here is to establish the optimal offer (amount of units) for a given product throughout 2023, as a byproduct of creating a demand predictor that uses historical data.

4..1 Model evaluation

First, we establish the metrics that are going to be used to evaluate, compare and select models.

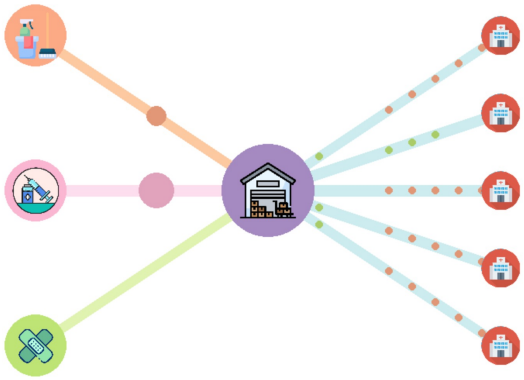

Figure 1: Image showing concatenation of 3 models (3 products), assuming that the distribution centers for each product are in the same location (graphical purposes)

4..1.1 Tweedie loss

The Tweedie loss, as presented in Zhou, He et al. [5], provides a way to measure goodness of fit of a forecasting model. It is robust to skewed distributions, as is the case in the dataset. The metric range is $(-\infty, 1]$, where $-\infty$ is a completely random model, 0 is a model that always predicts the empirical mean of a times series and 1 is an exact prediction.

4..1.2 Expense MAPE

To allow the less-technical users of the model to be able to interpret the error with its real-life repercussions, we report the mean absolute percent error in predicted expenses for 2023 by product. That is, the mean of the percentual absolute error between the real expenses and the predicted expenses for 2023.

4..2 Models considered

To allow for a diverse set of models, we have considered several families of predictors, specific forecasting models (Prophet, [6]), general regression models (GBMs, regularized polynomial regressions, Generalized Additive Models) and deep learning models (Seq2Seq [**?**], Temporal Fusion Transformers [4]). In the following, we report the best two results.

4..2.1 Boltzmann ensemble of GBMs

A Boltzmann ensemble is an ensemble that uses a Boltzmann distribution to weight the predictions of the models in the ensemble. The Boltzmann distribution is a probability distribution that assigns a probability to each model in the ensemble, and the probability of a model is proportional to its performance on the validation set. The Boltzmann distribution is defined as follows:

$$P(m) = \frac{e^{-\frac{1}{T}\cdot \text{loss}(m)}}{\sum_{m' \in M} e^{-\frac{1}{T}\cdot \text{loss}(m')}}$$

Thus, the probability of a model is inversely proportional to its loss, and the temperature T controls the variance of the distribution. The lower the temperature, the more the distribution is concentrated around the best model. We empirically choose a temperature of 0.1, which is a good compromise between variance and bias.

In terms of the individual models, we use a set of gradient boosted trees (XGBoost, CatBoost, HistGradientBoostedRegressors).

The results are shown in Table 1.

4..2.2 Temporal Fusion Transformers

The Temporal Fusion Transformer [4] is a novel attention-based architecture that combines high-performance multi-horizon forecasting with interpretable insights into temporal dynamics. To learn temporal relationships at different scales, TFT uses recurrent layers (LSTMs) for local processing and interpretable self-attention layers for long-term dependencies. TFT utilizes specialized components to select relevant features (variable selection networks) and a series of gating layers (with gated linear units) to suppress unnecessary components, enabling high performance in a wide range of scenarios.

We have employed our own implementation of the paper in PyTorch, and we have used the same hyperparameters as in the paper, that demonstrate a 36-69% improvement over other deep forecasting learners such as DeepAR and

NHiTS in standard benchmarks (see paper). The model has been trained in 1,5 GPU hours in an NVIDIA P100.

See the summary of the results in Table 1.

Model	Tweedie	Expense MAPE
Ensemble	0.70894	0.07146
TFT	0.16871	0.2293

Table 1: Metric results for presented models

5. Mathematical formulation of the model

5..1 Model description

5..1.1 Sets of indices

- $I = \{\text{index set for the product}\}, i \in I$

5..1.2 Constants

- c_i: monthly storage cost for a unit of product i.
- C^i_{max}: maximum quantity of product i that we can store.

5..1.3 Computed parameters

- $v^i(t)$: consumption velocity of product i at time t
- $\xi^i(t)$: Demand checkpoint of unified demand at time t.

5..1.4 Parameters of optimization

- β: Resilience factor, factor by which we multiply demand to increase supply chain resilience.
- P_{max} : Number of orders $\propto CO_2$ emissions, a proxy for environmental impact.

5..1.5 Variables

- $p^i(t)$: quantity of product i demanded at time $t \in \{1, \ldots, 12\}$
- $\delta(t)$: boolean (binary) variable to determine if there is an order at time $t \in \{1, \ldots, 12\}$

5..1.6 Auxiliary variables

Remark 1 *For simplicity and interpretability, we define this adjacent variable $s^i(t)$ corresponding to storage of product i at the beginning of the time unit t.*

- $s^i(t) = -\sum_{t'=1}^{t-1} v^i(t') + \sum_{t'=1}^{t} \delta(t') \cdot p^i(t')$

5..1.7 Constraints

Capacity constraint:

$$s^i(t) \leq C^i_{max} \quad \forall t = 1, \ldots, 12.$$

Sufficient inventory constraint:

$$s^i(t) + \sum_{t'=1}^{t-1} v^i(t') \geq \beta \cdot \sum_{t'=1}^{t} \xi^i(t') \quad \forall t = 1, \ldots, 12.$$

Environmental constraint (P_{max} orders):

$$\sum_{t=1}^{12} \delta(t) = P_{max}.$$

Variable domain constraint:

$$s^i(t) \geq 0 \quad \forall t = 1, \ldots, 12$$
$$p^i(t) \geq 0 \quad \forall t = 1, \ldots, 12$$
$$\delta(t) \in \{0, 1\} \quad \forall t = 1, \ldots, 12$$

5..1.8 Objective function

$f_i(\mathbf{p}) = \sum_t s^i(t) \cdot c_i$

5..2 Computation of parameters

5..2.1 Computing $\xi_i(t)$

$\xi_i(t)$ is computed by cumulating all the purchased quantities of the product i by all the hospitals predicted by our model in time step t.

5..2.2 Computing $v_i(t)$

$v_i(t) = \frac{\xi_i(t_-)}{t_+ - t_-}$, where $t_- = \max\{t' : 0 \leq t' \leq t, \xi_i(t') \neq 0\}$, and $t_+ = \min\{t' : t < t' \leq 12, \xi_i(t') \neq 0\}$, i.e. $v_i(t)$ represents the slope between the previous purchase and the next, therefore supposing uniform consumption between orders, and complete depletion from one order to the next (i.e. the demand is perfectly predicted).

6. Example: product 70130

We have used different values for β and P_{max} to observe the effects of different environmental and robustness restrictions on the optimal cost of storage. As expected, the more robust and the fewer orders allowed (i.e. the less environmental impact) lead to increased optimal costs. We also observe that it is significantly harder to have a lesser environmental impact than to be more robust.

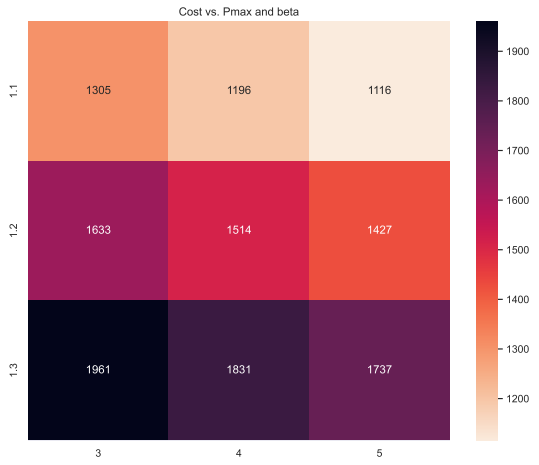

Figure 2: Heatmap of optimal costs in terms of β and P_{max}.

7. Conclusions and future work

In conclusion, our collaborative scheme and mathematical model offer a promising approach to optimize the storage and distribution of medical products in a unified demand scenario. The consideration of environmental impact and resilience scores in the optimization process aligns with the growing emphasis on sustainability and supply chain robustness. However, there are several avenues for improvement and future research that can enhance the effectiveness of our proposed solution.

One key area for improvement lies in refining the deep learning model used for predicting product demand. Our current approach relies on a predictive model for the purchase plan, which determines the quantity of units needed for the upcoming year. By incorporating more advanced deep learning techniques, such as recurrent neural networks or attention mechanisms, we can potentially enhance the accuracy of our demand predictions. This, in turn, would lead to more precise optimization of storage and distribution processes.

Furthermore, a deeper exploration into storage optimization processes could yield additional insights. Investigating advanced storage strategies, such as dynamic allocation algorithms or real-time inventory management, could contribute to further reducing costs and improving overall efficiency. Additionally, considering dynamic changes in demand patterns and adjusting storage strategies accordingly could enhance our model's adaptability to evolving scenarios.

In summary, future work should focus on enhancing the accuracy of demand predictions through advanced deep learning techniques and exploring advanced storage optimization strategies. By addressing these aspects, we can further advance the effectiveness of our proposed solution and contribute to the ongoing efforts in creating resilient, environmentally conscious supply chains for medical products.

References

[1] Sources of greenhouse gas emissions | US EPA. (n.d.). `https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions`

[2] L'ICS ESTALVIA 30,5 milions d'euros fent Compres Agregades. Institut Català de la Salut. (n.d.). `https://ics.gencat.cat/ca/detall/noticia/compres-agregades.html`

[3] Phuong JM, Penm J, Chaar B, Oldfield LD, Moles R. The impacts of medication shortages on patient outcomes: A scoping review. PLoS One. 2019 May 3;14(5):e0215837. doi: 10.1371/journal.pone.0215837. PMID: 31050671; PMCID: PMC6499468.

[4] Lim, Bryan & Arık, Sercan & Lo-

eff, Nicolas & Pfister, Tomas. (2021). Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting. 37. 10.1016/j.ijforecast.2021.03.012.

[5] Zhou, He, et al. Tweedie Gradient Boosting for Extremely Unbalanced Zero-Inflated Data. arXiv, 14 Nov. 2019.

[6] Sean J. Taylor, Benjamin Letham (2018) Forecasting at scale. The American Statistician 72(1):37-45 (https://peerj.com/preprints/3190.pdf)