Optimization of Inventory Allocation

Introduction

In 2016, e-commerce sales totaled an estimated $394.9 billion, accounting for 8.1 percent of total annual sales. This total was a 15 percent increase from 2015. Advances in technology and adoption of the internet has forced the retail industry to make dramatic shifts toward e-commerce. While this presents a tremendous opportunity for business growth, the cost associated with inefficiencies in supply chains makes optimally allocating inventory to fulfillment centers integral to retailers' success. Customer retention and loyalty requires the business to deliver products quickly and efficiently, so inventory allocation to fulfillment centers must account for two primary factors:

- item cost of delivering units from fulfillment centers to delivery addresses
- length of time required to fulfill orders to delivery addresses from fulfillment centers

While the first of these costs can be associated with a particular dollar value, the cost related to the second consideration is driven by customer disatisfaction with delayed product arrivals and is more difficult to quantify. Additionally, there are a host of other factors to be considered when optimizing inventory allocation. These include the cost of transferring units between fulfillment centers to respond to shifts in demand, as well as decisions about whether to fulfill online orders from stores instead of dedicated fulfillment centers. In the following sections, we propose a method to determine the proper allocation of inventory to fulfillment centers. Our method assumes that no transfers are made between fulfillment centers and that e-commerce orders are fulfilled only from dedicated fulfillment centers and no orders will be fulfilled by stores, however, potential future work could be explored by relaxing these assumptions.

[Typical Solution] How they work:

There are several products on the market for supply chain optimization. Many of those products are outfitted with the functionality for determining efficient allocation of inventory to distribution centers. We can decompose the common approach to optimizing inventory allocation into two primary tasks: predicting future inventory demand and then given this forecast, develop a plan to purchase and distribute products so as to meet the demands at each fulfillment center in the most cost efficient way possible.

Demand predictions are made by fitting statistical models using past demand as well as predicted future events to forecast future demand. Often, these forecasts are done using data that has been aggregated to a fairly high level, such as by week, subpopulation of customers, category of merchandise, etc, though new technical ability to store and access large amounts of data quickly allows demand forecasting to be done at a far more granular level, such as at the daily level, per item per customer. Safety stock is the extra stock kept on hand to mitigate rish of inventory depletion due to unpredictability in demand.

The *service level* refers to the probability of the period-specific demand quantity exceeding the inventory quantity on hand:

$$\alpha = Prob\left(\text{period demand} \leq \text{available inventory stock}\right).$$

Setting safety stock levels appropriately is critical to minimizing costs. Typically, the amount of safety stock is chosen so that the probability of not meeting customer demand is low, based on the predicted sales. Loss is incurred due to incorrect forecasts by when

- demand is overestimated, leading to excess inventory which will be wasted or discounted, or when
- demand is underestimated, resulting in missed sales opportunities and unsatisfactory customer experience.

Once they specify the service level and forecast sales, retailers determine the safety stock levels and the corresponding product *buy b* - the total amount of product to purchase for order fulfillment. Determining how this inventory is distributed is then a deterministic task - there is little to no uncertainty associated with which fulfillment center is most optimal for shipping a given order. The center that will fulfill each online order is determined using a set of rules which specify the logic necessary for minimizing

- the cost of shipping orders from fulfillment centers to delivery locations and
- the length of time required for products to be delivered to the customer.

Once this set of rules for determining fulfillment center territories, or which order locations correspond to which fulfillment center, is specified, then the task of allocating inventory is simply a matter of executing some straighforward bookkeeping. We do not seek to optimize this set of rules. Much of the difficulty in allocating the optimal amount of inventory to each fulfillment center lies in properly accounting for the uncertainty in the forecasted demand across each of the fulfillment center territories.

Typical Solutions: Why they suck

Despite the numerous existing applications offering solutions to the allocation problem, inventory allocation inefficiencies are still quite prevelant across retail supply chains. The difficulty in accounting for demand forecast uncertainty is the root of the underperformance of many tools on the market. Many of these tools overly simplify this uncertainty in various ways. Many rely on rigid assumptions about the distribution of future product demand for a given fulfillment center. For example, many universally accepted methods assume that safety stock is proportional to the standard deviation of the demand, which is assumed to follow a Normal distribution. It is also common in the existing market for software solutions to assume that the demand at a given fulfillment center is independent from week to week, i.e. that the demand in the current week contains useful information for predicting demand for the week to follow. In most practical situations, this assumption is invalid due to the innate dependency between time and demand. The classic example is the trendsetter: she purchases an item and after wearing it, the crowd follows suit and demand for a given product increases until it reaches a certain point at which the market is saturated. Invalid assumptions about the variability in demand inhibit the ability to adequately quantify the uncertainty with predicted demand. Characterizing this variability faithfully is integral to specifying the initial product buy and allocation to each fulfillment center.

Even when all the assumptions of a forecasting model can be verified to be reasonable, there is still uncertainty associated with both future predicted demand as well as the model we use to make these predictions. Most existing inventory allocation tools determine the best initial allocation of inventory to fulfillment centers by assuming that this prediction error doesn't exist. Unfortunately, the allocation quantities inherit the uncertainty in the demand forecast used to determine these quantities. Failure to account for the uncertainty can lead to suboptimal allocation of inventory, leading to loss incurred due to fulfillment centers with overstock or inability to fulfill customer orders where the demanded exceed projections.

In the section to follow, we will provide a simple illustration of the uncertainty associated with an inventory allocation specification. We propose adjusting for this uncertainty by employing a Bayesian approach to statistical modeling which allows retailers to consider all of the *what-ifs* when forecasting sales and allocating the corresponding appropriate inventory quantities.

Solution: skinny in 3 weeks!

Forecasting is not a new concept in ecommerce. Retailers have been using statistical models to forecast sales for years. Howevever, to our knowledge, most of the products which support optimal inventory allocation calculation are leveraging *Bayesian* methods: a collection of statistical tools which allow people to update their beliefs as they observe more data. Consider the following example: suppose a retailer is launching an email promotion for a particular product, and they believe that the campaign will result in a 40% conversion

rate - that is, for every 100 customers who receive the promotional ad, 40 customers will purchase the promoted product through the email link. They launch the campaign, and upon completion,

- 100,000 customers recieved a promotional email, and
- 10,000 customers made a purchase.

The retailer plans to launch a similar campaign for an analogous product. If you were asked to predict the coversion rate for the new campaign, Would you predict the same 40% conversion rate? Perhaps the Bayesian methods us to incorporate *prior belief* about products and the market as well as update this belief as we observe actual sales.

Let's shift our focus back to the problem of inventory allocation. To forecast online sales, we start with a set of households. Together, these households constitute the entire product market. While we could forecast purchases for each household in the market, forecasts are typically built on an aggregated level. For example, we might predict sales within political boundaries (*e.g.,* county or state) or within designated marketing areas. These aggregated geographic areas *partition* the market: each household belongs to exactly one of the specified regions, and together, the regions cover the entire market area. The first component of optimizing inventory allocation is to determine how much total inventory to buy. The value associated with a particular allocation scheme relies on

- the total demand,
- the total buy, and
- the allocation of the buy.

Denote the *true demand* during the pre-specified time period d^*. Retailers mark down prices to accelerate the rate of sales of inventory that hasn't been sold after a prespecified amount of time has passed. When the demand does not meet the buy, i.e. $d^* < b$, there are $b - d^*$ excess units to be sold. These units are sold at a price resulting in a lower AUR than during the planned sales period.

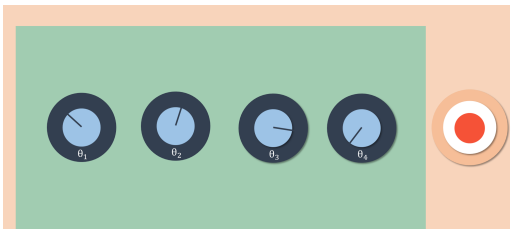
We assume that the AUR for excess units can be expressed using some function m. A typical choice for a function describing the relationship between how much overstock was purchased and the AUR specified for excess units is

$$m(b, d^*) = r \exp\{-\delta(b - d^*)\},$$

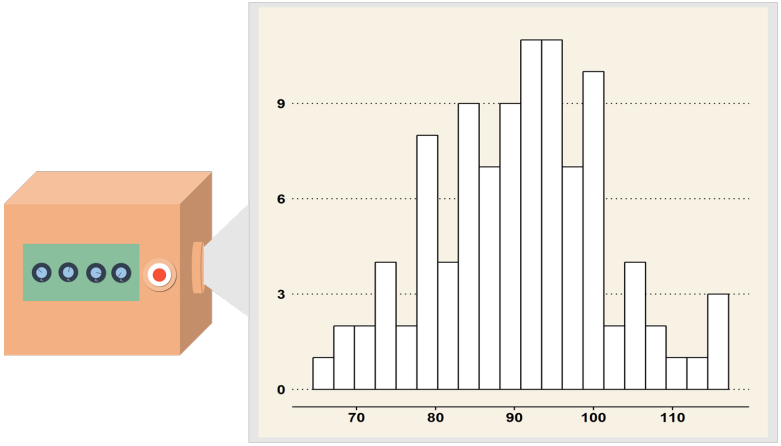
which is displayed in Figure~**??**. r is the AUR during the planned sales period, δ is a parameter controlling how much discounting is required to sell excess inventory.

The total revenue across all units sold is then

Imagine that I have a magic box with a magic button and a set of dials. When the button is pressed, the box generates values of the total sales of a certain product in each of K different regions for a specified period of time. The sales output of the box, though, is never the same The dials allow you to specify the "state of the market" in each region during the sales period. These dials control the specific relationship between influential factors, or *independent variables* and total sales of a particular product. For example, the box might have four dials which control how each of the seasons affect sales in each region. In certain southern regions, we might set these dials so that winter has a strong, positive impact on sales if, for instance, there is an influx of tourism during the winter months that leads to a bump in sales. In northern regions, winter weather may dampen retail activity, so we may set the dial to generate sales given that we expect, say, a 30% decrease during winter months. Imagine that on the box, there are also dials for specifying the effect of all the other relevant independent variables such as population demographics and marketing spend, as well as dials specifying the relationship between past sales and future sales.



In the context of statistical modeling, the dials control the values of the *model parameters*, $\theta_1, \theta_2, \ldots, \theta_p$. We denote the collection of parameters by $\boldsymbol{\theta}$. In a perfect world, we would simply adjust the dials to the correct settings - the settings which reflect the *true* nature of the relationship between each of the independent variables and sales, denoted $\mathbf{Y}$. The black box, however, is not a simple deterministic machine: holding the dials in the same place and pushing the magic button repeatedly would not result in the output of the same number again and again. This is due to the fact that we can never be sure that we have included a dial on the box for every predictor variable that impacts sales. Additionally, we are limited in our precision with which we can set the dials - we can never tune the dial to the *exact* true value, even if we know what that value is. Instead, if we repeatedly press the button, the box will generate a set of values for sales. The box will output sales values with higher frequency the more likely they are to occur given the conditions specified by the dials, giving us the ability to examine the *distribution* of sales. Figure~**??** shows what the result of sampling 100 values from the distribution of sales might look like.



If we knew exactly where to tune the dials in order to output values of future sales under the correct market conditions, then for each of the regions partitioning the product market, we could simply press the magic button many times and obtain a sample of future sales values we might observe for each day of the sales period. Denote the future sales for region r on day t by Y^*_{rt}. Then for a sales period lasting T total days,

$$\text{total predicted demand for region } r = \sum_{t=1}^{T} Y^*_{rt}$$

For simplicity of modeling, we assume that forecasting for the SKU of interest has been performed taking a Bayesian approach, so the retailer has posterior distributions for each of the parameters in the forecasting model. Denote the historical demand data from which the forecasting model was constructed $\mathbf{Y}$ and the historical independent variables used to predict demand $\mathbf{X}$. Denote the set of parameters in the model $\boldsymbol{\theta}$. Using a Bayesian approach, the forecasting model has a posterior distribution for $\boldsymbol{\theta}$,

$$\pi(\boldsymbol{\theta} \mid \mathbf{Y}, \mathbf{X}) \propto \pi(\mathbf{Y} \mid \boldsymbol{\theta}, \mathbf{X})\pi(\boldsymbol{\theta}).$$

For the purpose of forecasting future sales, we assume that future values of the independent variables are

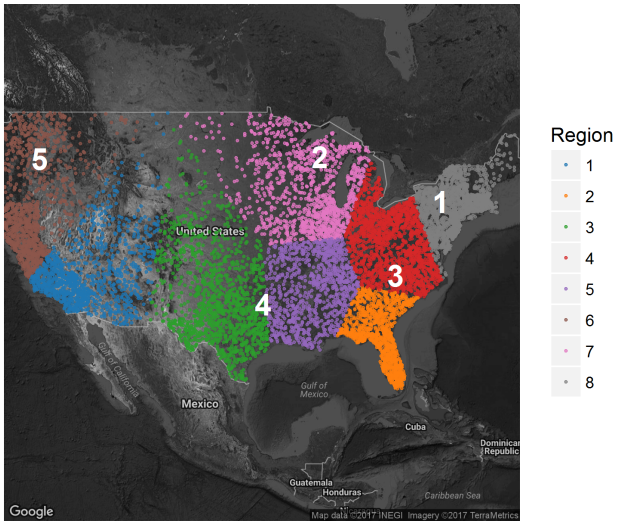
known. We denote those future values $\mathbf{X}^*$. Similarly, we denote future values of demand $\mathbf{Y}^*$. The predictive distribution of $\mathbf{Y}^*$ is

$$\pi(\mathbf{Y}^* \mid \mathbf{X}^*, \mathbf{Y}, \mathbf{X}) = \int_{\boldsymbol{\Theta}} \pi(\mathbf{Y}^* \mid \boldsymbol{\theta}, \mathbf{X}^*)\pi(\boldsymbol{\theta} \mid \mathbf{Y}, \mathbf{X})d\boldsymbol{\theta}$$

.

We assume that forecasting is performed over a finite set of time points, $1, \ldots, T$ within each region. The forecast in region k at time t is denoted Y^*_{kt}. The total demand across all time points and regions is denoted D^*, and

$$D^* = \sum_{t=1}^{T} \sum_{k=1}^{K} Y^*_{kt}.$$



After

- Output -> Outcome -> minimize loss -> increase ROI