Integration of Machine Learning and Optimization Models for a Data-Driven Newsvendor Problem with Random Yield

Bijan BIBAK
Department of Industrial Engineering, Koç University, Sariyer, Istanbul, Turkey
bbibak20@ku.edu.tr
Fikri KARAESMEN
Department of Industrial Engineering, Koç University], Sariyer, Istanbul, Turkey
fkaraesmen@ku.edu.tr

We investigate a data-driven lot-sizing problem within the random yield context, inspired by semiconductor manufacturing. In this setting, the yield rate of a production process is influenced by numerous observable features prior to making lot-sizing decisions. Additionally, demand variability is considered, which may also depend on various features. Addressing the data-driven lot-sizing problem is complex due to the reliance on a vast array of features with limited data. To navigate these complexities, we introduce methods that blend machine learning techniques with stochastic optimization. Leveraging both a publicly accessible semiconductor yield data set and a crafted synthetic data set, we assess various estimation and optimization strategies. Our findings reveal the substantial benefits of incorporating feature information for reducing costs. Moreover, the best approach for tackling the problem merges machine-learning based estimation techniques with the theoretical optimization principles specific to random yield inventory problems.

Key words: Inventory Control: Lot-sizing; Yield Uncertainty; Data-Driven Optimization; Machine Learning

1. Introduction

In many sectors, firms encounter substantial planning difficulties due to yield uncertainty, where the actual quantity produced or replenished may deviate from what was initially planned. This uncertainty in yield adds a layer of complexity to production and replenishment processes, making effective planning challenging. Historically, the research community in operations research has dedicated efforts to understanding and addressing these challenges of production and inventory planning under uncertain yields with significant contributions to the field. A survey by Yano and Lee (1995) has meticulously categorizes the existing literature, highlighting key discoveries and methodologies developed to tackle these issues.

The motivation for our research stems from a specific challenge faced in the semiconductor industry, known for its struggles with yield variability during wafer production. Despite advancements in technology and design, yield unpredictability remains a critical concern. Data from leading manufacturers indicate particularly low yield rates for some products during the initial stages of production, emphasizing the need for planning models that accommodate yield uncertainty amidst increasing product variety and decreasing lead times. With the advent of machine learning, predictive methods for yield have shown promise, suggesting that features related to design and produc-

tion environment could significantly enhance yield rate forecasts. Our study explores a data-driven approach for estimating and optimizing planning problems under yield uncertainty, using both real and synthetic data sets to test the efficacy of these methods. By addressing the complexities of lot-sizing problems affected by both yield and demand uncertainties, our approach demonstrates the practical value of integrating data-based decision-making strategies from a cost reduction standpoint.

2. The Problem and Models

We consider the following problem: a manufacturer has a random demand D to satisfy by some deadline. The manufacturing process is subject to random yield and whenever Q items are released, a random proportion turns out to be of acceptable quality. Let us denote by Y the random yield rate (the proportion of acceptable items). We assume that whenever Q items are released, QY is good for delivery. This is the well-known multiplicative (or proportional) model of yield uncertainty (Henig and Gerchak 1990). We assume that Y takes values between 0 and 1 and consider the following standard formulation each unit short of D incurs a penalty cost of b, and each unit above the demand has a cost of h. The objective is choose the lot size Q that minimizes the expected cost. This results in the following formulation:

$$\min_{Q} z(Q) = E\left[\left(b(D-QY)^{+} + h(QY-D)^{+}\right)\right] \tag{1}$$

The above is a well-studied problem in the random yield literature. The solution to the problem (1) with full distributional information on D and Y dates back to Shih (1980), Gerchak et al. (1988) and Henig and Gerchak (1990).

2.1. A Model with Yield and Demand Features

Our main focus in the paper is on the case where there are additional feature data that can possibly explain part of the variability in the random yield and demand. The feature information is available before determining the lot size Q. Let us assume that $\mathbf{X}$ is a vector consisting of r features that can potentially provide information on D and $\mathbf{U}$ is a vector consisting of s features that can provide information on Y. We can now choose the lot size conditional on the observed features, leading to:

$$\min z(Q) = E\left[\left(b(D|\mathbf{X}=\mathbf{x} - QY|\mathbf{U}=\mathbf{u})^{+} + h(QY|\mathbf{U}=\mathbf{u} - D|\mathbf{X}=\mathbf{x})^{+}\right)\right] \tag{2}$$

The above formulation is a direct extension of (1) and shows that if the conditional distributions are available, optimization proceeds as in the case without features. However, in many real applications, the number of features may be very large and the observed sample that is available may be relatively small and conditional distributions cannot be estimated easily. Next, we turn our attention to the data-driven case with a large number of features to develop viable solutions to (2).

3. Data-Driven Case with Many Features

Now, let us assume that we have a sample of data that includes the yield and demand realizations (d_i and y_i) along with corresponding set of feature values $\mathbf{X}_i$ and $\mathbf{U}_i$ for each observation. To provide an example of the potential challenges of data-driven optimization, a publicly available semiconductor data set has more than 500 features for each observation for a relatively small sample

of 1567 observations. This underlines the importance of model reduction and robust estimations for conditional uncertainty from a large number of features.

Some recent research considers a number of promising approaches for predictive analytics that combine estimation capabilities of machine learning methods with cost-based stochastic optimization (see for example Ban and Rudin (2019), Bertsimas and Kallus (2020))). To address the challenge of solving (2) with a large number of features, we employ and test a number of the proposed approaches along with some new problem-specific approaches. At the end of a comprehensive numerical study, We reach two main conclusions: i) feature information when used smartly leads to a significant decrease in the costs ii) the best performing method for the random yield problem combines the analytical solution of (2) with machine learning methods for esimating the conditional distribution.

4. Conclusions

We develop and systematically test a number data-driven strategies for calculating optimal lot sizes in manufacturing, taking into account uncertainties in yield and demand. Machine learning-based predictions are integrated with an expected cost minimization approach to derive rules for determining optimal lot sizes within a training dataset to learn the best lot-sizing rule for a given set of features.

A comprehensive version of the results can be found Bibak and Karaesmen (2024)

Acknowledgments

This research was supported by the Scientific and Technological Research Council of Türkiye through a TUBITAK 1001 Research Grant (project no: 123M871)

References

Ban, G., C. Rudin. 2019. The big data newsvendor: Practical insights from machine learning. *Operations Research* **67**(1) 90–108.

Bertsimas, D., N. Kallus. 2020. From predictive to prescriptive analytics. *Management Science* **66**(3) 1025–1044.

Bibak, B., F. Karaesmen. 2024. Integration of machine learning and optimization models for a data-driven lot sizing problem with random yield. *Available at SSRN 4760990* .

Gerchak, Y., R.G. Vickson, M. Parlar. 1988. Periodic review production models with variable yield and uncertain demand. *IIE Transactions* **20** 144.

Henig, I.M., Y. Gerchak. 1990. The structure of periodic review policies in the presence of random yield. *Operations Research* **38** 634–643.

Shih, W. 1980. Optimal inventory policies when stockouts result from defective products. *International Journal of Production Research* **18** 677–686.

Yano, C.A., H.L. Lee. 1995. Lot sizing with random yields: A review. *Operations Research* **43** 311–334.