Classification and Regression Tree Model to Predict the Probability of a Backorder in Uncertain Supply Chain

Gazi Md Daud Iqbal[#1], Matthew Rosenberger[*2], Lidan Ha[#3], Sadie Gregory[#4], Emmanuel Anoruo[#5]

[#]*College of Business, Coppin State University, 2500 W North Ave, Baltimore, MD 21216, USA*

[*]*Industrial Engineering, Worcester Polytechnic Institute, 100 Institute Rd, Worcester, MA 01609, USA*

[1]GIqbal@coppin.edu
[2]mhrosenberger@gmail.com
[3]LHa@coppin.edu
[4]SrGregory@coppin.edu
[5]EAnoruo@coppin.edu

Received July 30, 2023, **Accepted:** August 19, 2023, **Published Online:** August 31, 2023

Reviewers: Anonymous Peer Review

Citation: Iqbal, GMD., Rosenberger, M., Ha, L., Gregory, S., & Anoruo, E. (2023). Classification and Regression Tree Model to Predict the Probability of a Backorder in Uncertain Supply Chain. *International Journal of Supply Chain Management*, *12*(4), 1-8, https://doi.org/10.59160/ijscm.v12i4.6199

Abstract— **Supply chain uncertainties pose a massive and ever-present challenge for modern companies. These uncertainties can manifest in two contrasting scenarios: supply surplus, where companies have excess items, and supply shortages, where there is an insufficient quantity of goods. Each situation demands a different approach from businesses to adapt to the varying outcomes and maintain a competitive edge in the market. Product backordering is one of the important things that companies need to deal with in an uncertain supply chain. A backorder occurs when a customer-ordered product or service is not in stock or cannot be supplied immediately, and the customer has to wait. Companies striving for a balance in managing backorders. Machine learning models can help to determine the probability of a product being backordered. In this research, we develop Classification and Regression Tree (CART) model that uses previously known parameters to predict the likelihood of a product being backordered. We also use different model parameters to evaluate the accuracy of the model. Result shows that the developed model can help decision makers to identify the key factors that lead to a product backordering.**

Keywords— *CART, Product Backorder, Supply Chain, Rare Event, Predictive Modeling*

1. Introduction

Ref. [1] defined a supply chain as “a network of organizations and business processes for procuring raw materials, transforming these materials into intermediate and finished products, and distributing the finished products to customers” (p. 342). Supply chain management (SCM) practices connect suppliers, manufacturers, distributors at different levels, and logistics providers. Companies utilize SCM systems to manage the upstream and downstream relationships with these supply chain entities. The ultimate goal of SCM is to streamline the processes from raw material to consumption (or return) with minimum time and cost [1].

With product/service and data moving across organizational boundaries through different systems platforms in an increasingly expanding global marketplace, SCM faces a lot of challenges. At the macro level, the global, economic, and political environment can cause supply chain disruptions, uncertainties, delivery delays, and regulatory issues. For example, the impact of the disruptions we experienced during the COVID-19 pandemic is not over yet; and the Extensiv report [2] lists new challenges of “lack of warehouse space, consolidation within the industry, the ongoing labor crisis, and an uncertain economic climate to boot” evolving in 2023.

At the micro level, managing proper inventory levels based on customer demands, determining the best transportation logistics, and communicating across different technology platforms are just a few of the many complications involved, among which product backorder is one of the most common inventory management issues. A backorder occurs when a customer-ordered product or service is not in stock or cannot be supplied immediately, and the customer has to wait [3]. A backorder can be both a good thing and a bad thing. Backorders indicate healthy demand for a product while it might cause loss of customers.

When supply shortages occur, companies face elevated costs due to scarcity, leading to increased prices for their products or services. A classic example is the shortage of oil, which drives up production and raw material costs, subsequently raising gasoline prices. The domino effect extends to last-mile delivery, as higher fuel expenses translate to increased shipping fees, ultimately causing the prices of various goods to surge. Such price hikes can exert significant pressure on consumers and potentially hinder economic growth.

Conversely, when companies experience an excess supply of certain items, they often resort to reducing prices to clear surplus inventory and prevent waste. For instance, a cheese surplus can flood the market, driving down cheese prices, leading to its widespread use in various American dishes. While lower prices may appear advantageous to consumers, businesses must manage the repercussions of reduced profit margins and the potential challenge of recovering prices once the surplus is depleted.

Ref. [4] summarized the downsides of backorders as losing customers, losing market share to competitors, and increasing customer service complications. Ref. [5] listed five industries that have been undergoing supply chain delay challenges: bars and restaurants, aviation and air travel, electronics manufacturing, construction and building services, and hospitality. Several big household name stores like Costco, Target, and Walmart all experienced understock and/or overstock issues in recent years [6].

Given that the ultimate objective of SCM is to deliver customer-desired quality products or services in a timely and cost-effective manner, the backorder issue becomes a critical piece of the intertwined SCM challenges from different levels. Backorders are hard to predict and can be caused by a wide variety of factors, such as unusual customer demands, forecasting complexity and inaccuracy, shortage of source materials or products, logistics issues along the supply chain, inaccurate or outdated data, and other uncertainties [4,7].

Backorders can cause a bullwhip effect across a supply chain [1,4], leading to magnified inventory, warehousing, production, transportation, and other issues. One of the most intricate aspects of supply chain uncertainty is the bullwhip effect, a phenomenon where minor changes in consumer demand cause amplified fluctuations as they move up the supply chain, affecting wholesalers and manufacturers. The bullwhip effect arises due to information delays, order batching, price fluctuations, and inventory management practices. As a result, companies may find themselves grappling with inefficient inventory levels and increased carrying costs. The bullwhip effect can disrupt the balance between supply and demand, leading to stockouts or excess inventory, both of which negatively impact a company's bottom line. Overstocking is costly too and cannot be the solution for backorders. If the likelihood of the occurrence of a backorder can be predicted, better planning can be done for a product or service to flow through the supply chain to the customer cost-effectively without delays and thus improving the customer experience. Ref. [8] grouped the currently used backorder prediction models into two categories, “Classical machine learning classifiers” and “Deep learning-based predictive models,” and our research approach falls into the first category.

In this research, we use Classification and Regression Tree (CART) model to predict the probability of a product being back-ordered, i.e., a binary target variable, based on multiple independent variables. The independent variables, both binary and continuous, include current inventory level, transit time, amount of product in transit from source; forecast sales for the next 3, 6, and 9 months; sales quantity for the prior 1, 3, 6, and 9 month(s); minimum recommended stock amount; identified source issue, parts overdue from source; source performance for prior 6 and 12 months; amount of stock orders overdue, and risk flags. The rest of the paper is organized as follows.

In section II, we summarize literature related to product backorder prediction modeling. Section III presents a case study that includes the description of data source, data summary, and data processing. In section IV, we discuss the results. Section V describes the conclusion and future works.

2. Literature Review

Product backorder is a rare event. Rare events are defined as when something happens with dozens to thousands of times fewer ones than nonevents. They don’t usually occur very often but when they do that become severe. Because of the nature of the product backorder and its consequences, scientific and business communities are focusing on prediction of a product being backordered.

Ref. [10] investigate the applicability of advanced machine learning techniques, including neural networks, recurrent neural networks, and support vector machines, to forecasting distorted demand at the end of a supply chain (bullwhip effect). They compare their machine learning techniques with traditional methods including naïve forecasting, trend, moving average, and linear regression by using simulated supply chain data and actual Canadian Foundries orders data. Their findings suggest that advanced machine learning models give the best performance while forecasting accuracy was not statistically significant better than that of the regression model. Ref. [3] used Distributed Random Forest and Gradient Boosting Machine learning techniques to develop prediction models for probable backorder scenarios in the supply chain. They utilized a five-level metric to indicate the inventory level, sales level, forecasted sales level, and a four-level metric for the lead time. In their research, they list major probable backorder scenarios to facilitate business decisions. They show how their model can be used to predict the probable backorder products before actual sales take place. Ref. [11] predicted the product backorders using a new H2O automated machine learning algorithm approach to identify products at risk of backorders, use synthetic minority over-sampling technique (SMOTE) to synthetically improve dataset balance, and show Cost/Benefit Information using confusion matrix. This research optimized the model for expected profit and chose optimal cutoff value for the classification by trial-and-error method. Ref. [12] propose a framework for addressing estimation uncertainty that is applicable to any inventory model, demand distribution, and parameter estimator. Their framework involves four steps which are formulate the decision model in terms of the lead time demand distribution function, estimate the parameters and the distributions of their estimation errors efficiently, replace the true parameters with an appropriate function of the point estimates and estimation errors in the lead time demand distribution function, and use the expectation of the lead time demand distribution function with respect to the estimation errors as the new predictive lead time demand distribution in the decision model. They claim that their method can be applied to any inventory model using any demand process and parameter estimator, as long as the distribution of its estimation error can be derived or approximated. Ref. [13] used support vector regression (SVR) method to develop a demand forecast model for supply chain and then compare the result with RBF neural network method. This research shows that SVR is superior to RBF in prediction performance. Ref. [14] proposed a novel approach for supply chain risk management using Bayesian Belief Network based on the dependency among risk factors. They integrate their machine learning model into the supply chain network to model the risk propagation. Finally, dynamic backorder replenishment plan is developed based on the impact of risks. Ref. [15] examine how best to redistribute stock amongst several secondary warehouses in a two-echelon network. They developed a stochastic redistribution model to find the optimal allocation to each warehouse to minimize the expected waiting time in repair shops. They use data from Spanish Army to demonstrate the model's effectiveness. The results are satisfactory in terms of the number of positive cases, the reduction in waiting days and solution accuracy. Ref. [16] develop machine learning models to predict material backorders in inventory management, which is a common supply chain problem, and impacts inventory system service level and effectiveness. Their model will identify parts with the highest chances of shortage prior to its occurrence can present a high opportunity to improve an overall company’s performance. Ref. [8] proposes a model that uses a deep neural network to predict backorders which handles the data imbalance between backorders and filled orders with efficient techniques. Their model achieves a new state-of-the-art performance and outperforms some prominent classification models in terms of standard evaluation metrics and expected profit measure. Ref. [17] propose a new approach to decide on ordering policies of supply chain members in an integrated manner so that product backorder is minimized. They formulate supply chain ordering management problems using reinforcement learning model, and finally use Q-learning algorithm to solve reinforcement learning model. Results show that their approach provides better results than other known algorithms. Ref. [18] provide insights into the barriers of forecasting uncertain product demand in supply chain by focusing on the relative importance of the barriers for businesses, particularly the forecast practitioners and prospective forecast implementers.

In this research, we use a Classification and Regression Tree (CART) approach to predict the occurrence of a product being backordered based on previous known parameters, and presented important variables that influence product backordering. We also show different model parameters to evaluate the accuracy of the developed model. To the best of our knowledge, this is the first research that uses CART model to predict the probability of a product being backordered.

3. Case Study

In this section, we describe the data source, data summary, and methodology that was used to develop a prediction model.

3.1 Data Summary

The data set used in this research is collected from github [https://github.com/AasthaMadan/Product-Backorders/wiki/Product-back-orders-prediction], an online community platform for data scientists and machine learning enthusiasts.The data set contains 1290650 observations and has 22 variables. The response variable is a product that went on backorder. It contains the historical data for the 8 weeks prior to the week we are trying to predict. The variables name, their description, and values are shown in Table 1.

3.2 Data Imputation

We have a combination of continuous and binary variables, and there are a lot of missing values in the data set. The simplest and popular approach is to remove all the samples with missing values, but this ignores a vast amount of useful information and dramatically reduces the number of samples in the analysis which includes product backordering. Consequently, there may not be enough observations for modeling. Therefore, we need to impute dataset to preserve the use of as much data as possible. In missing data imputation, we assume that data are missing randomly. Linear regression is used to impute continuous missing values while logistic regression is used for binary variables using general strategies for data imputation discussed by previous researchers [19,20,21,22].

3.3. Methodology

In this research, we use the CART modeling approach. CART is a decision tree tool for creating a predictive model. CART starts from a root node and splits the dataset into two branches based upon least squares and cross validation. The splitting process continues until reaching terminal nodes. Each observation falls into exactly one terminal node based upon the tree logic [9]. CART partitions the factor variable space into regions and estimates the response in each region with a different constant. Within a specific region, the constant response estimates average the observations in that region. More regions are needed where the response changes frequently. CART uses recursive binary partitioning. In addition, CART yields variable importance scores that show the relative importance of each independent variable in the model.

The complexity parameter (cp) is used by CART to control the size of the decision tree and to select an optimal tree size. If the cost of adding another variable to the decision tree from the current node is above the value of cp, then tree building terminates. Specifically, tree construction terminates unless continuing to build the tree decreases the overall lack of fit by a factor of cp.

4. Results

In this section, we discuss the CART model to predict the probability of a product being backordered, and different model performance parameters.

4.1 CART Model

Figure 1 shows the probability of a product being backordered. The condition at the split gives the logic for the left branch of the split. For example, “national >= 2” in Figure 1 indicates that if current inventory level for the part is greater than or equal to 2, then we should follow the left branch of the tree logic; otherwise, we should follow the right branch. The leaves of the trees show the probability of a product being backordered in the next week

Table 1: Variable Description

Variable Name	Descriptions	Values
national_inv	Current inventory level for the part	continuous
lead_time	Transit time for product (if available)	continuous
in_transit_qty	Amount of product in transit from source	continuous
forecast_3_month	Forecast sales for the next 3 months	continuous
forecast_6_month	Forecast sales for the next 6 months	continuous
forecast_9_month	Forecast sales for the next 9 months	continuous
sales_1_month	Sales quantity for the prior 1-month time period	continuous
sales_3_month	Sales quantity for the prior 3-month time period	continuous
sales_6_month	Sales quantity for the prior 6-month time period	continuous
sales_9_month	Sales quantity for the prior 9-month time period	continuous
min_bank	Minimum recommend amount to stock	continuous
potential_issue	Source issue for part identified	0: No, 1: Yes
pieces_past_due	Parts overdue from source	continuous
perf_6_month_avg	Source performance for prior 6-month period	continuous
perf_12_month_avg	Source performance for prior 12-month period	continuous
local_bo_qty	Amount of stock orders overdue	continuous
deck_risk	Part risk flag	0: No, 1: Yes
oe_constraint	Part risk flag	0: No, 1: Yes
ppap_risk	Part risk flag	0: No, 1: Yes
stop_auto_buy	Part risk flag	0: No, 1: Yes
rev_stop	Part risk flag	0: No, 1: Yes
went_on_backorder	Product actually went on backorder.	0: No, 1: Yes

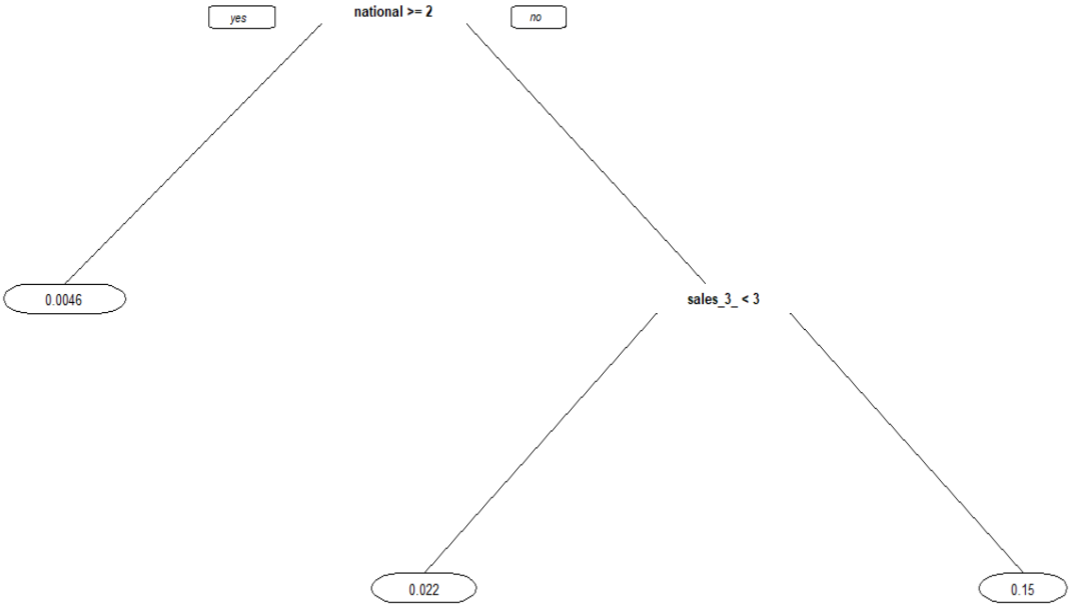

Figure 1: Probability of a product being backordered based on different features

given the aforementioned tree logic based upon the previous 8 weeks product variables explained in Table 1. The complexity parameter (cp) is used by CART to control the size of the decision tree and to select an optimal tree size. If the cost of adding another variable to the decision tree from the current node is above the value of cp, then tree building terminates. Specifically, tree construction terminates unless continuing to build the tree decreases the overall lack of fit by a factor of cp [23]. For this CART model, the cp value is 0.03, while cross-validation error is 0.71. This cross-validation error is used to better estimate the test error of the model.

Figure 2 shows the variables importance from the CART model. The sales quantity for the prior 3-month time period has the highest importance among all the variables. Only the amount of stock orders overdue has less than 10% importance level. Although the CART model considers 21 variables to develop the model, actually two variables were used in this model to construct the tree as shown in Figure 1.

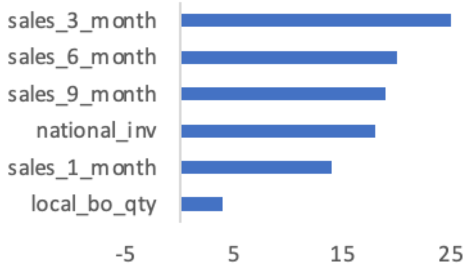

Figure 2: Variable importance from CART model

From Figure 1, we can say that if the current inventory level of a part is less than 2 and the sales quantity for the prior 3-month time period is greater than or equal to 3, then the probability that a product will be backordered is 15%. However, there is less than a 1% probability of a product being backordered if the current inventory level for a part is greater than or equal to two.

4.2 Model Performance

We split the dataset discussed in Section III randomly to 80% and 20% for training and testing, respectively. The training dataset is used to fit the CART Model, while the testing dataset is used to validate these models. Accuracy, sensitivity, specificity, and R-squared parameter values for the model are shown in Table 2. For the accuracy, sensitivity, and specificity metrics, we assume the model predicts the product backorder if it predicts the probability of a product being backordered is over 50%. Consequently, accuracy is defined as the model's ability to predict correctly whether a product is being backordered. Sensitivity is defined as the model’s ability to identify the occurrence of a product being backordered, while specificity is the ability to identify the non-occurrence of a product being backordered.

Table 2: Model Parameters

	Training	Testing
Percentage of backorder	0.85%	1.11%
Accuracy	0.99	0.98
Sensitivity	0.12	0.14
Specificity	0.98	0.96
CART R-square	0.04	0.03

The results show that the sensitivity is low while the specificity is quite high for both training and testing data set. This is because in this research we want to predict the probability of a product being backordered, which is rare (0.85% and 1.11% in training and testing data set, respectively). This is the same reason why the CART R-squared values for both training and testing data set are low.

5. Conclusion and Future Works

To tackle the challenges of supply chain uncertainties, companies employ several strategies. One such approach is maintaining safety stock - the additional inventory held as a buffer against unexpected demand spikes or supply disruptions. Safety stock acts as a cushion, enabling companies to respond promptly to changes in market conditions without incurring severe disruptions. However, keeping high levels of inventory harm companies. Therefore, companies need to make a balance between overstocking and customer service management. Companies can improve customer experience if the likelihood of the occurrence of a backorder can be predicted. In this research, we use classification and regression tree model to predict the occurrence of a product backorder. Result shows that the developed model can help decision makers to identify the key factors that lead to a product backordering and to find out the probability of a product being backordered.

Since the occurrence of a product backorder is rare, we will develop a new probability support vector machine approach to determine the probability. These new machine learning models will improve the results. Finally, we will compare the accuracy of the model with existing methods that are available in the literature.

References

[1] Laudon, Kenneth C., and Jane Price Laudon. Management information systems: Managing the digital firm. Pearson Educación, 2004.

[2] Extensiv, "2023 State of the Third-Party Logistics (3PL) Industry Report," https://www.extensiv.com/resource-library/report/state-of-the-third-party-logistics-industry-report. [Online; accessed 26-July-2023].

[3] Islam, Samiul, and Saman Hassanzadeh Amin. "Prediction of probable backorder scenarios in the supply chain using Distributed Random Forest and Gradient Boosting Machine learning techniques." Journal of Big Data 7 (2020): 1-22.

[4] D. Luther, "Backorders Defined: What It Is, Causes & Solutions," https://www.netsuite.com/portal/resource/articles/inventory-management/backorder.shtml . [Online; accessed 26-July-2023].

[5] E. Newton, "5 Industries Feeling the Strain of Supply Chain Delays," https://www.globaltrademag.com/5-industries-feeling-the-strain-of-supply-chain-delays/. [Online; accessed 29-July-2023].

[6] D. kline, "Costco Made a Huge Mistake Just Like Walmart and Target," https://www.thestreet.com/investing/costco-faces-same-problem-as-walmart-target. [Online; accessed 29-July-2023].

[7] Infosys BPM, "Supply chain challenges in 2023 and how to overcome them," https://www.infosysbpm.com/blogs/supply-chain/supply-chain-challenges-in-2023-and-how-to-overcome-them.html. [Online; accessed 26-July-2023].

[8] Shajalal, Md, Petr Hajek, and Mohammad Zoynul Abedin. "Product backorder prediction using deep neural network on imbalanced data." International Journal of Production Research 61, no. 1 (2023): 302-319.

[9] Loh, Wei-Yin. "Classification and regression trees." Wiley interdisciplinary reviews: data mining and knowledge discovery 1, no. 1 (2011): 14-23.

[10] Carbonneau, Real, Kevin Laframboise, and Rustam Vahidov. "Application of machine learning techniques for supply chain demand forecasting." European journal of operational research 184, no. 3 (2008): 1140-1154.

[11] Matt Dancho, “Predictive Sales Analytics: Use Machine Learning to Predict and Optimize Product Backorders”. https://www.business-science.io/business/2017/10/16/sales_backorder_prediction.html. Accessed on July 20, 2023

[12] Prak, Dennis, and Ruud Teunter. "A general method for addressing forecasting uncertainty in inventory models." International Journal of Forecasting 35, no. 1 (2019): 224-238.

[13] Guanghui, W. A. N. G. "Demand forecasting of supply chain based on support vector regression method." Procedia Engineering 29 (2012): 280-284.

[14] Shin, KwangSup, YongWoo Shin, Ji-Hye Kwon, and Suk-Ho Kang. "Development of risk based dynamic backorder replenishment planning framework using Bayesian Belief Network." Computers & Industrial Engineering 62, no. 3 (2012): 716-725.

[15] García-Benito, Juan Carlos, and María-Luz Martín-Peña. "A redistribution model with minimum backorders of spare parts: A proposal for the defence sector." European Journal of Operational Research 291, no. 1 (2021): 178-193.

[16] De Santis, Rodrigo Barbosa, Eduardo Pestana de Aguiar, and Leonardo Goliatt. "Predicting material backorders in inventory management using machine learning." In 2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI), pp. 1-6. IEEE, 2017.

[17] Chaharsooghi, S. Kamal, Jafar Heydari, and S. Hessameddin Zegordi. "A reinforcement learning model for supply chain ordering management: An application to the beer game." Decision Support Systems 45, no. 4 (2008): 949-959.

[18] Abou Maroun, Elias, Didar Zowghi, Renu Agarwal, and Babak Abedin. "Uncovering barriers in forecasting uncertain product demand in the supply chain." International Journal of Supply Chain Management 11, no. 6 (2022): 35-44.

[19] Palmer, Raymond F., and Donald R. Royall. "Missing data? Plan on it!." Journal of the American Geriatrics Society 58 (2010): S343-S348.

[20] Schneider, Tapio. "Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values." Journal of climate 14, no. 5 (2001): 853-871.

[21] Gao, Sujuan. "A shared random effect parameter approach for longitudinal dementia data with non-ignorable missing data." Statistics in Medicine 23, no. 2 (2004): 211-219.

[22] Candès, Emmanuel J., and Terence Tao. "The power of convex relaxation: Near-optimal matrix completion." IEEE Transactions on

Information Theory 56, no. 5 (2010): 2053-2080.

[23] Iqbal, Gazi Md Daud, Sadie Gregory, Jay M. Rosenberger, Muhammad Shah Alam, and Tom Mazzone. "Predicting Heatwaves Using Classification and Regression Trees." In IIE Annual Conference. Proceedings, pp. 67-72. Institute of Industrial and Systems Engineers (IISE), 2021.