IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received March 6, 2020, accepted March 21, 2020, date of publication March 24, 2020, date of current version April 7, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2983118

A Profit Function-Maximizing Inventory Backorder Prediction System Using Big Data Analytics

PETR HAJEK[1] AND MOHAMMAD ZOYNUL ABEDIN[2,3]
[1]Institute of System Engineering and Informatics, Faculty of Economics and Administration, University of Pardubice, 532 10 Pardubice, Czech Republic
[2]Collaborative Innovation Center for Transport Studies, Dalian Maritime University, Dalian 116026, China
[3]Department of Finance and Banking, Hajee Mohammad Danesh Science and Technology University, Dinajpur 5200, Bangladesh

Corresponding author: Mohammad Zoynul Abedin (abedinmz@yahoo.com)

This work was supported by the scientific research project of the Czech Sciences Foundation under Grant 19-15498S.

ABSTRACT Inventory backorder prediction is widely recognized as an important component of inventory models. However, backorder prediction is traditionally based on stochastic approximation, thus neglecting the substantial amount of useful information hidden in historical inventory data. To provide those inventory models with a big data-driven backorder prediction, we propose a machine learning model equipped with an undersampling procedure to maximize the expected profit of backorder decisions. This is achieved by integrating the proposed profit-based measure into the prediction model and optimizing the decision threshold to identify the optimal backorder strategy. We show that the proposed inventory backorder prediction model shows better prediction and profit function performance than the state-of-the-art machine learning methods used for large imbalanced data. Notably, the proposed model is computationally effective and robust to variation in both warehousing/inventory cost and sales margin. In addition, the model predicts both major (non-backorder items) and minor (backorder items) classes in a benchmark dataset.

INDEX TERMS Big data, inventory backorder, machine learning, prediction.

I. INTRODUCTION

In customers' purchasing pattern forecasting, it is discovered that consumers favor their demands to be backordered when inventory goes in shortfalls. In government bodies like military systems and distribution channels, this phenomenon mostly happens [1]. Moreover, it also occurs in monopolized commodities and luxury products. The customers who acquire goods from government control marketplaces are ready to stay for the precious object even if it is out of stock at that point of the day. Once the next supply arrives, the particular product order is issued; therefore, the demand in such cases does not misplace.

Prediction accurateness is a critical factor for, among other things, lessening the production costs and ensuring better inventory services. Precise prediction tasks are significant for inventory management operations. However, modelers should concentrate not only on reducing the prediction error but also on exploring the economic gain of the predictions. Moreover, inaccurate material backorder prediction may worsen the operations of inventory management and production systems. Li & Wang assert that erroneous predictions cost vendors as much as 10 percent of their revenue [2]. Therefore, a sound material prediction minimizes inventory backorder risk that is a rare event task in the accounting domain.

In accounting and inventory management literature, the existing experimental works focus on either inventory control or inventory planning issues. Bearing in mind about dissimilar systematic solutions of control, the earlier study tries to settle on its choice about when to order and how much quantity should order [3]. In this line of research, a recent study recommends an inventory control system to reduce the mean manufacturing expenditure, together with the material holding expenses and the backorder charges [2]. Afterward, a number of material planning mechanisms are proposed based on the highest life of the products [4]. Moreover, the majority of the analytical techniques suggested hitherto articulate the inventory management issues as a multi-objective optimization problem, that is to say, the order placement and

The associate editor coordinating the review of this manuscript and approving it for publication was Inês Domingues.

material holding expenses should keep to a minimum level, while the degree of utility (e.g., profit) needs to maintain as much as feasible. In the literature, expected backorders are usually stochastically approximated on the basis of the base stock level [5]. Recently, a stock-out prediction model was proposed that uses inventory levels at all nodes of multi-echelon networks [6]. Time series forecasting represents an alternative approach to predict the inventory level [7]. However, a substantial amount of information is hidden in inventory databases, such as prior stock amount, amount of stock overdue and risk flags indicating news events related to operational issues [8]. Applications of these big data are challenging problems in operations and supply chain management [9]–[11]. Indeed, big data analytics generate competitive advantages by mining important information from the high-volume databases. Big data applications mainly assist enterprises to formulate feasible business decisions and improve their strategy, operational effectiveness, supply chain sustainability and economic efficiency [12]. Moreover, big data analytics also assists in deep comprehend of enterprises' dynamics, escalates customer attachment, optimizes routine operations, and generates new sources of organizational profit [11]. These are the reasons why big data analytics has received increasing attention in supply chain management [11].

After reviewing many previous studies of inventory backorder prediction [2]–[4], [13]–[19], to the best of our knowledge, no big data-driven inventory backorder prediction model that maximizes profit function has been presented, which is a gap in the literature that this study bridges. Based on these experiences, this study aims to investigate the question of how a profit function-maximizing inventory backorder prediction system provides quantitative insights into the economic merit of optimal backorder strategy, particularly in the era of big data.

A difficulty emerges in this specific kind of machine learning classifier. In typical supply chain management, the number of goods which are in backorder (positive or minority class) carries entirely few instances relative to the number of lively products or non-backorder items (negative or majority class). This scenario is recognized as the class imbalance event. In many real-life prediction domains, for example, loan approval data modeling [20], corporate bankruptcy prediction [21] or credit card anomaly detection [22]; class skewed scenarios are prevalent. In class inequity data modeling, the minority positive instances are more significant than the majority class examples. Therefore, the principal objective of a class skewed event is to optimize a feasible tradeoff between the majority and minority instances which will exploit the contributions of the positive data class. The different costs associated with incorrect predictions of the backorder and non-backorder product is another problem that needs to overcome.

To address the abovementioned issues, this study proposes a novel system for predicting inventory backorders to optimize organizations' backorder strategy in terms of economic performance. The novelty of the proposed system is that it integrates data undersampling and a profit function-maximizing ensemble classifier, allowing for big inventory data utilization. The proposed model predicts inventory instances at risk of backorder before shortages happen, thus providing the inventory manager with an appropriate period of time to respond. By predicting backorder items before customers' orders placed, the manufacturing department can regulate their production, which in turn, it may reduce lead times and the cost efficiency of holding inventory. The inventory management predictive systems applied in this paper facilitate the utmost materials to acquire in the hands of corporate users or individual customers at the lowest possible charge to the production house. As a result, the outcome would be win-win; explicitly, organizations augment sales margin while end-users obtain to enjoy the goods they demand. To evaluate prediction performance, we develop a profit-based measure that takes into account the economic effects of managerial decisions based on the different cutoff points. We also show the implications of different inventory strategies on the economic performance of the inventory backorder prediction system. Eventually, this study asserts that the proposed prediction model outperforms the existing machine learning methods applied over an extremely imbalanced big data.

The rest of the paper follows five sections. Section II briefly reviews the related literature on inventory models with backorders, big data analytics in inventory management and the imbalanced class distribution problem. Section III outlines the proposed inventory backorder prediction model. In Section IV, the benchmark dataset is presented and used. Section V illustrates experimental results and their applications. Lastly, Section VI concludes the study.

II. RELATED LITERATURE

A. INVENTORY MODELS WITH BACKORDERS

An extensive literature integrates the prediction, pricing, and material control decisions of backorder systems; for outstanding reviews, see [23] and [24]. This line of empirical research dates to Rosling [25], with further refinements optimizing material backorder issues presented by many academicians.

Van Foreest *et al.* [26] proposed a base-stock policy derived reservation with an ongoing material review and favorable lead-times. Their model follows the modification of regular backorder policies using the Poisson principle. They claimed that their method significantly increases the inventory backorder rate, which minimizes total inventory costs.

Two integrated mixed-integer programming techniques for assembling a production routing problem with backordering were suggested [27]. The proposed hybrid algorithm was applied to solve a supply chain management problem by considering multiple item lot-sizing decisions and vehicle routing decisions to the points of sale. Their findings show that the combined heuristic provides better prediction results and training times than existing algorithms. Similarly, two conflicting objective functions, inventory level and the number of backorders, were combined in a reinforcement learning model to obtain optimal production/maintenance control policy [28].

Feng *et al.* [29] formulated the generalized additive model for the dynamic inventory backorder and pricing control problem. They asserted that their model overcomes the shortcomings of several demand models, including the frequently utilized base stock list price policy of existing studies. In their experiment, the base stock list price ensures optimality under a set of constraints, and the findings are validated by applying constrained maximum likelihood estimation.

Wee *et al.* [30] suggested a generalized production lot size model with backordering. Their experiments focused on the permissible shortage backordering and the impact of changing backordering costs. All parameters, including an expected net profit, backordering cost and percentage increase in profit, are optimized in their model. Sadeghi *et al.* [31] generalized a feasible solution to the material order size problem ensuring the maximum backorder quantities. The hybrid imperialist competitive algorithm was trained to generate their results. A cash flow-based profit-maximizing net present value model was derived by Ghiami and Beullens [32]. This model confirmed a feasible inventory policy that integrated backorder costs, sales costs, and holding costs.

Trapero *et al.* [33] applied parametric GARCH models and non-parametric kernel density estimation to generate safety material stock levels. They found that the normality assumption is more significant and the kernel density estimation is the most appropriate for shorter lead times, whereas parametric models are more suitable for longer lead times. To capture the temporal dynamics among multiple inventory level time series, a joint prediction model was proposed in [7]. Most recently, a machine learning-based model was developed for predicting stock-outs in multi-echelon supply chains [6]. This model employs a deep neural network using information from all the nodes in the multi-echelon network.

After reviewing related studies, this paper finds that the material backorder prediction based on big data characteristics and its profitability occurring from misclassification have not been considered in previous studies. Explicitly, big data imbalance scenarios are a vital issue that significantly degrades material backorder policy and has an adverse effect on inventory management. Based on these scenarios, this study reveals the skewed material backorder data traits that offer sound data-balancing techniques for establishing a feasible trade-off between costs and profits. As our inventory backorder prediction system is specifically designed for highly imbalanced big data in real-world scenarios, a brief summary of the theoretical background on the big data analytics in inventory management and imbalanced class distribution problem is provided in the next two subsections.

B. BIG DATA ANALYTICS IN INVENTORY MANAGEMENT

A variety of high-volume data is generated in ERP systems, including inventory size, lead times, historic orders or cost of placing the orders. Moreover, inventory levels vary depending on the diverse organizational and customer needs causing fluctuations in supply and demand [11]. Three types of big data analytics have been used to assist in inventory management, namely descriptive, predictive and prescriptive analytics [11], [34], [35]. Descriptive analytics was used to provide insight into what has happened in complex retail and multi-channel inventory systems. For example, the perceived performance of several material planning methods in different inventory settings was investigated using the dataset from more than two hundred manufacturing and distribution companies [36]. Predictive analytics is aimed at predicting what might happen, including the accurate forecasts of inventory needs. Such predictions, often based on customer demand, resulting in a substantial reduction in inventory costs. It was demonstrated that the accuracy of inventory prediction systems can be enhanced using the time series analysis of each component's demand [37]. A statistical multi-period prediction system was proposed by [38] to consider demand uncertainty and satisfy demand with minimal operating costs, which is achieved by incorporating a dynamic programming model for the optimal inventory policy determination. Gumus *et al.* [39] employed a neuro-fuzzy system to predict demand and lead time to deploy the inventory efficiently. Finally, prescriptive analytics seeks to provide the optimum solution among various alternatives, given the known restrictions. Wang and Lei [40] proposed an optimization system that assigns both suppliers and demand points to distribution centers to ensure that the total shipping costs and the number of unfulfilled orders are minimized. Network capacity and deadline were the constraints considered in their model. The optimal inventory policy was found for a multi-product multi-echelon system that considers stochastic demand and batching [41]. In their system, both the service level and ordering constraint was respected. Another optimization model using genetic algorithms was developed to find the best inventory policy for perishable items in a three-stage supply chain [42].

In summary, previous studies have demonstrated the effectiveness of utilizing big data analytics in various inventory management problems. However, several issues must be considered when dealing with big data in this domain [9]. First, the data volume and variety, closely associated with their heterogeneity and statistical biases, require developing suitable analytics models. This, in turn, leads to high computational costs and instability of traditional statistical methods when dealing with big inventory data. Therefore, more adaptive and robust models must be developed to overcome these problems [43].

C. IMBALANCED CLASS DISTRIBUTION PROBLEM

Considering the significance of class imbalance issues, numerous methods have been applied to raise the accuracy of the trained algorithms. Skewed data learning algorithms can be categorized into three groups [44]. On the first hand, algorithmic approaches propose a new learning technique or amend existing algorithms to intensify and concentrate

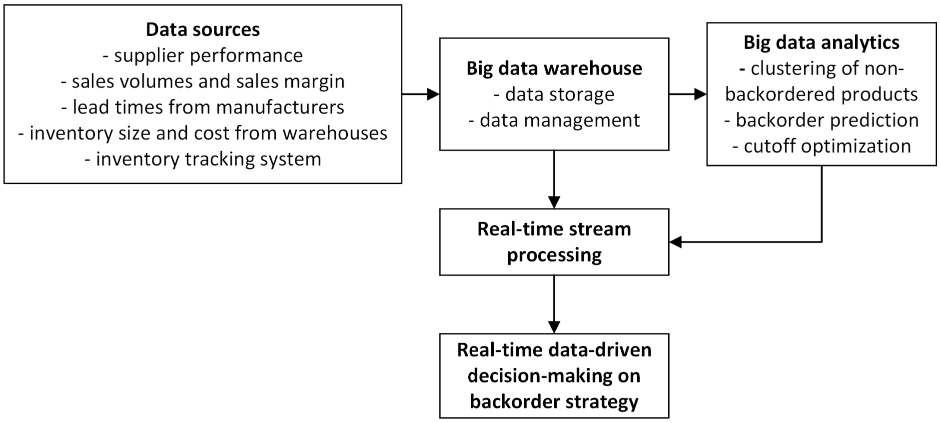

FIGURE 1. Big data analytics framework for inventory backorder management.

on the implication of positive classes [45]. Then, data-level solutions apply sample pre-processing techniques to generate a balanced database; these deal with undersampling and oversampling methods [21]. The third case combines both these schemes and assigns different misclassification costs to each group in the training stage [22]. In addition, ensembles (combinations) are designed to augment the performance of a single algorithm using numerous learning models and merging their results into a single class label [46]. Specifically, data-level solutions are extensively applied in the literature since they propose a skewed data pattern before training the algorithms. Moreover, sample pre-processing techniques are self-supporting learning systems compared with classifiers. Finally, data-level solutions integrating into ensembles of classifiers outperform other techniques [47].

As for the sample pre-processing technique, data-level solutions can be differentiated into oversampling [48], [49] and undersampling methods [50], [51]. The oversampling method increases the number of positive instances in minority samples to generate a balanced training set. SMOTE [48] is the best-known oversampling algorithm used widely in the literature. SMOTE generates artificial positive instances in minority examples by choosing some of the adjacent positive neighbors of a minority object (called S) and produces new positive class instances along the lines between S and all adjacent positive neighbors. SMOTE outperforms random oversampling techniques based on its well-known traits. Having solved the overfitting problems, it lessens unequal sample allocation in datasets. However, it may cause overgeneralization as it ignores majority class objects and generates artificial positive data in minority class samples.

Alternatively, the random undersampling (RUS) method aims to lower the number of instances with the majority class, as there are more instances of one group than the other group in the extreme class allocation problem. If $T =$ {training example set}, $N =$ {majority class instances} and $P =$ {minority class instances}, then an extreme class allocation problem reduces the skewed allotment of N and P by decreasing the size of N. As databases and their features have been progressively growing, the undersampling method should thus be a better option than its oversampling counterpart [47].

Undersampling techniques decrease the size of majority instances to generate a rebalanced sample size. The RUS method can remove potentially meaningful information from the majority class example set, which is a major shortcoming of this data sampling approach. To allow more significant instances in the majority class, the clustering approach has therefore been proposed as an undersampling technique [44], [50], [52]. The central idea of the clustering procedure is to combine identical instances into equivalent clusters, and the data points in their respective clusters are dissimilar based on their various illustrations. Moreover, clustering-based undersampling (CBUS) techniques [44], [52] divide majority or minority instances using the k-means algorithm [53], and the existing literature typically either fixes the number of clusters (k) or k is equivalent to the number of minority class objects in the dataset. The k-means methodology trained on the majority (or minority) instances subsequently generates k cluster centers. These cluster centers are applied to restore the whole dataset. Accordingly, the experimental datasets are rebalanced owing to the identical number of instances for both classes.

III. MODEL FOR INVENTORY BACKORDER PREDICTION USING BIG DATA ANALYTICS

A. BIG DATA ANALYTICS FRAMEWORK

The critical role of big data analytics in inventory backorder management is highlighted in Figure 1. Firstly, relevant data sources should be identified at the level of items (orders). These data sources include data from suppliers, sales transactions and warehouses [9]. Distributed data sources are the cause of data variety in inventory management. Weekly time buckets should be used for the data because higher fill-rates (without backorders) can be achieved for lower inventory

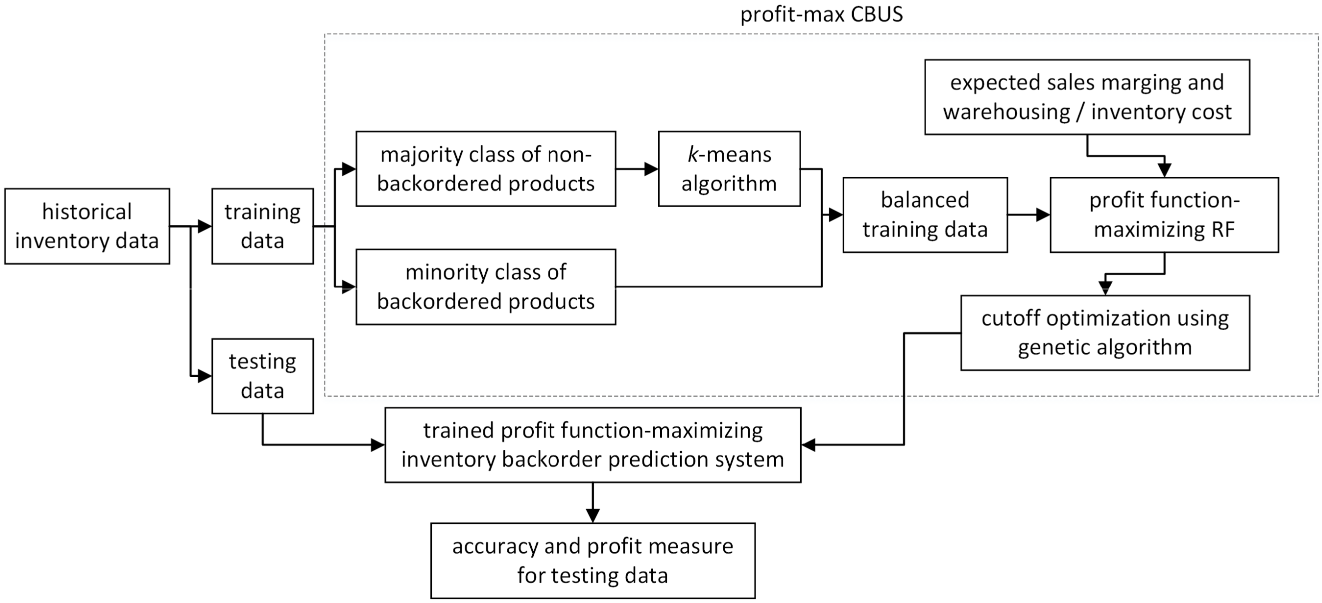

FIGURE 2. **Model for profit function-maximizing inventory backorder prediction.**

levels [54]. Data collection at a more disaggregated level and frequent updates increase data volume and velocity, respectively. As a result, data are generated for individual orders and weekly time buckets. Data are recorded via Internet of Things (IoT) devices and passed to the big data warehouse using a suitable transmission mechanism (via IoT gateway cloud). In the big data warehouse component, raw input data are stored and organized for efficient processing. Storage virtualization is recommended to provide a scalable interface to real-time stream processing and big data analytics [55]. Big data analytics for backorder prediction is the core component of the framework. Note that the proposed profit-function-maximizing inventory backorder prediction system represents a combination of predictive and prescriptive analytics. On one hand, it is used to perform the one-week-ahead prediction of backorder/non-backorder items. On the other hand, it provides a decision-maker with an optimal backorder strategy using a profit-based measure. To increase the accuracy of the backorder prediction system, training data should be selected using a clustering algorithm so that the backorder and non-backorder classes of items are balanced. Then, a machine learning algorithm should be employed to perform the prediction task. To increase the efficiency of the decision support system, traditional classification performance metrics should be replaced by a profit-based objective function in the training process. In addition, the decision support system must be able to project the outcomes of different backorder strategies represented by different cutoff points. Big data analytics should be deployed in real-time to generate outputs for real-time data-driven decision-making on backorder strategy. Therefore, low training times of the backorder prediction system are required to enable real-time stream processing. The details on the big data analytics component are presented in the following subsection.

B. PROFIT FUNCTION-MAXIMIZING INVENTORY BACKORDER PREDICTION SYSTEM

This section defines the model proposed for profit function-maximizing inventory backorder prediction (Figure 2). This model modifies CBUS to maximize the expected profit of backorder decisions. To achieve this objective, it is necessary to employ an algorithm that accurately predicts the need for inventory backorders. To optimize the economic effects of backorder decisions, a profit-based measure is integrated into the prediction algorithm. In addition, to help decision-makers easily recognize the optimal decision cutoff, a search procedure is used to maximize the expected profit.

The CBUS method was proposed to overcome the problem of the trade-off between prediction performance and complexity present in earlier approaches to the imbalance problem [44], [52]. To achieve high prediction performance while bounding time complexity, the k-means algorithm is first used to cluster the instances in the minority class. For each cluster, the method selects the same number of instances from the majority class to balance the training dataset.

A classifier is then trained on the balanced training dataset [44] or cluster-specific classifiers are used to maximize accuracy on the given data subsets [52]. As a result, the time complexity of the method is bound by the number of instances in the minority class. To further improve the effectiveness of this method, we modify it in two directions. First, the C4.5 decision tree classifier was used in original CBUS approaches [44], [52]. To overcome the potential problem

TABLE 1. Confusion matrix for inventory backorder prediction.

Prediction/Target	Positive (non-backorder item)	Negative (backorder item)
Positive	TP ($b_1 = 0$)	FP ($c_0 = 0$)
Negative	FN (c_1 = warehousing and inventory cost)	TN (b_0 = sales margin)

of overfitting, we propose replacing the C4.5 decision tree classifier with the Random Forest classifier, which solves this problem by introducing an ensemble of diverse base decision trees trained on the training data subsets. The prediction performance of the Random Forest classifier is further improved by randomly sampling a feature subset. Thus, feature selection is embedded directly into the classifier and the performance on imbalanced big data can be improved [56]. The second modification is the use of profit function maximization as the objective function of the Random Forest classifier. The problem of maximizing classification accuracy in original CBUS approaches is that it assumes that the costs associated with type I errors (false positive; the wrong prediction of a backorder item) and type II errors (false negative; the wrong prediction of an item that is not on the backorder list) are equal. However, this is an erroneous assumption in the inventory backorder prediction model because different benefits and costs are produced by different classifications in a confusion matrix (Table 1). The benefit from TP classification is zero because this item was not sold due to lack of demand, whereas items in the TN category produce a profit from revenue (sales margin). Similarly, FN items generate warehousing and inventory costs (demand is not present), while no cost is generated when incorrectly classifying the items as FP.

To consider these benefits and costs in the objective function of the Random Forest classifier, we use a profit-based classification measure π defined as follows [57]:

$$\pi = b_0 \times \mathrm{TNR} - c_1 \times \mathrm{FNR}, \tag{1}$$

where TNR = TN / (TN + FP) and FNR = FN / (FN + TP). For the sake of simplicity and improve the interpretability of the results, b_0 and c_1 are assumed to be constant over all the items in the dataset. Another advantage of this measure is that it can be expressed in relative terms (as a percentage of the item price), which enables the easy comparison of the results obtained by different prediction systems over different datasets.

The modified CBUS can be defined as follows:

1. Cluster the instances from the minority (backorder) class using the k-means algorithm into k clusters.
2. Undersample the majority (non-backorder) class by selecting the same number of instances as in the clusters of minority class instances. The selection is performed using the Euclidean distance to the cluster centers. If the same instance from the majority class is assigned to multiple clusters, the closest cluster is selected and additional instances are selected to balance the remaining clusters.
3. Train the Random Forest classifier to maximize the profit-based classification measure π on the balanced training dataset.
4. Optimize the cutoff value. The optimal cutoff is determined using the genetic algorithm as the value with the maximum profit measure π.

The computational complexity of the modified CBUS is given by the sum of the k-means algorithm complexity $O({n_{\mathrm{minor}}}^{m\times(k+1)} \times \log(n_{\mathrm{minor}}))$ and the classifier complexity $O(n_{\mathrm{trees}} \times m_s \times n \times \log(n))$, where n is the number of instances, n_{minor} is the number of instances in the minority class, k is the number of clusters produced by the k-means algorithm, m is the number of features, m_s is the number of randomly sampled features by the Random Forest and n_{trees} is the number of decision trees in the Random Forest. The complexity of the original CBUS approaches is $O({n_{\mathrm{minor}}}^{m\times(k+1)} \times \log(n_{\mathrm{minor}}) + k \times m \times n_{\mathrm{minor}} \times \log(n_{\mathrm{minor}}))$ [52] and $O({n_{\mathrm{major}}}^{m\times(k+1)} \times \log(n_{\mathrm{major}}) + m \times n \times \log(n))$ [44], respectively. Therefore, the classifier complexity in the proposed model increases by about n_{trees}. However, it can still be considered to be computationally effective because compared with C4.5, no pruning is necessary to avoid overfitting and m_s can also substantially decrease complexity when high feature reduction is achieved.

Overall, the pseudo-code of the proposed profit-max CBUS algorithm can be expressed as follows:

Algorithm 1 Profit-Max CBUS

Input: training set T, $T = N \cup P$, the number of clusters k, sales margin b_0, warehousing and inventory cost c_1, the number n_{trees} of decision trees in RF
Output: trained profit function maximizing prediction system Cluster P into k clusters;
For $j = 1$ to k {
  Select n_{minor} instances from N using minimum distance from the j-th cluster center;
}
For $b = 1$ to n_{trees} {
  Create an m_s-dimensional bootstrap replicate T_b from the undersampled T;
  Construct a base learner L_b on T_b by maximizing π;
}
Combine base profit function maximizing learners L_b, $b = 1,2,\ldots, n_{\mathrm{trees}}$ into RF by majority voting;
Determine the optimal cutoff value of RF by maximizing π;

IV. DATA

This study validates the profit function-maximizing inventory management policy by applying a real-world material backorder big dataset. It comes from the Kaggle's contest, namely *Can You Predict Product Backorders*? (https://github.com/rodrigosantis1/backorder_prediction). It is a highly skewed dataset having 1:137 imbalance ratios.

It carries 13,981 positive samples (i.e., the material goes on backorder), while 1,915,954 negative samples are not in the backorder position. It also has 22 attributes, including a class variable. The data traits reveal that about 99.28% of inventory is ready for sale and only a minor proportion (0.72%) goes on backorder.

The example dataset belongs to the historical sample for the eight weeks before the week that this study is trying to classify. The instances were acquired as a weekly snapshot at the beginning of each week. The dataset includes the following attributes: item identification (stock keeping unit, sku; used to track inventory levels), present inventory size (*national_inv*), registered transit time (*lead_time*), item quantity in transit (*in_transit_qty*), sales forecasts (*forecast_3_month, forecast_6_month, forecast_9_month*), prior sales volumes (*sales_1_month, sales_3_month, sales_6_month, sales_9_month*), minimum amount of stock recommended (*min_back*), source issue for item identified (*potential_issue*), amount of overdue from source (*pieces_past_due*), prior source performance (*perf_6_month_avg, perf_12_month_avg*), amount of stock overdue (*local_bo_qty*), risk flags (*deck_risk*, operating entities constraint *oe_constraint*, production part approval process risk *ppap_risk*, *stop_auto_buy, rev_stop*) and a response variable (*went_on_backorder*). Table 2 presents the basic descriptive statistics of the dataset. The dataset was pre-processed through attribute standardization and missing value imputation using the fuzzy k-means algorithm, as recommended by [58] under general assumptions.

TABLE 2. Basic descriptive statistics of the dataset.

Attribute	Description	Frequency (nominal) / Mean value (numeric)
sku	item identification	1,929,935 distinct values
national_inv	present inventory size	496.57
lead_time	registered transit time	7.88
in_transit_qty	item quantity in transit	43.06
forecast_3_month	sales forecasts for the next 3, 6 and 9 months	178.54
forecast_6_month		345.47
forecast_9_month		506.61
sales_1_month	sales volumes for the prior 1, 3, 6 and 9 months	55.37
sales_3_month		174.66
sales_6_month		341.57
sales_9_month		523.58
min_back	minimum amount of stock recommended	52.78
potential_issue	source issue for item identified	989/1,928,946
pieces_past_due	amount of overdue from source	2.02
perf_6_month_avg	source performance for prior 6 and 12 months	0.78
perf_12_month_avg		0.78
local_bo_qty	amount of stock overdue	0.65
deck_risk	risk flags	435,453/1,494,482
oe_constraint		292/1,929,643
ppap_risk		232,552/1,697,383
stop_auto_buy		1,859,391/70,544
rev_stop		839/1,929,096
went_on_backorder	item went on backorder or not	13,981/1,915,954

V. EMPIRICAL EVALUATION

This section illustrates the adoption of the proposed profit function-maximizing inventory backorder prediction system.

A. EXPERIMENTAL SETTING

The experiments were performed to empirically demonstrate both the economic effectiveness of the proposed system and its prediction performance compared with state-of-the-art approaches to the imbalanced class distribution problem. Consistently with previous studies, we consider the area under the ROC (Receiver Operating Characteristics) to evaluate prediction performance. In addition, the proposed profit-based measure is used for economic evaluation. To obtain reliable performance estimates, the data were divided into training and testing sets using 10-fold cross-validation.

In the first step, the backorder instances (minority class) in the training sets were clustered using the k-means algorithm (the Statistics and Machine Learning Toolbox™function kmeans was used to find the solution). As indicated by [52], to avoid model overfitting it is advisable to use a small number of clusters in the k-means algorithm. Following this recommendation, we set the number of clusters to two. As shown below, we also examined a larger number of clusters, but the performance was not improved. In the second step, the corresponding non-backorder instances were assigned to the two clusters using the minimum Euclidean distance to the cluster centers. In the third step, the Random Forest classifier was trained to maximize the profit-based classification measure π. To determine the values of the parameters in (1), we used the industry average values (without financials) reported in [59] and [60]. Specifically, the sales margin was set to $b_0 = 8.01$ % [59] and the warehousing and inventory cost to $c_1 = 25\%$ [60]. Hence, the profit-based classification measure π was calculated as,

$$\pi = 0.0801 \times \text{TNR} - 0.25 \times \text{FNR}. \tag{2}$$

This measure was also used as the fitness function for all the compared methods.

B. EXPERIMENTAL RESULTS

Figure 3 and Figure 4 depict the effect of the number of clusters on the performance of the proposed prediction system on the testing data. To compare the performance of the system with competitive methods, alternative classifiers were used, including traditional Logistic Regression (LR) and k-Nearest Neighbor (kNN) classifiers, as well as the classifiers used in earlier studies of CBUS [44], [52]: C4.5 Decision Tree, Support Vector Machine (SVM) and Multi-Layer Neural Network (NN). For all the classifiers, we used their

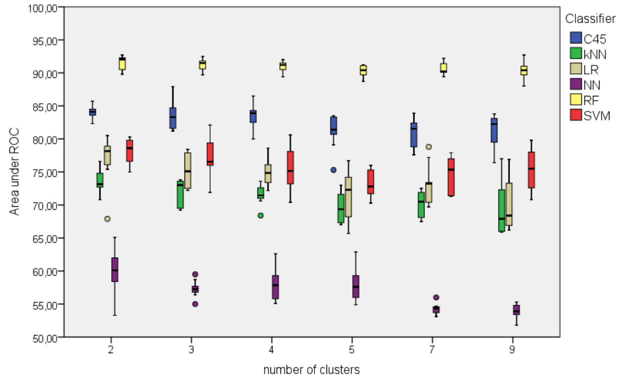

FIGURE 3. The effect of the number of clusters in *k*-means on the performance of the classifiers in terms of area under ROC.

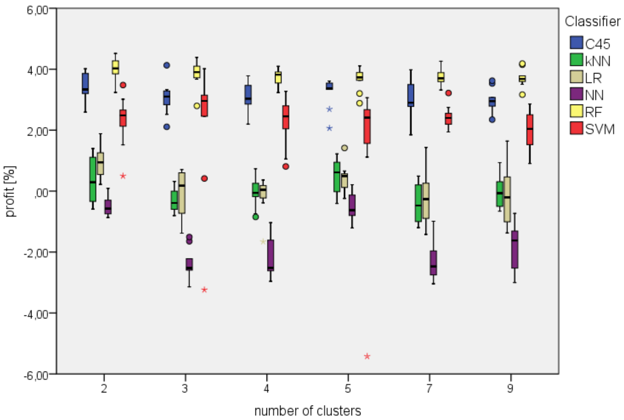

FIGURE 4. The effect of the number of clusters in *k*-means on the performance of the classifiers in terms of profit.

implementation in the Weka 3.8.1 program environment due to the simple implementation of the profit-based measure. Figure 3 and Figure 4 show that the best performance was obtained by the Random Forest classifier, with an average area under ROC of 91.57 and expected profit $\pi = 4.00\%$ for the two clusters. The C4.5 decision tree was the second-best choice, whereas the remaining classifiers performed relatively poorly in terms of both performance measures. In this first set of experiments, the decision cutoff was set to 0.5 for the sake of comparability. In other words, it was not optimized for any of the tested classifiers.

To statistically compare the results, we employed the Wilcoxon signed-rank test. The results of this test confirmed two points. First, significant differences were found between the performance of the Random Forest classifier and those achieved by the remaining classifiers at $P < 0.05$ for both performance measures irrespective of the number of clusters. Second, the Random Forest performed significantly better when only two clusters were used to produce a balanced training dataset.

To demonstrate the robustness of the proposed model to variation in both sales margin and warehousing/inventory cost, the expected profit was calculated for different values of b_0 and c_1, respectively. More precisely, we examined the sensitivity of the profit-based measure to both sales margin $b_0 = \{4\%, 6\%, \ldots, 12\%\}$ for $c_1 = 25\%$ and warehousing /inventory cost $c_1 = \{15\%, 20\%, \ldots, 35\%\}$ for $b_0 = 8\%$. These experiments were performed to simulate real-world inventory management scenarios. The results in Figure 5 show that the profit-max CBUS with Random Forest performed the best under all the scenarios considered here, providing a positive profit function even for low values of sales margin.

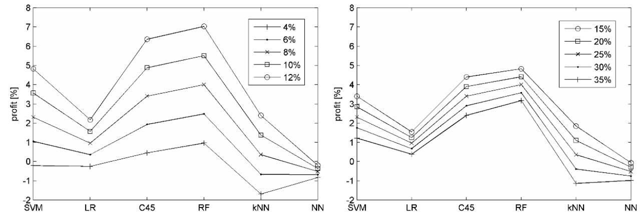

FIGURE 5. Robustness of the classifiers to variation in sales margin (left) and warehousing/inventory cost (right).

To demonstrate the robustness of the proposed model to variation in both sales margin and warehousing/inventory cost, the expected profit was calculated for different values of b_0 and c_1, respectively. More precisely, we examined the sensitivity of the profit-based measure to both sales margin $b_0 = \{4\%, 6\%, \ldots, 12\%\}$ for $c_1 = 25\%$ and warehousing /inventory cost $c_1 = \{15\%, 20\%, \ldots, 35\%\}$ for $b_0 = 8\%$. These experiments were performed to simulate real-world inventory management scenarios. The results in Figure 5 show that the profit-max CBUS with Random Forest performed the best under all the scenarios considered here, providing a positive profit function even for low values of sales margin. Interesting behavior patterns can be observed for *k*NN and NN. The performance of *k*NN regarding the profit function is highly sensitive to the change in c_1 and b_0 values. In contrast to *k*NN, the performance of NN remains stable regardless of the effect of the profit function components. Similar behavior can be observed in previous studies on cost-sensitive learning. The high sensitivity of *k*NN can be explained by the fact that the k nearest neighbors represent only a small subset of the training data space [61]. Alternative k values or distance functions for the identification of the nearest neighbors are recommended to control this behavior [61]. The consistently poor performance of NN can be attributed to the use of the undersampling method. Indeed, training cost-sensitive NNs with undersampling is reportedly worse compared with the other skewed data learning methods, such as oversampling or threshold moving [62].

To further examine the effect of CBUS on the performance of the classifiers, we compared the results from the first set of experiments with those obtained using alternative approaches to the imbalanced class distribution problem, namely SMOTE [48], RUS [47], EasyEnsemble [63] and XGBoost [64]. The original dataset (without any manipulation) was used as the baseline approach. The balanced training datasets obtained by SMOTE and RUS were classified using the same classifiers as the proposed profit-max CBUS. Experiments with SMOTE (with the five nearest neighbors) and RUS were performed in Weka 3.8.1. EasyEnsemble employs C4.5 as its base classifier by default and its implementation in the Keel 3.0 program environment with four bags (and 10 classifiers in each bag) was used in this study. XGBoost was trained using 100 base classifiers with a maximum depth of 5 in the Knime 3.7.1 open platform program environment.

Table 3 shows that XGBoost and RUS+RF performed statistically similar to the proposed prediction system in terms of area under ROC, suggesting that (1) ensemble learning was more effective than individual classifiers and (2) undersampling was more effective than oversampling or no manipulation.

TABLE 3. Area under ROC for the compared approaches to the imbalanced class problem.

Classifier	profit-max CBUS	no manipul.	SMOTE	RUS	Easy-Ensemble
C45	84.01±0.94	69.75±4.33	79.02±2.69	64.70±1.79	89.76±1.06
*k*NN	73.59±1.56	65.51±1.10	66.99±7.04	67.27±1.93	-
LR	77.08±3.39	71.50±1.55	71.17±1.45	79.58±1.51	-
NN	60.12±3.14	55.18±1.83	56.27±1.72	57.14±1.86	-
SVM	78.32±1.75	52.71±1.76	62.29±7.04	75.96±5.07	-
RF	**91.57±0.96**	89.04±1.16	89.91±0.76	90.89±1.09*	-
XGBoost	-	90.79±0.88*	-	-	-
C45	84.01±0.94	69.75±4.33	79.02±2.69	64.70±1.79	89.76±1.06

* significantly similar at $P < 0.05$ as the best performer (in bold).

Regarding the profit-based measure presented in Table 4, no other approach achieved statistically similar performance to that of the proposed prediction system. Indeed, most classifiers performed well only in the majority class, resulting in poor profit-based performance. Only EasyEnsemble and XGBoost performed well in terms of the profit measure, providing additional evidence for the effectiveness of ensemble-based approaches in predicting inventory backorders.

TABLE 4. Expected profit [%] for the compared approaches to the imbalanced class problem.

Classifier	profit-max CBUS	no manipul.	SMOTE	RUS	Easy-Ensemble
C45	3.40±0.43	-14.82±0.41	-12.09±0.61	-8.10±0.62	2.45±0.70
*k*NN	0.49±0.52	-16.45±0.20	-15.56±0.29	-10.22±0.83	-
LR	0.96±0.46	-16.96±0.06	-16.74±0.11	1.16±0.57	-
NN	-0.52±0.29	-17.00±0.00	-17.00±0.00	-6.61±4.87	-
SVM	2.44±0.34	-14.73±1.47	-9.68±4.26	0.83±3.33	-
RF	**4.00±0.38**	-15.71±0.31	-12.25±0.46	-10.17±0.67	-
XGBoost	-	2.15±0.76	-	-	-
C45	3.40±0.43	-14.82±0.41	-12.09±0.61	-8.10±0.62	2.45±0.70

To show the computational effectiveness of the proposed method, training times for the compared methods are presented in Table 5. Note that only the best methods (in terms of profit) were selected from all the approaches to the imbalanced class problem. The Wilcoxon signed-rank test indicates

TABLE 5. Training times for the compared approaches.

Classifier	Training time (in seconds)
SMOTE+SVM	66.61±1.72
RUS+LR	10.89±0.45
EasyEnsemble	13.92±0.42
XGBoost	11.39±0.42
profit-max CBUS	**0.76±0.01**

that our method was significantly faster than the compared approaches.

Recall that the decision cutoff was not optimized in the previous experiments (it was set to 0.5 by default). However, as demonstrated in related business decision support systems [20], [57], the optimization of the decision cutoff may substantially improve the performance of classifiers. This can be demonstrated on the example of one of the training/testing data partition. Figure 6 shows the trade-off between the TNR and FNR associated with different decision thresholds. The optimal backorder strategy can be found between two extremes, namely "backorder everything" for cutoff=0 (at the cost of high warehousing and inventory costs) and "backorder nothing" for cutoff=1 (at the cost of a low sales margin). To find the optimal strategy, the profit-based measure can be used, as indicated in Figure 7. For this specific data partition, the optimal cutoff value was greater than 0.5, suggesting that a more conservative backorder strategy is preferable in this case.

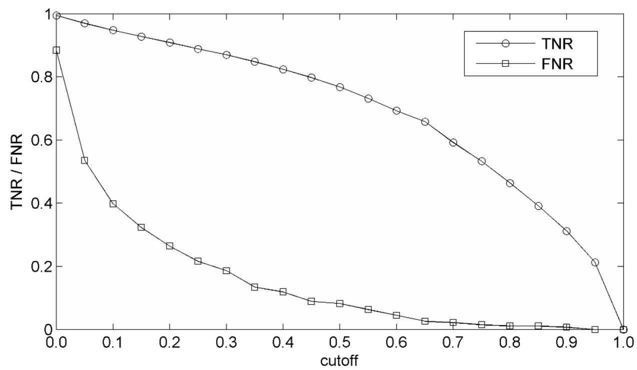

FIGURE 6. The effect of the cutoff on TNR and FNR.

To assist decision-makers with the selection of the cutoff, a search procedure was used to maximize the profit-based measure. More precisely, a standard genetic algorithm (with a population of 20 individuals, crossover rate of 0.8 and mutation rate of 0.01; trained in the xlOptimizer environment) was employed to search the space of cutoff points using the profit measure π as the fitness function. The optimal cutoff point over the 10 training datasets was 0.548±0.076, ranging from 0.401 to 0.661. Figure 8 shows that the average expected profit obtained after this cutoff optimization procedure increased to 4.14%, significantly outperforming the

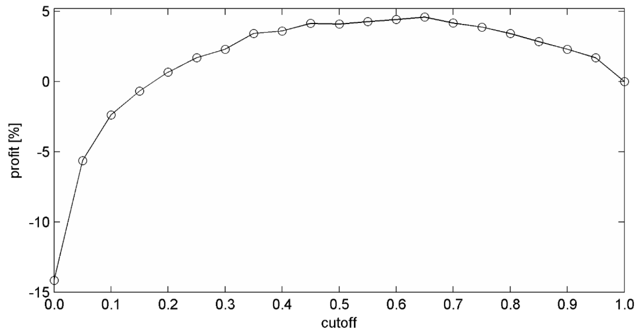

FIGURE 7. The effect of the cutoff on expected profit.

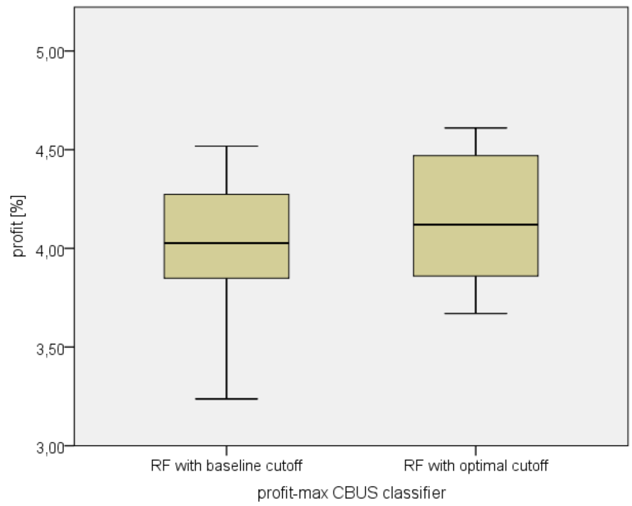

FIGURE 8. The effect of decision cutoff optimization.

prediction system with the baseline cutoff (at $P < 0.05$ using the Wilcoxon signed-rank test).

To test the robustness of the profit-based CBUS classifier to variations in present inventory size and registered transit time, the values of both variables were converted to deciles and average profit measure was calculated for each decile (Figure 9). As expected, higher profit measure can be achieved for larger values of present inventory size and registered transit time because the percentage of backorders decreases with increasing inventory size and lead time, respectively.

Based on the above experimental results and discussion, the study findings are as follows:

a) Irrespective of the number of clusters, Random Forest derived inventory backorder prediction system generates robust results that provide area under ROC curve of 91.57 and 4% of economic benefits. In contrast, NN was the worst choice in terms of the two criteria.

b) The varying sales margin and inventory cost ascertain the reliability of the proposed profit-max CBUS technique. The proposed system is also robust to variations in present inventory size and registered transit time.

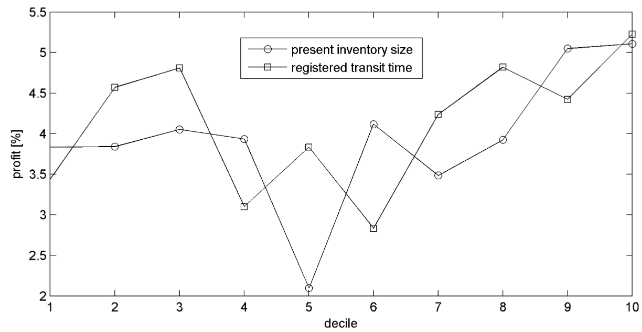

FIGURE 9. Robustness of the profit-max CBUS classifier to variations in present inventory size and registered transit time.

c) Undersampling is the most useful big inventory data balancing system in the current settings compared with other competitors, such as SMOTE and EasyEnsemble, indicating that the proposed profit-max CBUS enhances the competence of our prediction system.

d) The proposed prediction system also ensures the training efficiencies by generating the lowest computational time.

Thereby, the current application provides quantitative insights into the profit function-maximizing inventory backorder prediction system using big data analytics. It adds several novel points in the production and accounting domains to show the financial merits of optimal backorder strategy, and it provides a roadmap about how a firm can generate monetary gains by implementing an effective material backorder forecasting system.

VI. CONCLUSION

In contemporary inventory management systems, it is crucial to develop and optimize all stages of production as markets are extremely competitive. Material backorder is an intrinsic trait of inventory systems under unpredictable demand and pertinent supply risk. Inventory-associated profit and cost constitute a major part of managing supply chain profitability, highlighting the interest in inventory backorder prediction problems. In related literature, irrespective of the rich body of studies in inventory management that inspects different aspects of material backorders, shortages or stock-outs, no studies have thus far pointed out the data-driven prediction of inventory backorders based on profitability maximization.

Based on the above scenarios, a profit function-maximization methodology was incorporated into an inventory backorder prediction system that optimizes the economic effects of backorder decisions. The proposed approach followed four steps. First, a modified version of the CBUS method was established that applies the k-means algorithm to balance inventory backorder instances. Second, a profit-based classification measure was ascertained that trades off the inventory backorder benefits and costs. After that, machine learning algorithms were trained on the balanced training dataset or cluster-specific classifiers were used to maximize the accuracy of the given data subsets. Finally, a genetic algorithm-based search procedure was finally applied to optimize the profitability measure.

To compare the performance of the system with other methods, alternative machine learning classifiers were used. To further examine the effect of CBUS on the performance of the classifiers, we compared the results obtained by the proposed prediction model with those given by alternative approaches to the imbalanced class distribution problem. The profit function-maximizing methodological setup was validated using a real-world inventory backorder imbalanced big dataset. Hence, we demonstrated the effectiveness of the proposed approach.

Our findings suggest that in the absence of decision cutoff optimization, the best performance is shown by the Random Forest classifier, while the C4.5 Decision Tree is the second-best choice. This finding is consistent with experimental findings on imbalanced big data in other business domains. Specifically, Random Forest significantly outperformed SVM in insurance big data analysis [56], as well as C4.5 Decision Tree in credit scoring [20] and customer churn prediction [65].

Considering the CBUS performance over the classifiers, XGBoost and RUS+RF show statistically similar performance, signifying that ensemble learning is more efficient than individual learners and undersampling is more successful than its counterparts. Furthermore, the profit measurement shows that EasyEnsemble and XGBoost perform well, generating additional evidence for the usefulness of ensemble-based approaches for predicting inventory backorders. For the problem discussed in this study, the feasible cutoff point is above 0.5, advocating that a more conservative backorder strategy is desirable for the company. Lastly, applying the cutoff optimization procedure, the profit measure under the inventory backorder prediction system increased compared with the default cutoff threshold. This finding suggests that the expected profit from a backorder strategy can be optimized using the proposed system.

The proposed profit function-maximizing inventory backorder prediction system may add several insights into the accounting and inventory management literature. This study adopts the profit metric as an appraisal measure for the validation process rather than as an intention to be optimized in the training of the classifier, which corresponds to a significant contribution, particularly in the inventory backorder prediction domain. In addition, our methodology contributes by modifying CBUS-based imbalance learning and integrates it into the profit metric, which overcomes the skewed data problem by maximizing its profitability. The findings of this study can assist managers and stakeholders to formulate a balanced decision on inventory backorder assessment based on dissimilar data traits. Further, the method applied in this study not only enhances the accuracy and efficiency of supply chain operations but also supports stable production systems.

The experiments performed in this study were investigated using static inventory policies, assuming stationary customer demand. However, as indicated in the big data analytics framework for inventory backorder management, the inventory policy can be easily changed in real-time due to the computational efficiency of the proposed backorder prediction system. In addition, the use of the IoT-based system enables effective inventory control under a dynamic environment of changing customer orders [66]. Further studies, which use longer time-series data, will need to be undertaken to investigate the proposed system under dynamic inventory policy. A huge volume of spatiotemporal data generated by the IoT-based system represents another significant challenge posed upon machine learning models. Other notable source for usable datasets represents the Retail Product Stockouts Prediction problem available on GitHub.

Future research should also aim to replicate the current methodology but applying additional example sets to diverse inventory portfolios. The research could also consider the sensitivity of the profit objective. Moreover, additional investigation is to be done in this area, specifically in terms of the assumptions made about how the revenue/loss factors of expected profit are generated as well as how they vary over prediction horizons. The proposed profit metric can also be further elaborated. For example, both the sales margin and the warehousing and inventory cost can be calculated in an item-dependent way (i.e., for each item separately) to improve its accuracy. Furthermore, the compensations of customers for their pending orders can be incorporated [67]. Finally, it might be promising to evaluate the performance of genetic algorithm-derived classifiers using other advanced optimization models such as salp swarm algorithm or grey wolf optimizer.

REFERENCES

[1] A. Srivastav and S. Agrawal, "Multi-objective optimization of hybrid backorder inventory model," *Expert Syst. Appl.*, vol. 51, pp. 76–84, Jun. 2016.

[2] D. Li and X. Wang, "Dynamic supply chain decisions based on networked sensor data: An application in the chilled food retail chain," *Int. J. Prod. Res.*, vol. 55, no. 17, pp. 5127–5141, Sep. 2017.

[3] T.-Y. Lin, "An economic order quantity with imperfect quality and quantity discounts," *Appl. Math. Model.*, vol. 34, no. 10, pp. 3158–3165, Oct. 2010.

[4] A. H. Tai, Y. Xie, W. He, and W.-K. Ching, "Joint inspection and inventory control for deteriorating items with random maximum lifetime," *Int. J. Prod. Econ.*, vol. 207, pp. 144–162, Jan. 2019.

[5] Y. Rong, Z. Atan, and L. V. Snyder, "Heuristics for base-stock levels in multi-echelon distribution networks," *Prod. Oper. Manage.*, vol. 26, no. 9, pp. 1760–1777, Sep. 2017.

[6] A. Oroojlooyjadid, L. Snyder, and M. Takáč, "Stock-out prediction in multi-echelon networks," 2017, *arXiv:1709.06922*. [Online]. Available: http://arxiv.org/abs/1709.06922

[7] Q. Zhou, R. Han, T. Li, and B. Xia, "Joint prediction of time series data in inventory management," *Knowl. Inf. Syst.*, vol. 61, no. 2, pp. 905–929, Nov. 2019.

[8] R. B. de Santis, E. P. de Aguiar, and L. Goliatt, "Predicting material backorders in inventory management using machine learning," in *Proc. IEEE Latin Amer. Conf. Comput. Intell. (LA-CCI)*, Nov. 2017, pp. 1–6.

[9] T. M. Choi, S. W. Wallace, and Y. Wang, "Big data analytics in operations management," *Prod. Oper. Manag.*, vol. 27, no. 10, pp. 1868–1883, 2018.

[10] R. Addo-Tenkorang and P. T. Helo, "Big data applications in operations/supply-chain management: A literature review," *Comput. Ind. Eng.*, vol. 101, pp. 528–543, Nov. 2016.

[11] G. Wang, A. Gunasekaran, E. W. T. Ngai, and T. Papadopoulos, "Big data analytics in logistics and supply chain management: Certain investigations for research and applications," *Int. J. Prod. Econ.*, vol. 176, pp. 98–110, Jun. 2016.

[12] R. Dubey, A. Gunasekaran, S. J. Childe, T. Papadopoulos, Z. Luo, S. F. Wamba, and D. Roubaud, "Can big data and predictive analytics improve social and environmental sustainability?" *Technol. Forecasting Social Change*, vol. 144, pp. 534–545, Jul. 2019.

[13] S. W. Shinn and H. Hwang, "Optimal pricing and ordering policies for retailers under order-size-dependent delay in payments," *Comput. Oper. Res.*, vol. 30, no. 1, pp. 35–50, Jan. 2003.

[14] S. Minner, "A note on how to compute economic order quantities without derivatives by cost comparisons," *Int. J. Prod. Econ.*, vol. 105, no. 1, pp. 293–296, Jan. 2007.

[15] M. Y. Jaber and M. A. Rosen, "The economic order quantity repair and waste disposal model with entropy cost," *Eur. J. Oper. Res.*, vol. 188, no. 1, pp. 109–120, Jul. 2008.

[16] H. Mokhtari, "Economic order quantity for joint complementary and substitutable items," *Math. Comput. Simul.*, vol. 154, pp. 34–47, Dec. 2018.

[17] A. A. Taleizadeh, D. W. Pentico, M. Aryanezhad, and S. M. Ghoreyshi, "An economic order quantity model with partial backordering and a special sale price," *Eur. J. Oper. Res.*, vol. 221, no. 3, pp. 571–583, Sep. 2012.

[18] J.-T. Teng, J. Min, and Q. Pan, "Economic order quantity model with trade credit financing for non-decreasing demand," *Omega*, vol. 40, no. 3, pp. 328–335, Jun. 2012.

[19] P.-S. Ting, "Comments on the EOQ model for deteriorating items with conditional trade credit linked to order quantity in the supply chain management," *Eur. J. Oper. Res.*, vol. 246, no. 1, pp. 108–118, Oct. 2015.

[20] M. Papouskova and P. Hajek, "Two-stage consumer credit risk modelling using heterogeneous ensemble learning," *Decis. Support Syst.*, vol. 118, pp. 33–45, Mar. 2019.

[21] L. Zhou, "Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods," *Knowl.-Based Syst.*, vol. 41, pp. 16–25, Mar. 2013.

[22] J. Jurgovsky, M. Granitzer, K. Ziegler, S. Calabretto, P.-E. Portier, L. He-Guelton, and O. Caelen, "Sequence classification for credit-card fraud detection," *Expert Syst. Appl.*, vol. 100, pp. 234–245, Jun. 2018.

[23] H. Jalali, R. Carmen, I. Van Nieuwenhuyse, and R. Boute, "Quality and pricing decisions in production/inventory systems," *Eur. J. Oper. Res.*, vol. 272, no. 1, pp. 195–206, Jan. 2019.

[24] M. Soysal, M. Çimen, S. Belbağ, and E. Toğrul, "A review on sustainable inventory routing," *Comput. Ind. Eng.*, vol. 132, pp. 395–411, Jun. 2019.

[25] K. Rosling, "Optimal inventory policies for assembly systems under random demands," *Oper. Res.*, vol. 37, no. 4, pp. 565–579, Aug. 1989.

[26] N. D. Van Foreest, R. H. Teunter, and A. A. Syntetos, "Base-stock policies with reservations," *Omega*, vol. 81, pp. 48–56, Dec. 2018.

[27] N. Brahimi and T. Aouam, "Multi-item production routing problem with backordering: A MILP approach," *Int. J. Prod. Res.*, vol. 54, no. 4, pp. 1076–1093, Feb. 2016.

[28] A. S. Xanthopoulos, A. Kiatipis, D. E. Koulouriotis, and S. Stieger, "Reinforcement learning-based and parametric production-maintenance control policies for a deteriorating manufacturing system," *IEEE Access*, vol. 6, pp. 576–588, 2018.

[29] Q. Feng, S. Luo, and D. Zhang, "Dynamic inventory–pricing control under backorder: Demand estimation and policy optimization," *Manuf. Service Oper. Manage.*, vol. 16, no. 1, pp. 149–160, Feb. 2014.

[30] H. Wee, J. Yu, and M. Chen, "Optimal inventory model for items with imperfect quality and shortage backordering," *Omega*, vol. 35, no. 1, pp. 7–11, Feb. 2007.

[31] J. Sadeghi, S. M. Mousavi, and S. T. A. Niaki, "Optimizing an inventory model with fuzzy demand, backordering, and discount using a hybrid imperialist competitive algorithm," *Appl. Math. Model.*, vol. 40, nos. 15–16, pp. 7318–7335, Aug. 2016.

[32] Y. Ghiami and P. Beullens, "Planning for shortages? Net present value analysis for a deteriorating item with partial backlogging," *Int. J. Prod. Econ.*, vol. 178, pp. 1–11, Aug. 2016.

[33] J. R. Trapero, M. Cardós, and N. Kourentzes, "Empirical safety stock estimation based on kernel and GARCH models," *Omega*, vol. 84, pp. 199–211, Apr. 2019.

[34] G. C. Souza, "Supply chain analytics," *Bus. Horiz.*, vol. 57, pp. 595–605, Sep./Oct. 2014.

[35] S. Tiwari, H. M. Wee, and Y. Daryanto, "Big data analytics in supply chain management between 2010 and 2016: Insights to industries," *Comput. Ind. Eng.*, vol. 115, pp. 319–330, Jan. 2018.

[36] P. Jonsson and S.-A. Mattsson, "Inventory management practices and their implications on perceived planning performance," *Int. J. Prod. Res.*, vol. 46, no. 7, pp. 1787–1812, Apr. 2008.

[37] M. Downing, M. Chipulu, U. Ojiako, and D. Kaparis, "Advanced inventory planning and forecasting solutions: A case study of the UKTLCS chinook maintenance programme," *Prod. Planning Control*, vol. 25, no. 1, pp. 73–90, Jan. 2014.

[38] P. Chhaochhria and S. C. Graves, "A forecast-driven tactical planning model for a serial manufacturing system," *Int. J. Prod. Res.*, vol. 51, nos. 23–24, pp. 6860–6879, Nov. 2013.

[39] A. T. Gumus, A. F. Guneri, and F. Ulengin, "A new methodology for multi-echelon inventory management in stochastic and neuro-fuzzy environments," *Int. J. Prod. Econ.*, vol. 128, no. 1, pp. 248–260, Nov. 2010.

[40] G. Wang and L. Lei, "Integrated operations scheduling with delivery deadlines," *Comput. Ind. Eng.*, vol. 85, pp. 177–185, Jul. 2015.

[41] W. J. Guerrero, T. G. Yeung, and C. Guéret, "Joint-optimization of inventory policies on a multi-product multi-echelon pharmaceutical system with batching and ordering constraints," *Eur. J. Oper. Res.*, vol. 231, no. 1, pp. 98–108, Nov. 2013.

[42] A. Dolgui, M. K. Tiwari, Y. Sinjana, S. K. Kumar, and Y.-J. Son, "Optimising integrated inventory policy for perishable items in a multi-stage supply chain," *Int. J. Prod. Res.*, vol. 56, nos. 1–2, pp. 902–925, Jan. 2018.

[43] A. L'Heureux, K. Grolinger, H. F. Elyamany, and M. A. M. Capretz, "Machine learning with big data: Challenges and approaches," *IEEE Access*, vol. 5, pp. 7776–7797, 2017.

[44] W.-C. Lin, C.-F. Tsai, Y.-H. Hu, and J.-S. Jhang, "Clustering-based undersampling in class-imbalanced data," *Inf. Sci.*, vols. 409–410, pp. 17–26, Oct. 2017.

[45] H. Yu, X. Yang, S. Zheng, and C. Sun, "Active learning from imbalanced data: A solution of online weighted extreme learning machine," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 4, pp. 1088–1103, Apr. 2019.

[46] S. Ren, B. Liao, W. Zhu, Z. Li, W. Liu, and K. Li, "The gradual resampling ensemble for mining imbalanced data streams with concept drift," *Neurocomputing*, vol. 286, pp. 150–166, Apr. 2018.

[47] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, Boosting-, and hybrid-based approaches," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 4, pp. 463–484, Jul. 2012.

[48] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[49] L. Abdi and S. Hashemi, "To combat multi-class imbalanced problems by means of over-sampling techniques," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 1, pp. 238–251, Jan. 2016.

[50] S.-J. Yen and Y.-S. Lee, "Cluster-based under-sampling approaches for imbalanced data distributions," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 5718–5727, Apr. 2009.

[51] B. Krawczyk, M. Galar, Ł. Jeleń, and F. Herrera, "Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy," *Appl. Soft Comput.*, vol. 38, pp. 714–726, Jan. 2016.

[52] N. Ofek, L. Rokach, R. Stern, and A. Shabtai, "Fast-CBUS: A fast clustering-based undersampling method for addressing the class imbalance problem," *Neurocomputing*, vol. 243, pp. 88–102, Jun. 2017.

[53] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-means clustering algorithm," *J. R. Stat. Soc. Ser. C, Appl. Stat.*, vol. 28, no. 1, pp. 100–108, 1979.

[54] J. R. D. Rego and M. A. D. Mesquita, "Demand forecasting and inventory control: A simulation study on automotive spare parts," *Int. J. Prod. Econ.*, vol. 161, pp. 1–16, Mar. 2015.

[55] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," *IEEE Access*, vol. 2, pp. 652–687, 2014.

[56] W. Lin, Z. Wu, L. Lin, A. Wen, and J. Li, "An ensemble random forest algorithm for insurance big data analysis," *IEEE Access*, vol. 5, pp. 16568–16575, 2017.

[57] T. Verbraken, C. Bravo, R. Weber, and B. Baesens, ''Development and application of consumer credit scoring models using profit-based classification measures,'' *Eur. J. Oper. Res.*, vol. 238, no. 2, pp. 505–513, Oct. 2014.

[58] J. Luengo, S. García, and F. Herrera, ''On the choice of the best imputation methods for missing values considering three groups of classification methods,'' *Knowl. Inf. Syst.*, vol. 32, no. 1, pp. 77–108, Jul. 2012.

[59] A. Damodaran. (2019). *Margins by Sectors*. Accessed: Mar. 21, 2019. [Online]. Available: http://pages.stern.nyu.edu/~adamodar/New_Home_Page/datafile/margin.html

[60] P. P. J. Durlinger, ''Inventory and holding costs,'' Durlinger Consultancy, Posterholt, The Netherlands, Tech. Rep. WP.04.2012, 2015.

[61] S. Zhang, ''Cost-sensitive KNN classification,'' *Neurocomputing*, early access, Aug. 2019, doi: 10.1016/j.neucom.2018.11.101.

[62] Z.-H. Zhou and X.-Y. Liu, ''Training cost-sensitive neural networks with methods addressing the class imbalance problem,'' *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 1, pp. 63–77, Jan. 2006.

[63] X.-Y. Liu, J. Wu, and Z.-H. Zhou, ''Exploratory undersampling for class-imbalance learning,'' *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 2, pp. 539–550, Apr. 2009.

[64] T. Chen and C. Guestrin, ''XGBoost: Reliable large-scale tree boosting system,'' in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794.

[65] Y. Xie, X. Li, E. W. T. Ngai, and W. Ying, ''Customer churn prediction using improved balanced random forests,'' *Expert Syst. Appl.*, vol. 36, no. 3, pp. 5445–5449, Apr. 2009.

[66] M. Tu, M. K. Lim, and M. F. Yang, ''IoT-based production logistics and supply chain system-Part 2: IoT-based cyber-physical system: A framework and evaluation,'' *Ind. Manag. Data Syst.*, vol. 118, no. 1, pp. 96–125, 2018.

[67] I. Masudin, F. R. Jannah, D. M. Utama, and D. P. Restuputri, ''Capacitated remanufacturing inventory model considering backorder: A case study of indonesian reverse logistics,'' *IEEE Access*, vol. 7, pp. 143046–143057, 2019.

PETR HAJEK was born in Duchcov, Czech Republic, in 1980. He received the B.S. and M.S. degrees in economic policy and the Ph.D. degree in system engineering and informatics from the University of Pardubice, Pardubice, Czech Republic, in 2003 and 2006, respectively.

From 2006 to 2012, he was a Senior Lecturer with the Institute of System Engineering and Informatics. Since 2012, he has been an Associate Professor with the Institute of System Engineering and Informatics, Faculty of Economics and Administration, University of Pardubice. He is the author of three books and more than 90 articles. His research interests include soft computing, machine learning, and economic modeling. He was a recipient of the Rector Award for Scientific Excellence, in 2018 and 2019, respectively, and six best paper awards by the international scientific conferences. He is an Associate Editor of four journals.

MOHAMMAD ZOYNUL ABEDIN received the B.B.A. and M.B.A. degrees in finance from the University of Chittagong, Chittagong, Bangladesh, and the D.Phil. in investment theory from the Dalian University of Technology, Dalian, China.

He is currently a Postdoctoral Fellow with the School of Maritime Economics and Management, Dalian Maritime University, Dalian. He is also an Associate Professor with the Department of Finance and Banking, Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh. His research interests include business analytics and computational finance.