Research Article

Classification Performance for Making Decisions about Products Missing from the Shelf

Dimitris Papakiriakopoulos and Georgios Doukidis

ELTRUN, Department of Management Science and Technology, Athens University of Economics and Business, 47 Evelpidon and Lefkados Street, 113 62 Athens, Greece
ELTRUN, Department of Management Science and Technology, Athens University of Economics and Business, 47 Evelpidon and Lefkados Street, 113 62 Athens, Greece

Correspondence should be addressed to Dimitris Papakiriakopoulos, dpap@aueb.gr

The out-of-shelf problem is among the most important retail problems. This work employs two different classification algorithms, C4.5 and naïve Bayes, in order to build a mechanism that makes decisions about whether a product is available on a retail store shelf or not. Following the same classification methods and feature spaces, we examined the classification performance of the algorithms in four different retail chains and utilized ROC curves and the area under curve measure to compare the predictive accuracy. Based on the results obtained for the different retail chains, we identified certain approaches for the development and introduction of such a mechanism in different retail contexts.

1. Introduction

Supply chain management is "an integrative philosophy to manage the total flow of a channel from the earliest supplier of raw materials to the ultimate customer" [1]. However, empty shelves are not unusual in retail stores. Empirical researches in grocery retail industry argue that "out-of-shelf" (OOS) is still a frequent phenomenon [2]. The term "out-of-shelf" is used to describe the situation where consumers do not find the product they wish to purchase on a supermarket shelf, although it is included in the product mix of the store.

The intensity of the OOS varies among retailers and their stores, depending on different factors (e.g., advertisement, seasonality, geographic location). However, an average of 5–10 percent of the products are missing from the shelf [2, 3]. The impact of the OOS problem on sales and profitability has also been recognized; the retailer loses 3.3 percent of the total sales and the product supplier loses up to 4 percent of sales, which in monetary values is interpreted in 3 million Euros a year for an average-size retailer [2, 4]. Other negative

implications are the waste of time for the staff to resolve consumers' out-of-shelf issues and the reduction of the effectiveness for promotional items.

To this end, the out-of-shelf problem is considered as an area with high opportunities, and both retailers and product suppliers are willing to employ management practices and tools in order to downsize the problem and gain the benefits of keeping filled shelves. In doing so, it is important to automatically monitor the on-shelf product availability. A mechanism is thus required that discriminates products in two mutually exclusive groups (classes): (a) products that are available on shelf and (b) products missing from the shelf. The objective of this study is to employ certain classification algorithms as a means of detecting products missing from the shelf in different retail chains. The study is exploratory and applies the same research method in four different retail chains to identify implementation obstacles and therefore suggests approaches that could increase performance. The cornerstone of the study is a mechanism that detects products missing from the shelf. The employed detection mechanism examines all the products on a daily basis, trying to identify "unexpected" changes on the sales pattern for every item merchandized by retail stores. When irregularities are found, the mechanism classifies the products as "out-of-shelf" and reports the findings to the store manager and other participants in the supply chain.

The remainder of this paper is organized as follows. Section 2 briefly presents related works regarding the out-of-shelf problem and the classification algorithms adopted in this work. Section 3 describes the methodology followed and the data available, while Section 4 presents the empirical findings and the results of the study. Finally, Section 5 concludes the paper.

2. Relevant Works

2.1. The Out-of-Shelf Problem

The bullwhip effect in supply chains leads inventories either to overstock or to stock out, due to a variety of reasons [5]. The out-of-shelf extends the stock-out case; if a product is stocked out in a store, it would also not be available on the shelf. However, it is possible for a product to be available in the backroom facilities of the store, but unavailable on the shelf, because the store personnel did not manage to replenish the shelf.

The causes for products missing form the shelf fall into two broad categories: (a) business processes of the store (e.g., poor shelf replenishment process) and (b) combined upstream problems in retail supply chain (e.g., inaccurate forecasting methods). Moreover, empirical studies have shown that the out-of-shelf problem differentiates across different days of the week, different retail stores, and different product categories [2]. The utilized detection mechanism is not able to determine the exact causes, but focuses on the effects regardless of the product category and the retail outlet.

Most of the relevant research efforts are identified in the area of marketing and consumer behavior, addressing the consumers' reaction to the out-of-shelf product. It is not surprising that all studies agree on the negative impact of out-of-shelf on brand, store loyalty, and satisfaction [6–9]. In the long run, consumers were found to have the "momentum" even to switch to another retail outlet if they are constantly not able to find the products they are looking for [4].

Currently, the practice of measuring out-of-shelf is empirical and based on physical store audits. A researcher visits the retail store and reviews on-shelf product availability once (or twice) a day, either by counting the available items on the shelf or by checking if the

item was (or not) found. Since a retail store usually merchandises more than 5.000 different products (also called stock keeping units (SKU)), the process is limited to a small number of selected products (e.g., a hundred). The duration is also limited and may not exceed two weeks. The high cost, the limited time, and the number of items monitored are important barriers for the wide adoption of such process in examining the out-of-shelf problem. Moreover, the accuracy of physical store audits is an important issue for two reasons. The periodic review of product availability provides information for a limited timeslot (e.g., when the researcher is found in front of the shelf) and cannot draw conclusions about the whole day with certainty. Finally, the conduction of store audits affects the replenishment rate of the store's employees, due to the Hawthorne effect. As a result, the reported out-of-shelf rate is expected to be slightly lower.

To this end, physical store audits are costly procedures with an uncertain result. Both product suppliers and retail chains are aware of the aforementioned limitations and are motivated to employ automated procedures with the expectation of getting reliable and uniform results.

A previous work on the out-of-shelf problem utilized classification methods in the context of a single retail chain, pointing out that it is possible to detect the products missing from the shelf with increased accuracy [10]. An automatic mechanism that detects products missing from the shelf needs to answer the following: (a) which is the available information, (b) how is the information transformed to features (independent variables), and (c) which is the induction mechanism that discriminates the product into the two classes (e.g., EXIST and OOS). With this work we extend the number of retail chains in order to identify important and common issues and finally draw approaches from the use of an automatic mechanism for out-of-shelf product detection.

2.2. Classification Algorithms

Classification (also called supervised learning) is a significant issue in the machine learning community. The objective is to deduce classifiers from existing and known training data in order to classify unknown examples. Classifiers are functions which, given any valid input, assign the example to a specific class. Based on the algorithm used, a classifier might be a hyperplane, a decision tree, or even a probabilistic expression. Many classification algorithms have been proposed (e.g., naïve Bayes, random tree, and multilayer Perceptron, etc.). In this work, we are focusing on two important and widely used algorithms. The first is naïve Bayes (NB) and the second is C4.5 [11].

NB utilizes the so-called Bayes theorem to produce classifiers. The classifier C has the form $C : X \rightarrow Y$, where X includes all the independent variables (or also called features) and Y is the class label. The application of the Bayes theorem for a specific class $Y = y_i$ can be represented as

$$P(Y = y_i \mid X = x_j) = \frac{P(X = x_j \mid Y = y_i)P(Y = y_i)}{\sum_k P(X = x_j \mid Y = y_k)P(Y = y_k)}, \tag{2.1}$$

where x_j denotes the j-possible vector for X. The naïve Bayes algorithm uses the available training data to estimate $P(X \mid Y)$ and $P(Y)$, and through this process any new example x that belongs to X could be classified based on the estimation made during the training process. naïve Bayes algorithm assumes conditional independence of the independent variables that

belong to the space of X. Through this assumption all the naïve Bayes classifiers are using the following equation:

$$P(Y = y_i \mid x_1 \cdots x_n) = \frac{P(Y = y_i) \prod_j P(X_j \mid Y = y_i)}{\sum_k P(Y = y_k) \prod_j P(X_j \mid Y = y_k)}.\qquad(2.2)$$

Although feature's independence is a barrier for the literal use of the naïve Bayes, few studies have reported high classification accuracy [12–14].

Another approach for classification problems is provided through the use of decision trees. A decision tree algorithm partitions the available training data into disjoint groups, until all the data belong to a particular class of the problem. A decision tree is a hierarchical sequential structure, consisting of internal nodes that are expressed as values of the features and the leaf nodes that are made up of the class label. C4.5 is one of the most well-studied algorithms proposed by Quinlan [11]. It generates a decision tree using gain ratio criteria for selecting a feature that splits the available data into smaller data sets and therefore participates as an internal node in the decision tree.

The machine learning community has also suggested methods to compare different classification algorithms. Recently, the receiver operating characteristic (ROC) curves have attracted the attention of researches as a means for comparing different classification algorithms. Instead of comparing only the accuracy of the classification, ROC curves provide a graphical illustration of the trade-off between the costs and benefits of a given model [15]. ROC produces a two-dimensional curve where the true-positive rate (TPR) is plotted against the false-positive rate (FPR = 1–TPR) for different cutoff points.

Scholars on the field suggest that when comparing different classification algorithms, the best is the one having the higher ROC curve [16]. On the one hand, when the ROC curve climbs rapidly towards the upper-left corner of the graph, it means that the rate of the true positive detection is much higher than the rate of the false-negative detections. On the other hand, when the progress of the ROC curve is close to the diagonal path, it is understood that every improvement occurring in the true-positive rate is neutralized by a high false-positive rate. Therefore, a ROC curve close to the diagonal line implies that the classification mechanism is characterized by a "random guess" and could not be utilized for predictions. Instead of comparing visual charts, the literature suggests the adoption of the measure called area under curve (AUC), which is calculated as the definite integral of the area between the ROC curve and the x-axis.

3. Methodology

3.1. Research Method

For the purpose of our study, three different retail stores from four different retail chains were selected to participate in the experimentations. The retail stores were suggested by the managers of each corresponding retail chain. In addition, we decided to examine the out-of-shelf problem in vital product categories: (a) coffee (b) chocolates (c) detergents and (d) shampoo and hair care. Although the detection mechanism could cover all different products, the participants expressed the willingness to focus on significant product categories.

Experts from the supplier side suggested 25 different products for every product category, resulting in a list of 100 different items. The list was used to conduct physical

store audits that lasted for two weeks, with the purpose of identifying and measuring the extent of out-of-shelf problem in every different store and retail chain. From a classification perspective, the results of the physical store audit form the labeled examples that could be utilized by the classification methods used for training and testing purposes.

From all retail chains, we acquired 6-month historical data depicting the sales and replenishment transactions of every store. These data were the basis to build the feature space. Note that all the retail chains do not maintain the same data sources; thus, the features finally used for every retail chain were subject to the available data.

By combining the physical store audit with constructed feature space for every retailer, we obtained a usable learning set from the classification algorithms (NB and C4.5). The learning set was used for training and testing purposes, with 10-fold cross-validation process. A comparison of the classification algorithms has been made using the ROC curves. Using the ROC curves as a driver of comparison, important performance issues were identified and the approaches were suggested to bypass limitations and problems.

3.2. Data Availability

The use of classification algorithms relies on the available data. We asked from all the retail chains to identify relevant data collections from their internal information systems. In detail, the requested data sources were as follows:

(i) point-of-sales data is the source that describes the daily sales of the store at SKU level;

(ii) order collection is the requests placed from the store to the central warehouse facilities of the retailing chain;

(iii) delivery quantities include the items sent to the store through the central warehouse;

(iv) promotion plan is a calendar describing the in-store promotion activities held in every store of the retail chain;

(v) product data is the list of products merchandized by the retail chain;

(vi) product assortment is the list of the active products currently available at a specific store.

Through the utilization of the aforementioned sources, a number of features have been extracted as Table 1 depicts. Most of them describe the sales of a product (S_1–S_{11}); others provide estimations about inventory levels of the product (I_1–I_4) and so forth. The type of feature is either numerical (Num) or nominal (Nom).

The features presented in Table 1 were calculated for all the products on a daily basis. Some of them required the existence of historical data (e.g., S_2). During the transformation of the source data into features, new records were added daily and we decided to remove a day from the tail of the historical data, forming a sliding window. Keeping a long history of data in retailing could be a misleading option because the demand of the product is subject to promotional effects and seasonality. In contrast, other features provide an instantaneous picture of the product within the day of issue and do not rely on a certain data history. For example, S_9 presents the day's lag between the most recent sale and the day of issue. Having a fast moving item with important sale's latency is an indicator that the product might not be on the shelf therefore is useful to maintain such a feature.

Table 1: Features developed based on the different sources.

Source	Feature	Description	Type
POS data	S_1	Items sold for the specific day	Num
POS data	S_2	Average number of items sold until the current date	Num
POS data	S_3	Standard deviation of items sold	Num
POS data	S_4	Average items sold for a specific day of week	Num
POS data	S_5	Standard deviation of items sold for a specific day of week	Num
POS data	S_6	Average sales frequency	Num
POS data	S_7	Standard deviation of sales frequency	Num
POS data	S_8	Fast moving index for the day	Num
POS data	S_9	The lag days between the last sale and the current date	Num
POS data	S_{10}	The average duration (in days) between two consecutive sales	Num
POS data	S_{11}	The standard deviation of the duration (in days) between two consecutive sales	Num
Deliveries—Pos Data	I_1	Estimation of inventory level	Num
Orders and Deliveries	I_2	A backorder flag	Nom
Deliveries	I_3	The time interval between the last delivery and the current day of issue	Num
Orders	I_4	Store replenishment method	Nom
Promotion plan Products	P_1	The product is coming from the regular line or it has advertisement properties (e.g., price discount, promotional activity)	Nom
Products	P_2	Groups products either as new or as regular	Nom
POS Data	PC_1	Presents seasonality of the whole product category	Num
Products and POS data	PC_2	Examines the product market share within the product category it belongs to	Num
	C_1	The size of the store	Nom
	C_2	The day of the week	Nom

4. Empirical Analysis

4.1. Retail Chain Profile and Out-of-Shelf Measurements

The common characteristic of all retail chains is the awareness about the out-of-shelf problem and the motive to increase the service level to the consumer, through the employment of an automatic mechanism for the detection of products missing from the shelf. However, retail chains have a different organization structure, maintain different information systems, and have different business processes regarding the store ordering and shelf replenishment.

Retail Chain 1 (RC1) is a national retailer operating 170 stores across the country. RC1 mainly supports small-and medium-size stores and the relevant information system

Table 2: Data availability per retail chain.

Data collection	RC1	RC2	RC3	RC4
Point-of-sales	√	√	√	√
Orders	√	√	√	√
Deliveries		√		√
Promotion plans		√	√	√
Products	√	√	√	√
Product assortment		√	√	√

focuses on the facilitation of store's ordering business process. Retail Chain 2 (RC2) is a multinational retailer with more than 240 stores. Several information systems support the complex organization structure containing different retail store types (e.g., cash & carry, hypermarket). Retail Chain 3 (RC3) is a national retailer, with 74 stores, most of them located in urban areas. The information systems of RC3 have been developed in-house and are tightly integrated with the business processes. Finally Retail Chain 4 (RC4) is a major national retailer present in other countries. The information system supporting store's replenishment and ordering processes is maintained by an external company.

RC3 and RC4 have developed relevant practices to tackle the out-of-shelf problem. In detail, RC3 has employed a business process that is triggered every morning before the store opens. According to the process employed, the store manager reviews shelf availability and when an empty shelf is found, he/she uses a hand scanner to collect all the missing products from the shelf. Afterwards, he/she moves to the backroom facilities of the store and initializes the shelf replenishment process for the product codes stored in the hand scanner. In the case that the store manager could not find any remaining items in the backroom for a specific product, he/she places the product code in the ordering list of the shop. Finally, RC4 has developed a novel information system to support the replenishment process of the store by taking into account several parameters like promotion items, sale's forecast, and inventory status. The RC4's information system is maintained by an external company, specializing in the provision of services within the retail sector.
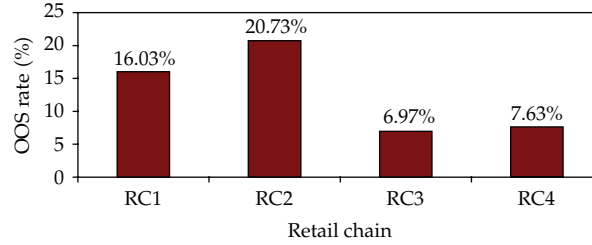
As discussed earlier, we examined on-shelf availability for 100 different products within the scope of all retail chains. For a two-week period, a researcher visited the store at the same time every day and examined whether the products included in the list were available on the shelf or not. By aggregating these physical counts, it was possible to calculate the OOS rate for every retail chain. The next figure illustrates the extension of the OOS problem for the different retail chains.

Based on the empirical data, RC1 and RC2 were found to have an increased out-of-shelf rate, while RC3 and RC4 show a significant decrease. We argue that the underlying reason justifying the significant differences among the retail chains is that RC1 and RC2 have not yet adopted any specific countermeasure to avoid the OOS problem; thus, it is not surprising that they have experienced "unacceptably" high levels of OOS. Other reasons that explain the high OOS rates for RC1 and RC2 are related to strategic and organizational decisions that have been made by every retail chain. For example, RC1 is operating with convenient stores that have limited storage capacity; thus, OOS rates tend to be higher. In general, retailers agree that an acceptable level of OOS rate should be under 2%.

At this point it is possible to identify the dilemma that a retail chain with high OOS rates faces. On the one hand, a retail company could decide to hire extra personnel (in order to increase the fill rate of the shelf) or increase inventory levels across the supply chain (with

Table 3: AUC measure for different retail chains.

Algorithm	AUC-RC1	AUC-RC2	AUC-RC3	AUC-RC4
C4.5	0.751	0.837	0.758	0.689
NB	0.686	0.774	0.692	0.685

Figure 1: Out-of-shelf measure for different retail chains.

the purpose of absurd demand fluctuations) and thus expect to provide service of a higher level to the consumer by increasing the on-shelf availability. On the other hand, a retailer could adopt a technological infrastructure to measure the problem and align the business process related to the shelf/store replenishment. The latter is perceived as a cost-effective and flexible decision, while the former implies a radical organizational redesign with increased implementation cost. Therefore, we expect that retailers with significant OOS rates will tend to invest on technological infrastructures to tackle the problem.

4.2. Classification Performance

The classification algorithms C4.5, and NB, were used with the corresponding learning set of each retail chain. The learning set derived was based on the available data provided by every retail chain, as Table 2 depicts. RC2 and RC4 were able to provide all the requested data collections, while RC1 and RC3 did not keep information for some data collections (e.g., Deliveries). As a result, the feature space of RC1 and RC3 did not include at least the I_1, I_2, and I_3 features (see Figure 1).

Every learning set had about 3.500 on-shelf availability examples, covering both classes of the problem (EXIST and OOS). The learning sets were used with 10-fold cross-validation for training and testing purposes. Based on the measurements obtained by this process, we evaluated ROC curves for every retail chain.

The results of the ROC curves (see Figure 2) revealed that in most of the cases C4.5 dominates the NB algorithm. As can be seen from the figure, the C4.5 ROC curve is higher than the NB's, indicating that the C4.5's classification performance is better than the NB's at different cutoff points.

A comprehensive way to compare ROC curves is conducted through the area under-the-curve (AUC) measure. AUC is useful because it aggregates performance across the entire range of cutoff points. The higher the AUC, the better, with 0.5 indicating random performance and 1.0 denoting perfect performance, thus perfect discrimination. The extracted AUC measures for the two classification algorithms and for all retail chains are presented in Table 3 .

The best predicative performance found was of C4.5. The NB algorithm achieved relatively poor AUC measure, compared with the results of C4.5. The absence of previous
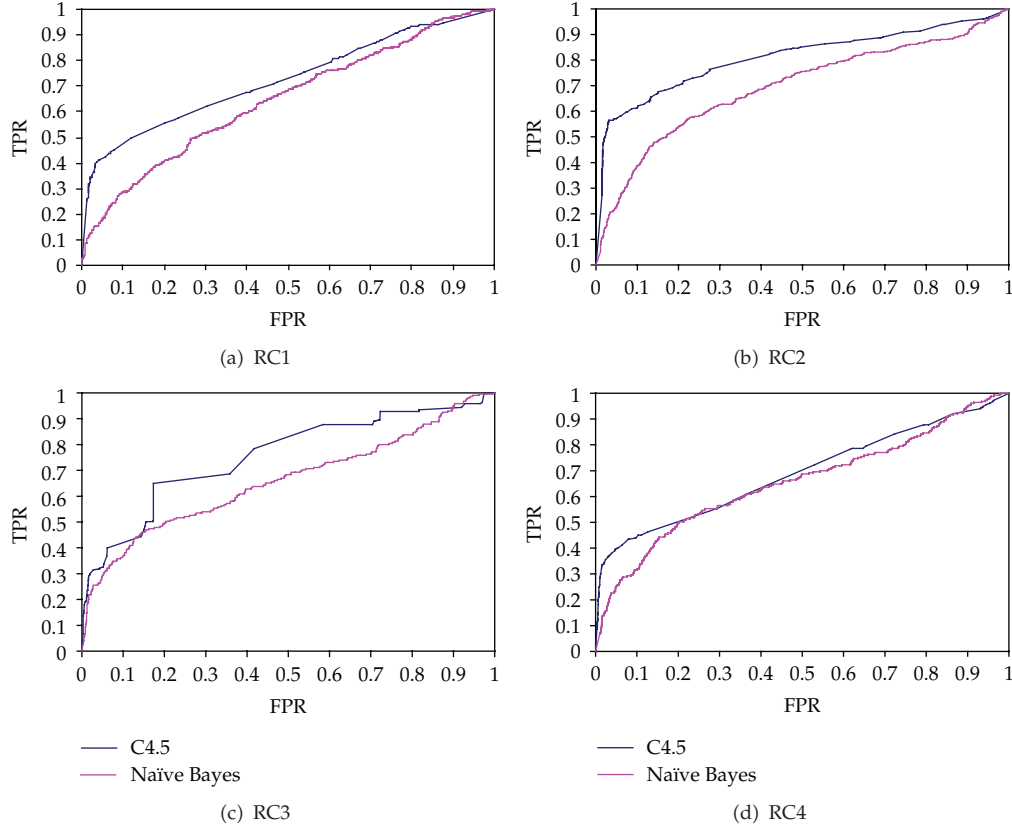
Figure 2: ROC curves for different retail chains.

studies on the detection of products missing from the shelf using AUC curves does not allow a comparative discussion. However, the AUC measure depicts the classification accuracy of the selected algorithms as a whole, regardless of the ability to detect only the products missing from the shelf. In essence, AUC is a good measure for selecting a classification algorithm, but it could not support the decision to select a classification algorithm, when only one class of the problem receives the attention of the research.

Based on the results, it is suggested that an automatic detection mechanism should be developed with C4.5 (or another decision tree algorithm that could perform better than the suggested algorithm). The RC2 was the retailer with the highest OOS rate and achieved the highest classification performance, as compared to the other retail chains (see Figure 2(c) and Table 3). The corresponding ROC curve evolves very fast, and for the threshold of FPR < 0.1, the value of TPR is close to 0.6, meaning that an important amount of OOS cases could be detected accurately. In other words, the rate of correct detections is six times higher than the rate of wrong detections; therefore the application of C4.5 in the specific setting is very promising. Similar conclusions could be drawn based on the AUC measure for the RC2. Moreover, the high data availability of RC2 facilitates the feature space and therefore, allows the decision tree to expand in different ways during the training phase.

The performance of the classification algorithms should be expected to be lower when used in real life environment, due to two important problems. The first one is the class imbalance of the OOS problem that hinders the performance of classifiers. Class imbalances

refer to the situations where the minor class is significantly lower than the major. A recent study by He and Garcia [20] shed light on the severity of the class imbalance, stating that existing widely used techniques for evaluating the classifiers are inadequate and the problem of lowering the predicative accuracy is escalated especially when testing the classifiers that refer to the minority class. On the one hand, having notably more examples for the major one class introduces high bias towards it during the training. On the other hand, the minor class receives a lower number of classifiers with substantial difficulties to evaluate when testing the predictive accuracy. As a result, the classification process produces classifiers that do not accurately predict the minority class [17]. For example, the classification performance yielded by using the same algorithm for the RC3 and RC4, which have the smallest OOS rate, is lower compared to the results obtained for RC1 and RC2. In particular, the RC3 and RC4 encounter the negative effects of the aforementioned problem, while RC1 and RC2 could achieve better classification accuracy because the classifiers of the minor class have more chances to learn accurately.

The second problem has to do with the quality and completeness of the data used, which affects the values of the used features and consequently the performance of the algorithms. A notable example regarding the quality of the data raised for the features related to the inventory records, also discussed in the relevant literature as "inventory inaccuracy" [18]. In more detail, it is very difficult to maintain perfect inventory records at the store level due to various sources of error (e.g., shoplifting, damage of the products during the transportation, etc.). Measuring the inventory levels is therefore not a reliable feature and could not precisely depict the exact situation that occurs for all products in every store. From the completeness of the data source perspective, few inconsistencies were found especially when trying to find which products were active (merchandized) by the store, because in all retail chains the product assortment was not updated by store managers.

4.3. Discussion and Future Research

Finding products that are missing from the shelf with automated means is a very interesting task in the retail sector. The managers are aware of the resource consumption process of store audits; therefore, they would accept automated solutions utilizing existing data and infrastructures. Based on the discussions we had, they suggested that even a mechanism detecting the out-of-shelf product with a 50% accuracy is still valuable, because it could be activated every day with a very small cost, covering all the stores of the retail chain. Moreover, they perceived such a system as an infrastructure to implement in-store business process for the daily review of product shelf availability.

From a classification algorithm perspective, the out-of-shelf is a challenging problem, because (a) the minority class (OOS) has a relatively limited presence in the training/test sets, (b) the class of every example could change from day to day, and (c) the information quality of the data does not always meet the real conditions of the store. The class imbalance problem and the issue of data quality have attracted the attention of scholars in the field [16]. However, we did not find any relative study regarding the change of the class label for the same example over time, which is an important obstacle when performing classification tasks.

As presented in the previous section, the RC2 could "easily" detect a significant number of products missing from the shelf by using a C4.5 decision tree. Retail chains with lower out-of-shelf rates should be able to guarantee the quality of the information sources used. For example, RC3 was found to have the smallest number of inconsistent records, but

at the same time the information quality provided was considered to be high. Using the ROC curve of RC3, it is easy to see that with an FPR < 0.25 it is possible to obtain TPR = 0.64, which approximates the area found RC2. These two solutions could therefore be considered as equivalent. However, in RC3 the AUC measure is lower when compared to RC2, implying that the expected detection opportunities in RC3 are lower than in the case of RC2.

Based on the comparison of the same algorithms to different retail chains, it is possible to identify approaches that could facilitate the process of developing an automatic mechanism for detecting products missing from the shelf. Our study depicted that retail chains facing high out-of-shelf rates and that keep extended data sources have the opportunity to sufficiently build a mechanism for detecting products missing from the shelf. In more detail, we state the following.

(i) Lack of specific management attention usually leads to a high out-of-shelf rate as found in cases RC1 and RC2.

(ii) The RC1 case revealed that having a retail chain with increased out-of-shelf product rate and limited data sources offers good opportunities to employ classification algorithms and discover products missing from the shelf.

(iii) The RC2 case was found to be the most promising one in the successful adoption of the proposed mechanism. The calculated AUC measure was the highest, implying that the specific retail chain would gain important benefits through the adoption of the proposed classification mechanism.

(iv) The RC3 case suggested that even with a low out-of-shelf rate, the opportunity to correctly identify missing products is still at relatively high levels due to information quality. RC3 was considered a well-organized and capable retail chain at a national level with innovative and powerful information systems. Therefore they face the opportunity of capitalizing existing data infrastructures and develop an accurate detection mechanism.

(v) The RC4 case suggests that keeping all business records (regardless of the accuracy) could pave the ground in the adoption of an intelligent system monitoring on-shelf availability. However, the opportunity in this case is considered relatively low, based on the obtained results of the AUC measure.

An important question is whether the aforementioned findings could be generalized efficiently in the context of every retail chain. Further field experimentation is required, where additional classification algorithms would be utilized. Few and rigorous classification algorithms are available in the literature (e.g., Random Forest) and should be incorporated in a large field experimentation. In addition, the inclusion of new retail stores and product categories is an experimental decision that should be considered carefully. By scaling up the experimentation, the generalization ability of the findings is expected to emerge.

The classification performance of all retail chains is expected to be increased by the employment of ensemble methods, which have emerged as a powerful way for improving the predictive performance, by combining many weak classifiers to produce a stronger one. Bagging [19] is a popular ensemble method and could be useful for developing classifiers addressing the out-of-shelf problem. The RC1, RC2, and RC4 ROC curves are found to have an extensive area, where the curve is almost linear. This means that the classification ability of the algorithm has already been "perfected" and the remaining detections are subject to a random guess. Thus, the classifiers employed are weak, and therefore the use of ensemble methods are expected to increase the ability of correct detections and reshape the ROC curves.

It is also noticeable that only a small fraction of the produced classifiers could be transferred between different retail chains. In more detail, we build classifiers in the context of one retail chain and test their predictive accuracy in the available data of a different retail chain. We found that the accuracy level was lower than 15%, implying that there are not rules of thumb governing the out-of-shelf problem. However, we found that the transferability of classifiers is possible between different product categories and stores within the context of the same retail chain.

This is of no surprise because the data used for training the classifiers reflected important decisions made by every retail chain. RC1 and RC3 maintain the same SKU code for a product and its variants (e.g., promotional items), leading to a very precise estimate for all the features ranging from S_1 to S_{11}. The remaining retail chains follow a different category management approach by assigning different SKU codes for a product and its variants. As a result, the frequent changes in a product affect the sales history of the product through the introduction of a long period without selling an item due to the availability of a variant product, leading to unreliable measures regarding sales features. To this end, decisions in the category management process shape the data management practices of a retail chain and consequently form the structure of the data. The use of identical data sources with the same measurement process leaded to different classifiers, which reflected the data management practices followed by the retail chain.

Introducing the same detection mechanism to different retail chains emerged some important questions for further study. The first question is whether to build a centralised or distributed detection mechanism, where every retail store would have its isolated set of classifiers. Initially, with this work, we envisioned a centralised mechanism, but the prospect of distributed data mining approach is challenging. The second question reveals the management practices that need to be adopted to capitalize the capabilities of the automatic detection and the integration of the proposed tool with existing business processes and management corrective actions.

5. Conclusions

We examined the introduction of classification methods in retailing in order to automatically identify products missing from the shelf. From a technical perspective it is possible to use such a mechanism in all retail chains, but the results are significantly different due to several factors, such as the data quality and availability, classification algorithms' performance with class imbalance and extension of the out-of-shelf problem in the context of the retail chain.

Although the suggested approach is promising, different decisions need to be made when introducing the mechanisms into different retail chains. Moreover the generalization of the results and the introduction of the mechanism to unknown retail stores and product categories should be continuously reviewed and validated in order to ensure the predictive performance of the mechanism.

References

[1] M. C. Cooper, L. M. Ellram, J. Gardner, and A. Hanks, "Meshing multiple alliances," *Journal of Business Logistics*, vol. 18, no. 1, pp. 67–89, 1997.

[2] T. Gruen and D. Corsten, *A Comprehensive Guide to Retail Out-of-stock Reduction in the Fast Moving Consumer Goods Industry*, Grocery Manufacturers of America, Food Marketing Institute, National Association of Chain Drug Stores, The Procter & Gamble Company, University of Colorado, Colorado Springs, Colo, USA, 2008.

[3] T. Gruen, D. Corsten, and S. Bharadwaj, *Retail Out-of-Stocks: A Worldwide Examination of Extent Causes and Consumer Responses*, The Food Institute Forum (CIES, FMI, GMA), 2002.

[4] E. T. Andersen, G. J. Fitzsimons, and D. Simester, "Measuring and mitigating the costs of stockouts," *Management Science*, vol. 52, no. 11, pp. 1751–1763, 2006.

[5] H. Lee, V. Padmanabhan, and S. Whang, "The bullwhip effect in supply chains," *Sloan Management Review*, vol. 38, no. 3, pp. 93–102, 1997.

[6] M. Emmelhainz, L. Emmelhainz, and J. Stock, "Consumer responses to retail stock-outs," *Journal of Retailing*, vol. 67, no. 2, pp. 138–147, 1991.

[7] W. Verbeke, P. Farris, and R. Thurik, "Consumer response to the preferred brand out-of-stock situation," *European Journal of Marketing*, vol. 32, no. 12, pp. 1008–1027, 1998.

[8] G. J. Fitzsimons, "Consumer response to stockouts," *Journal of Consumer Research*, vol. 27, no. 2, pp. 249–266, 2000.

[9] D. Corsten and T. Gruen, "Stock-outs cause walkouts," *Harvard Business Review*, vol. 82, no. 5, pp. 26–28, 2004.

[10] D. Papakiriakopoulos, K. Pramatari, and G. Doukidis, "A decision support system for detecting products missing from the shelf based on heuristic rules," *Decision Support Systems*, vol. 46, no. 3, pp. 685–694, 2009.

[11] J. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1993.

[12] P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss," *Machine Learning*, vol. 29, no. 2-3, pp. 103–130, 1997.

[13] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine Learning*, vol. 29, no. 2-3, pp. 131–163, 1997.

[14] M. Mozina, J. Demsar, M. Kattan, and B. Zupan, "Nomograms for visualization of naive Bayesian classifier," in *Knowledge Discovery in Databases: PKDD*, pp. 337–348, Springer, Berlin, Germany, 2004.

[15] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[16] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[17] N. Japkowicz and S. Stephen, "The class imbalance problem: a systematic study," *Intelligent Data Analysis Journal*, vol. 6, no. 5, pp. 429–449, 2002.

[18] Y. Kang and S. B. Gershwin, "Information inaccuracy in inventory systems: stock loss and stockout," *IIE Transactions*, vol. 37, no. 9, pp. 843–859, 2005.

[19] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[20] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.