Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Predicting product return volume using machine learning methods

Hailong Cui[a,*], Sampath Rajagopalan[a], Amy R. Ward[b]

[a] *Marshall School of Business, University of Southern California, Los Angeles, CA 90089, USA*
[b] *Booth School of Business, University of Chicago, Chicago, IL 60637, USA*

ARTICLE INFO

Article history:
Received 25 June 2018
Accepted 22 May 2019
Available online xxx

Keywords:
Analytics
Online returns
Predictive model
Variable selection
LASSO
Machine learning

ABSTRACT

In 2015, U.S. consumers returned goods worth $261 billion and the return rates for online sales sometimes exceeded 30%. Manufacturers and retailers have an interest in predicting return volume to address operational challenges in managing product returns. In this paper, we develop and test data-driven models for predicting return volume at the retailer, product type and period levels using a rich data set comprised of detailed operations on each product, and retailer information. The goal is to achieve a good prediction accuracy out of sample. We consider main effects and detailed interaction effects models using various machine learning methods. We find that Least Absolute Shrinkage and Selection Operator (LASSO) yields a predictive model achieving the best prediction accuracy for future return volume due to its ability to select informative interaction terms out of more than one thousand possible combinations. The LASSO model also turns in consistent performance based on several robustness tests and is easy to implement in practice. Our work provides a general predictive model framework for manufacturers to track product returns.

1. Introduction

In the retail industry consumer returns create a significant and costly issue for manufacturers and retailers. In 2015, consumers returned goods worth $261 billion out of $3.3 trillion sold in the U.S. (Economist, 2014) and returned goods worth $642.6 billion out of $14.53 trillion sold worldwide. Furthermore, more than half of all returns may not be resold for full price, which results in substantial financial losses (Cheng, 2015). The rise in online sales (Insider, 2017) has led to increasing return rates, sometimes exceeding 30% (Rudolph, 2016). The issue is particularly problematic for online retailers that offer extensive product variety and customization options, making it more difficult to resell the returned items.

Firms have an interest in developing a model for predicting return volume and understanding the underlying factors associated with it for several reasons. These can be broadly grouped into two categories: (i) operational issues and (ii) financial issues. Returns can cause significant operational and logistics challenges because firms have to devote resources such as staff and space to process returns, identify if an item should be resold or disposed, etc. Moreover, if returned items have to be repaired, then this may negatively impact the workflow in the production process depending on the product type and workload levels. Financially, return volume helps provide an estimate of the cost or loss due to returns. So, from a financial and operational perspective, understanding return volume is beneficial not just to the firm in our study, but also any firm facing a large return volume in the e-commerce era. Finally, anticipating return volume of a product type at a retailer in each period may be valuable in taking actions to reduce returns. For example, Jet.com offers a discount to customers who opt out of free returns (https://jet.com/help-center/faq). Rather than offering such discounts for all product types, a manufacturer may be able to optimize its discounting strategy by targeting such discounts to specific product types offered at certain retailers that are likely to have higher returns. Such discounts may also be restricted to certain periods, for instance during the Christmas season when returns are higher.

The primary objective of this paper is to build a good data-driven model to predict return volume in the future. Our study was done in collaboration with a leading manufacturer of automotive accessories in the U.S., referred to as Company *A* hereafter, whose real name is withheld due to confidentiality reasons. Company *A* primarily sells three types of products (seat cover, dash cover, car cover) through various online retailers. Customers can choose a car accessory matched to a specific car model year, color, fabric type and also customize the product with logos, specialized prints, etc. Customers place orders online via the retailer's website and the firm makes each item to order and ships it in around a week. Customer return policies, whether strict or liberal, are dictated by the retailer. However, all returns are handled by Company

[*] Corresponding author.
E-mail addresses: hailongc@marshall.usc.edu (H. Cui), raj@marshall.usc.edu (S. Rajagopalan), amy.ward@chicagobooth.edu (A.R. Ward).

https://doi.org/10.1016/j.ejor.2019.05.046
0377-2217/© 2019 Elsevier B.V. All rights reserved.

A, not the retailers, who serve only as a sales channel. Due to the extensive variety offered, the odds are low that a product returned by one customer will be demanded by another one within, say, a few months. However, the firm checks each returned item to assess whether to put it back in stock (if it is not defective and is likely to be resold) or to dispose it. Many items are disposed or sold at a steep discount and cause a significant financial loss to the company, because extracting value from returned product is difficult in the focal firm of our study. This is unlike for other product categories such as appliances where a substantial part of the value may be recovered from a resale.

Our predictive models are built using a large, detailed data set from Company *A* on every item that was sold and/or returned over 39 months. For each item, the firm provided us with data on order date, retailer, product type, production process details (including dates when each process step was completed and who worked on it), ship date, and return date. We also obtained aggregate information on production levels, workers, inspection policies, and return policies of retailers.

We first develop a baseline main effects model for predicting return volume using four factors deemed to be important based on our initial understanding of returns at Company *A*: sales volume, time, retailer and product type. Return volume is generally proportional to sales volume although the proportion (return rate) may vary by product type, retailer and over time. The retailer variable captures factors such as return policies, type of consumers who visit a retailer, which could influence return volume. For example, more impulsive purchases may generate higher return volume and such impulsive consumers may be more likely to visit certain retailers. Return volume may also vary by product type because certain product types may have greater fit issues (e.g., seat cover) or certain product types may be prone to defects in the manufacturing process. Finally, return volume may vary over time; for example, consumers may be more impulsive during holiday purchases. We then explore how additional variables such as manufacturing workload levels, process quality checks, production process complexity, and product personalization (e.g., logos) improve prediction performance–this leads to a full main effects model with a much larger set of independent variables.

In addition to the main effects models discussed above, we also investigate whether adding second and third order interaction effects improves prediction. For instance, returns may be increasing over time at a particular retailer. Alternatively, the presence of personalized logos may impact returns for some products but not others or inspection may be valuable in reducing defects/returns for some products but not others. The incorporation of aforementioned interaction effects, however, results in a large number of predictor variables and warrants the need for a robust variable/model selection methodology. The traditional methods such as best subset selection, forward/backward selection (see Hocking, 1976 for a review) are not applicable in our study, because we are in a high dimensional setting due to having more predictors than observations as will be clear later.

To address this issue, we use four high-dimensional machine learning methods: Least Absolute Shrinkage and Selection Operator (LASSO, Tibshirani, 1996), LARS-OLS hybrid (Efron, Hastie, Johnstone, & Tibshirani, 2004), Smoothly Clipped Absolute Deviation (SCAD, Fan & Li, 2001), and Elastic Net (Zou & Hastie, 2005) that can yield sparse models. In addition, we explore two notable tree-based machine learning methods, Random Forest (Breiman, 2001) and Gradient Boosting (Friedman, 2001) to capture possible complex non-linear structure in the data to improve prediction accuracy.

Our contributions can be summarized as follows. We use various machine learning methods to build and test predictive models for return volume using a real data set that is comprehensive and includes retailer, product type, and process related variables. We show that in our setting retailer effects are stronger than product effects in predicting returns, and the baseline main effects model utilizing sales, time, product and retailer effects achieves a fair prediction performance. We consider higher order interaction effects (e.g., product type and retailer) and apply various convex and concave regularization methods for variable/model selection to derive a sparse predictive model that strongly improves prediction accuracy. In our study, we find that the optimal Elastic Net model coincides with the LASSO model, which achieves smaller prediction errors in the test data compared to all the alternative methods. It shows robust performance with similar prediction accuracy both in the training and test data.

The remainder of the paper is organized as follows. In Section 2 we review the relevant literature. Section 3 discusses the empirical setting and the predictor variables derived from the data. In Section 4 we sequentially build main effects models. Section 5 explores higher-order interaction models with high-dimensional variable/model selection methods, and tree-based statistical machine learning methods. We provide results from robustness checks in Section 6, and discuss key take-aways from our study and future research direction in Section 7. Additional details can be found in the Online Appendix.

2. Literature review

Our study is related to several streams of literature, one of them being the marketing literature on consumer product returns. Hess and Mayhew (1997) are among the earliest to empirically study product returns, and provide a model to predict the timing of returns for an apparel direct marketer. Janakiramana, Syrdalb, and Freling (2016) show that return policy leniency overall increases purchases more than returns. Anderson, Karsten, and Duncan (2009) identify a considerable variation in the value of returns across customers and product categories at a mail-order catalog company. Petersen and Kumar (2009) determine the firm-customer exchange process factors that help explain product return behavior and the consequences of product returns on future customer and firm behavior. Urbanke, Kranz, and Kolbe (2015) propose a returns prediction system for an online retailer to explore the impact of different levers on the likelihood of returns, for example, by artificially increasing delivery time to deter a consumer from purchase. The above-mentioned research is mostly concerned with factors impacting consumer's decision to keep or return a product, and studies returns from the perspective of retailers. In contrast, our paper develops data-driven predictive models for a manufacturer, which sells online an extensive range of product variants and is primarily concerned with the cost and logistics of handling returns, and is interested in predicting consumer returns due to various reasons (e.g. defect in product, consumer's change of mind).

Another relevant stream is the operations literature on returns, which is primarily focused on analytical studies of product returns, for example, returned component reuse (DeCroix G, Song, & Zipkin, 2009), restocking fees (Shulman, Coughlan, & Savaskan, 2011), return policies (Su, 2009, Altug, 2016, Shang, Pekg̈un, Ferguson, & Galbreth, 2017a, Ülkü & Gürler, 2017), remufacturing (Calmon & Graves, 2017; Cerag, Ferguson, & Toktay, 2016), optimal retail assortment under consumer returns (Alptekinoğlu & Grasas, 2017), return strategy and pricing in a dual-channel supply chain (Li, Li, Sethi, & Guan, 2017). In contrast, our work is empirical in nature and focused on predicting returns. There are some works related to predicting returns for the purpose of remanufacturing. For example, Toktay (2003) reviews a few forecasting methods (e.g., using past sales and returns), and Tsiliyannis (2018) presents a stochastic method for real-time forecasting of product returns in remanufacturing. Our study focuses on data-driven prediction models for returns and is not concerned with remanufacturing, because the products in our study are made of fabrics that are not reused to make new products.

Some recent works in operations also conduct empirical studies on returns from the perspective of retailers and consumers. For example, Shang, Ghosh, and Galbreth (2017b) analyze both the return policy drivers from the retailer's perspective and the return policy value from the consumer's perspective, and show that the value of a full refund policy to consumers may not be as large as one might expect. Akturk, Ketzenberg, and Heim (2018) study omnichannel retailing in the context of a national jewelry retailer and suggest that introducing ship-to-store increased cross-channel customer returns of online purchases to physical stores. In contrast to these works, we study product returns from the perspective of a manufacturer.

Our paper is also related to the literature on predictive analytics, which are applied to a variety of settings. In the operations research literature, however, there is still a relatively low volume of analytics-orientated studies (Mortenson, Doherty, & Robinson, 2015). Here are two examples: Oztekin, Al-Ebbini, Sevkli, and Delen (2018) develop a hybrid methodology to predict and explain the quality of life for patients undergoing a lung transplant, and Sevim, Oztekin, Bali, Gumus, and Guresen (2014) develop an early warning system to predict currency crises using artificial neural networks, decision trees, and logistic regression.

We utilize high-dimensional machine learning methods, which have become increasingly popular and widely used in areas such as genomics, neuroscience, social media analysis and high-frequency finance. They are not only utilized to obtain a good prediction accuracy, but also to address variable/model selection problems associated with various challenges such as noise accumulation, spurious correlation, scalability and stochastic errors. We refer readers to Fan, Lv, and Qi (2011); Varian (2014) for excellent reviews, and Hastie, Tibshirani, and Friedman (2009); James, Witten, Hastie, and Tibshirani (2017) for in-depth treatments of high-dimensional methods. One of the best known high-dimensional methods is LASSO or otherwise known as L_1 regularization, which has recently received some attention in the operations management literature for predictive models, and here are some recent examples. Ma, Fildes, and Huang (2016) apply LASSO to select key explanatory variables from a high dimensional data set for demand forecasting for SKU retail sales. Martinez, Schmuck, Pereverzyev, Pirkerc, and Haltmeier (2018) use LASSO as one of the methods to develop a machine learning framework for customer purchase prediction in a non-contractual setting. Bertsimas, O'Hair, Relyea, and Silberholz (2016) use LASSO as one of the methods to predict the outcomes of clinical trials. In the context of a Red Cross fund-raising campaign, Ryzhov, Han, and Bradic (2016) employ LASSO to logistic regression models to identify key interactions between designs (e.g., the presence or absence of a free gift) and various donor segments. In contrast to the aforementioned literature, we study product returns, and extract features common to many product manufacturers from our data set. Furthermore, we consider higher-order interaction terms (e.g., product type and retailer) and initialize multiple machine learning methods (LASSO, SCAD, Elastic Net, Random Forest, Gradient Boosting) to derive and calibrate predictive models.

3. Empirical setting

In this section, we first introduce the operational context and process in our study in Section 3.1. Section 3.2 describes the data, and Section 3.3 discusses how we derive predictor variables from the data.

3.1. Operational context and process

Company *A* manufactures car accessories (seat covers, car covers and dash covers) at two factory locations – in California and Mexico. Consumers place orders at an online retailer for a specific vehicle (make, model, year, trim level) and a specific product type. A consumer can choose the fabric, color, design and whether to have a personalized logo. Sometimes, she may request multiple products in the same order. The order information is transferred to Company *A* which manufactures and ships the product to the consumer in about a week.

Once Company *A* receives an order, it is released into production immediately unless there is substantial pending work. The entire manufacturing process can be broadly categorized into three stages. The *pre-sewing stage* comprises of printing the order label, which contains detailed instructions for each of the subsequent operations, cutting of fabrics by computerized cutting machines, and placement of cut pieces in a plastic bag with the order label to be routed for sewing. During the *sewing stage*, sewers perform tasks such as joining, binding, embroidering and adding logos, which vary with the product types. In the last *post-sewing stage*, finished goods are inspected, packed and shipped to the consumer. There are some random inspections by the floor supervisor in the pre-sewing and post-sewing stages. Most products are exclusively made in the U.S. while some are produced jointly in both factories.

Returns. A product may be returned within the return period for refund, or beyond the return period for repair or replacement under warranty. The return policy is set by each retailer and return policies determine aspects such as time within which a return has to be made for a refund, allowing returns to physical stores, etc. All the returns are handled by Company *A*, not the retailers. If a returned product is confirmed to be in a resalable condition, it is first kept in the warehouse depending on the available space, and then disposed if it is not sold within a certain time window. If a returned product is for repair or replacement, the manufacturer needs to allocate resources to meet the request. Therefore, the returns handling increases operational overhead in staffing and resources.

3.2. Data description

In our study, we focus on three main product types that comprise over 90% of sales volume for Company *A*: seat cover, car cover and dash cover. They are sold to consumers in the United States and most sales are through 13 retailers. There was a small fraction (0.17%) of the products that were returned and resold during our study period, but they were excluded during data cleaning because they comprised of a negligible fraction of the sales. This gives us a data set containing 331,390 products sold between July 2012 and September 2015, for a total of 39 months denoted as periods $t = 1, \ldots, 39$. In Fig. 1, we show how sales, returns (in units) and return rates change over 39 periods in the focal firm in our study. One can observe that both returns as well as return rates fluctuate over time.

Though aggregate return volume in each period is of first-order interest, Company *A* is also interested in predicting return volume by each product type and each retailer for the following reasons. First, the revenue loss varies among different product types, for example, the revenue loss due to refund of seat cover is often higher than one for dash cover. Second, the ability to resell the returned product may also vary with the product type. Third, since different product types have different manufacturing processes, when products are returned for repair or replacement under warranty, the increased operational workload and costs may also be quite different. Fourth, the same seat cover may lead to a larger revenue loss due to return for refund through a retailer with a liberal re-

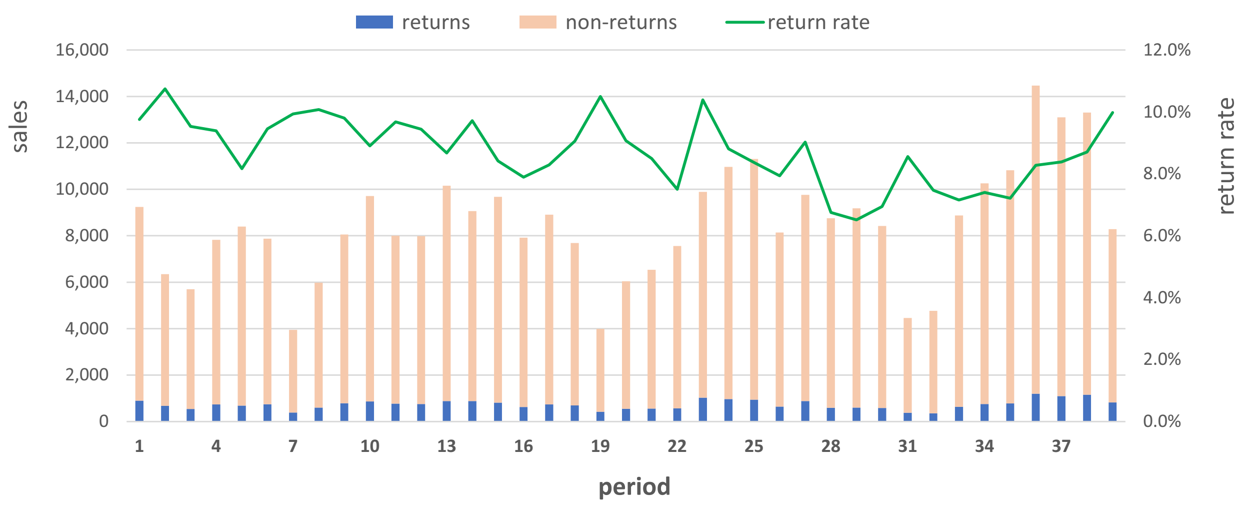


Fig. 1. Sales, returns and return rates over 39 months, between July 2012 and September 2015.

Table 1
A sample operational data set.

ReleaseDate	ScanDate	OrderNo	SerialNo	Product	Retailer	Logo	Operation	EmpID	Loc	FName	LName	ReturnDate
5-13-2015	5-13-2015	149038902	L49325331	SCxxxx	xxxx.com	USC001	Printing labels	xx01	US	xxxx	xxxx	7-10-2015
5-13-2015	5-14-2015	149038902	L49325331	SCxxxx	xxxx.com	USC001	Cutting fabrics	xx12	US	xxxx	xxxx	7-10-2015
5-13-2015	5-15-2015	149038902	L49325331	SCxxxx	xxxx.com	USC001	Joining	xx53	US	xxxx	xxxx	7-10-2015
5-13-2015	5-15-2015	149038902	L49325331	SCxxxx	xxxx.com	USC001	Binding	xx53	US	xxxx	xxxx	7-10-2015
5-13-2015	5-18-2015	149038902	L49325331	SCxxxx	xxxx.com	USC001	Logo	xx26	US	xxxx	xxxx	7-10-2015
5-13-2015	5-18-2015	149038902	L49325331	SCxxxx	xxxx.com	USC001	Random Inspection	xx45	US	xxxx	xxxx	7-10-2015
5-13-2015	5-18-2015	149038902	L49325331	SCxxxx	xxxx.com	USC001	Packing	xx34	US	xxxx	xxxx	7-10-2015

turn policy, but may result in less costly return for repair through a retailer with a 30-day return policy.

There are many possible reasons why a particular product may be returned and understanding them is helpful when building a predictive model. Next, we provide a sample operational level data set obtained from Company *A* (see Table 1) to provide a context to discuss these reasons. Some values in the table are hidden or modified to preserve confidentiality. The data used in the study includes both retailer level and operational level data sets obtained from the company. The retailer level data contains the return policy and name of each retailer.

This example shows 5 different operations performed on a product with serial number L49325331 in order number 149038902. We can identify the product type–seat cover–by the prefix of the product, and observe that this particular product was released into production and the order label was printed on Wednesday, 5/13/2015. The next day, an employee read the label to find out which fabrics to pick up from inventory, and cut the fabric in preparation for sewing according to the specification. On Friday, 5/15/2015, an employee with ID xx53 performs two types of sewing operations–Joining and Binding. On Monday, 5/18/2015, a different sewer adds a logo and finishes this product. A floor supervisor (usually a sewer with more than 10 years of experience) performs a random inspection on the product, does not notice any defects, and she passes the product to an employee who packs the product. Depending on the time of day, this product is shipped to a consumer on the same day or next day. The product is returned to the factory on 7/10/2015 as shown in the ReturnDate column.

Which factors may have contributed to this return? This product was ordered just before the summer. There may be a seasonality effect for returns if the product fades during the summer heat. Product type may influence returns; for example, seat covers usually cost much more than dash covers, so a customer is more likely to go through the trouble of returning seat covers. Seat covers might also have higher defects because of more difficult sewing operations. Retailer xxxx.com in Table 1 is found to have a 60 days return policy in the retailer level data set, and the product is returned between 30 days and 60 days, so the return may have been because the consumer changed his/her mind or due to a defect in the product. The production processes and resources used in manufacturing the product may cause defects and in turn result in returns. The example in Table 1 shows that two sewers performed three sewing tasks on this product, and we may ask if single tasking (three sewers, each doing a single task) leads to lower defects and thus lower returns. It is also possible that the consumer may have simply changed her mind, and historical returns may partly capture this factor. We could also check whether an order of multiple products lead to higher (or lower) returns. In the next subsection, we explore the specific predictor variables used in the study in more detail.

The *predicted* variable in the study is $Returns_{tij}$, that is, how many units of product type i will be returned out of sales through retailer j in period t. As is clear from Table 1, every item that is returned can be tracked, and the details about retailer, product and production process characteristics for the item are available. While we collected data until October 2016, we analyze returns only for sales until Sep 2015 in our study so that almost all returns, which occur within one year, are captured.

3.3. Predictor variables

Our focus in this subsection is to identify the variables that can help predict the volume of returns defined as $Returns_{tij}$, where the subscripts t, i, j denote the period, product type and retailer respectively. As potential predictors of returns, we consider production process and resources, multi-product order and historical returns in addition to aforementioned factors–sales, time, product type, retailer. We list the predictor variables reflecting these factors in Table 2 along with their definitions. In the following, we discuss each of these predictor variables and the justification for using them in the prediction models.

Please cite this article as: H. Cui, S. Rajagopalan and A.R. Ward, Predicting product return volume using machine learning methods, European Journal of Operational Research, https://doi.org/10.1016/j.ejor.2019.05.046

Table 2
Definition of response and predictor variables for $t = 1, \dots, 39$, $i = 1, 2, 3$, $j = 1, \dots, 13$.

Category	Variable name	Definition
Response variable	$Returns_{tij}$	Number of returns out of sales of product type i via retailer j in period t.
Sales effect	$Sales_{tij}$	Number of sales of product type i via retailer j in period t.
Time effect	$Year_t$	Sales year from 2012, 2013, 2014 to 2015, used to capture time trend.
	$Month_t$	Dummy variable for the month corresponding to period t, used to capture seasonality.
Product effect	$Product_i$	Dummy variable for product type i.
Retailer effect	$Retailer_j$	Dummy variable for retailer j ranked by volume.
Production process and resources	$SewerCnt_{tij}$	Average number of sewers per product among $Sales_{tij}$.
	$SewingTaskCnt_{tij}$	Average number of sewing tasks per product among $Sales_{tij}$.
	$SewingDays_{tij}$	Average number of days in sewing stage per product among $Sales_{tij}$.
	$BacklogDays_{tij}$	Average number of backlog days before production per product among $Sales_{tij}$.
	$ProductionDays_{tij}$	Average number of days in entire production process per product among $Sales_{tij}$.
	$Workload_t$	Average number of finished products per employee in period t.
	$JointProduction\%_t$	Fraction of products jointly manufactured in period t.
	$CustomFabric\%_{tij}$	Fraction of products using customized fabrics among $Sales_{tij}$.
	$Logo\%_{tij}$	Fraction of products ordered with special logos among $Sales_{tij}$.
	$InspectionPreSewing\%_{tij}$	Random inspection rate before sewing among $Sales_{tij}$.
	$InspectionPostSewing\%_{tij}$	Random inspection rate after sewing among $Sales_{tij}$.
Multi-product effect	$MultiProduct\%_{tij}$	Fraction of orders with two or more products among $Sales_{tij}$.
Historical returns	$LaggedReturns_{tij}$	Observed returns in period t from sales in periods $t-4, t-5, t-6$.

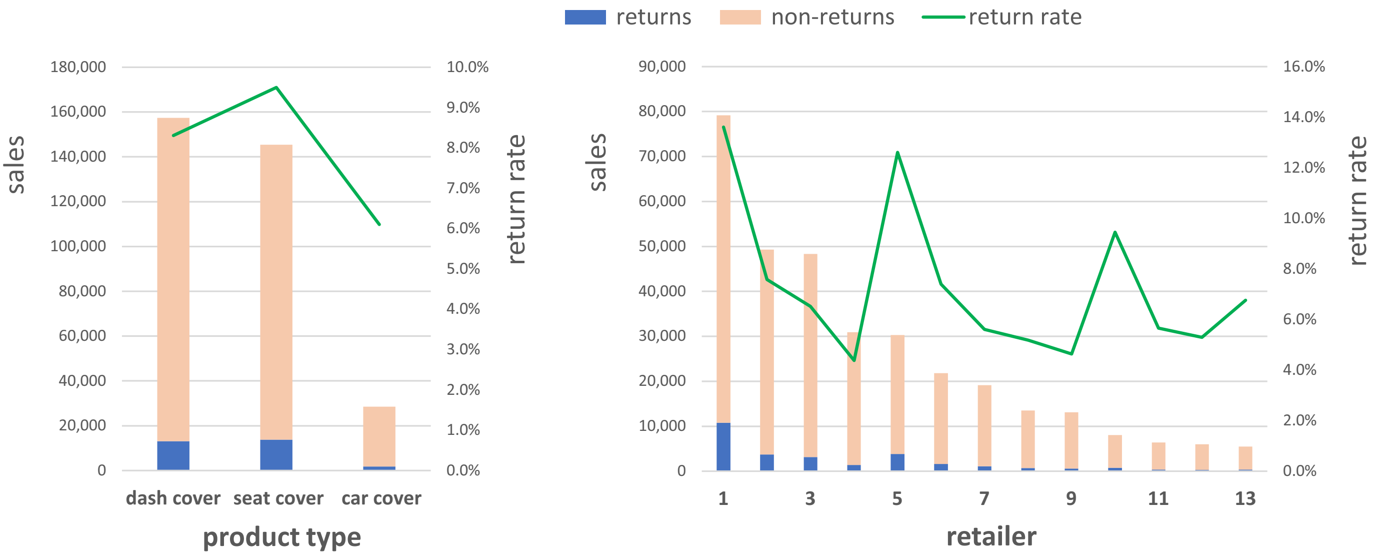


Fig. 2. Sales, returns and return rates by product type (left) and by retailer (right) over all periods.

Sales. As sales volume goes up, return volume is likely to increase. So, we include sales as a predictor.

Time. Similar to Anderson et al. (2009) we consider two types of time effects: trend effect and monthly fixed effect (e.g. consumers may purchase products more impulsively during holiday seasons resulting in higher returns.) The trend effect is captured by the predictor variable $Year_t \in \{2012, 2013, 2014, 2015\}$, which depends on the period t; for example, $t = 1, \dots, 6$ implies 2012, $t = 7, \dots, 18$ implies 2013, and so on, recalling that the dataset begins in July 2012. The monthly fixed effect is captured by the dummy variable $Month_t \in \{January, \dots, November, December\}$, which depends on the month the period t corresponds to; for instance, $Month_1$ indicates July, $Month_2$ equals August, and so on. We provided the returns pattern over the 39 periods in Fig. 1. As an alternative to using year and month variables, we de-trended and de-seasonalized the data before estimating the models but we found little difference (see Online Appendix A1 for details).

Product. The three product types in our study–seat cover, dash cover and vehicle cover–have distinct characteristics, and may have product specific fixed effects that influence returns. For example, a slightly larger vehicle cover may be considered fine by consumers, but a slightly loose seat cover may be regarded as defective and returned. We provided the returns for each product type in Fig. 2, and found that the return rate of car cover was much lower than that for the other two products. In a study of an online retailer, Anderson et al. (2009) show variations in returns among different product categories. Therefore, we could expect that product effects are important in predicting return volume.

Retailer. Based on previous research on the impact of return policies on sales and returns (e.g., Janakiramana et al., 2016 and references therein) and given different returns policies of retailers in our study, we would expect return volume to vary with retailers. In addition, there may be other important factors associated with a retailer that influence returns. Retailers with different return policies may attract different types of customers who exhibit different return behaviors. The number of brick-and-mortar stores of a retailer may have an impact on returns because more locations may make it easier to return a product: a warehouse club in our study operates hundreds of stores, whereas an auto specialty store has thousands of locations where a consumer may return a prod-

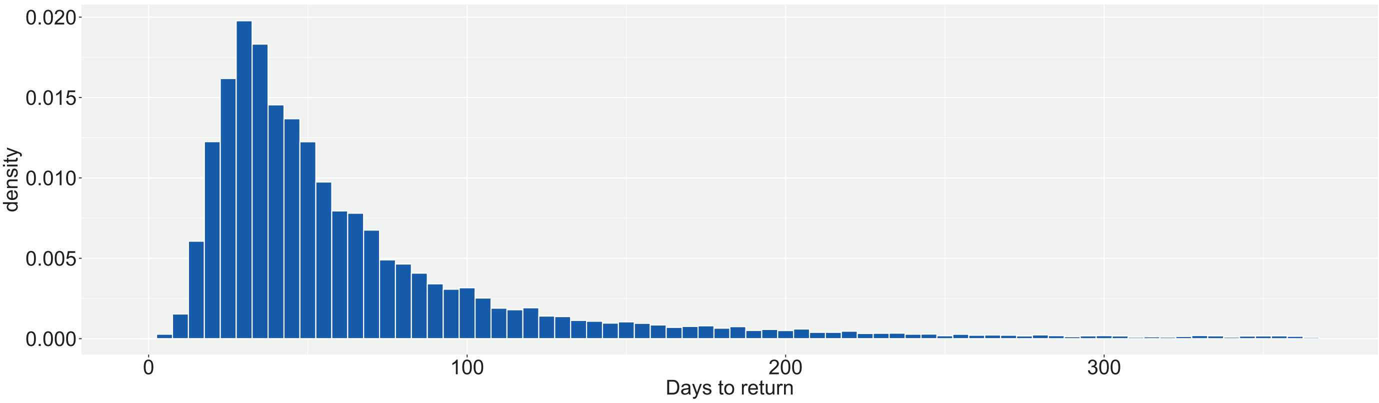


Fig. 3. A histogram of days to return for products returned within a year.

uct. Therefore we expect that the retailer's return policy and other characteristics may play an important role in returns, which are both captured by retailer indicator variables, as we showed earlier in Fig. 2. In our analysis, we also explored incorporating variables to reflect different dimensions of return policies such as return period, in-store return acceptance, in addition to the fixed retailer effect but we did not find the additional variables helpful in increasing the prediction accuracy.

Production processes and resources. Unlike prior studies of returns which considered only product and consumer purchase related factors, we consider various factors related to the production process that may contribute to returns due to defects. We measure most of the variables in this category as an average over each period, product and retailer combination. We track the average number of sewers and sewing tasks by $SewerCnt_{tij}$ and $SewingTaskCnt_{tij}$ for each product, to understand if more sewers or fewer sewing tasks per sewer (e.g. grouping some sewing operations into one task) lead to lower defects, and therefore are useful in predicting returns. We use $SewingDays_{tij}$ as a proxy to the actual sewing time, and we may expect this predictor to be negatively correlated with returns, that is, if sewers spend more time in sewing a product, the quality of the product may increase and the likelihood of returns may decrease. $BacklogDays_{tij}$ and $ProductionDays_{tij}$ reflect different aspects of the manufacturing process and may be useful in predicting returns. We are interested in whether the monthly workload of the factory and the fraction of items jointly produced in Mexico and US impact returns, and measure these effects by $Workload_t$, $JointProduction\%_t$ for each period. We investigate the effect of special customization on returns by $CustomFabric\%_{tij}$ for the usage of in-house custom fabric, and $Logo\%_{tij}$ for personalizing products with custom-logos, and the effect of random quality inspections by floor supervisor on returns before and after the sewing stage by $InspectionPreSewing\%_{tij}$ and $InspectionPostSewing\%_{tij}$.

Multi-product effect. Anderson et al. (2009) do not find a significant effect of multiple product purchase on returns in their study, but suggest that this may not always be true and in other applications, return rates may depend on whether a consumer purchased single or multiple products in an order. In our study, about 9% of products are sold as part of an order with two or more customized products. We utilize $MultiProduct\%_{tij}$ to capture and investigate this effect.

Historical returns. We are interested in capturing new trends in returns that may not be captured by the aforementioned predictor variables. For example, recent returns data may include a change in consumer's purchase and return behaviors, such as increased usage of smartphones to make purchases (Martin, 2018) which may be more impulsive and lead to higher returns than average online orders. In addition, using lagged values to predict current period's value is not uncommon (for example, see Wilms, Basu, Bien, & Matteson, 2017). We use a predictor variable called $LaggedReturns_{tij}$ to capture this effect, using historical returns data. To do this, we consider three criteria: (1) we want the returns status of historical data to be accurate, (2) we want to capture recent returns trend, 3) we want to reduce noise in low-volume returns. To satisfy these requirements, we define $LaggedReturns_{tij} := Returns_{t-6,ij} + Returns_{t-5,ij} + Returns_{t-4,ij}$ to predict $Returns_{tij}$. Next we discuss each of these points.

Table 3
Summary statistics of response and predictor variables. Data size $N_{total} = 1360$ represents the number of data points in our model because we model the return volume per product type, per retailer, per time period.

Variable	Min	Mean	Median	Max	SD
Returns	0.00	21.05	6.00	509.00	49.53
Sales	1.00	243.70	99.00	2744.00	406.10
SewerCnt	1.00	3.28	3.27	6.75	1.14
SewingTaskCnt	1.00	3.61	3.86	8.75	1.33
SewingDays	1.00	1.99	1.80	8.00	0.77
BacklogDays	0.00	0.48	0.17	6.00	0.83
ProductionDays	2.00	7.07	6.97	15.00	3.23
Workload	410.00	1033.40	1117.40	1832.80	428.41
JointProduction%	0.00%	19.05%	0.42%	100.00%	29.50%
CustomFabric%	0.00%	5.41%	0.00%	73.84%	12.33%
Logo%	0.00%	2.74%	0.00%	72.73%	6.84%
InspectionPreSewing%	0.00%	0.07%	0.00%	26.07%	1.00%
InspectionPostSewing%	0.00%	11.97%	0.00%	116.67%	24.96%
MultiProduct%	0.00%	7.22%	2.97%	99.69%	16.82%
LaggedReturns	0.00	58.37	16.00	1490.00	148.17

In the above Fig. 3, we observe that about 90% of returns occur within the first 4 months, thus the returns status largely becomes finalized within four months after sales of an item. For this reason, when we consider historical returns data, we want to go back at least 4 months, in other words, we want to use $Returns_{t-n,ij}$ for $n \geq 4$ for the prediction of $Returns_{tij}$ to satisfy our first criterion. Returns data from more than half a year ago may not reflect recent returns trend, and for this reason we propose to use $Returns_{t-n,i,j}$ for $n \leq 6$ to meet our second criterion. To reduce noise coming from low-volume returns, we aggregate historical returns over three months to satisfy our third criterion, instead of picking one of the variables $Returns_{t-4,ij}$, $Returns_{t-5,ij}$, $Returns_{t-6,ij}$. Please see Online Appendix A2 on details on how we construct the variable *LaggedReturns* for the test periods.

We provide summary statistics of predictor variables in Table 3. One can notice a large variation among returns (and sales) ranging from 0 to 509 (and 1 to 2,744) depending on the period, product type and retailer. It requires an average of 3 to 4 sewers to finish a product and some sewers perform more than one task; the average time spent in the sewing stage is less than 2 days. Each employee, on average, finishes 1,033.40 products per period, as seen in *Workload*. When a product fails a post-sewing random inspection, it is sent back for rework and may be inspected again by the supervisor, which leads to a maximum possible 200% post-sewing inspection rate and explains the highest inspection rate of 116.67% for some period, product type and retailer. The *LaggedReturns* are aggregated over three months and its summary statistics show a similar pattern as the ones for *Returns*, suggesting that historical returns may be an important predictor variable.

Recall from Section 3.2 that our data contains 331,390 products sold between July 2012 and September 2015, thus 331,390 represents the total number of transactions (items purchased). As we discussed at the beginning of this subsection, we are modeling $Returns_{tij}$–the *volume* of returns per product type, per store, per time period. Thus, the index "tij" represents one data point, therefore we have a much smaller set of data points, $N_{total} = 1,360$. Since $t = 39$ periods, $i = 3$ product types, and $j = 13$ retailers, one might assume that $N_{total} = 39 \times 3 \times 13 = 1,521$. The reason $N_{total} = 1,360 < 1,521$ is due to missing data points, that is, no sales and hence no returns for certain "tij" values (see Figure A1 in Online Appendix A5 for details).

4. Main effects models

In this section, we explore various models to predict return volume using some or all of the variables listed in Table 2. We would ideally like a parsimonious model that has good predictive power and can be easily implemented in practice. Then, a natural question is–can we use only sales, time, product type and retailer information to build a model with reasonable prediction performance? We consider this model to be the baseline main effects model in our study, because these variables can be readily derived from a company's ERP system and an operations manager can easily implement this predictive model using a spreadsheet, without the effort required to incorporate the remaining predictor variables. In Section 4.1, we sequentially add the predictor variables discussed above, and then present in Section 4.2 the results of model fit (training set) and prediction (test set) for all the models.

4.1. Model specification

In the following simple model (1), we first use the predictor, *Sales*, to predict how many units out of such sales will be returned.

$$Returns_{tij} = \beta_0 + \beta_1 Sales_{tij} + \epsilon_{tij} \tag{1}$$

In the next model (2) we explore whether trend and seasonality effects help increase the prediction of return volume.

$$Returns_{tij} = \beta_0 + \beta_1 Sales_{tij} + \beta_2 Year_t + \beta_3^t Month_t + \epsilon_{tij} \tag{2}$$

We now want to check the effects of product types and retailers on the prediction of return volume, but we don't know which effect is more important in predicting returns. Thus, we first investigate the product effect in model (3) and retailer effect in model (4) individually, and then both effects in the baseline main effects model (5) simultaneously.

$$\begin{aligned} Returns_{tij} &= \beta_0 + \beta_1 Sales_{tij} + \beta_2 Year_t + \beta_3^t Month_t \\ &\quad + \beta_4^i Product_i + \epsilon_{tij} \end{aligned} \tag{3}$$

$$\begin{aligned} Returns_{tij} &= \beta_0 + \beta_1 Sales_{tij} + \beta_2 Year_t + \beta_3^t Month_t \\ &\quad + \beta_5^j Retailer_j + \epsilon_{tij} \end{aligned} \tag{4}$$

$$\begin{aligned} Returns_{tij} &= \beta_0 + \beta_1 Sales_{tij} + \beta_2 Year_t + \beta_3^t Month_t + \beta_4^i Product_i \\ &\quad + \beta_5^j Retailer_j + \epsilon_{tij} \end{aligned} \tag{5}$$

To simplify the exposition, we adopt column vector notation (shown in bold). For example, $\boldsymbol{U}$ consists of the remaining 13 variables shown in Table 2, and we add it to obtain the full main effects model (6) below.

$$\begin{aligned} Returns_{tij} &= \beta_0 + \beta_1 Sales_{tij} + \beta_2 Year_t + \beta_3^t Month_t + \beta_4^i Product_i \\ &\quad + \beta_5^j Retailer_j + \boldsymbol{\beta_6^T U} + \epsilon_{tij} \end{aligned} \tag{6}$$

For all the models (1) to (6) above, we considered a family of power transformations (Box & Cox, 1964) for our response variable $Returns_{tij}$ to maximize normality before fitting models to the data; however, we did not find a need for a transformation.

4.2. Results from main effects models

To evaluate our models, we assign the first 33 periods (7/2012 to 3/2015) to be our training set, and the last 6 periods (4/2015 to 9/2015) to be our test set. We discuss alternative ways of splitting the data into training and test sets in our robustness checks in Section 6. For models (1) to (6) we present in Table 4 the model fit in training set, and the prediction performance in test set.

In our study, we measure the model fit and prediction performance by MSE (equivalently, Prediction Error) and R^2. For the purpose of model comparison and selection for main effects models, we focus on adjusted R^2 and Akaike Information Criterion (AIC). Alternative measures may include Mallow's C_p, Bayesian Information Criterion (BIC), Risk Inflation Criterion (RIC), and predicted R^2. In the OLS setting, C_p and AIC are proportional to each other, therefore only one is needed. BIC and RIC generally place a heavier penalty on models with more variables compared to AIC (see Foster & George, 1994.) Predicted R^2 method is identical to leave-one-out cross-validation (CV, Stone, 1974), and is asymptotically the same method as AIC for model selection (Stone, 1977).

We observe that the simplest model (1) yields $R^2 = 0.864$ for training set, and $R^2 = 0.836$ in test set for predicting returns over the future 6 months. This shows that the predictor variable sales by itself explains a large portion of variability in return volume, and that sales is a good predictor of return volume as expected. In the subsequent models (2) to (5), when we sequentially add time, product type, retailer, product type and retailer effects to model (1), we see that in both training and test sets, R^2 increases and MSE decreases indicating that we achieve not only an increasingly better model fit in the training set, but also a higher prediction accuracy in the test set. We also find that knowing the retailer is more important than knowing the product type when predicting returns, by comparing models (3) and (4). Our largest model (6)–with 13 additional predictors in $\boldsymbol{U}$ compared to model (5)–obtains the lowest $AIC = 9315.068$, the highest Adjusted $R^2 = 0.918$; and if we adopt conventional model selection methods, we may consider model (6) to be the best main effects model for training set. This model also has the best prediction performance in test set.

Model (6) may have one potential drawback for use in practice, because the required overhead from data collection and analysis for many of the variables in $\boldsymbol{U}$ is substantial. By comparison, our second best model (5) may be preferred by Company *A*, because it only requires sales, time, product type and retailer information, thus it is much easier to implement by operations managers. For these reasons, we consider models (5) and (6) as the two best main effects models; and to address the aforementioned trade-off between models (5) and (6), we first provide in Table 5 on the next page the coefficient, standard deviation and significance levels for all the main effects models. We notice that two predictor variables–*MultiProduct%* and *LaggedReturns*–show up as significant at 0.001 level in model (6).

A natural question that arises is whether one should choose one or both of these significant predictors from model (6) to add

Please cite this article as: H. Cui, S. Rajagopalan and A.R. Ward, Predicting product return volume using machine learning methods, European Journal of Operational Research, https://doi.org/10.1016/j.ejor.2019.05.046

Table 4
Prediction performance of 6 main effects models trained on 33 periods and tested on 6 periods.

		Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Training set	Number of predictors	1	13	15	25	27	40
	Data size	1140	1140	1140	1140	1140	1140
	R^2	0.864	0.867	0.869	0.908	0.912	0.921
	Adjusted R^2	0.863	0.865	0.867	0.906	0.910	0.918
	AIC	9864.255	9860.589	9843.836	9459.401	9410.657	9315.068
	MSE	333.524	325.527	319.655	224.186	214.050	192.395
Test set	Data size	220	220	220	220	220	220
	R^2	0.836	0.838	0.844	0.888	0.895	0.923
	MSE	402.506	396.787	381.472	275.189	257.709	189.386

Table 5
OLS results for main effects models in the training set. Standard errors are shown in parentheses.

Predictor variables	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Sales	0.119 (0.001)***	0.119 (0.001)***	0.121 (0.001)***	0.111 (0.001)***	0.113 (0.001)***	0.099 (0.002)***
Year		−2.168 (0.703)**	−2.248 (0.698)**	−2.323 (0.596)***	−2.449 (0.584)***	−2.146 (1.406)
January		7.509 (2.640)**	7.629 (2.619)**	6.795 (2.207)**	7.091 (2.160)**	5.503 (2.401)*
February		5.656 (2.597)*	5.734 (2.576)*	5.064 (2.169)*	5.224 (2.122)*	4.484 (2.401)
March		2.480 (2.603)	2.533 (2.582)	2.462 (2.172)	2.589 (2.125)	3.205 (2.267)
April		0.290 (2.820)	0.278 (2.797)	0.329 (2.353)	0.320 (2.301)	1.807 (2.371)
May		4.459 (2.820)	4.436 (2.797)	4.568 (2.353)	4.540 (2.301)*	6.475 (2.374)**
June		1.432 (2.833)	1.431 (2.810)	1.739 (2.364)	1.757 (2.312)	4.322 (2.417)
July		−1.130 (2.544)	−1.191 (2.523)	−0.566 (2.123)	−0.631 (2.077)	2.388 (2.132)
August		2.051 (2.555)	2.135 (2.535)	2.217 (2.132)	2.387 (2.086)	3.777 (2.308)
September		0.716 (2.530)	0.667 (2.509)	0.847 (2.111)	0.800 (2.064)	1.792 (2.517)
October		−1.315 (2.530)	−1.296 (2.509)	−1.332 (2.111)	−1.282 (2.064)	−0.404 (2.437)
November		−2.971 (2.530)	−2.977 (2.509)	−2.802 (2.111)	−2.786 (2.065)	−2.065 (2.497)
Product$_1$ (dash cover)			−6.137 (1.373)***		−8.287 (1.142)***	−9.894 (1.900)***
Product$_2$ (seat cover)			−2.608 (1.374)		−4.688 (1.147)***	−6.956 (3.590)
Retailer$_1$				29.670 (2.810)***	29.760 (2.748)***	18.930 (2.893)***
Retailer$_2$				−10.770 (2.216)***	−11.460 (2.170)***	−9.767 (2.558)***
Retailer$_3$				−17.520 (2.228)***	−18.300 (2.183)***	−15.050 (2.152)***
Retailer$_4$				−16.300 (2.193)***	−16.810 (2.147)***	−15.840 (2.072)***
Retailer$_5$				5.938 (2.212)**	5.499 (2.165)*	6.958 (2.313)**
Retailer$_6$				−4.459 (2.208)*	−4.566 (2.159)*	−3.672 (2.113)
Retailer$_7$				−6.106 (2.175)**	−6.380 (2.128)**	−5.026 (2.085)*
Retailer$_8$				−4.334 (2.172)*	−4.530 (2.125)*	−3.662 (2.065)
Retailer$_9$				−4.281 (2.171)*	−4.447 (2.124)*	−4.465 (2.033)*
Retailer$_{10}$				0.202 (2.425)	1.304 (2.380)	0.425 (2.288)
Retailer$_{11}$				−0.662 (2.271)	0.007 (2.223)	0.276 (2.136)
Retailer$_{12}$				−1.001 (2.771)	−1.073 (2.710)	0.163 (2.814)
Workload						0.000 (0.003)
SewerCnt						−1.591 (2.301)
SewingTaskCnt						1.097 (2.384)
SewingDays						−1.431 (0.853)
BacklogDays						0.545 (0.714)
ProductionDays						0.097 (0.253)
CustomFabric%						−8.025 (4.538)
Logo%						−6.793 (8.682)
InspectionPreSewing%						35.990 (40.230)
InspectionPostSewing%						−1.052 (7.213)
JointProduction%						1.721 (4.338)
MultiProduct%						−10.980 (3.107)***
LaggedReturns						0.052 (0.005)***

***, **, and * denote statistical significance at the 0.001, 0.01 and 0.05 confidence levels respectively.

to model (5) to achieve prediction performance close to model (6) yet without the added overhead for implementation. This question then further invites another question–what is the best model out of all possible models based on the 40 variables in model (6)? Finding the best subset of predictor variables can be onerous due to the potential model space size $2^p - 1$. The best subset selection method also suffers from a lack of stability (Breiman, 1996).

The model selection issue becomes even more challenging if we want to explore the effect of interaction terms to improve model performance, in which case we have a larger number of predictors in the model, and thus an even larger number of potential models. Even though the model performance is our primary concern, we are also interested in selecting a stable and sparse model for predicting returns and to implement it in practice. In the next Section 5, we discuss multiple machine learning methods to address the aforementioned challenges in model selection and to achieve our goal of obtaining a parsimonious and stable model.

5. Incorporating interaction effects

The main objective of this section is to explore higher order interaction terms to see if adding such terms to the main effects model helps improve prediction performance. To do this, we first specify a model with a large number of interaction terms in Section 5.1, which creates a challenge in fitting the traditional OLS model to the data. To overcome this challenge, we introduce LASSO in Section 5.2 as a method to select predictor variables, and present the prediction performance of two models selected

Table 6
2-way and 3-way interaction terms.

2-way interactions	3-way interactions	3-way interactions
Sales × *Year*	*Sales* × ***Product*** × *Year*	***Product*** × *Year* × *LaggedReturns*
Sales × ***Month***	*Sales* × ***Product⊗Month***	***Product*** × *Year* × *MultiProduct%*
Sales × ***Product***	*Sales* × ***Product⊗Retailer***	***Product⊗Month*** × *LaggedReturns*
Sales × ***Retailer***	*Sales* × ***Product*** × *LaggedReturns*	***Product⊗Month*** × *MultiProduct%*
Sales × *LaggedReturns*	*Sales* × ***Product*** × *MultiProduct%*	***Product*** × *LaggedReturns* × *MultiProduct%*
Sales × *MultiProduct%*	*Sales* × ***Retailer*** × *Year*	***Retailer*** × *Year* × *LaggedReturns*
Product × *Year*	*Sales* × ***Retailer⊗Month***	***Retailer*** × *Year* × *MultiProduct%*
Product⊗Month	*Sales* × ***Retailer*** × *LaggedReturns*	***Retailer⊗Month*** × *LaggedReturns*
Product⊗Retailer	*Sales* × ***Retailer*** × *MultiProduct%*	***Retailer⊗Month*** × *MultiProduct%*
Product × *LaggedReturns*	*Sales* × *LaggedReturns* × *Year*	***Retailer*** × *LaggedReturns* × *MultiProduct%*
Product × *MultiProduct%*	*Sales* × *LaggedReturns* × ***Month***	*LaggedReturns* × *MultiProduct%* × *Year*
Retailer × *Year*	*Sales* × *LaggedReturns* × *MultiProduct%*	*LaggedReturns* × *MultiProduct%* × ***Month***
Retailer⊗Month	*Sales* × *MultiProduct%* × *Year*	
Retailer × *LaggedReturns*	*Sales* × *MultiProduct%* × ***Month***	
Retailer × *MultiProduct%*	***Product⊗Retailer*** × *Year*	
LaggedReturns × *Year*	***Product⊗Retailer⊗Month***	
LaggedReturns × ***Month***	***Product⊗Retailer*** × *LaggedReturns*	
LaggedReturns × *MultiProduct%*	***Product⊗Retailer*** × *MultiProduct%*	
MultiProduct% × *Year*		
MultiProduct% × ***Month***		

by LASSO. Then, in Section 5.3, we provide an interpretation of the "more sparse" LASSO model and its prediction performance for some high volume retailer product pairs. Finally, in Section 5.4 we employ alternative methods and show a comparison of prediction performance among such methods.

5.1. Model specification

We restrict our attention to the following three categories of higher order interaction terms that may help improve prediction performance of the full main effects in model (6). As in model (6), we use vector notations $\boldsymbol{Month} = (January, \ldots, December)^T$, $\boldsymbol{Product} = (Product_1, \ldots, Product_3)^T$ and $\boldsymbol{Retailer} = (Retailer_1, \ldots, Retailer_{13})^T$.

I. Quadratic effect: We are interested in the effect of $Sales^2_{tij}$ because we have shown that sales is a strong predictor of return volume in all the main effects models, and our residual analysis indicated a possible quadratic effect of sales on returns.

II. 2-way interaction effects: The 2-way interactions of predictor variables may help predict returns more accurately. For example, we can ask: Does the effect of sales volume on return volume vary depending on the retailer? The interaction term $Sales \cdot \boldsymbol{Retailer}$ addresses this question. For another example, historical returns are shown to be significant and have positive correlation with returns (see Table 5). However, the effect of historical returns could increase or decrease over time and this would get discovered by including the interaction term $LaggedReturns \cdot Year$.
The question that arises is, which interaction terms should we consider? We begin with our original set of predictors *Sales*, *Year*, ***Month***, ***Product***, ***Retailer*** used in model (5), and additionally consider the two predictors found to be significant at the 0.001 level in model (6)–*MultiProduct%*, *LaggedReturns*; then, we utilize all the 2-way interactions among these predictor variables to create a new set of predictor variables shown in Table 6. Because ***Month***, ***Product*** and ***Retailer*** are each a vector of 12, 3 and 13 dummy variables, we can see, for example, $\boldsymbol{Product} \otimes \boldsymbol{Retailer}$ (here ⊗ denotes Cartesian product) creates 39 interaction terms. Vector $\boldsymbol{V}$ includes all the 2-way interaction terms in Table 6, and one can easily verify there are 337 predictor variables in $\boldsymbol{V}$.

III. 3-way interaction effects. We use 3-way interactions to achieve higher accuracy in returns prediction, and also to address questions such as: Is return rate of any particular retailer predicted to increase over time? Because return rate equals returns divided by sales, the interaction term $Sales \cdot \boldsymbol{Retailer} \cdot Year$ addresses this question. We consider potentially meaningful 3-way interactions among the above-mentioned key predictors–*Sales*, *Year*, ***Month***, ***Product***, ***Retailer***, *MultiProduct%*, *LaggedReturns*. Vector $\boldsymbol{W}$ includes all such 3-way interaction terms shown in Table 6, and one can verify that there are 1336 predictor variables in $\boldsymbol{W}$.

Finally, we specify a new model in Eq. (7) below, which includes all the interactions discussed above.

$$\begin{aligned} Returns_{tij} &= \beta_0 + \beta_1 Sales_{tij} + \beta_2 Year_t + \beta_3^t Month_t \\ &\quad + \beta_4^i Product_i + \beta_5^j Retailer_j + \boldsymbol{\beta_6^T U} \\ &\quad + \beta_7 Sales^2_{tij} + \boldsymbol{\beta_8^T V} + \boldsymbol{\beta_9^T W} + \epsilon_{tij} \end{aligned} \tag{7}$$

This model has a total of $p = 1{,}717$ predictors, which comprise of 43 main effects, 1 quadratic term, 337 2-way interactions, and 1,336 3-way interactions. Recall that in the main effects model (6) we manually dropped 3 binary dummy variables *December*, $Product_3$, $Retailer_{13}$ to remove redundancy, however we no longer do this for model (7) and will instead let LASSO select predictor variables for us. For these reasons, we have 43 main effects now, increased from 40 in model (6). Recall that we have a training set data size $N = 1,140$ and thus we have $p > N$, and we are effectively in a high-dimensional setting. For this reason, some type of penalty is needed to reduce the number of variables used in the model, and most importantly to perform a proper model selection and address problems such as spurious correlation, noise accumulation, and non-uniqueness of the linear model solution. To do these, in Sections 5.2 and 5.3 we employ LASSO, which enjoys an appealing theoretical property–the oracle inequalities (Bickel, Ritov, & Tsybakov, 2009), meaning that under some conditions LASSO can recover the true model with high likelihood, given that the true model is sparse. Since most of the predictor variables in (7) are interaction terms, and not all of them are likely to play an important role in predicting returns, we expect sparse predictive models to fit our data well. To simplify the exposition, henceforth we use

Please cite this article as: H. Cui, S. Rajagopalan and A.R. Ward, Predicting product return volume using machine learning methods, European Journal of Operational Research, https://doi.org/10.1016/j.ejor.2019.05.046

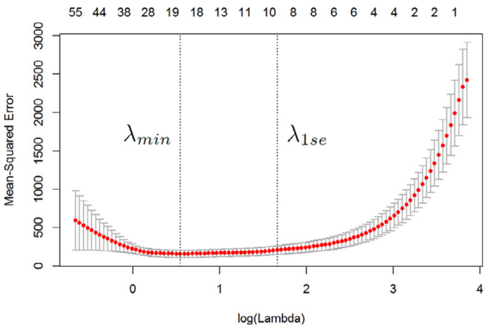


Fig. 4. Selecting λ based on 10-fold CV.

Table 7
Prediction performance of 2 LASSO models.

		Model 7a	Model 7b
Training set	Number of predictors provided	1,717	1,717
	Data size	1,140	1,140
	Tuning parameter type	λ_{min}	λ_{1se}
	Tuning parameter value	1.809	4.378
	Number of predictors selected	19	9
	R^2	0.958	0.934
	MSE	103.754	161.908
Test set	Data size	220	220
	R^2	0.928	0.930
	MSE	177.815	171.824

$\boldsymbol{X}^{\boldsymbol{l}} = (x_{l1}, \ldots, x_{lp})^T$ to denote all the predictor variables, and y_l to denote the response variable in (7) for $l = 1, \ldots, N$.

5.2. LASSO analyses

In this subsection, we focus on LASSO which solves the following convex optimization problem (8) to obtain the LASSO estimate $\hat{\boldsymbol{\theta}}^{Lasso}$ for data $(\boldsymbol{X}^l, y_l)$, $l = 1, \ldots, N$.

$$\hat{\boldsymbol{\theta}}^{Lasso} = \underset{(\theta_0, \boldsymbol{\theta}) \in \mathbb{R} \times \mathbb{R}^p}{\arg\min} \left[\frac{1}{2N} \sum_{l=1}^{N} \left(y_l - \theta_0 - \sum_{m=1}^{p} x_{lm} \theta_m \right)^2 + \lambda \sum_{m=1}^{p} |\theta_m| \right]. \tag{8}$$

Here λ is a tuning parameter, which assigns an appropriate level of penalty, and continuously shrinks coefficients, and can force some to go to exactly zero to obtain a sparse model (see §3 in Hastie et al., 2009). Therefore, it is vital to obtain good values of λ for successful model selection, and in our study we utilize the k-fold cross-validation (CV) method. Generally, $k = 5$ or 10 works well in practice and for our study, we use the option $k = 10$ which is widely used in recent statistics literature. The reason we use CV instead of other measures such as AIC for model selection is to estimate the prediction error in the test set and compare different models (see Section 5 in James et al., 2017)

Notice that without the L_1 penalty term $\lambda \sum_{m=1}^{p} |\theta_m|$, the LASSO problem (8) reduces to a standard OLS problem. With the the L_1 penalty term, the problem (8) is still a convex optimization problem, and this optimization problem can be solved efficiently using cyclical coordinate descent algorithms. It is important that all the predictors be standarized to have mean 0 and standard deviation 1, so that the different scales of variables do not impact the optimization problem. To implement LASSO with 10-fold CV, we utilized R-package *glmnet* from Friedman, Hastie, and Tibshirani (2010).

We show a plot of cross-validation fit with respect to different values of λ in Fig. 4, where the two dotted vertical lines correspond to model fit for two tuning parameters: λ_{min} (left) and λ_{1se} (right). Here λ_{min} denotes the tuning parameter value that achieves the minimum mean CV error, whereas λ_{1se} represents a larger tuning parameter which leads to the smallest model such that mean cross-validated error is within one standard error of the minimum. In our study, we have $\lambda_{min} = 1.809$ and $\lambda_{1se} = 4.378$ and, using these penalty values LASSO gives reduced models 7a and 7b in Table 7, which have only 19 and 9 predictors selected respectively out of a total of 1,717 predictors in Eq. (7). We find that the LASSO models 7a and 7b outperform the two best main effects models (5), (6) with respect to MSE and R^2 in both training and test sets; and the "more sparse" model 7b exhibits a more robust performance with a smaller gap in MSE between training and test sets (161.908 vs. 171.824), when compared to the "less sparse" model 7a (103.754 vs. 177.815.) Hence, we focus on the "more sparse" model 7b referenced as *the LASSO model* hereafter.

In Table 8, we show the prediction performance of model (5), model (6), and the LASSO model (all trained on periods 1 to 33) in each test period from 34 to 39, and also over the entire 6 test periods. One can observe that if we apply these predictive models to the entire 6 test periods (34 to 39), then the LASSO model reduces the MSE of the baseline main effects model (5) by $\frac{257.709-171.824}{257.709} = 33.3\%$, and the MSE of the full main effects model (6) by $\frac{189.386-171.824}{189.386} = 9.3\%$. However, if we apply such predictive models only in the first test period 34, then LASSO model has a more pronounced advantage in prediction performance by reducing the MSE of model (5) by $\frac{206.460-56.030}{206.460} = 72.8\%$, and the MSE of model (6) by $\frac{126.230-56.030}{126.230} = 55.6\%$.

To understand the robustness of the above finding, we conducted further analyses by using the data from periods 1 to $p-1$ to build the model, and then tested the model in period p for each prediction period $p \in \{35, 36, 37, 38, 39\}$. The results are provided in Table 9. The LASSO model reduces the MSE of model (5) by $\frac{144.358-126.085}{144.358} = 12.7\%$ to $\frac{206.46-56.03}{206.46} = 72.9\%$ (average 43.0%) and reduces the MSE of model (6) by $\frac{126.424-126.085}{126.424} = 0.3\%$ to $\frac{126.230-56.03}{126.230} = 55.6\%$ (average 29.7%).

Because forecasting models are often implemented on a rolling horizon basis, it may be advisable to update the LASSO model every month and use it for the first test period for a more accurate prediction. Updating the LASSO model is easier than one might expect. The afrontmentioend R-package *glmnet* that implements LASSO is well-documented. All we need to change are the variable names and the data set names. It took 10 seconds to train and test the LASSO model in our desktop computer (Windows 10, Intel Core i7-8770 CPU, 8Gb RAM 2666MHz, 1Tb 970 Samsung Evo SSD). So, we believe model updating is not going to be onerous from a cost or time perspective.

5.3. Selected predictors and their performance

We first provide the 9 predictor variables and their coefficients in the LASSO model in Table 10. These can be interpreted as the strongest effects identified by LASSO to predict returns, and have been shown to achieve robust prediction performance both in the training and test data. It is noteworthy that the LASSO model chosen is very easy to implement in practice to predict future returns, as one only needs to compute *Sales*, *LaggedReturns*–which can be easily done in a spreadsheet–and use them along with *Year* and two retailer dummy variables *Retailer*$_1$, *Retailer*$_5$.

We first note that the coefficients of all 9 predictor variables are positive, and also that only two main effects *Sales*, *LaggedReturns* are selected by LASSO, which is quite different from main effects models (see Table 5.) Also notice that the LASSO coefficient of *Sales* is 0.0313, reduced from 0.099 in model (6) shown in Table 5. A key reason is that the effect of *Sales* is broken down and absorbed

Table 8
Prediction performance of three models (trained on periods 1 to 33) for each of the 6 future periods.

	Period	34	35	36	37	38	39	Over the entire 6 test periods
MSE	Model 5	206.460	331.218	243.908	163.228	308.388	297.050	257.709
	Model 6	126.230	242.053	202.965	93.578	224.846	251.791	189.386
	LASSO	56.030	156.162	86.369	131.564	276.216	313.843	171.824
R^2	Model 5	0.833	0.748	0.945	0.941	0.896	0.833	0.895
	Model 6	0.903	0.816	0.954	0.966	0.924	0.859	0.923
	LASSO	0.960	0.881	0.980	0.952	0.907	0.824	0.930

Table 9
Prediction performance for each of the 6 future periods with models updated every period.

	Period	34	35	36	37	38	39
MSE	Model 5	206.46	319.520	252.153	144.358	282.045	256.877
	Model 6	126.23	204.381	209.460	126.424	217.225	246.960
	LASSO	56.030	103.170	100.651	126.085	192.630	223.589
R^2	Model 5	0.833	0.756	0.943	0.948	0.905	0.856
	Model 6	0.903	0.844	0.953	0.954	0.927	0.861
	LASSO	0.960	0.921	0.977	0.954	0.935	0.874

Table 10
Predictors and coefficients selected by LASSO w.r.t. λ_{1se}.

Predictor category	Predictors selected	Coefficients
main effects	*Sales*	3.13E-02
	LaggedReturns	3.12E-02
quadratic effect	$Sales^2$	2.80E-05
	Sales × *Year*	1.31E-06
	$Sales \times Retailer_1$	9.34E-03
2-way interactions	$Sales \times Retailer_5$	7.61E-03
	Sales × *LaggedReturns*	6.89E-08
	LaggedReturns × *Year*	2.67E-06
3-way interactions	$Sales \times Year \times Retailer_1$	8.46E-07

into quadratic term $Sales^2$, as well as 2-way and 3-way interaction terms which include *Sales*. As the coefficients of *Sales* and $Sales^2$ are both positive, the LASSO model shows that the predicted returns are convex and increasing in sales. This suggests that the likelihood of returns increases for each incremental unit sold. One possible explanation is that production process capacity limitations are pushed as sales increase (for example, sewers are pressured to work faster) leading to more errors and thus returns.

We then observe the 2-way interaction terms selected by the LASSO model. *Sales* × *Year* suggests that as time goes by, sales has a greater effect on returns, that is, we can expect higher return rates year over year. This may be because as competition increases, retailers become more lax in accepting returns to please customers and keep them from going to competitors (e.g., see Montaldo, 2019). Recall that only $Retailer_1$ (auto specialty store with a strong bargaining power) and $Retailer_5$ (warehouse club) carry liberal return policies, and the two interaction terms $Sales \times Retailer_1$ and $Sales \times Retailer_5$show that for the same sales, these two retailers contribute more to returns than other retailers. *Sales* × *LaggedReturns* indicates that when historical returns are higher, sales generates higher returns, i.e. higher return rate. Also we see from *LaggedReturns* × *Year* that the impact of historical returns increases over time.

We now pay attention to the only 3-way interaction terms selected by the LASSO model: $Sales \times Year \times Retailer_1$. This shows that return rate of retailer 1, which carries a liberal return policy, has increased over the years more than for others. In a way, sales through this retailer are becoming “more problematic”.

We end this subsection by comparing the prediction performance for the LASSO model vs. models (5), (6). We do this for the retailer and product pairs that have higher observed returns, because that is where all three models show the best prediction performance. Fig. 5 shows how well the three models predict returns for the future 6 periods for top 6 retailer-product pairs which account for 68.3% of total return volume. When compared to models (5), (6), we find that the LASSO model predicts returns closer to the observed returns when they are high enough. It is also interesting to observe that for $Retailer_6$, $Product_1$ the LASSO model predicts returns substantially better than models (5), (6) for the first test period, as is consistent with our discussion in the last paragraph of Section 5.2. For low volume returns, the three models overall do not provide nearly as good a prediction performance. This is expected because such low volume returns have higher coefficient of variation. For completeness, we provide prediction results for all retailer product pairs for future 6 months in Fig. A1 in Online Appendix A5.

5.4. Alternative methods to LASSO

Our goal in this subsection is to fit three alternative high-dimensional statistical models, and two tree-based statistical machine learning methods to our data and compare their prediction performance against LASSO. The results are provided in Table 11.

5.4.1. High-dimensional machine learning methods

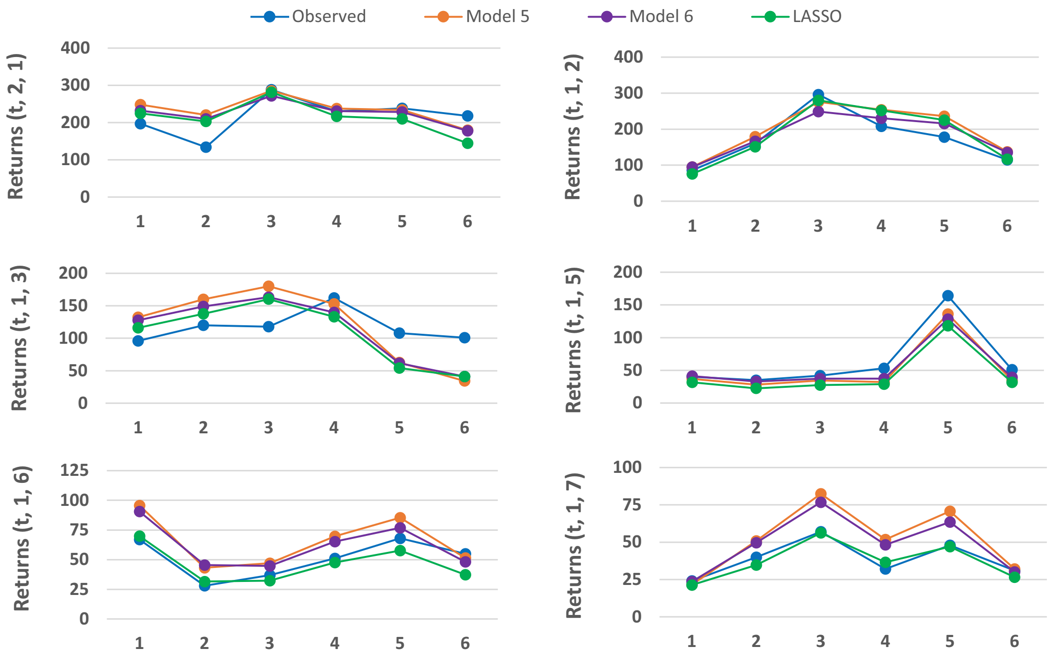
In this subsection, we discuss three alternative high-dimensional statistical models. In our setting, they all select 9 predictors which we discuss in detail below.

- The first model we consider is called LARS-OLS hybrid (Efron et al., 2004) or OLS post-LASSO (Belloni & Chernozhukov, 2013). This model involves two stages: in the first stage, it utilizes LASSO to select predictor variables only; in the second stage, it uses OLS to estimate the coefficients of the selected predictor variables–also known as debiasing. By construction, this method uses exactly the same 9 predictor variables as LASSO (see Table 10). The rationale for this approach is to reduce bias in the coefficients estimated by LASSO, because LASSO tends to shrink the nonzero coefficients towards zero compared to OLS. Similar to Efron et al. (2004), in our study we find that the LARS-OLS hybrid method achieves a smaller MSE in the training data set, compared to the LASSO model; however, it results in a large MSE in the test set.
- Recall that LASSO is a convex regularization method with L_1 penalty. Next, we consider a concave regularization method–Smoothly Clipped Absolute Deviation (SCAD), which yields nearly unbiased estimators (Fan & Li, 2001). Both LASSO and SCAD solve the following penalized least-squares problems to achieve sparse solutions:

$$\min_{(\theta_0,\boldsymbol{\theta})\in\mathbb{R}\times\mathbb{R}^p}\left[\frac{1}{2N}||\boldsymbol{y}-\theta_0\mathbf{1}-\boldsymbol{X\theta}||_2^2+||p_{\boldsymbol{\lambda}}(\boldsymbol{\theta})||_1\right] \qquad (9)$$

where we use compact notations for data $(\boldsymbol{X}, \boldsymbol{y})$, where $\boldsymbol{X}\in\mathbb{R}^{N\times p}$ is a design matrix, $\boldsymbol{y}\in\mathbb{R}^N$ is a response vector, and $p_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = p_{\boldsymbol{\lambda}}(|\boldsymbol{\theta}|) = (p_{\boldsymbol{\lambda}}(|\theta_1|), p_{\boldsymbol{\lambda}}(|\theta_2|), \dots, p_{\boldsymbol{\lambda}}(|\theta_p|))^T$ is a

Please cite this article as: H. Cui, S. Rajagopalan and A.R. Ward, Predicting product return volume using machine learning methods, European Journal of Operational Research, https://doi.org/10.1016/j.ejor.2019.05.046



Horizontal axis denotes test period t = 1 to 6. Vertical axis is return volume in period t, product i, retailer j.

Fig. 5. Observed vs. predicted returns for some retailer product pairs (see colors online).

Table 11
Prediction performance of alternative methods against LASSO.

		LASSO	LARS-OLS hybrid	SCAD	Elastic Net[a]	Random Forest	Gradient Boosting
Training set	Data size	1140	1140	1140	1140	1140	1140
	R^2	0.934	0.961	0.947	0.934	0.990	0.996
	MSE	161.908	94.888	129.674	161.908	25.036	10.606
Test set	Data size	220	220	220	220	220	220
	R^2	0.930	0.912	0.874	0.930	0.878	0.882
	MSE	171.824	216.610	308.920	171.824	298.208	290.564

[a] Note: Optimal Elastic Net is reduced to LASSO.

penalty for a vector of regression coefficients. Here, the penalty function $p_{\boldsymbol{\lambda}}(t)$ is defined on $t \in [0, \infty)$ indexed by possibly more than one tuning parameter in $\boldsymbol{\lambda}$. It can be observed that LASSO is a special case of (9) with one tuning parameter $\lambda \geq 0$ and a penalty function $p_{\boldsymbol{\lambda}}(t) := p_{\lambda}(t) = \lambda t$. On the other hand, SCAD utilizes two tuning parameters $\lambda \geq 0$, $\gamma > 2$ and a penalty function defined as $\rho'_{\lambda,\gamma}(t) = \lambda\left\{I(t \leq \lambda) + \frac{(\gamma\lambda - t)_+}{(\gamma-1)\lambda} I(t > \lambda)\right\}$, which leads to

$$p_{\boldsymbol{\lambda}}(t) := p_{\lambda,\gamma}(t) = \left[\lambda t - \frac{(t-\lambda)^2}{2(\gamma-1)} I(t > \lambda)\right] I(t \leq \lambda\gamma) + \frac{1}{2}\lambda^2(\gamma+1) I(t > \lambda\gamma).$$

Notice that $p_{\lambda,\gamma}(t)$ is concave in t, and as $\gamma \to \infty$, the penalty functions of SCAD and LASSO coincide for $t > 0$. To solve (9) for SCAD, one may search for the optimal values for λ, γ over two-dimensional space, however it is shown that $\gamma \approx 3.7$ is shown to be a robust choice (see §2.1 in Fan & Li, 2001). In our study, we considered $\gamma \in [3.5, 4.0]$ in R-package *ncvreg* (Breheny & Huang, 2011) with 10-fold CV, and find that the choice of γ has little impact on MSE in the training and test sets, and we choose $\gamma = 3.9$ which leads to 9 predictor variables, which is not surprising (see a simulated example comparing SCAD against LASSO in Table 3 in Fan, Feng, & Wu, 2009). Like LARS-OLS hybrid, SCAD outperforms LASSO in the training data; however, LASSO still achieves smaller MSE in the test set in our study.

- A notable generalization of LASSO is the Elastic Net (Zou & Hastie, 2005) which makes a compromise between the L_1 (LASSO) and L_2 (Ridge) penalties, and solves the following optimization problem:

$$\min_{(\theta_0, \boldsymbol{\theta}) \in \mathbb{R} \times \mathbb{R}^p} \left[\frac{1}{2N}||\boldsymbol{y} - \theta_0 \mathbf{1} - \boldsymbol{X}\boldsymbol{\theta}||_2^2 + \lambda\left(\alpha||\boldsymbol{\theta}||_1 + \frac{(1-\alpha)}{2}||\boldsymbol{\theta}||_2^2\right)\right] \tag{10}$$

Notice that (10) involves two tuning parameters $\alpha \in [0, 1]$ and $\lambda \geq 0$, and it is reduced to LASSO when $\alpha = 1$, and to Ridge regression when $\alpha = 0$. In some simulation studies (e.g., Table 1, Table 2 in Zou & Hastie, 2005), Elastic Net is shown to outperform both LASSO and Ridge. To implement Elastic Net for our study, for a range of α values between 0 and 1, we utilize the R-package *glmnet* with 10-fold CV to derive predictive models with respect to λ_{1se} as we do for LASSO. We find that $\alpha = 1$ leads to a model with the minimum MSE in the training set, therefore conclude that the optimal Elastic Net is reduced to LASSO, thus also selects 9 predictors.

5.4.2. Tree-based machine learning methods

We next consider two statistical machine learning methods based on trees, namely Random Forest (Breiman, 2001) and Gradient Boosting (Friedman, 2001), which may further improve prediction performance when complex, nonlinear structures are present in the data. Both of these methods are rooted in the ensemble idea, that is, producing multiple trees which are then combined to create a single model to improve the prediction accuracy. Unlike the aforementioned high-dimensional statistical methods that yield sparse predictive models, the ensemble models based on Random Forest and Gradient Boosting can be difficult to interpret despite potential improvement in prediction accuracy.

- **Random forest.** To produce multiple trees, this method employs bootstrap (Efron, 1979) to create B training samples, and then grows a random-forest tree from each bootstrapped sample until a certain minimum node size (e.g., 5) is reached. Often the prediction performance improves sharply in the beginning as the number of trees increases, but the performance stabilizes when we have enough trees.

  We begin our analysis by using 500 trees denoted by $B = 500$ as recommended in Hastie et al. (2009). Growing the tree is done as follows: in each and every step of split in the tree, randomly choose m predictors out of all p predictors as split candidates, and then use only the best one out of these m predictors. A key idea behind the Random Forest method is the random sampling of split candidates, which results in a fresh sample of m predictors chosen at each split. Thus, the choice of m impacts the prediction performance and m becomes a tuning paramater and this parameter is optimized by cross-validation. We follow Bertsimas et al. (2016) and consider m values among $\{\lfloor\frac{1.5^2 p}{3}\rfloor, \lfloor\frac{1.5^1 p}{3}\rfloor = 859, \lfloor\frac{p}{3}\rfloor, \ldots, \lfloor\frac{1.5^{-15} p}{3}\rfloor = 1\}$. We use the minimum node size 5 which ensures no split when growing the tree if the node size is less than 5 (see an example in Breiman, 2001).

  To avoid a potential overfitting issue, we conducted further numerical experiments with larger values of minimum node size from 5 to 50 similar to Fig. 15.8 in Hastie et al. (2009) but did not see an improvement in prediction performance in the training set (see Table A4 in Online Appendix A3). We also experimented with B from 500 to 2,500 but the prediction performance in the training set did not improve and remained quite similar (see Table 5 in Online Appendix A3). We did not consider tuning another potential parameter, tree depth, because it is controlled by the minimum node size–the larger the minimum node size, the shallower the trees.

  We implemented Random Forest in the R-package *randomForest* (Liaw & Wiener, 2002) and found the optimal $m = 859$ by 10-fold CV. This method fits the training data set well, in particular, it achieves a very high $R^2 = 0.990$. In the test set, however, it obtains lower prediction accuracy as compared to LASSO.
- **Gradient boosting.** This alternative tree-based machine learning method, in contrast to Random Forest, grows trees in a sequential way to reduce bias as described next. We first build up to B trees each with $d + 1$ terminal nodes to the training data set, then repeatedly update each tree by adding a shrunken version of the new tree by a factor, $\nu \in (0, 1)$. Note that the new tree is fit on the current residuals, not on the response variables. At each iteration, a fraction ζ of the training data is sampled without replacement which is further used to grow the next tree. Because this is done in an adaptive way, the new tree depends on the previous tree. Finally, we combine all the trees to obtain an ensemble model as we did for Random Forest.

  As discussed above, we consider four key parameters to fit a Gradient Boosting model: d, ν, ζ, B (see more details in §10.12 in Hastie et al., 2009) and implemented Gradient Boosting in the R-package *gbm* (Ridgeway, 2007). The parameter d is often called depth of the tree; ν is referred to as shrinkage parameter or learning rate; ζ is the subsampling rate; the optimal number of iterations B^* is determined by cross validation, which is also denoted as early stopping in some literature (e.g., see Yao, Lorenzo Rosasco, & Andrea Caponnetto, 2007; Zhang & Yu, 2005).

  We use $d = 1$ because our set of predictor variables already includes higher-order interaction terms. We begin model fitting by exploring a range of small shrinkage parameters $\nu \in [0.0005, 0.1]$ as suggested by Hastie et al. (2009) with default subsampling rate, $\zeta = 0.5$ which tends to work well though perhaps is not optimal. For each ν we allow the algorithm to grow up to 5000 trees and for this ν value we derive the corresponding optimal B^* by 10-fold CV. For example, for $\nu = 0.1$ we find that the best cross-validation iteration is obtained in 1234 iterations, therefore the corresponding $B^* = 1234$. We find that in our data set $\nu < 0.01$ leads to a poorer model fit in the training data sets. Therefore, we focus on ν values between 0.01 and 0.1 and further calibrate the model by tuning both parameters ν, ζ. We provide details on the model calibration in Tables A6 and A7 in Online Appendix A4.

  The best calibrated model with respect to the training data set is obtained by $\nu = 0.1, \zeta = 0.4, B^* = 1140$ which achieves MSE $=$ 10.606 and $R^2 = 0.996$; however, in the test set it obtains MSE $=$ 290.564 and $R^2 = 0.882$. This gradient boosting model performs slightly better than Random Forest in our study but does not outperform LASSO in the test set.

To summarize, we find that the models with unbiased estimators (LARS-OLS hybrid, SCAD) fit the training set better than LASSO, but yield lower prediction accuracy in the test set, which suggests bias-variance trade-off as discussed above. The reduction of Elastic Net to LASSO indicates that the true underlying model is much more likely a sparse model as we expected at the end of Section 5.1. It is not surprising that the two tree-based methods (Random Forest, Gradient Boosting) fit the training set very well, however, their prediction performance in the test set suggests likely overfitting issues of these complex models despite the optimal choice of tuning parameters. In our setting, the LASSO model shows up as the best because it obtains the smallest MSE in the test set among all the models considered, and also because its MSE in the training and test sets are very close, which suggests that the LASSO model is a robust choice.

6. Robustness checks

In our main analysis, we focused on training our predictive models on 33 periods $\{1, \ldots, 33\}$ and then testing the prediction performance for the 6 future periods $\{34, \ldots, 39\}$. In this section, we evaluate the robustness of three predictive models: models (5), (6) and the LASSO model on different test sets. In our study, we focus on empirically investigating whether the LASSO model consistently selects the same 9 predictors as in the main analysis shown in Table 10, and also check the sensitivity of prediction performance on such test sets. In Section 6.1, we construct different test sets that contain 6 randomly chosen periods, for example, {3, 13, 25, 34, 35, 37}. Next, in subsection 6.2, we check the robustness of the length of test period using alternative test sets with 5 and 7 future periods. For both of these scenarios, we compare the results with our finding in the main analysis.

6.1. Test on 6 random periods

In this subsection, we construct a test set by randomly choosing 6 periods over the set of periods $\{1, \ldots, 39\}$, and assign the remaining 33 periods to the training set. We repeat this process 4 times to create 4 such test sets to check if different random test

Table 12
Comparison of 4 cases of random test periods against the main analysis for models 5 and 6.

		Main analysis		Case 1		Case 2		Case 3		Case 4	
		model 5	model 6	model 5	model 6	model 5	model 6	model 5	model 6	model 5	model 6
Training set	Data size	1,140	1,140	1,145	1,145	1,146	1,146	1,152	1,152	1,154	1,154
	R^2	0.912	0.921	0.910	0.919	0.911	0.922	0.914	0.924	0.913	0.924
	MSE	214.050	192.395	224.295	202.027	226.203	196.866	212.314	186.034	219.111	190.382
Test set	Data size	220	220	183	183	256	256	256	256	256	256
	R^2	0.895	0.923	0.917	0.941	0.912	0.922	0.897	0.912	0.901	0.912
	MSE	257.709	189.386	188.735	134.561	175.970	156.695	248.364	213.197	209.560	187.039

Table 13
Comparison of 4 cases of random test periods against the main analysis for the LASSO model.

		Main analysis	Case 1	Case 2	Case 3	Case 4
Training set	Data size	1,140	1,145	1,146	1,152	1,154
	λ_{1se}	4.378	5.273	4.619	5.245	4.824
	R^2	0.934	0.924	0.932	0.925	0.928
	MSE	161.908	189.366	172.608	183.696	180.537
Test set	Data size	220	215	214	208	206
	R^2	0.930	0.944	0.936	0.916	0.924
	MSE	171.824	128.335	128.143	204.272	160.598

sets lead to substantially different results. We obtain test periods {3, 13, 25, 34, 35, 37} for case 1, {10, 15, 19, 27, 28, 34} for case 2, {4, 8, 15, 25, 29, 38} for case 3, {2, 5, 17, 27, 30, 33} for case 4. For each of the 4 cases, we build three predictive models using model (5), (6) and the LASSO model on the corresponding training sets, then test their prediction performance in each test set. The results are given in Table 12 for models (5) and (6), and Table 13 for the LASSO model.

First, we find that the LASSO model selects the same 9 predictors in cases 1 to 4 as the ones found in the main analysis. This is consistent with the theory of LASSO's consistent variable selection with high probability that was established under the irrepresentable conditions (Meinshausen & Buhlmann, 2006; Zhao & Yu, 2006). Second, we observe that the prediction performance in the 4 cases shows some fluctuation for different test periods. Third, we find that the LASSO model outperforms both main effects models in the training and test sets in all 4 cases as in the main analysis. The advantage of the LASSO model over model (6) is not as substantial, when compared to model (5). For instance, the LASSO model reduces the test MSE in model (5) by 33.3% in the main analysis, and between 17.8% and 32.0% among 4 cases. By comparison, the LASSO model reduces the test MSE in model (6) by 9.3% in the main analysis, and between 4.2% and 18.2% among the 4 cases.

6.2. Test on 5 and 7 future periods

Our goal in this subsection is to understand if the choice of training set periods impacts the predictive models and their performance. To do this, we consider two alternative analyses: train all three predictive models on periods $\{1, \ldots, 34\}$ and test their performance on periods $\{35, \ldots, 39\}$; then also train on periods $\{1, \ldots, 32\}$ and test on periods $\{33, \ldots, 39\}$. We show the results in Table 14.

Similar to the main analysis, we find that the LASSO model continues to select the same set of 9 predictors in the two new test sets containing 5 and 7 future periods. This indicates that the "sparse model" selection by LASSO is consistent and robust with respect to the different test periods we considered.

Now we compare prediction performance using the measure MSE. In the main analysis, the choice of LASSO over models (5) and (6) achieved a reduction in the test MSE by 33.3% and 9.3% respectively. When we train our models on 34 periods and test on 5 periods, the LASSO model reduces the test MSE of model (5) by 15.8%. However, it increases the test MSE of model (6) by 10.4%. This indicates that LASSO does not outperform model (6) when the MSE was computed over the 5 test periods. Nonetheless, if we only consider the prediction performance in the first test period (trained on periods 1 to 34, tested on period 35 only) we can see from Table 8 in §5.2 that LASSO still beats model (6) and reduces the MSE by $\frac{204.381-103.170}{204.381} = 49.5\%$. In summary, even though the full main effects model (6) is competitive in prediction performance compared to the LASSO model based on MSE computed over the entire set of test periods, we find that the LASSO model consistently outperforms both main effects models in the first few periods. This suggests that the LASSO model is the better choice when the predictive model is implemented on a rolling horizon basis.

7. Discussion

In this paper, we developed data-driven models to predict future return volume and applied them to a firm that sells car accessories with large product variety. We explored various factors that could help predict return volume–sales, time, product features, retailer, production process and resources, multi-product order and historical returns, which are not exclusive to the particular firm in our study. We evaluated various main effects models and an interaction effect model with convex and concave regularization, to build prediction models that have good prediction accuracy and are easy to implement in practice and also provide a preliminary understanding of important factors associated with return volume. We also considered tree-based machine learning methods to improve on the prediction accuracy. We found that LASSO was effective in selecting a small number of interaction terms, which are useful in prediction, out of a large number of possible candidates. LASSO identified a parsimonious model that achieved the highest accuracy in predicting future return volume.

Sales show up as an important predictor across all models. In the LASSO model, we find that the predicted return volume is convex and increasing in sales, suggesting that the likelihood of returns increases for each incremental unit sold. One possible explanation is that production process capacity limitations are pushed as sales increase (for example, sewers are pressured to work faster) leading to more errors and thus returns. The LASSO model also

Please cite this article as: H. Cui, S. Rajagopalan and A.R. Ward, Predicting product return volume using machine learning methods, European Journal of Operational Research, https://doi.org/10.1016/j.ejor.2019.05.046

Table 14
Comparison of 5 and 7 test periods against the main analysis (6 periods) for three models.

		Main analysis			Test on 5 periods			Test on 7 periods		
		model 5	model 6	LASSO	model 5	model 6	LASSO	model 5	model 6	LASSO
Training set	Data size	1140	1140	1140	1177	1177	1177	1104	1104	1104
	λ_{1se}	NA	NA	4.378	NA	NA	5.484	NA	NA	5.321
	R^2	0.912	0.921	0.934	0.911	0.921	0.922	0.914	0.922	0.926
	MSE	214.050	192.395	161.908	213.599	190.106	188.451	214.041	192.990	182.958
Test set	Data size	220	220	220	183	183	183	256	256	256
	R^2	0.895	0.923	0.930	0.903	0.926	0.918	0.886	0.914	0.927
	MSE	257.709	189.386	171.824	259.609	197.879	218.545	259.903	197.624	166.056

suggests that as time goes by, sales has a greater effect on returns, that is, we can expect higher return rates year over year.

Historical returns and retailer fixed effects are found to be useful in predicting returns in all models. Historical returns capture recent trends in returns that are due to both defects and non-defects. This variable may capture the effect of trends in manufacturing defects and/or consumers' impulse purchasing behaviors in recent months. Retailers may have different return policies which may influence returns due to non-defects, but not returns due to defects. The LASSO model selects only two retailers (auto specialty store and warehouse club) with liberal return policies in predicting returns, and only when these two retailers interact with other effects.

The two-way interactions between retailers and sales indicate that both these retailers are predicted to yield higher returns rates than others. In addition, the three-way interaction term shows that the auto specialty store, but not the warehouse club, is predicted to have increasing return rates over time. This should be of concern to Company A because this retailer accounted for 24.0% of total sales, and had the highest return rate at 13.6% between July 2012 and September 2015. Since returned products at Company A cannot be resold in most cases, they will have to carefully weigh the cost of such a high return volume from this retailer in determining the terms of trade with this retailer. They should also explore the reasons for the high returns with the retailer.

It was expected ex ante that product type would be useful in predicting returns. The rationale was that the three products–dash cover, seat cover and car cover–have distinct features, designs, purchase/usage characteristics and they go through different manufacturing processes that may have varying defect rates. In fact, the results from the main effects models (see Table 5) show that dash cover products lead to significantly lower return volume than vehicle cover products. The LASSO model, however, did not select product type in predicting return volume. Similarly, the LASSO model did not select any of the production process and resources related variables, which were ex ante expected to be related to defects and useful in predicting returns. This suggests that the returns due to defects in the firm in our study may not be substantial, which is consistent with the finding in a study (Lawton, 2008) that only about 5% of returns are due to true defects.

We noticed in our analysis that products with logos or other types of customization appeared to have lower returns. However, we do not find them to be useful in prediction in any of the models, likely because sales of such highly customized products are very small at Company A. However, the lower return rate suggests that perhaps Company A could consider promoting such products to consumers.

The inability to distinguish returns due to defect from other returns is one of the limitations of our study. Unfortunately, Company *A* does not systematically track the reasons for returns. If we are able to accurately identify returns due to defects, then we could ask interesting questions related to defects: whether defects differ substantially among different products, whether production process and resources are important in predicting defects, etc.

An important issue for future research is the potential costs and benefits of retailer return policies. For example, do we want retailers to strictly enforce return policy or should we allow retailers to be liberal in their return policy and in enforcing it? With stricter return policies and enforcement, there may be fewer returns, however, this may discourage consumers from making purchases. The trade-off between sales and returns is not immediately clear. A field experiment, which modifies return policies of select retailers and tracks subsequent changes in returns, may be able to address this question. When we can accurately extract defects among returns, we may be able to address some system design questions, by analyzing effects of production process and resources on defects. For example, should we increase random inspections by supervisors, if we find that such random inspections reduce defects? Another example is – do workers who speed up produce greater defects resulting in higher returns?

Acknowledgment

We would like to express our sincere gratitude to Richard Vidgen and three anonymous referees who provided critical feedback that helped us significantly improve the manuscript.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ejor.2019.05.046

References

Akturk, M. S., Ketzenberg, M., & Heim, G. R. (2018). Assessing impacts of introducing ship-to-store service on sales and returns in omnichannel retailing: A data analytics study. *Journal of Operations Management, 61*, 15–45.
Alptekinoğlu, A., & Grasas, A. (2017). When to carry eccentric products? Optimal retail assortment under consumer returns. *Production and Operations Management, 23*(5), 877–892.
Altug, M. S., & Aydinliyim, A. (2016). Counteracting strategic purchase deferrals: the impact of online retailers return policy decisions. *Manufacturing & Service Operations Management, 18*(3), 376–392.
Anderson, E. T., Karsten, H., & Duncan, S. (2009). The option value of returns: Theory and empirical evidence. *Marketing Science, 28*(3), 405–423.
Belloni, A., & Chernozhukov, V. (2013). Least squares after model selection in high--dimensional sparse models. *Bernoulli, 19*(2), 521–547.
Bertsimas, D., O'Hair, A., Relyea, S., & Silberholz, J. (2016). An analytics approach to designing combination chemotherapy regimens for cancer. *Management Science, 62*(5), 1511–1531.
Bickel, P. J., Ritov, Y., & Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics, 37*(4), 1705–1732.
Box, G. E. P., & Cox, D. R. (1964). An analysis of transformation. *Journal of the Royal Statistical Society, Series B., 26*(2), 211–252.
Breheny, P., & Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics, 5*(1), 232–253.
Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics, 24*(6), 2350–2383.
Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Calmon, A. P., & Graves, S. C. (2017). Inventory management in a consumer electronics closed-loop supply chain. *Manufacturing & Service Operations Management, 19*(4), 568–585.

Cerag, P., Ferguson, M., & Toktay, L. B. (2016). Extracting maximum value from consumer returns: Allocating between remarketing and refurbishing for warranty claims. *Manufacturing & Service Operations Management, 18*(4), 475–492.

Cheng, A. (2015). Consumers return $642.6 billion in goods each year. *MarketWatch, Inc.*. Accessed April 15, 2018, http://www.marketwatch.com/story/consumers-return-6426-billion-in-goods-each-year-2015-06-18.

DeCroix G, A., Song, J. S., & Zipkin, P. H. (2009). Managing an assemble-to-order system with returns. *Manufacturing & Service Operations Management, 11*(1), 144–159.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*(1), 1–26.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics, 32*(2), 407–499.

Fan, J., Feng, Y., & Wu, Y. (2009). Network exploration via the adaptive LASSO and SCAD penalties. *The Annals of Applied Statistics, 3*(2), 521–541.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal American Statistical Association, 96*(456), 1348–1360.

Fan, J., Lv, J., & Qi, L. (2011). Sparse high-dimensional models in economics. *Annual Review of Economics., 3*, 291–317.

Foster, D. P., & George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics, 22*(4), 1947–1975.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software., 33*(1).

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer-Verlag.

Hess, J. D., & Mayhew, G. E. (1997). Modeling merchandise returns in direct marketing. *Journal of Interactive Marketing, 11*(2), 20–35.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics, 32*(1), 1–49.

Insider, B. (2017). *National retail federation estimates 8–12% US e-commerce growth in 2017*. Accessed April 15, 2018, http://www.businessinsider.com/national-retail-federation-estimates-8-12-us-e-commerce-growth-in-2017-2017-2.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning*. Springer-Verlag.

Janakiramana, N., Syrdalb, H. A., & Freling, R. (2016). The effect of return policy leniency on consumer purchase and return decisions: A meta-analytic review. *Journal of Retailing, 92*(2), 226–235.

Lawton, C. (2008). The war on returns. *Wall Street Journal*. (May 8) D1, Accessed April 15, 2018, https://www.wsj.com/articles/SB121020824820975641.

Li, G., Li, L., Sethi, S. P., & Guan, X. (2017). Return strategy and pricing in a dual-channel supply chain. *International Journal of Production Economics*. doi:10.1016/j.ijpe.2017.06.031.

Liaw, A., & Wiener, M. (2002). Classification and regression by random-forest. *R News, 2*(3), 18–22. http://CRAN.R-project.org/doc/Rnews/.

Ma, S., Fildes, R., & Huang, T. (2016). Demand forecasting with high dimensional data: The case of SKU retail sales forecasting with intra- and inter-category promotional information. *European Journal of Operational Research, 249*(1), 245–257.

Martin, C. (2018). *Mobile millennials: 63% shop on smartphones every, day, 53% buy in stores*. Accessed April 15 https://www.mediapost.com/publications/article/282639/mobile-millennials-63-shop-on-smartphones-every.html.

Martinez, A., Schmuck, C., Pereverzyev, S., Pirkerc, C., & Haltmeier, M. (2018). A machine learning framework for customer purchase prediction in the non-contractual setting. *European Journal of Operational Research*. doi:10.1016/j.ejor.2018.04.034.

Meinshausen, N., & Buhlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics, 34*(3), 1436–1462.

Montaldo, D. L. (2019). Target bucks the trend with a new super lenient return policy. *Manufacturing and Technology News*. Accessed April 15, 2018, https://www.thebalance.com/target-s-new-return-policy-939859.

Mortenson, M. J., Doherty, N. F., & Robinson, S. (2015). Operational research from Taylorism to Terabytes: A research agenda for the analytics age. *European Journal of Operational Research, 241*(3), 583–595.

Oztekin, A., Al-Ebbini, L., Sevkli, Z., & Delen, D. (2018). A decision analytic approach to predicting quality of life for lung transplant recipients: a hybrid genetic algorithms-based methodology. *European Journal of Operational Research, 266*(2), 639–651.

Petersen, J., & Kumar, V. (2009). Are product returns a necessary evil? Antecedents and consequences. *Journal of Marketing, 73*(3), 35–51.

Ridgeway, G. (2007). *Generalized boosted models: A guide to the GBM package*. Accessed April 15,2018, http://www.saedsayad.com/docs/gbm2.pdf.

Rudolph, S. (2016). E-commerce product return statistics and trends. *Business 2 Community*. Accessed April 15, 2018, http://www.business2community.com/infographics/e-commerce-product-return-statistics-trends-infographic-01505394.

Ryzhov, I. O., Han, B., & Bradic, J. (2016). Cultivating disaster donors using data analytics. *Management Science, 62*(3), 849–866.

Sevim, C., Oztekin, A., Bali, O., Gumus, S., & Guresen, E. (2014). Developing an early warning system to predict currency crises. *European Journal of Operational Research, 237*(3), 1095–1104.

Shang, G., Ghosh, B. P., & Galbreth, M. (2017b). Optimal retail return policies with wardrobing. *Production and Operations Management, 26*(7), 1315–1332.

Shang, G., Pekgün, P., Ferguson, M., & Galbreth, M. (2017). How much do online consumers really value free product returns? Evidence from eBay. *Journal of Operations Management, 53–56*, 45–62.

Shulman, J. D., Coughlan, A. T., & Savaskan, R. C. (2011). Managing consumer returns in a competitive environment. *Management Science, 57*(2), 347–362.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B., 36*(2), 111–147.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series B., 39*(1), 44–47.

Su, X. (2009). Consumer returns policies and supply chain performance. *Manufacturing & Service Operations Management, 11*(4), 595–612.

The Economist (2014). *The business of reselling returned shop items: What happens to all the goods shoppers don't want*. Accessed April 15, 2018, https://www.economist.com/news/business/21710855-what-happens-all-goods-shoppers-dont-want-business-reselling-returned-shop-items.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B., 58*(1), 267–288.

Toktay, L. B. (2003). Forecasting product returns. business aspects of closed-loop supply chains. In V. D. R. Guide, & L. N. Van Wassenhove (Eds.), *International management series. vol 2*. Carnegie Bosch Institute.

Tsiliyannis, C. A. (2018). Markov chain modeling and forecasting of product returns in remanufacturing based on stock mean-age. *European Journal of Operational Research, 271*(2), 474–489.

Ülkü, M. A., & Gürler, U. (2017). The impact of abusing return policies: A newsvendor model with opportunistic consumers. *International Journal of Production Economics, 203*, 124–133.

Urbanke, P., Kranz, J., & Kolbe, L. M. (2015). Predicting product returns in e-commerce: The contribution of mahalanobis feature extraction. *Proceedings of the thirty sixth international conference on information systems*.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives, 28*(2), 3–28.

Wilms, I., Basu, S., Bien, J., & Matteson, D. (2017). Interpretable vector autoregressions with exogenous time series. *NIPS symposium proceedings*. http://arxiv.org/abs/1711.03623.

Yao, Y., Lorenzo Rosasco, L., & Andrea Caponnetto, A. (2007). On early stopping in gradient descent learning. *Constructive Approximation, 26*(2), 289–315.

Zhang, T., & Yu, B. (2005). Boosting with early stopping: Convergence and consistency. *The Annals of Statistics, 33*(4), 1538–1579.

Zhao, P., & Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research, 77*, 2541–2563.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B, 67*(2), 301–320.

Please cite this article as: H. Cui, S. Rajagopalan and A.R. Ward, Predicting product return volume using machine learning methods, European Journal of Operational Research, https://doi.org/10.1016/j.ejor.2019.05.046