Improved Decision Support for Product Returns using Probabilistic Prediction

Dirar Sweidan
School of Informatics
University of Skövde, Sweden
Dept. of Information Technology
University of Borås, Sweden
dirar.sweidan@his.se
dirar.sweidan@hb.se

Ulf Johansson
Department of Computing
Jönköping University
Jönköping, Sweden
ulf.johansson@ju.se

Beatrice Alenljung
School of Informatics
University of Skövde
Skövde, Sweden
beatrice.alenljung@his.se

Anders Gidenstam
Dept. of Information Technology
University of Borås
Borås, Sweden
anders.gidenstam@hb.se

***Abstract*—Product returns are not only costly for e-tailers, but the unnecessary transports also impact the environment. Consequently, online retailers have started to formulate policies to reduce the number of returns. Determining when and how to act is, however, a delicate matter, since a too harsh approach may lead to not only the order being cancelled, but also the customer leaving the business. Being able to accurately predict which orders that will lead to a return would be a strong tool, guiding which actions to be taken. This paper addresses the problem of data-driven product return prediction, by conducting a case study using a large real-world data set. The main results are that well-calibrated probabilistic predictors are essential for providing predictions with high precision and reasonable recall. This implies that utilizing calibrated models to predict some instances, while rejecting to predict others can be recommended. In practice, this would make it possible for a decision-maker to only act upon a subset of all predicted returns, where the risk of a return is very high.**

***Index Terms*—Product Return, Decision Support, Probabilistic Predictions, Calibration, Predict with Rejection.**

I. Introduction

Many retailers are adopting an easy and hassle-free product returns policy and service to positively impact purchases [1] and to improve customer engagement and satisfaction [2]. However, the cost of managing returned online orders has increased significantly and has become a key challenge for online retailers [3]. For example, in 2022, the total worth of returns in the United States was $816 billion [4], nearly matching the U.S. public school budget [5] and almost double the 2020 return expenses [6]. This caused substantial financial losses that affected the economic sustainability of the industry and impacted the environment through harmful emissions [7]–[9]. As a consequence, companies have started applying strategies to reduce product returns and introduced more restrictive policies. For example, de-marketing [10] is a strategy where companies discourage customers temporarily or permanently from demanding to minimize returns. Besides, retailers may reject the transaction as a restrictive policy in some extreme cases [11].

This research is a part of the industrial graduate research school in digital retailing (INSiDR) at the University of Borås, funded by The Swedish Knowledge Foundation, grants nr. 20160035, 20170215.

Generally speaking, in order to successfully apply business strategies and to developing applicable policies, companies need reliable information for planning and decision-making [1]. Therefore, companies have started to employ intelligent systems that can analyze data from e-commerce to produce such information to solve various industrial problems, enriching productivity and enhancing the overall business outcome [11]. Providing data-driven decision support regarding the problem of product returns became viable when companies started to analyze sales data and their customers' shopping behaviors to evaluate the likelihood of returns using machine learning and statistics [11]–[13].

However, there are not enough studies and resources on how to predict product returns to define general patterns for decision support of this complex matter [13]. There are suggestions to focus on extreme cases with a high likelihood of returns [11], which is effective since predictions are made to obtain only a subset of the increased risk of return. However, this implies that prediction accuracy that evaluates the overall predictions for all classes should be replaced by precision to assess such a specific group of high risk of a return. Thus, aiming for increased precision with a reasonable recall within those top predictions is beneficial so retailers can confidently take action. This confidence can be achieved by employing Probabilistic Predictions, where the model produces probabilities over class labels [14]. However, not all machine learning algorithms can provide probabilities that reflect reality, so the probabilistic predictors must be well-calibrated to provide valuable information [15]. Some algorithms, such as Logistic Regression, are supposed to be well-calibrated off the shelf, while others need external calibration, like Platt Scaling [16].

This paper presents a case study applying a data-driven approach to predict product returns to support decision-making. In particular, probabilistic predictors and how they can support decisions on whether to take actions or not on predicted returns are investigated.

II. Background

A. Product Returns

A considerable problem in digital retailing is the amount of product returns, that is, when customers, after purchase, send items back to the seller, which has costly effects on the supply chain and negatively impacts the environment [9]. The problem is spread worldwide due to the easy return policies [1] accompanied by the product price, demand, and quality [13]. In order to gain insights into the problem and to support decision-makers in formulating new policies and strategies to reduce the number of product returns and obtain customer satisfaction, data analytics and predicting returns have become fundamental pillars [11].

The initial research concerning the problem of product returns focused mainly on optimizing product return policies [17] and forecasting return volume afterward [18], which provided some solutions to the problem from managerial and industrial perspectives. However, the number of returns was increasingly growing, and predicting returns volumes was not compelling enough since it primarily used time-series data to predict the timing and quantity of returns within a given system based on past sales and returns information for manufacturing [19]–[21]. Furthermore, the early solutions did neither apply to online transactions nor provide decision support to retailers, especially for single returns from specific orders. Therefore, predictive modeling, typically using machine learning techniques with online transaction data, has become essential for classifying product or order-level returns to support decision-making [12].

The task is specifically to utilize historical data to classify orders directly as either returns or not. However, when no historical data is available, i.e., when a customer makes the first order, predicting likely returns can be challenging, unlike predicting returns from customers who have made more than one purchase. Those are different problems because, with access to historical data, it is possible to analyze more attributes to identify potential factors that could influence the customer's behaviors. The problem of predicting returns is, therefore, complex, and it is still hard to obtain high accuracy. Some studies suggest targeting extreme cases, i.e., groups of orders with a high likelihood of returns, rather than predicting all returns from the entire data [11]. In other words, to analyze all predictions and to identify the orders with products that have a high risk of being returned implies that prediction accuracy has no more effective since it measures prediction performance on the entire data set, including both returns and no-returns. Precision plays, in such cases, a crucial role in evaluating the extent to which predictions are accurate within the specific group of orders.

B. Probabilistic Prediction and Calibration

In probabilistic prediction, many classifiers, in addition to predicting the class label of the target variable, also produce probability distributions of the different classes available to express the predictive models' level of uncertainty [14], [22]. However, because of the nature of the data, the problem tackled, and the modeling algorithm, probability estimates do not always match the ground truth [15]. For example, predictions are underconfident in models built with ensemble-based algorithms, e.g., random forest (RF), while models built with the decision tree (DT) are overconfident [23]. Models must therefore be well-calibrated. Thus, a classifier should produce probabilities that accurately reflect the true probability of the target class, i.e., the event must occur in reality equally as the chance of occurring. Mathematically, $\mathrm{p}(p(c_j|p^{c_j}) = p^{c_j}$, where p^{c_j} is the probability estimate for class j [24].

Calibration is a mapping from model scores to probabilities. A primary model (predictor) trains on training data and produces probability estimates of two disjoint sets, i.e., the test data and calibration data. A secondary model (calibrator) is then trained on a 2-D data set, i.e., the calibration set's probabilities produced by the primary model associated with its actual class labels. The secondary model then takes the test set probabilities produced by the primary model as inputs and predicts (calibrates) those probabilities [24].

With Platt scaling [25], the calibration method assumes a logistic relationship between the labels and the probabilities. Thus, a logistic regression model fits a model with a sigmoid function, the calibration probabilities, and the actual labels to adjust the predicted probabilities and align with the actual frequencies of the positive class [22].

C. Related Work

Regarding product returns prediction, few previous studies suggest methods or algorithms focused on predicting returns from online transactions to support decision-making [13]. However, there are some.

Li et al. [26] and Zhu et al. [27] proposed hyper-graph representations that can effectively model the rich information in the historical purchase and return records. They presented two different local graph-cutting algorithms based on random-walk to identify similar historical baskets shopped by other customers to produce predictions. However, these studies did not incorporate explicit product attributes or customer information, which could lead to policies targeting the wrong customers or making inappropriate decisions.

Another study by Joshi et al. [28] produced a product return predictive model by combining network theory and machine learning concepts. Given data about a customer, a brand, and the size of the product, the model can make a return prediction based on past purchase and return data. The proposed model tolerates a few false alarms (falsely predicted returns) since the high recall. However, the algorithm would suffer from the cold-start problem when no historical data is available for customers or products [28].

In our previous study [29], a proof-of-concept was performed using a random forest model to predict product returns and identify critical attributes correlated with returns.

The proposed model had acceptable accuracy but, most notably, high precision and good recall for the top 1% to 5% return predictions. It means that the model is sufficiently helpful by retaining a shallow false-positive rate coupled with a reasonable detection rate that could potentially be used to confidently target a group of customers likely to make returns and support for forming better policies or making better decisions.

All solutions above have a machine-learning perspective, applying various techniques to predict product returns with the purpose of minimizing returns through new strategies, policies, or making better decisions. However, the studies differ regarding the running approach, conditions, data sets, and setups, making them hard to replicate and compare.

III. Method

A. The Data

The data set is from a well-known Swedish online retailer specializing in fashion. It is relatively information-rich, such as which products customers included in their orders and exactly which products were returned. The data set also has detailed information about all previous purchases and returns made by the customer.

In the current study, the data was first pre-processed and engineered with new features concerning customer purchasing history, then segmented into two types of customers to be able to analyze separate product return problems. The first segment, New Customers, includes those with no previous orders, thus, there is no purchase or return history. The Multi-purchase Customers is the second segment, consisting of those with at least one previous order, i.e., historical attributes are present. In Table I, the data is summarized, including a brief description, data types, and a $(*)$ mark for features in the New Customers segment.

The data include features relevant to orders, customers, and products. Firstly, the *order attributes* are, for instance, the Number of products in the order, Number of items in the order, which concern the products and the different items of each product, respectively. The Order value in Swedish krona. The Order day: WD, FR, SA, and SU stand for the day the order was placed, that is, weekday, i.e., Monday to Thursday, Friday, Saturday, and Sunday. The Order time: D, E, and N denote the time category, i.e., daytime 07-18, evening 19-23, and night 00-06. The Delivery way: Dw1, Dw2, ... Dw5 are five different categories of the way of delivery. Secondly, regarding *product attributes*, for each order, the previous sales and returns of every item in the order were used to calculate the return rate for the items. Therefore, three attributes were generated from this: the Mmaximum return rate is the highest individual return rate, the Mean return rate is the average return rate of the items, and the Modified mean return rate was calculated in the same way as the mean return rate but using Laplace estimates instead of relative frequencies for the return rates of the items. Thirdly, the *customer attributes* are, for example, the gender, age intervals, country, total previous returns value, and average order items.

The data set has two data types: numerical, such as the number of previous returns, the value of previous returns, and the number of items in the order. It also includes binary indicators, such as if an order contains the same products in different sizes or colors. The second data type comprises nominal values, e.g., the customer's gender, the order's day, and the order's way of delivery. Nominal values are encoded in the pre-processing stage, and some numerical attributes, such as age, are aggregated into categorical binary columns.

The two segments have orders of 2042 unique product items placed by 22483 customers between November 2013 and February 2020. The New Customer's data size is 20262 instances and 38 features, and a target variable indicates the Return class label, which occupies 30% of the data. The size of Multi-purchase Customers' data is 15738 instances and 56 features, and the target variable where the Return class occupies 37% of the data.

B. Modeling and Calibration Techniques

This study applied two methods widely used in data mining and applications to predict the likely returns: the Logistic Regression (LR) and the Gradient Boosting (GB) models and the calibration technique, Platt Scaling.

Logistic Regression is a generalized linear model that fits a logistic function to its outputs to model the probability of the dependent variable (usually binary) based on the values of the independent variables [30]. It estimates the coefficients of the predictor variables that represent the strength and direction of the relationship between each variable and the outcome. LR models produce well-calibrated probability estimates off-the-shelf [31].

While Gradient boosting is an algorithm that combines a series of simple models, typically decision trees, to create a stronger model and make more accurate predictions [32]. It iteratively fits a new model to the previous model's residuals. The model then predicts the residual of the current model and adds it to the previous model's prediction to improve the overall prediction. It repeats the process until convergence, i.e., attempting to reduce the remaining error using gradient descent optimization to minimize the loss function, which measures the difference between predictions of the target variable and its actual values. It used 300 estimators for the GB model and the default parameter setting for the LR.

C. Testing Protocol

Each of the New Customers' and Multi-purchase Customers' data segments were partitioned into a train, a calibration, and a test set while maintaining the sequence of the data, i.e., first orders are the oldest, and the last orders are the newest, as follows: two principal sets, *Training*, 75%, that is further partitioned into 80% *True Training* for model building and 20% *Calibration* for model calibration, and the remaining part, 25%, is a *Hold-out* for final testing.

TABLE I: The New Customers set includes only starred features. The Multi-purchase Customers set includes all features.

ID	Description	∗	Type	ID	Description	∗	Type
	Order Attributes				***Customer Attributes***		
34	Number of items in the order	∗	Int.	1 -3	Gender: M, F, Unknown	∗	{0,1}
35	Number of products in the order	∗	Int.	4 -10	Age: 0-17, 18-25, 26-33, 34-40, 41-48, 49-56, 57+	∗	{0,1}
36	Price of most expensive items	∗	Float	11-15	Country: SE, FI, NO, DK, UK	∗	{0,1}
37	Different-size same items	∗	{0,1}	16	Number of previous orders		Int.
38	Different-color same items	∗	{0,1}	17	Total value of previous orders		Float
39	Order value in SEK	∗	Float	18	Number of previous returns		Int.
40	Order discount in SEK	∗	Float	19	Total value of previous returns		Float
41	Order Freight cost	∗	Float	20	Biggest value of previous returns		Float
42-45	Order day: WD, FR, SA, SU	∗	{0,1}	21	Customer's return rate		Float
46-48	Order time: D, E, N	∗	{0,1}	22	Days since the first order		Int.
49-53	Delivery way: Dw1, Dw2, ..., Dw5	∗	{0,1}	23	Days since the latest order		Int.
				24	Days since the first return		Int.
				25	Days since the latest return		Int.
	Product-level Attributes			26	Average value of previous orders		Float
54	Items' maximum return rate	∗	Float	27	Average number of products of previous orders		Float
55	Items' average return rate	∗	Float	28	Average number of items of previous orders		Float
56	Items' modified mean return rate	∗	Float	29	Average expensive product price of previous returns		Float
				30	Average previous discounts		Float
				31	Average previous returns value		Float
	Target Attribute			32	Average quantity ratio of previous returns		Float
57	At least one item in the order returned	∗	{0,1}	33	Frequency of return reasons		Float

To evaluate the predictive performance, the confusion matrix, together with the accuracy, precision, recall, and the area under the ROC curve, are used. Furthermore, the expected calibration error was used to measure how well the predicted probabilities of the test data are calibrated with the true frequencies of class labels [33].

D. Experimental Setups

Since this study aims to support decision-making regarding the product returns problem by providing accurate and confident predictions, building a well-calibrated predictive model on the given data sets is essential. Therefore, probabilistic predictors with the LR and GB techniques were used to obtain probabilities and evaluate their alignment with the actual frequencies of returns, leading to developing a well-calibrated model that provides confident predictions where the retailer can target upon, i.e., making better and informed decisions.

Two experiments were conducted, one "standard classification" and one "classification with a reject option". In the standard classification experiment, since LR provides well-calibrated probabilities off the shelf, a model was built using the whole data but the test set to predict the test orders in each customer segment. At the same time, only the GB model uses a calibration set. Thus, a GB model was built with the true training set of each segment and used the calibration set with Platt Scaling to calibrate the probability estimates of the test set orders. Predictive performance and probabilities estimates were then evaluated with the various evaluation metrics.

In the second experiment, i.e., classification with a reject option, probabilities resulting in the first experiment were analyzed by investigating if the average top probabilities of the return class orders met the precision scores on various confidence levels. Thus, to provide actionable product return predictions, true and false positives were evaluated, where the predictive model should be at its most confident state. In other words, when the average probability estimates of specific returns is high, and at the same time the model is well-calibrated, particular return predictions are analyzed where the average probability estimates meet certain confidence levels, i.e., 99% 95%, 90%, until 70%. Thus, the confidence levels must reflect the empirical accuracy/precision. Such analysis must be helpful to predict returns of extreme cases, namely when purchases are made with very high probability and confident return intentions where retailing companies can take immediate action. This implies that the precision of predictions within a high probability range is considerably more important than the recall [11] and can be used as a confidence of such predictions.

IV. Results

A. Calibration

Fig. 1 and 2 show the LR and GB models' reliability diagrams and their expected calibration errors on the new customers' segment. Since LR naturally produces well-calibrated probabilities off the shelf, unsurprisingly, Fig. 1 shows that it has a curve that mostly lies on the diagonal, demonstrating well-calibrated probabilities with a very small ECE of 0.014. The GB model in Fig. 2 seems to be well-calibrated to some extent, with an ECE of 0.026. However, the ECE did not improve after calibration but was still too close to the state before calibration, with an ECE of 0.030.

Fig. 3 and 4 show the calibration results in the multi-purchase customers' segment. The LR seems weekly calibrated for the higher levels of probabilities with an ECE

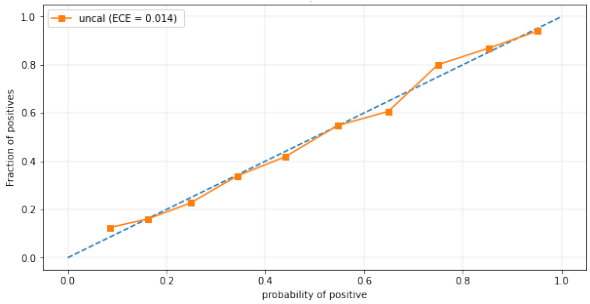

Fig. 1: A Logistic Regression model's reliability diagram trained on the New Customer's segment.

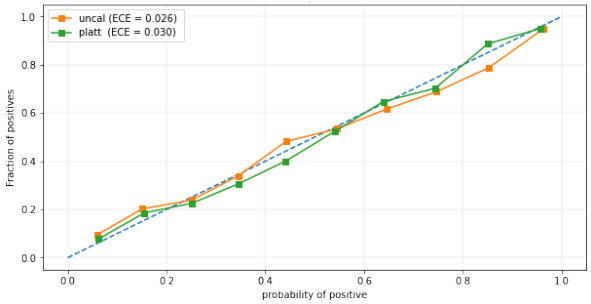

Fig. 2: A Gradient Boosting model's reliability diagram trained and calibrated on the New Customer's segment.

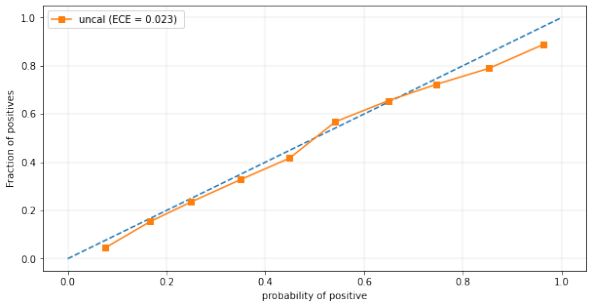

Fig. 3: A Logistic Regression model's reliability diagram trained on the Multi-purchase Customer's segment.

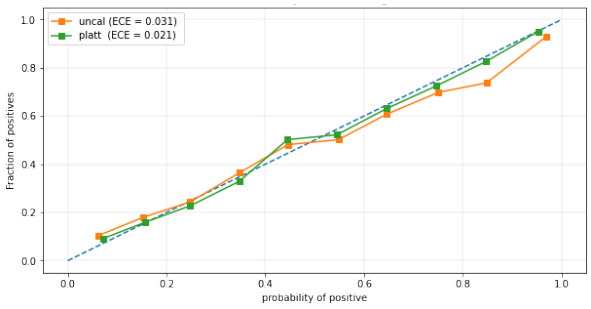

Fig. 4: A Gradient Boosting model's reliability diagram trained and calibrated on the Multi-purchase Customer's segment.

of 0.023. While the reliability diagrams of the GB model before and after calibration revealed that with Platt Scaling, the calibration improved the model slightly, and the expected calibration error lowered from 3% to 2%. As seen in Fig. 4, the green reliability diagram is broadly lying on the diagonal line, meaning that average probabilities in each interval correspond to the fraction of returned orders to a great extent.

B. Prediction Performance

A detailed evaluation of the predictive performance of the models is demonstrated in Table II and Table III, i.e., the area under the ROC curve (AUC), confusion matrix's TP, FP, TN, FN, and the related accuracy (ACC), precision (PRE), and recall (REC) in the different customer segments.

Table II displays that all models demonstrate good and similar performance in the new customers' segment. Still, the confusion matrix shows that the calibrated GB has the best scores with the highest number of true positives, a fairly good accuracy of 74%, a reasonable recall of 30%, and the lowest number of false positives with a relatively good sensitivity of 71%, which is promising in predicting positive alarms in this segment of customers who just made their first purchase.

Table III, on the other hand, reveals that the performance of the uncalibrated GB is the best among other models in this segment of multi-purchase customers, with a 37% return rate compared with 30% of the new customers segment.

C. Classification with a Reject Option

Table IV and Table V present the calibrated GB model's different subsets of predictions on new and multi-purchase customer segments testing data, respectively. As can be seen, Table IV demonstrates the well-calibrated state of the GB model on the new customers' data, where the mean probabilities match the empirical precision scores to a great extent, except for the 99% confidence level. This can be explained by the minimal number of true positive predictions on that level with just one false. In this segment, the model can confidently predict, e.g., 16% of actual returns with 90% confidence, which in the end, is a business matter in which a retailer decides the level to which he/she will act on.

Comparing this result with the multi-purchase customer's segment predictions analysis, Table V displays a high match between the mean probabilities and precision scores in all levels. This provides a reliable foundation for retailers to act upon highly confident predictions while disregarding others. For instance, out of the actual returns, 21% are predicted with 88% confidence. When aiming for a higher recall rate, it often means accepting a lower confidence level in the predictions. However, having access to a well-calibrated probabilistic predictor enables decision-makers to directly evaluate the trade-off between precision and recall. This assessment empowers the implementation of strategies where the decision-maker focuses exclusively on the returns that are predicted with the highest levels of confidence.

TABLE II: Model performance on the New Customers set.

Mdl	ACC	AUC	PRE	REC	TP	TN	FP	FN
LR	0.737	0.707	0.712	0.285	458	3276	1147	185
GB	0.738	**0.719**	**0.719**	0.284	456	**3283**	1149	**178**
GB$_{cal}$	**0.739**	0.710	0.711	**0.298**	**479**	3266	**1126**	195

TABLE III: Model performance on the Multi-purchase Customers set.

Mdl	ACC	AUC	PRE	REC	TP	TN	FP	FN
LR	0.705	0.725	0.673	0.434	649	2124	847	315
GB	**0.725**	**0.774**	**0.691**	**0.499**	**746**	2106	**750**	333
GB$_{cal}$	0.718	0.768	**0.691**	0.468	700	**2126**	796	**313**

TABLE IV: Calibrated GB model predictions on the New Customers set with a rejection option.

Mean Prob.	TP	FP	PRE	REC
0.99	16	1	0.94	0.01
0.95	187	10	0.95	0.12
0.90	259	28	0.90	0.16
0.85	314	57	0.85	0.20
0.80	372	96	0.79	0.23
0.75	434	151	0.74	0.27
0.70	479	195	0.71	0.30

TABLE V: Calibrated GB model predictions on the Multi-purchase Customers set with a rejection option.

Mean Prob.	TP	FP	PRE	REC
0.99	9	0	1.00	0.01
0.95	201	11	0.95	0.13
0.90	315	41	0.88	0.21
0.85	408	84	0.83	0.27
0.80	506	138	0.79	0.34
0.75	613	211	0.74	0.41
0.70	700	313	0.69	0.47

Hence, it would be very hard to use a model to predict returns since the precision will be too low to act on all predictions. By using well-calibrated probabilistic predictors, the trade-off between precision and recall to select a sweet spot can be assessed. Using such a model with a reject option is therefore recommended.

D. Feature Analysis

As shown in Fig. 5, together with the names of the features and their importance shown in Tables VI and VII, a few features seem important even though it can easily infer different indicators based on the business case.

However, An important finding is that order with the same items in different sizes is the main critical indicator for likely returns in both segments, which a retailer can confidently act upon. Other indicators differ between the data sets. For example, the historical feature, namely, the Customer return rate, is a leading indicator in the multi-purchase customers set. Additionally, the product-level features, e.g., the mean and max product return rate, have useful indicators in identifying the likely returns from the product-level return history.

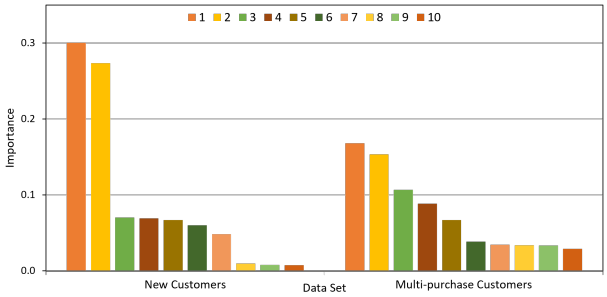

Fig. 5: The importance score of the top ten features numbered in tables VI and VII.

TABLE VI: New Customer segment's most importat features.

∗	Feature	Imp.	∗	Feature	Imp.
1	Order value	30 %	6	Mod. mean ret. rate	6 %
2	Different-size items	27 %	7	Items mean ret. rate	5 %
3	Most expensive item	7 %	8	Num. of products	1 %
4	Num. of items	7 %	9	Male customer	1 %
5	Items max ret. rate	7 %	10	Order discount	1 %

TABLE VII: Multi-purchase Customer segment's most important features.

∗	Feature	Imp.	∗	Feature	Imp.
1	Customer ret. rate	17 %	6	Total order value	4 %
2	Order value	15 %	7	Items max ret. rate	3 %
3	Different-size items	11 %	8	Most expensive item	3 %
4	Avg. ret. quantity rate	9 %	9	Days since latest ret.	3 %
5	Num. of items	7 %	10	Total returns value	3 %

V. Concluding Remarks

Most online retailers offer free product return services to increase sales and consumer satisfaction, even though the costs of product returns are growing. Due to this, excessive returns can negatively impact the economy, including increased costs for retailers and consumers, waste, and disruption to the supply chain. Along with these impacts, the problem expands to affect the purity of the environment.

This paper presents a case study that addressed the problem of product returns and investigated product return prediction through analysis of actual returns data sets from a machine-learning perspective with the purpose of supporting decision-making. The study confirms that an item with different sizes and the presence of products with a high return rate in an order are vital indicators of returns. Furthermore, the customer return rate is another strong indicator for those with a multi-purchase history. From a technical perspective, the study found that probabilistic predictors are essential tools for data-driven decision-making. Thus, probabilities can be used as confidence on top of predictions but must be well-calibrated. The result showed that the use of calibrated models to predict a small group of instances with high and confident probabilities and reject predicting others is the most viable option.

The decision situation could be further formalized for future work using utilities, i.e., costs and benefits of wrong and correct predictions and outcomes, to allow for automated decision-making.

REFERENCES

[1] Wood, Stacy L. "Remote purchase environments: The influence of return policy leniency on two-stage decision processes." Journal of Marketing Research 38, no. 2 (2001): 157-169.

[2] Cassill, Nancy L. "Do customer returns enhance product and shopping experience satisfaction?." The International Review of Retail, Distribution and Consumer Research 8, no. 1 (1998): 1-13.

[3] Petersen, J. Andrew, and V. Kumar. "Are product returns a necessary evil? Antecedents and consequences." Journal of Marketing 73, no. 3 (2009): 35-51.

[4] National Retail Federation (NRF). "2022 Consumer Returns in the Retail Industry". Appriss Retail, 2022. Retrieved 30/06/2023, from https://nrf.com/research/2022-consumer-returns-retail-industry.

[5] National Center for Education Statistics. (2023). Public School Expenditures. Condition of Education. U.S. Department of Education, Institute of Education Sciences. Retrieved 30/06/2023, from https://nces.ed.gov/programs/coe/indicator/cmb.

[6] National Retail Federation (NRF). "Consumer Returns in the Retail Industry 2020". Appriss Retail, 2020. Retrieved 30/06/2023, from https://nrf.com/research/customer-returns-retail-industry.

[7] Optoro. "2022 Impact Report". Optoro Inc, 2022. Retrieved 30/06/2023 from, https://www.optoro.com/returns-content/environmental-impact.

[8] Economist.com. 'The Business of Reselling Returned Shop Items'. The Economist, 2016. Retrieved 30/06/2023, from https://www.economist.com/business/2016/11/24/the-business-of-reselling-returned-shop-items.

[9] Edwards, Julia B., Alan C. McKinnon, and Sharon L. Cullinane. "Comparative analysis of the carbon footprints of conventional and online retailing: A "last mile" perspective." International Journal of Physical Distribution & Logistics Management 40, no. 1/2 (2010): 103-123.

[10] Levy, Sidney J. Brands, consumers, symbols and research: Sidney J Levy on marketing. Sage, 1999. pp.75.

[11] Urbanke, Patrick, Johann Kranz, and Lutz Kolbe. "Predicting product returns in e-commerce: the contribution of mahalanobis feature extraction." (2015).

[12] Kedia, Sajan, Manchit Madan, and Sumit Borar. "Early bird catches the worm: Predicting returns even before purchase in fashion E-commerce." arXiv preprint arXiv:1906.12128 (2019).

[13] Ambilkar, Priya, Vishwas Dohale, Angappa Gunasekaran, and Vijay Bilolikar. "Product returns management: a comprehensive review and future research agenda." International Journal of Production Research 60, no. 12 (2022): 3920-3944.

[14] Niculescu-Mizil, Alexandru, and Rich Caruana. "Predicting good probabilities with supervised learning." In Proceedings of the 22nd international conference on Machine learning, pp. 625-632. 2005.

[15] Dormann, Carsten F. "Calibration of probability predictions from machine-learning and statistical models." Global ecology and biogeography 29, no. 4 (2020): 760-765.

[16] Sweidan, Dirar, and Ulf Johansson. "Probabilistic Prediction in scikit-learn." In The 18th International Conference on Modeling Decisions for Artificial Intelligence, On-line (from Umeå, Sweden), September 27-30, 2021. 2021.

[17] Padmanabhan, Venkata, and Ivan PL Png. "Manufacturer's return policies and retail competition." Marketing Science 16, no. 1 (1997): 81-94.

[18] Toktay, L. B. "Forecasting product returns, in business aspects of closed-loop supply chains." Carnegie Mellon University Press. Pittsburgh, Pennsylvania (2003).

[19] Canda, Alexandra, Xue-Ming Yuan, and Feng-Yu Wang. "Modeling and forecasting product returns: An industry case study." In 2015 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), pp. 871-875. IEEE, 2015.

[20] Ma, Jungmok, and Harrison M. Kim. "Predictive model selection for forecasting product returns." Journal of Mechanical Design 138, no. 5 (2016): 054501.

[21] Cui, Hailong, Sampath Rajagopalan, and Amy R. Ward. "Predicting product return volume using machine learning methods." European Journal of Operational Research 281, no. 3 (2020): 612-627.

[22] Feng, Runhai. "Improving uncertainty analysis in well log classification by machine learning with a scaling algorithm." Journal of Petroleum Science and Engineering 196 (2021): 107995.

[23] Johansson, Ulf, Tuwe Löfström, and Henrik Boström. "Calibrating probability estimation trees using venn-abers predictors." In Proceedings of the 2019 SIAM International Conference on Data Mining, pp. 28-36. Society for Industrial and Applied Mathematics, 2019.

[24] Zadrozny, Bianca, and Charles Elkan. "Transforming classifier scores into accurate multiclass probability estimates." In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 694-699. 2002.

[25] Platt, John. "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods." Advances in large margin classifiers 10, no. 3 (1999): 61-74.

[26] Li, Jianbo, Jingrui He, and Yada Zhu. "E-tail product return prediction via hypergraph-based local graph cut." In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 519-527. 2018.

[27] Zhu, Yada, Jianbo Li, Jingrui He, Brian Leo Quanz, and Ajay A. Deshpande. "A Local Algorithm for Product Return Prediction in E-Commerce." In IJCAI, pp. 3718-3724. 2018.

[28] Joshi, Tanuj, Animesh Mukherjee, and Girish Ippadi. "One size does not fit all: Predicting product returns in e-commerce platforms." In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 926-927. IEEE, 2018.

[29] Sweidan, Dirar, Ulf Johansson, and Anders Gidenstam. "Predicting returns in men's fashion." In Developments of Artificial Intelligence Technologies in Computation and Robotics: Proceedings of the 14th International FLINS Conference (FLINS 2020), pp. 1506-1513. 2020.

[30] King, Gary, and Langche Zeng. "Logistic regression in rare events data." Political analysis 9, no. 2 (2001): 137-163.

[31] Sweidan, Dirar, and Ulf Johansson. "Probabilistic Prediction in scikit-learn." In The 18th International Conference on Modeling Decisions for Artificial Intelligence, On-line (from Umeå, Sweden), September 27-30, 2021. 2021.

[32] Friedman, Jerome H. "Stochastic gradient boosting." Computational statistics & data analysis 38, no. 4 (2002): 367-378.

[33] Nixon, Jeremy, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. "Measuring Calibration in Deep Learning." In CVPR workshops, vol. 2, no. 7. 2019.