Learning Reasons for Product Returns on E-Commerce

Miriam Farber, Slava Novgorodov, Ido Guy
Meta
mfarber@meta.com, slavanov@meta.com, idoguy@acm.org

Abstract

In the rapidly evolving landscape of e-commerce, product returns have become a significant economic burden for businesses, where the reasons for returns may vary from wrong sizing and defective products to simply no longer needing the purchased product. This paper presents, to the best of our knowledge, the first comprehensive study of the complexities of product returns across a variety of e-commerce domains, focusing on the task of predicting the return reason. We propose a supervised approach for predicting return likelihood and the underlying return reason. We test our approach over a real-world dataset from a large e-commerce platform.

Keywords: e-commerce, product return, return reason prediction

1. Introduction

Due to the rapid growth of the e-commerce industry in the past years, online selling has become very trending. E-commerce platforms deal with many technological problems such as recommendations and personalization, search, product categorization, content generation, and various logistic aspects such as inventory optimization and delivery. The e-commerce supply chain is becoming more complex as organisations are both expanding their businesses geographically and increasing their supplier base to continue their growth. Consumers are frequently ordering and returning items (the return rate may vary from 5% to up to 60% (Zhu et al., 2018; Cullinane et al., 2019; Li et al., 2018), dependent on product category, returns policy and other reasons). Moreover, some buyers will not make a purchase if there is no return policy and will prefer sellers that provide comfortable and fair return policy (Hjort and Lantz, 2016).

Managing product returns in e-commerce is an important problem in the past years due to several main factors. The first is financial impact, since high return rates can significantly impact a retailer's bottom line. Returns lead to additional costs in terms of restocking, and potential loss of saleable inventory, which can erode profit margins. Second factor is customer satisfaction. A smooth return process is crucial for maintaining customer satisfaction and trust. Negative experiences with returns can lead to loss of customer loyalty and negative word-of-mouth, impacting future sales. This factor includes resource allocation and inventory management. Handling returns requires time, labor, and infrastructure, diverting resources from other essential business operations. An efficient return management system is needed to minimize these resource demands. Moreover, high return rates can disrupt inventory management and forecasting, making it more challenging for retailers to maintain optimal stock levels and meet customer demand. Last but not the least factor is the environmental impact. Frequent returns contribute to higher carbon emissions due to increased transportation needs for reverse logistics. Additionally, returned items may end up in landfills if they cannot be resold, contributing to waste and pollution. Hence, by investing effort in dealing with the problem of product returns, e-commerce businesses can improve their financial performance, enhance customer satisfaction, optimize resource allocation, maintain better inventory management and even reduce their environmental impact.

The problem of products returns can be seen as part of a wider field of reverse logistics. In general, reverse logistics is the process of managing the flow of goods from the point of consumption back to the point of origin for various purposes such as returns, repairs, recycling, or disposal. In the context of e-commerce, reverse logistics primarily deals with the management of product returns. Managing reverse logistics effectively in e-commerce requires a combination of efficient processes, technology, and partnerships. By addressing these challenges, retailers can minimize the financial and environmental impact of returns, improve customer satisfaction, and maintain optimal inventory levels.

In this paper we present a deep-dive study of the complexities of product returns across a variety of e-commerce domains, focusing on the task of predicting the return reasons. To the best of our knowledge we are the first to extensively study this problem in general e-commerce setting, in opposite to previous works that focus on specific domains or specific reasons. We propose an ensemble-based machine learning approach for predicting return likelihood and the underlying return reason. We showcase the performance of our proposed approach over real-world dataset of product transaction from a large e-commerce platform.

The Seventh Workshop on e-Commerce and NLP (ECNLP 7), pages 1–7

2. Related Work

Many works studied product returns in e-commerce, however, in general, the problem of product return prediction in e-commerce has not attracted much attention from the data mining community, despite the large amount of data available from historical purchase and return records (Li et al., 2018). A line of papers that is most related to our work, focus on predictive analytics using machine learning methods (e.g., (Fuchs and Lutz, 2021; Ma and Kim, 2016; Urbanke et al., 2015)). These works apply advanced data mining and machine learning techniques to predict the likelihood of product returns and in some cases try to predict the return reason. These predictive models can help businesses identify high-risk customers or products, allowing for proactive interventions to reduce return rates. For example, Urbanke et al. (Urbanke et al., 2015) use feature extraction to generate a large set of features that are originated from various categorical variables such as return history, preferred payment method and device information from which the returned product was originally ordered. Some information is available only after the customer finishes the transaction, hence this methods limits the ability to take proactive actions. Other works focus on improvements in product information, images, and descriptions, which can reduce return rates by ensuring that customers have a clear understanding of what they are purchasing. The works mentioned above focus on prediction of the return event (binary classification). In our paper we focus on the more fine-grained task of predicting a return reason out of large list of possible reasons.

Moreover, while in our paper we work on various e-commerce domains, some of the papers (e.g. (Seewald et al., 2019; Nestler et al., 2021; Kedia et al., 2019)) focus on fashion, where the return rate may reach up to 60% (Zhu et al., 2018; Cullinane et al., 2019; Li et al., 2018). One of the most popular reason for returns in fashion is wrong size(Nestler et al., 2021). To deal with the size-related returns, some works propose methods that unify sizes across different platforms (e.g., (Du et al., 2019)) and help users to choose the correct size on any platform. Other works (e.g., (Abdulla and Borar, 2017)) proposed personalized size recommendations, or other innovative tools to prevent the return event (Castelblanco DÃaz, 2021). Many works try to proactively predict the return event, e.g. Kedia et al. (Kedia et al., 2019) that proposes a method to predict the chance that the customer will return the product even before the order is completed. It uses deep neural network model that uses latent size and fit features of the product and the customer. As mentioned above, in this work we do not focus on any specific domain, but provide a solution for various domains that predicts the return event and the return reason for a specific product, given only information about the product.

Other line of works (e.g., (Hjort and Lantz, 2016; El Kihal et al., 2021; Ambilkar et al., 2022)) study the connection between the returns and the return policies: These studies focused on understanding the impact of different return policies on consumer behavior, sales, and returns. They observe that factors such as return time windows, restocking fees, and return shipping costs, can affect the customer satisfaction and minimizing return rates.

Finally, returns management in e-commerce can be viewed as part of a larger problem of reverse logistics management. The reverse logistics optimization research (e.g., (EL HACHIMI et al., 2018; Sandhya and Kumara, 2020)) focuses on improving the efficiency and cost-effectiveness of reverse logistics processes, such as transportation, inspection, refurbishment, and disposition of returned products. The goal is to minimize the financial and environmental impact of returns.

3. Dataset and characteristics

We start by describing the dataset utilized in this work. Our dataset is obtained from one of the largest e-commerce platforms, covering 618240 products across 2928 categories from 26 domains. Each entry in the dataset is associated with a transaction, with an indication regarding whether it was resulted in return. The data is split 50/50, with 309120 of the entries resulted in return and 309120 did not result in return. In case of a return, the customer can choose one of predefined 13 options as the return reason, whose distribution is presented in Table 1. Customers can also include free-form text elaborating on the return reason, and $\sim$10% chose to do so. In addition, the following information regarding the products and transactions is provided: Textual features (product name, category and description), numerical features (product price and quantity) and categorical features (product size, transaction country, transaction platform, coarse platform category).

4. Returns prediction

In this paper we deal with two types of prediction tasks: Binary prediction - whether the product is going to be returned, and Multiclass prediction - Predicting a return reason for products that were returned. Here we analyze two types of return reasons lists - an extensive list consisting of all available return reasons (see Section 3), and a concise list consisted of common 5 reasons. To examine the contribution of the different features, for each task we train several models, some of them utilize

Table 1: Distribution of return reason across our dataset

Return reason	Percentage
Too small	21.0%
Too large	15.4%
Item quality not as expected	10.3%
Not needed anymore	9.9%
Inaccurate description	9.2%
Did not like the style	8.7%
Bought by mistake	7.7%
Defective item	5.7%
Damaged item	4.1%
Wrong item received	3.9%
Did not like the color	2.2%
Found better price	1.2%
Item not compatible	0.6%

only textual features while others utilize the entire range of features. For the latter type of tasks, an analysis of the features importance is provided.

4.1. Methods

As described in Section 3, our dataset consists of several types of features: textual, categorical and numerical. To train a range of models on a tabular data utilizing those different types, we use AutoGluon (Erickson et al., 2020). This is an AutoML package that trains common types of classification models (including tree-based, neural networks and transformers), and performs model selection and hyperparameters tuning. The models are then combined to produce an ensemble model that provides the final predictions (Shi et al., 2021).

We use *TabularPredictor* with multimodal support [1] to train models that utilize numerical, categorical and textual features. Models that utilize only textual features are trained using *TextPredictor* [2]. For both TabularPredictor and TextPredictor, transformers-based model ELECTRA (Clark et al., 2020) with the hyperparameters specified in footnote 1 is used to train classification task on the textual features. Textual features are concatenated with a separator between each pair. In TabularPredictor, the categorical and numerical features are fed into tree based models like XGBoost as well as well as neural networks (see Figure 1 on page 3 in (Erickson et al., 2020) for details), and the final model is formed from an ensemble of these models together with the Electra model mentioned above.

In our experiments, the data is split randomly into training (70%), validation (15%) and test (15%), with a distinct set of products belonging to each. Results are reported on the test set.

Table 2: Multi-class performance (full set of reasons).

Model features	Accuracy	Macro F1
Product Category	0.255	0.119
Product name	0.319	0.202
Product description	0.322	0.220
All textual features	0.337	0.223
All features (ensemble)	0.352	0.249

4.2. Experiments and Results

The binary model for predicting a return, based on ensemble with all features, reached 0.942 ROC AUC on the test set, with 0.876 F1 score and 0.877 accuracy. This accuracy reduces a bit to 0.866 when utilizing textual features only. This is on par or above with previously reported results for this task (Zhu et al., 2018; Urbanke et al., 2015; Li et al., 2018).

As the binary classification task is already well studied (see above and Section 2) and achieves high accuracy, we move to the more challenging task of predicting the specific return reason. For this purpose, we limit the data only to entries that resulted in a return, and build a classifier to predict the return reason. First we perform the experiments on the full list of 13 return reasons using the models described in Section 4.1. Table 2 summarizes the results over different facets of the products as features. When limiting to a single textual feature, performance is higher when using the product description, compared to using its name only and, in turn, category. It makes sense as the description contains richer data compared to the other 2 fields, and is directly tied to some of the return reasons (e.g "inaccurate description"). Using all the 3 textual features yields further performance boost, and using all available features via an ensemble reaches the highest performance at 0.352 accuracy.

To gain a deeper understanding of the roles the different features play, we display feature importance in Table 3. The importance of each feature is measured via the impact on model's accuracy when fixing the rest of the features and permuting the entries of the given feature[3]. We can see that product description (which intuitively contains the most rich information about the product) is the most important feature, followed by product category. In fact, all features except of country and quantity have significant importance (p-value smaller than 0.01).

[1] https://auto.gluon.ai/0.4.0/tutorials/text_prediction/multimodal_text.html

[2] https://auto.gluon.ai/0.0.15/tutorials/text_prediction/beginner.html

[3] https://auto.gluon.ai/stable/api/autogluon.tabular.TabularPredictor.feature_importance.html

Table 3: Feature importance (full set of reasons).

Feature	Importance	p-value
Description	0.061	0.000007
Inferred Model Category	0.035	0.0002
Size	0.021	0.00008
Page category name	0.017	0.001
Checkout product	0.014	0.00003
Name	0.011	0.002
Price	0.007	0.0004
Country	0.001	0.15
Pack Quantity	0.001	0.08

Since our data is based mostly in the US and the majority of product quantities is 1, these features become somewhat redundant, which explains their low importance. The high importance of the textual features is also reflected in the fact that the model that was trained on textual features only is not significantly inferior compared to the best performing model.

The relatively low accuracy, even of the best performing model, can be explained by the following factors: 1) The data is highly unbalanced, with some of the return reasons having very few entries in the dataset (see Table 1), making it much harder to infer those. 2) Some of the return reasons require much deeper familiarity with the customer or the product journey which is not present in the data we have ("item not needed anymore", "bought by mistake", "wrong item received", etc). 3) The predefined list of return reasons provided by the e-commerce platform includes many subjective and also not so well defined/overlapping reasons (e.g "item not compatible" vs "item quality not as expected", and also "defective item" vs "damaged item", etc). Customer's confusion is also demonstrated in the free text responses that they provide, which sometimes are not aligned with the reason they picked from the list. To demonstrate the model's difficulty to distinguish between "overlapping" reasons, we examined the confusion matrix. Consider the following two return reasons: "found better price" and "not needed anymore". These reasons overlap, since if the customer found the same item in a better price, then they don't need this item anymore. In our dataset the latter reason is 8 times more common than the former. Thus, unsurprisingly, 24% of the test samples who belong to "found better price" category were labeled as "not needed anymore" by the model (most common label for this category). Similar phenomenon occurs for the classes "item not compatible" and "defective item". The latter is 9 times more common that the former, and is labeled as such by the model in 40% of the cases that belong to "item not compatible".

Table 4: Distribution of return reason across our dataset, when limiting to 5 common return reasons

Wrong size	Quality	No need	Description	Defective
50.9%	14.4%	13.9%	12.8%	8.0%

Table 5: Multi-class performance (5 common reasons).

Model features	Accuracy	Macro F1
All textual features	0.648	0.464
All features (ensemble)	0.656	0.463

4.2.1. Predicting common return reasons

To alleviate some of the issues above, we filtered a more concise list of 5 common return reasons: *Wrong size* - union of too small and too large, item quality not as expected (*Quality*), not needed anymore (*No need*), inaccurate description (*Description*), and defective item (*Defective*). Their distribution is presented in Table 4.

Table 5 shows the performance of the prediction model when limiting the data to these reasons. It is substantially higher than in the previous task, indicating they could be distinguished more effectively, with accuracy reaching 0.656 using the ensemble with all features. Note that this is significantly better than the baseline of choosing the most common class (wrong size), which according to Table 4 would have reached an accuracy of 0.509. In Table 6 we detail the precision and recall over each of the 5 reasons.

To provide more insights into the ensemble model , we depict in Figure 1 the components of the ensemble (trained on all features), showing the score of each one on the validation set, as well as the score of the ensemble on the validation set. The weights of each component within the ensemble are as follows: XGBoost: 0.26, NeuralNetTorch: 0.05, LightGBMLarge: 0.32, TextPredictor: 0.37 (see https://auto.gluon.ai/ for details about these models). This demonstrates the high

Table 6: Precision and recall of the ensemble classifier across each of the 5 reasons.

Reason	Precision	Recall
Wrong size	0.776	0.978
Quality	0.396	0.253
No need	0.421	0.339
Description	0.463	0.374
Defective	0.531	0.346

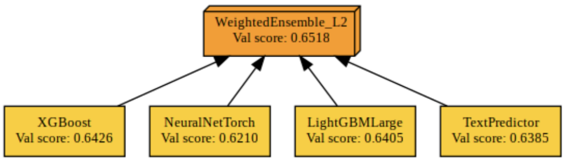

Figure 1: Composition of the top performer ensemble model, predicting 5 return reasons

importance of the textual features, with the TextPredictor model receiving the highest weight among the list.

Next we display in Table 9 the confusion matrix between the classes. Naturally as "wrong size" is the largest class (and significantly larger compared to the others), some of the data that belongs to the other classes is predicted as "wrong size". Other than that, the diagonal elements are the largest in each row, meaning that for each class, the largest bucket predicted is indeed the class itself.

Table 7: Distribution of return reason across the 5 most common domains.

Reason	Clothing	Home	J&W	H&B	Electronics
Wrong size	78.1%	0.7%	43.1%	4.7%	9.5%
Quality	8.9%	29.7%	21.6%	35.9%	12.7%
No need	6.9%	30.2%	16.3%	30.8%	26.7%
Description	4.4%	27.3%	12.1%	15.3%	17.9%
Defective	1.7%	12.2%	6.8%	13.3%	33.2%

4.2.2. Cross-domain analysis

As mentioned in Section 3, our dataset spans a variety of e-commerce domains. We set out to compare product return behavior and predictability across different domains. To this end, we considered the 5 most common domains in our dataset, which account overall for 85% of the return instances. As Table 7 shows, the distribution of return reasons varies substantially over these domains. Particularly, the distribution within the Clothing domain, where most of the previous work has focused, as mentioned in Section 2, is largely different than within other domains, reinforcing the need to study product return behavior across multiple domains. As might be expected, the majority of returns in the Clothing domain (nearly 80%), are due to wrong size. The only other domain where “wrong size” is the most common reason is Jewelry & Watches, but to a lesser extent than in Clothing. In the Home and Health & Beauty domains, wrong size is a rare return reason. In Home, “quality not expected”, “not needed anymore” and “inaccurate description” are the most common reasons. In Electronics, “defective item” is the most common.

After observing the notable differences in return reasons across e-commerce domains, we set out to examine the performance differences of our ensemble classifier across domains. Table 8 summarizes these results. The performance in the Clothing domain is noticeably different than in all other domains. Accuracy reaches 0.79, higher than any other domain, whereas macro F1 is the lowest among all domains. This is due to particularly strong performance of the classifier for the “wrong size” reason, at the expense of the performance for other reasons. In fact, on the Clothing domain, the accuracy of the ensemble model yield an uplift of only 1% compared to a majority baselines always deeming the reason as “wrong size” (see Table 7). Yet, the uplift in Macro F1 is more substantial, and, as discussed, the overall performance of the model across all categories is substantially higher than the majority baseline.

Table 8: Performance of the ensemble model across the 5 most common domains.

Reason	Accuracy	Macro F1
Clothing	0.790	0.243
Home	0.410	0.334
Jewelry & Watches	0.472	0.314
Health & Beauty	0.456	0.363
Electronics	0.505	0.432

For the other four domains, results are more similar across reasons, which yields a more balanced trade-off between the accuracy and macro F1 metrics. For the Electronics domain, macro F1 is the highest, while accuracy is second best among the 5 domains. Table 10 demonstrates the precision and recall across the 5 reasons for Electronics. It can be seen that precision and recall are fairly high for three of the reasons: “wrong size”, “defective item”, and “not needed anymore”. It is especially interesting to observe the performance for “wrong size” which account for only 9.5% of the Electronics returns (Table 7). This may indicate that the model learns to generalize this reason from other categories, where it is more frequent (e.g., Clothing). We leave further exploration of cross-domain transfer learning for future work.

Table 10: Precision and recall of the ensemble classifier over the Electronics domain across each of the 5 reasons.

Reason	Precision	Recall
Wrong size	0.554	0.864
Quality	0.293	0.119
No need	0.518	0.587
Description	0.238	0.099
Defective	0.552	0.716

Table 9: Confusion matrix for the 5 classes prediction model. Rows represent GT label and columns represent model's prediction. Rows are normalized.

	Wrong size	Quality	No need	Description	Defective
Wrong size	0.978	0.008	0.007	0.007	0.000
Quality	0.391	0.252	0.168	0.139	0.048
No need	0.309	0.158	0.338	0.128	0.064
Description	0.245	0.147	0.166	0.373	0.066
Defective	0.156	0.118	0.176	0.202	0.345

5. Conclusions

In this work, we study the problem of product returns in e-commerce. To the best of our knowledge, we are the first to systematically investigate the underlying reasons for returns and aims to predict in e-commerce in general, as opposed to focusing on specific domains. In this paper we proposed an ensemble-based machine learning approach for predicting return likelihood and the underlying return reason. The proposed method was tested over real-world dataset of product transactions from a large e-commerce platform.

6. References

G Mohammed Abdulla and Sumit Borar. 2017. Size recommendation system for fashion e-commerce. In *KDD workshop on machine learning meets fashion*.

Priya Ambilkar, Vishwas Dohale, Angappa Gunasekaran, and Vijay Bilolikar. 2022. Product returns management: a comprehensive review and future research agenda. *International Journal of Production Research*, 60(12):3920–3944.

Tatiana Alexandra Castelblanco DÃaz. 2021. Innovative tools for the prevention of product returns in e-commerce.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Sharon Cullinane, Michael Browne, Elisabeth Karlsson, and Yingli Wang. 2019. Retail clothing returns: A review of key issues. *Contemporary operations and logistics: Achieving excellence in turbulent times*, pages 301–322.

Eddie SJ Du, Chang Liu, and David H Wayne. 2019. Automated fashion size normalization. *arXiv preprint arXiv:1908.09980*.

Hajar EL HACHIMI, Mourad OUBRICH, and Omar SOUISSI. 2018. The optimization of reverse logistics activities: a literature review and future directions. In *2018 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD)*, pages 18–24. IEEE.

Siham El Kihal, Namig Nurullayev, Christian Schulze, and Bernd Skiera. 2021. A comparison of return rate calculation methods: Evidence from 16 retailers. *Journal of Retailing*, 97(4):676–696.

Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. 2020. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*.

Kevin Fuchs and Oliver Lutz. 2021. A stitch in time saves nine-a meta-model for real-time prediction of product returns in erp systems. In *ECIS*.

Klas Hjort and Björn Lantz. 2016. The impact of returns policies on profitability: A fashion e-commerce case. *Journal of Business Research*, 69(11):4980–4985.

Sajan Kedia, Manchit Madan, and Sumit Borar. 2019. Early bird catches the worm: Predicting returns even before purchase in fashion e-commerce. *arXiv preprint arXiv:1906.12128*.

Jianbo Li, Jingrui He, and Yada Zhu. 2018. 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*, pages 519–527.

Jungmok Ma and Harrison M Kim. 2016. Predictive model selection for forecasting product returns. *Journal of Mechanical Design*, 138(5):054501.

Andrea Nestler, Nour Karessli, Karl Hajjar, Rodrigo Weffer, and Reza Shirvany. 2021. Sizeflags: reducing size and fit related returns in fashion e-commerce. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3432–3440.

S Sandhya and SA Vasantha Kumara. 2020. System and method to predict the number of returns in a supply chain: A reverse logistics case study. *IUP Journal of Operations Management*, 19(3):16–24.

Alexander K Seewald, Thomas Wernbacher, Alex Pfeiffer, Natalie Denk, Mario Platzer, Martin Berger, and Thomas Winter. 2019. Towards minimizing e-commerce returns for clothing. In *ICAART (2)*, pages 801–808.

Xingjian Shi, Jonas Mueller, Nick Erickson, Mu Li, and Alex Smola. 2021. Multimodal automl on structured tables with text fields. In *8th ICML Workshop on Automated Machine Learning (AutoML)*.

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

Yada Zhu, Jianbo Li, Jingrui He, Brian Leo Quanz, and Ajay A Deshpande. 2018. A local algorithm for product return prediction in e-commerce. In *IJCAI*, pages 3718–3724.