SizeFlags: Reducing Size and Fit Related Returns in Fashion E-Commerce

Andrea Nestler*‡
andrea.nestler@zalando.de
Zalando SE
Berlin, Germany

Nour Karessli*
nour.karessli@zalando.de
Zalando SE
Berlin, Germany

Karl Hajjar*†
karl.hajjar@polytechnique.edu
Paris-Saclay University
Paris, France

Rodrigo Weffer*
rodrigo.weffer@zalando.de
Zalando SE
Berlin, Germany

Reza Shirvany*‡
reza.shirvany@zalando.de
Zalando SE
Berlin, Germany

ABSTRACT

E-commerce is growing at an unprecedented rate and the fashion industry has recently witnessed a noticeable shift in customers' order behaviour towards stronger online shopping. However, fashion articles ordered online do not always find their way to a customer's wardrobe. In fact, a large share of them end up being returned. Finding clothes that fit online is very challenging and accounts for one of the main drivers of increased return rates in fashion e-commerce. Size and fit related returns severely impact 1. the customers experience and their dissatisfaction with online shopping, 2. the environment through an increased carbon footprint, and 3. the profitability of online fashion platforms. Due to poor fit, customers often end up returning articles that they like but do not fit them, which they have to re-order in a different size. To tackle this issue we introduce SizeFlags, a probabilistic Bayesian model based on weakly annotated large-scale data from customers. Leveraging the advantages of the Bayesian framework, we extend our model to successfully integrate rich priors from human experts feedback and computer vision intelligence. Through extensive experimentation, large-scale A/B testing and continuous evaluation of the model in production, we demonstrate the strong impact of the proposed approach in robustly reducing size-related returns in online fashion over 14 countries.

CCS CONCEPTS

• **Information systems** → **Data mining**; *Recommender systems*; *Clustering and classification*; • **Mathematics of computing** → *Probabilistic algorithms*; • **Applied computing** → *Online shopping*.

KEYWORDS

Size and Fit; Fashion e-commerce; Bayesian model

ACM Reference Format:
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 and Data Mining (KDD '21), August 14–18, 2021, Virtual Event, Singapore.* ACM, Singapore, 9 pages. https://doi.org/10.1145/3447548.3467160

1 INTRODUCTION

Article returns are critical to any retail industry where customers return those articles which they find unsatisfactory for various reasons. Customers in turn receive a full or a partial refund for their returns according to the retailer's specific return policy. Over the past years, customer returns have been growing at a significant rate reaching up to 50% increase year over year for certain categories [4, 29]. Article return rates vary greatly among retail industries, categories, brands, and distribution channels [3, 19, 23]. However, the highest return rates are in online fashion apparel with 25 – 40% overall return rates that can reach up to 75% for specific categories and brands [3, 22]. The underlying reason is that fashion apparel involves complex factors such as size, fit, color, style, taste and unquantifiable factors such as "it's not me" [3]. Among these reasons, poor size and fit is cited as the number one factor in online fashion returns [24]. On one hand, physical examination of fashion articles to assess their size, fit, fabric, design and pairing with other fashion articles is crucial for customers in their order decisions [12]. On the other hand, in online fashion customers order garments and shoes without the possibility of trying them on and this crucial sensory and visual feedback is delayed to the unboxing experience. The absence of "feel and touch" experience therefore leads to major uncertainties in the buying process and to the hurdle of returning articles. As such, many customers either hesitate to place an order, or opt for various strategies for reducing the uncertainty, such as ordering multiple sizes or colors of the same article and then returning those that did not match their criteria - even more so for fashion categories and brands they are less familiar with.

Escalating the problem for size and fit, fashion articles suffer from significant sizing variations [2] due to: different sizing systems (Alpha, Numeric, Confection); coarse definition of size systems (S,

*All authors contributed equally
‡ Corresponding Authors: andrea.nestler@zalando.de, reza.shirvany@zalando.de
† Work done while at Zalando SE

KDD '21, August 14–18, 2021, Virtual Event, Singapore

ACM ISBN 978-1-4503-8332-5/21/08…$15.00
https://doi.org/10.1145/3447548.3467160

M, L for garments); country conventions (EU, FR, IT, UK); different specifications for the same size according to the brand; *vanity sizing* [33] where brands deliberately adapt their nominal sizes to target segment of customers based on age, sportiness, etc.; and different ways of converting a local size system to another. Moreover, customers' perception of size and fit for their body remains highly personal and subjective which influences what the right size is for each customer. The combination of the aforementioned factors leaves the customers alone to face a highly challenging problem of determining the right size and fit during their order journey. The problem of size and fit has a major impact on the environment, and the profitability of online fashion retailers. A major issue in sustainability is the substantial carbon footprint incurred in the logistic process of returning articles [31]. Online fashion retailers are strongly inclined to offer lenient return policies to lower customer perceptions of uncertainty [35] which inherently leads to an increase in returns. As customer expectations around fit and sustainability evolve, online fashion retailers have to quickly adapt and find innovative techniques to conclusively reduce the high size and fit related return rates. In recent years, an emerging body of multi-disciplinary research has been proposed with the aim of enabling customers to find something that fits them from the first time as discussed in the related work section [1, 5–11, 13, 14, 16–18, 20, 21, 26–28, 30, 32, 36].

In this work we introduce a powerful approach, called SizeFlags, to provide customer agnostic size advice with strong impact on decreasing size and fit related returns in fashion e-commerce. This approach naturally benefits from the advantages of Bayesian methods - modeling uncertainty and the use of priors. Considering the lack of size and fit expert labeled data, we rely on large-scale weakly annotated data from the returns process. More specifically, first we leverage the crowd's subjective and noisy return reason feedback (which is highly influenced by individuals perception of size and fit) as an input signal to determine the fit of an article. We thus construct a binomial model based on the return behaviour of articles and then extend it to a fully Bayesian setting which integrates rich priors from both human expert feedback and computer vision intelligence. We present multiple versions of our model including the baseline algorithm launched for textile and shoe categories in 2017, and the fully Bayesian SizeFlags algorithm launched in early 2020, currently in use for millions of articles ordered over 14 countries.
The contributions of this work are: (1) We introduce the Bayesian SizeFlags framework to determine the fit behaviour of articles based on subjective noisy and anonymized return data from customers, (2) We integrate, for the first time to our best knowledge, human fashion expert feedback as well as computer vision based cues as rich priors for tackling the size advice problem and demonstrate these priors strongly contribute in reducing size and fit related returns - by addressing the challenging cold-start problem for the thousands of new articles appearing on shopping platforms every day and (3) we demonstrate with extensive experimental results obtained through A/B testing and continuous in-production evaluation, how each part of our model contributes to reducing size and fit related returns.

We note that, evaluating the impact of size and fit recommendations remains highly challenging in the literature due to multiple underlying factors specific to this problem space; on one hand the "true" size of a customer is often unknown, remains subjective for each customer, and can vary greatly by external factors including life changing events impacting a customer's physical body, and/or mindset around what fits best. On the other hand, looking at the problem from the article side, the right size for a customer is not a unique quantity and varies greatly both within and across hundreds of brands, in different sizing systems, in different countries, and for different fashion categories. Therefore, within this work along side introducing our novel size advice approach, we also make a substantial attempt (a first to the best of our knowledge) to establish the state-of-the-art baseline, and rigorously assess the impact of size recommendations with respect to *reducing size-related returns* in online fashion.

2 RELATED WORK

The customers' struggle of finding the right size while shopping online is a well known challenge in the fashion industry [1, 2, 5–14, 16–18, 20, 21, 24, 26–28, 30, 32, 33, 36]. [5, 6] address issues related to customer returns in online fashion retailing and discuss their implications while [32] studies customer-based preventive article return management. [36] suggests a crowdsourcing approach where customers get suggestions on garments based on matching the existing articles in their wardrobes to others in the community. With the aim of unifying the different sizes across size systems and brands, [9] propose a method to automate size normalization to a common latent space through the use of articles ordered data, generating mapping of any article size to a common space in which sizes can be better compared. More recently, there has been emerging research addressing the problem of personalized size recommendation for online fashion retailers [1, 7, 11, 13, 20, 21, 26–28]. Given the order history of a customer (or personal customer data such as age, weight, height, etc.), these methods predict for that customer which size of an article would fit best. Such size recommendation systems personalized to the customer have proved high value in supporting customer decision and enhancing their experience on the platform with respect to size and fit with strong impact on the customer conversion. However, [30] shows that most customers who receive a correct size recommendation would not buy the size recommended due to their individual fit preferences. Additionally, those methods have to explicitly deal with the potentiality of having multiple customers, and thus multiple size and fit preferences behind a single account so to present the right recommendation to its target customer behind the account. However, it should be stated here that to the best of our knowledge, the positive impact of such systems on reducing size-related returns has not yet been demonstrated. We further develop on this matter in section 4 by implementing and A/B testing recent well-known personalized recommendations described in [11] and assess their potential impact on size-related returns.

From a radically different angle, other methods [8, 10, 14] try to estimate a customer's body shape using computer vision (2D or 3D modelling). Although such approaches can efficiently create realistic looking avatars, the claimed virtual fitting experience cannot accurately portray the actual garment fit on the individual complex body. What is more, such virtual try-on methods require personal data from customers such as images in tight clothing, age, gender, weight, height, etc. From a different angle, [16] suggests

learning body-aware embeddings to recommend clothing which complements a specific body type using mined attributes with visual features for clothing, and estimated body shapes. However the customer body shape estimation suffers from the same personal data requirement as the works mentioned above. Relaxing the constraint on having personal customer data, [17] proposes to focus on articles only and suggests using article images to predict how likely a given article is to have a general size issue.

Looking into reducing returns, [18] proposes a method to predict the probability that a customer will return a specific article before the order is placed. Although, the deep neural network architecture introduced is able to capture some latent size and fit information about a customer, the used article features are not size and fit specific (even though the authors acknowledge that a major part of the returns are due to poor size or fit). When an order is found likely to be returned, the suggested approach tries to limit that order with candidate mechanisms including placing more burden on the customer by prohibiting a return, introducing an additional fee, or stopping the order altogether.

In this paper, we aim to both support customers in their size and fit decisions—without imposing any burden of personal data or limitations on their article orders—and reduce size and fit related returns. To that end, we present a Bayesian model which uses size-related return data from customers to learn which articles would fit normally and which would exhibit a size issue. Unlike previous work [1, 7, 11, 13, 20, 21, 26–28], our approach extensively leverages the size-related return rates of articles to greatly focus on modeling articles sizing behaviour. Although the (weakly annotated and subjective) return data from customers is leveraged in the model, the latter is agnostic to the specific customer who places the order and thus by design does not provide a customer with a personalized size recommendation; instead it informs them about article specific sizing characteristics. Thanks to this strategy we create an algorithmically driven customer experience that presents a very low cognitive load by purposefully focusing our approach towards the articles themselves and the underlying manufacturing and brand characteristics, in contrast to personalized size recommendations [1, 7, 11, 13, 20, 21, 26–28] which are emotionally engaging for the customers by trying to convey to them "what their size is"—strongly increasing the chances of a correct recommendation being dismissed altogether [30]—our approach rather focuses on the articles themselves and supports customers to make an informed decision on which size to order given the inferred size and fit characteristics of a given article. This in turn, gives customers the flexibility to adapt their choice based on their fit preference instead of having to cope with a rigid recommendation engine telling them what their size is. Therefore, in contrast to [11, 20, 28] which evaluate the performance of their methods solely based on the accuracy of the personalized size recommendations (i.e. whether the system is good at inferring a customers' size), in this work we go a huge step forward and evaluate the impact of our system on the end-to-end size and fit journey; beyond onsite and unboxing stages and at the size-related return level. Detecting size issues on an individual article level also alleviates other challenges like having multiple users behind one account and having different size systems which [1, 7, 11, 18, 26, 27] have to explicitly deal with.

3 APPROACH

The SizeFlags model we present is fully Bayesian and aims at modeling the behaviour of the size-related returns of an article using some prior information about the article as well as orders and returns data. Figure 1 shows the high level overview of the approach. We use a Binomial distribution for the likelihood and a Beta distribution for the prior. As we will discuss below, prior information can come from various sources of different complexity to help shape the prior Beta distribution. Among the more sophisticated ways, we will discuss obtaining prior information about the fit of an article before observing any return using advanced computer vision deep learning models, such as [17], to process the images of the article.

Let us first simplify the complex problem of determining the size and fit of fashion articles into a 3-class classification problem {no size issue, too small, too big} as in [26, 27], where we have summarized the possible size issues of articles into two categories: 'too small' (indicating the article runs smaller than usual e.g. too short, shoulder too tight, etc) and 'too big' (indicating the article runs bigger than usual e.g. too loose, sleeves too long, etc). To determine whether an article has a size issue or not we use a Bayesian model based on a binomial likelihood. In practice, we use two Bayesian models to predict separately if an article is 'too big' or not, and if it's 'too small' or not. In this context, an article is classified as 'no size issue' if it is tagged by neither of the two Bayesian models. Although the events 'too big' and 'too small' are not independent for a given article, we choose to model them in this way for simplicity and leave the extension to a unified multinomial model for future work. To make things easier for the reader, in all that follows we will present a single model to tag whether an article has a size issue or not, although in practice we use two separate models to tag 'too big' versus 'not too big' and 'too small' versus 'not too small'.

In subsection 3.1 we introduce the binomial likelihood model and explain in subsection 3.2 how this model can be used to predict size issues (raise a size issue flag). In subsection 3.3 we discuss the cold-start problem related to having new articles on the platform, and describe in subsection 3.5 how the incorporation of priors, making the model fully Bayesian, can help tackle this problem.

3.1 The Binomial Likelihood

We start by formalizing the problem, given an article $a \in C$ that belongs to a fashion category $C \subset C_{all}$ (*e.g.* jeans, shirts, shoes, etc.). We define its observed size-related return rate $srr(a)$ as the ratio of the article returns for reasons 'too big' (or similarly 'too small') k out of the number of the total article orders n. We denote the true size-related return rate of an article as $srr_{true}(a)$. This value is unknown, but we know that the higher the number of article orders n, the more confident we are that $srr(a)$ is close to $srr_{true}(a)$: $srr_{true}(a) = \lim_{n\to\infty} srr(a)$. Similarly to [17], we consider two main elements:

(1) Sales period. Depending on the period during which an article is sold, customers' return behaviour and article's ratio of size issues in a category might vary.

(2) Fashion category. Each fashion category C poses different size and fit challenges. Therefore, we compute the mean

$$\pi = \text{mean}(\{srr(a)\}_{a\in C})$$

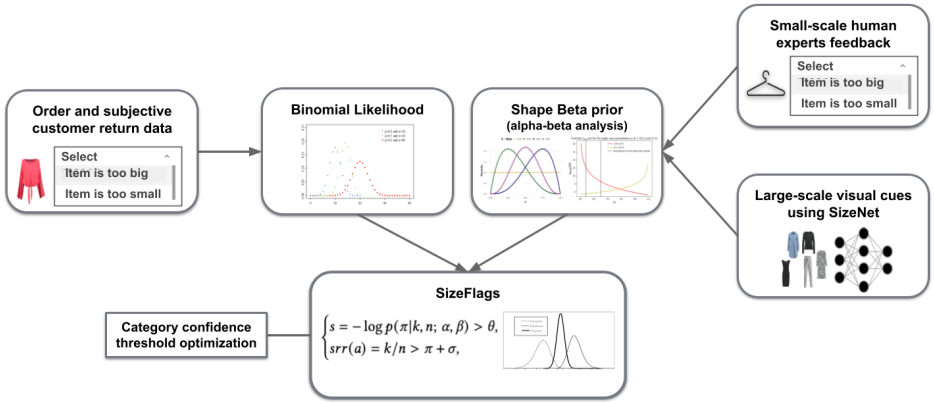

Figure 1: High level overview of the SizeFlags approach.

and the standard deviation

$$\sigma = \text{std}(\{srr(a)\}_{a \in C})$$

of size-related return rate of all articles in the category.

The article a could be considered to have a size issue if $srr_{true}(a) > \pi + \sigma$ when the return rates are calculated over the same articles ordered period of time. Thus, as in [17], we use a binomial law to assess the confidence in the class prediction. The probability of observing k size-related returns over a total of n orders for that article a is modelled by the binomial likelihood

$$p(k|n,\pi) = \binom{n}{k} \pi^k (1-\pi)^{n-k}. \tag{1}$$

This likelihood is maximized when the ratio of k over n is equal to π. In other words, when k is the expected number of size-related returns when drawing n samples from a Bernoulli distribution of parameter π. The estimator becomes more confident whether the k returns were actually drawn with probability π as more article orders are observed. As n grows larger, the likelihood will be close to 1 if the ratio k/n is close to π, and close to 0 if it diverges from it. On the other hand, for low values of n, the estimator is more uncertain and tends to weigh uniformly all possible values of k. We use the score s defined in [17], associated to the observation of k returns out of n article orders given π based on the negative log-likelihood:

$$s = -\log\, p(k|n,\pi). \tag{2}$$

This score s is a positive number and for large values of n, the score (2) scales linearly with respect to n with the divergence rate of the article size-related return rate probability distribution in comparison to its category. For a detailed analysis of the score behavior, we refer the reader to [17].

3.2 Raising a size issue flag

For a given article a for which we have observed k returns out of n article orders, and given π, we can query the binomial likelihood $p(k,n|\pi)$ defined in (1), or alternatively its negative log defined in (2), to know if this article has a size issue or not. A high likelihood means that the observed $srr(a) = k/n \approx \pi$ likely has a "normal" sizing behaviour. On the other hand, a low likelihood could mean one of two things: either $srr(a)$ is too high compared to π ($srr(a) \gg \pi$) and has therefore a size issue, or it is too low ($srr(a) \ll \pi$) then the article is behaving better than average in its category. Ideally, if we had an infinite number (or more realistically a very large number) of orders n for a given article a, then the observed $srr(a)$ would tend to the "true" return rate $srr_{true}(a)$ of the article a, and we would need only compare this value to π to decide if the article indeed has a size issue or not. Then we could raise a "size issue" flag if

$$srr_{true}(a) \geq \pi + \sigma, \tag{3}$$

where σ is the standard deviation of $srr(a)$ over the category C. However, in a realistic setting, using the condition of (3) is not enough to decide whether an article has a size issue or not. Indeed, for an article with a low number of orders (say $n = 10$), the condition could be verified but we do not have enough article orders to make sure that $srr(a)$ is actually close to $srr_{true}(a)$. This is where the binomial likelihood comes into play, as it is able to account for that uncertainty. As mentioned in subsection 3.1, this likelihood tends to be relatively flat for low values of n, resulting in probabilities which are not too low for any value of k. We set a very low error bound $\varepsilon > 0$ as an upper bound on $p(k,n|\pi) < \varepsilon$, and use this in combination with the condition of (3) to determine whether an article has a size issue or not. The condition on the likelihood will ensure that the binomial model is sure enough that $srr(a)$ is "abnormal" compared to what is observed over the whole category C, and that is not just due to lack of article orders. The condition of (3) will tell us in turn that the return rate $srr(a)$ is abnormally high, resulting in a size issue flag, and not abnormally low compared to the rest of the category (which would mean a very well fitting article).

Using ε as an upper bound on the likelihood translates into a lower bound threshold $\theta = -\log(\varepsilon) > 0$ on the score s from (2). Combining the two conditions described above into a single statement leads to the following joint condition for raising a size issue flag:

$$\begin{cases} s = -\log p(k|n,\pi) \geq \theta, \\ srr(a) = k/n \geq \pi + \sigma. \end{cases} \tag{4}$$

3.3 Parameter setting and cold-start problem

With the two joint conditions in (4), the question is now how to set the threshold $\theta = -\log(\varepsilon)$ that ensures high confidence. Here are 2 ways to set an appropriate parameter θ for the algorithm:

(1) If the goal is to identify the x most problematic articles (e.g. 5% of a category C), then we can set the steering parameter θ by calibrating this threshold on past data.
(2) If we want to raise a sizing flag with high certainty, then $\varepsilon > 0$ must be chosen very small. In this case, ε can be set for example equal to machine epsilon ε_{mach} (e.g. in single precision $\varepsilon_{mach} = 2^{-23} \approx 1.19e^{-07}$).

In calibrating this threshold, we run through the risk of taking too long until the estimator achieve the required confidence level. This risk is particularly relevant in fashion e-commerce, where articles generally have a short lifetime on the platform and return data is delayed due to logistic constraints. Therefore, in the discussed approach, the article might run out of stock before a sufficient amount of orders and returns is observed. This directly leads to customers dissatisfaction and increased environmental and financial costs as more returns directly translates to a higher carbon footprint and a higher logistic costs. Moreover, with this high confidence bar, problematic articles with low number of orders might never get flagged during their whole lifetime on the platform. This means that customers are not informed about many articles that have size issues.

3.4 Threshold optimization

Setting a conservative threshold θ makes the algorithm more robust to noise and increases the confidence. This hinders the algorithm's performance as too high numbers of article orders and returns (n and k) are necessary to raise a size issue flag, interfering with the main goal of the algorithm. To tackle this cold-start problem, we propose to choose a more conservative $\varepsilon_{\min}$ in the beginning to get stable and accurate sizing flags first. Based on this $\varepsilon_{\min}$ and $\theta_{\max} = -\log(\varepsilon_{\min})$ selected as a starting point, a better threshold $\theta^\star$ for (4) can be determined by solving an optimization problem. Taking into consideration that different fashion categories exhibit different sizing challenges and return behaviour, consequently we search for optimal threshold per category.

We show now how to get an optimized threshold $\theta^\star = \theta^\star(C)$ using historical data in order to obtain almost the same size issue flags as with $\theta_{\max}$, only much faster. We select $\theta^\star > 0$ according to the following problem:

$$\theta^\star = \underset{\theta \in (0, \theta_{\max}]}{\arg\min}\ \theta \quad \text{subject to} \quad \frac{N(C,\theta) - S(C,\theta)}{N(C,\theta)} \leq \varepsilon_1, \tag{5}$$
$$\frac{N(C,\theta) - N(C,\theta_{\max})}{N(C,\theta_{\max})} \leq \varepsilon_2, \quad \frac{N(C,\theta) - S(C,\theta)}{N(C,\theta_{\max}) - S(C,\theta_{\max})} \leq \varepsilon_3.$$

We denote by $N(C,\theta)$ the number of articles $a \in C$ flagged by the algorithm with parameter θ and by $S(C,\theta)$ (the stability) the number of flags raised whose value do not change across the time period considered. Thus, the difference $N(C,\theta) - S(C,\theta)$ can be seen as the number of unstable flags. The second constraint restricts the number of additional flags compared to the baseline. The third constraint translates that the ratio of unstable flags with parameter θ compared to the same ratio with the baseline $\theta = \theta_{\max}$ must be restricted by ε_3. Parameters ε_i are thus values set to ensure the flags raised using a threshold θ are similar in number and quality to those raised by baseline. Due to the monotony behavior of the constraints, $\varepsilon_1, \varepsilon_2 \in (0,1)$ and $\varepsilon_3 \geq 0$.

In order to be able to reproduce our results, we would like to point out that we have solved the optimization problem (5) by first discretizing equidistantly in $\theta \in (0, \theta_{\max})$ and then choosing $\theta^\star$ with the smallest target-functional. We have set $(\varepsilon_1, \varepsilon_2, \varepsilon_3) = (0.2, 0.05, 1.5)$ in order to generate sufficiently stable results.

3.5 Full Bayesian Model with priors

To overcome the limitation (cold-start problem) exposed in subsection 3.3, we propose the addition of a prior to our binomial likelihood model in order to introduce "expert information" to shape the binomial distribution even before any order is observed for an article. The expert information can come from different sources; we focus on Human expert feedback from fashion models and Sizing information extracted from visual cues of articles (SizeNet). We use the Beta law as a prior distribution on $srr(a)$ as it is the conjugate distribution to the Binomial law. In this full Bayesian setting, we model each individual article as having a true, but unknown size-related return rate. We choose to model this return rate as a random variable R following a Beta distribution $R \sim Beta(\alpha, \beta)$. Given a value of $R = r$ for an article and n orders , the likelihood of observing k returns for this article follows a Binomial distribution of parameter r, $p_{binom}(k|n,r) = \binom{n}{k} r^k (1-r)^{n-k}$. The density function of the Beta distribution with parameters $\alpha, \beta > 0$ is given by $p_{\text{beta}}(r|\alpha,\beta) = B(\alpha,\beta)^{-1} r^{\alpha-1} (1-r)^{\beta-1}$ where $B(\alpha,\beta)$ is the value of the Beta function at (α,β). In this Bayesian setting, the full joint likelihood over variables k, r is given by

$$p(k, r|n, \alpha, \beta) \quad = \quad p_{\text{binom}}(k|n,r) p_{\text{beta}}(r|\alpha,\beta).$$

The Binomial law and the Beta law being conjugates, the posterior distribution

$$p(r|k,n;\, \alpha,\beta) = B(k+\alpha, n-k+\beta)^{-1} r^{k+\alpha-1} (1-r)^{n-k+\beta-1}$$

over the true return rate r of an article is also a Beta distribution of parameters $(k+\alpha-1,\ n-k+\beta-1)$.

Raising a size issue flag: In this new Bayesian setting, we replace the condition on the Binomial likelihood being small enough by a condition on a point estimate of the posterior distribution. The new joint conditions for raising a size issue flag now become:

$$\begin{cases} s_{\text{posterior}} = -\log p(\pi|k,n;\, \alpha,\beta) \geq \theta, \\ srr(a) = k/n \geq \pi + \sigma, \end{cases} \tag{6}$$

where $\theta = \theta^\star$ can be obtained using the threshold optimization in subsection 3.4 and α and β are article-specific parameters which are set using prior information about articles. Using this posterior distribution which incorporates expert knowledge through the prior distribution helps raise flags faster as discussed below.

Prior distribution analysis: In our modeling, the values of α and β are defined as ≥ 1 so that the prior distribution is not skewed around the extreme values 0 and 1. Since for $\alpha = \beta = 1$ the prior represents a uniform distribution, we choose $\alpha, \beta \geq 1$ with at least one of them > 1. $p_{\text{beta}}(r|\alpha,\beta)$ therefore becomes a unimodal distribution.

For the first condition of (6) it is essential that the parameters α and β are reasonably restricted by $\alpha \in [1, \alpha_{\max}]$ and $\beta \in [1, \beta_{\max}]$ where the upper boundaries strongly depend on the threshold θ. Here $\alpha_{\max}$ and $\beta_{\max}$ may not be chosen too large to ensure a natural balance between p_{binom} and p_{beta} in condition (6): raising a flag

should not rely too strongly on the prior information only through p_{beta}. The prior plays a big role especially at the beginning when we do not have any sales and return data, i.e. n = 0 and k = 0. Therefore the optimization of $\alpha_{\max}$ and $\beta_{\max}$ is modeled for this edge case. With $p_0(\pi, \alpha, \beta) = -\log p(\pi|0, 0; \alpha, \beta)$ we impose the following conditions to get $\alpha_{\max}$ and $\beta_{\max}$:

$$\begin{cases} \alpha_{\max} &= \arg\min_{\alpha>1} \left|\max_{\pi\in\Pi} \left\{p_0(\pi, \alpha, 1)\right\} - \theta\right|, \\ \beta_{\max} &= \arg\min_{\beta>1} \left|\min_{\pi\in\Pi} \left\{p_0(\pi, 1, \beta)\right\} + 1\right|. \end{cases} \tag{7}$$

The inner optimization problems $\pi^\star_\alpha = \arg\max_\Pi \{p_0(\pi, \alpha, 1)\}$ and $\pi^\star_\beta = \arg\min_\Pi \{p_0(\pi, 1, \beta)\}$ ensures that $p_0(\pi^\star_\alpha, \alpha_{\max}, 1) = \delta_\alpha \geq p_0(\pi, \alpha_{\max}, 1)$ and $p_0(\pi^\star_\beta, 1, \beta_{\max}) = \delta_\beta \leq p_0(\pi, 1, \beta_{\max})\ \forall \pi \in \Pi$. By optimizing α to achieve θ at $\pi^\star_\alpha$, we ensure that p_0 for all other $\pi \in \Pi$ will be lower than $\delta_\alpha \approx \theta$. For β it's the opposite: By optimizing β to achieve -1 at $\pi^\star_\beta$, we ensure that p_0 for all other $\pi \in \Pi$ will be greater than $\delta_\beta \approx -1$.

One challenge in solving the optimization problems (7) is to find a suitable choice of the fixed parameters $\theta > 0$ and $\Pi \subset [0, 1]$: In subsection 3.3 we described how a suitable θ can be chosen.

The interval $\Pi = \Pi(C)$ depends on C and specifies the area in which the actual parameter π is very likely to be located. π can change over time because the category can evolve, so an interval is required here, in which π usually is. Let's define fixed initial parameters C^{init} as the initial state of the category, $\pi^{init} = srr(C^{init})$ and $\sigma^{init} = \text{std}(\{srr(a)\}_{a\in C^{init}})$ then we define

$$\Pi = [\pi^{init} - \sigma^{init}, \pi^{init} + \sigma^{init}].$$

Exemplification: Assuming that $\Pi = [0.08, 0.3]$ and $\theta = 15$. Under the condition that both $\alpha \geq 1$ and $\beta \geq 1$ must be integers, we get $\alpha_{\max} = 8$ and $\beta_{\max} = 3$ as solution of the optimization problem (7). As you can see in Figure 2, the functions $s_{\text{beta}}(\pi, 8, 1)$ and $s_{\text{beta}}(\pi, 1, 3)$ form an upper and lower boundary for $s_{\text{beta}}(\pi, \alpha, \beta)$ $\forall\ \alpha \in [1, 8]$ and $\beta \in [1, 3]$. All values $\alpha \in [1, 8]$ and $\beta \in [1, 3]$ can now be used for the algorithm, depending on the prior information of an article $a \in C$.

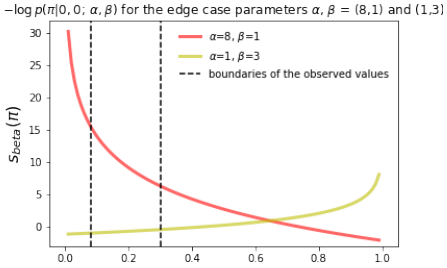

Figure 2: Prior information for $\alpha_{\max} = 8$ and $\beta_{\max} = 3$.

3.6 Extended Full Bayesian Model

The fully Bayesian model discussed in subsection 3.5 requires additional information to be integrated as prior knowledge within the algorithm. In this section we present two extensions which leverage two different sources of expert knowledge: First we explain how to integrate feedback from human models, and afterwards we show how to use previous work on extracting sizing information from article images [17] to incorporate prior information about articles even before any order is observed. The combination of these extensions allows us to tackle the cold-start problem in a robust and efficient manner, constituting our final SizeFlags algorithm. Finally, we give an overview of the whole algorithm written as pseudo code.

Human expert feedback: To incorporate expert knowledge from human feedback we ask fashion models with relatively standard body sizes to try on articles before they are activated on the platform to give us early size and fit feedback. For the purpose of this paper, we constrain the feedback to be one of {"certain size issue", "potential size issue", "good fit"}. Multiple human models can fit the same article and based on the aggregated feedback we define article-specific prior parameters (α, β). If the aggregated feedback points strongly in the direction of a size issue, we chose values which strongly favor high return rates, if the aggregated feedback is more uncertain we favor more mildly high return rates, and if the aggregated feedback points towards a good fit we favor low return rates.

Size and fit visual cues: Using the full Bayesian model presented in subsection 3.5 with human feedback does help alleviate the cold-start problem but it does not totally suppress the issue. In fact, a well trained group of human models can only fit a very limited part of the ever changing online assortment every week (less than 1%), which means only a few articles have prior information. For our second extension, we thus propose to use size and fit visual cues to obtain prior information on large-scale and use it in our full Bayesian model. Although the literature is very rich on computer vision techniques for various fashion problems, there are only few works considering fashion images for the size and fit problem (see [16, 17] and references therein). We use our fully Bayesian model by incorporating sizing information extracted from article images using SizeNet [17]. We use the student part in which a backbone convolutional neural network is used to extract visual features from article images and multi-layer perceptron learns article sizing behaviour. We convert the output of the network into prior values (α, β) of the Beta distribution. More specifically, the output is a size issue probability and, similarly to human expert feedback, we chose values which strongly favor high return rates for articles that have high size issue probability.

To give the reader an overview of the SizeFlags algorithm, a summary in the form of a pseudo code is provided in Algorithm 1.

4 EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Challenges in Evaluating Size and Fit Recommendations

Evaluating size and fit recommendations remains highly challenging due to multiple underlying factors specific to this problem space, in particular:

(1) Starting with the customer, the "true" size of a customer is often unknown, remains subjective for each customer, may depend on the context and the occasion for which a customer is shopping, and can vary greatly by external factors such as fashion trends, seasonality, and life changing events impacting a customer's physical body, and/or mindset around what fits best.

Algorithm 1 SizeFlags

Require: Category $C = \{a_1, \ldots, a_N\}$ contains N articles a_i with attributes $\{k_i$ size-related returns, n_i number of orders$\}$

Initialize
- $srr(a_i) = k_i/n_i \; \forall a_i \in C$
- $\pi = \text{mean}(\{srr(a_i)\}_{a_i \in C})$
- $\sigma = \text{std}(\{srr(a_i)\}_{a_i \in C})$

for each $a_i \in C$ **do**
  if $srr(a_i) \geq \pi + \sigma$ **then**
    use prior information to set $(\alpha_i, \beta_i) = (\alpha(a_i), \beta(a_i))$
    if $-\log p(\pi | k_i, n_i; \alpha_i, \beta_i) > \theta$ **then**
      a_i has size issue: raise sizing flag
    else
      a_i has probably no size issue: don't raise sizing flag
    end if
  else
    a_i has no size issue: don't raise sizing flag
  end if
end for

(2) Looking at the problem from the article side, the right size for a customer is not a unique quantity and varies greatly both within and across brands, in different sizing systems, in different countries, and for different fashion categories.
(3) In online fashion, there is a significant delay between the time that a size and fit recommendation is provided to a customer, and the feedback signal coming from the customer once they have actually tried the recommended size on. Through the lens of size and fit related returns, in order to evaluate the quality and effect of a recommendation, one needs to wait for several days if not weeks for the customer's return to reach the platform.
(4) From the customer experience point of view, customer satisfaction from a size and fit recommendation is not solely based on the recommendation's quality itself but is often intertwined with the physical characteristics and the experienced "feel" of the shoe or the garment once it is worn. This satisfaction is time-dependent and may change radically after wearing an article for a few weeks, or after a washing experience.

Considering the above challenges, previous work [11, 20, 27] has been mainly focused on evaluating the quality of the recommendations *not* on their impact on reducing size-related returns, but rather on a combination of customer based metrics through A/B tests [34] or continuous evaluation frameworks. Such customer based metrics include monitoring a change in customers' conversion rate, in the number of products added to the cart, in the revenues per visit, in selection orders (where a customer orders the same article in multiple sizes), and in reorders (where a customer returns an article and reorders it in a different size). In the same spirit, the literature also considers the customer's acceptance of the recommendations (i.e. how often they order in a recommended size), and accuracy of such recommendations (i.e. how often they keep or return a recommended size when they have accepted or not that recommendation). For more details on this direction, readers are invited to see [11, 20, 27] and references therein. Within this work, we propose to go one critical step ahead and rigorously assess, for the first time to the best of our knowledge, the quality of the recommendations on their impact in reducing the size and fit returns.

Table 1: Name abbreviations of the proposed models

V_0	Binomial model on order and return data
V_{HF}	Only human feedback
V_{Base}	Baseline (order and return data + human feedback prior)
V_{SN}	Baseline + size and fit visual cues prior
V_{TH}	Baseline + optimized thresholds
SizeFlags	Baseline + size and fit visual cues prior + optimized thresholds

In what follows, we first establish the state-of-the-art baseline and then evaluate our models, summarized in Table 1, from the binomial model launched for textile and shoe categories all the way to the full Bayesian SizeFlags launched in 2020 over 14 European countries with various local and mix size systems.

4.2 Establishing state-of-the-art benchmark

In online fashion reducing size-related returns is achieved through an algorithmically driven user experience, where different machine learning and recommendation approaches deliver a size advice to the customers and aim to drive their behaviour towards selecting the size that fits them the first time [25]. Within this domain, controlled live A/B testings [34] are widely used as a proven mechanism to benchmark the effect of such algorithmically driven user experiences on a set of metrics of interest. Therefore, to assess our approach we first establish our baseline to be the personalized size-recommendation in online fashion introduced in [11], a solid state-of-the-art within this domain. Following [11] a customer receives a recommendation on which size to order for a given article based on their order and return history. Two successive live A/B tests were performed respectively on (1) women and men shoes and (2) women and men textile with over 300k customers per group. Using controlled and live A/B test settings, we randomly divided the customers into two groups; a control group that received no size advice and a test group with personalized advice delivered by the approach from [11]. The results showed positive financial impact with a significant increase in conversion rate (+2.1%), more products added to the cart (+1.8%), and increased revenues per visit (+2.1%), however, no statistically significant impact on reducing size-related returns were observed (<0.5% relative reduction).

4.3 Evaluating our SizeFlags Models

A/B test on shoes (V_0): To evaluated our binomial model (denoted V_0 in Table 1), we performed a first A/B test on women and men shoes in 2017. We provided the size advice derived from the size issue predictions of V_0 to the test group and no advice to the control group, both groups with 720k customers. The A/B test results demonstrated that *srr* was significantly reduced by the provided size advice (3.8% relative reduction). Interestingly, the A/B test also demonstrated that V_0 has a clear effect on size-related selection orders (customer orders two or more different sizes of the same article

at the same time): the size advice 'too small' led to a size-related selection order increase of 11.1%, whereas 'too big' led to an increase of 19.0%, hinting at potentially different customer perception on whether a smaller or bigger size is a low risk order or flattering.

A/B test on textile (V_0): A second A/B test was performed for V_0 on textile in 2017. Each group with over 180k customers. In addition to all the usual textile categories such as dresses, trousers, etc. the sub-categories 'beach & lingerie' and 'sportswear' were also included in this group. The A/B test results demonstrated that the size-related return rate was significantly reduced thanks to the provided size advice by 4.3% for 'too small' and 6.6% for 'too big' size flags. Interestingly, we observe that customers react stronger to the size advice for articles that are flagged as too big.

4.4 Continuous evaluation

Outside of A/B testing, estimating the causal effects of size advice is a non-trivial exercise as very little is known about the individual customer behaviour before ordering an article with size advice. A/B tests provide an accurate snapshot of the intervention effect, however they are detrimental to customers' experience when continually performed since, by design, a group of customers end up with the relatively less attractive experience for the duration of the test. To tackle this challenge, here we first introduce an efficient approach for continuously monitoring the impact of our models (outside of A/B tests) and next present the experimental results of our models under the continuous evaluation regime.

As discussed in [15], the nearest neighbor Difference-in-Differences (DiD) matching estimator method is widely used in impact evaluation studies. We adopt DiD since it is able to compare the evolution over time t of the size-related return rate srr in two groups of treatment Ω_T and control Ω_C, with parallel trends: $\Omega_T \subset C$ that covers all articles with size advice in a given category C and $\Omega_C \subset C$ that counterparts articles which have never had a size issue flag. From Ω_C we can find $\Omega_C^\star$, with say the ten ($k = 10$) nearest neighbors within the same category C of article a where $a \in \Omega_T$ sold during the same time period. We use Euclidean distance over four continuous covariates to determine the nearest neighbours; (1) general return rate (including non size-related reasons), (2) price, (3) discount rate, and (4) return rate for the return type 'unknown', as they have proven to be the four most common confounders in this context. Let's denote $srr(\Omega_C^\star)$ as the average srr of its elements and $t_{flag}(a)$ as the point of time when an article a received a sizing flag. Using $\Omega_C^\star$ we are now able to estimate the average effect

$$srr_{effect} = |\Omega_T|^{-1} \sum_{a \in \Omega_T} \left(\frac{\gamma_{a,t_0} - \gamma_{a,t_1}}{srr(a|t < t_{flag}(a))} \right)$$

of the size issue flags where

$$\gamma_{a,t_0} = srr(a|t < t_{flag}(a)) - srr(\Omega_C^\star|t < t_{flag}(a)),$$
$$\gamma_{a,t_1} = srr(a|t > t_{flag}(a)) - srr(\Omega_C^\star|t > t_{flag}(a)).$$

In our estimations $t < t_{flag}(a)$ denotes a period of six weeks before the size issue flag was raised for an article a and $t > t_{flag}(a)$ covers a period of six weeks after the size issue flag was raised.

4.4.1 Impact of human expert feedback: We focus here on the version V_{Base} described in subsection 3.6 which was launched in 2017 in order to tackle the cold-start problem by incorporating human expert size and fit feedback as a prior to our model. We utilize the continuous evaluation method DiD presented in subsection 4.4 to estimate the impact of this prior on the size-related returns. To isolate and demonstrate the impact of this prior, we first focus on V_{HF} flags which represent the subset of flags raised thanks to the human expert feedback ($V_{HF} \subset V_{Base}$). Afterwards, we will discuss the reached impact from the whole V_{Base}.

Impact of V_{HF}: Table 2 summarizes the strong positive impact of V_{HF} in terms of srr reduction observed between September and December 2019. For the evaluation, 2678 articles were used that received a size advice too big or too small thanks to the human feedback prior. Our samples consist of 469 textile articles with 27773 orders and 4411 size-related returns and 2219 shoe articles with 53951 orders and 7846 size-related returns. As for the impact on the cold-start problem, it was observed that incorporating this prior leads to 33% faster flagging, which entails that customers are informed about size issues much earlier, resulting in fewer size-related returns as shown in Table 2.

Table 2: Estimated impact of V_{HF}

category	srr reduction	avg. returns saved/article
Shoes	4.40 %	5.68
Textile	8.15 %	5.87

Impact of V_{Base}: Table 3 summarizes the strong positive impact of V_{Base} in terms of srr reduction observed between March 2019 and February 2020. The results are the average calculated effect for all articles $a \in \Omega_T$. The resulting group is much larger and consists of 10704 textile articles with 645667 orders and 120057 returns and 3625 shoes with 503849 orders and 80339 size-related-returns. The estimated impact also remains in line with the results obtained from our A/B tests.

Table 3: Estimated impact of V_{Base}.

category	srr reduction	avg. returns saved/article
Shoes	5.65%	2.29
Textile	5.01%	1.50

4.4.2 Impact of SizeNet and threshold optimization: We evaluated the impact of V_{SN} and V_{TH} along with the full SizeFlags model in comparison to the reference model V_{Base}. We ran these versions simultaneously over a 1 month period following the launch of the SizeFlags in 2020. The data for the simultaneous evaluation was limited to textile articles, the size of the dataset was 311851 articles and 25% of them are flagged in the shared group. To assess the impact, we focus on the two key metrics illustrated in Figure 3. (1) Article orders $n(a)$: the number of necessary article orders before a flag is raised, (2) Article returns $ret(a)$: the number of necessary article returns before a flag is raised. Since the evaluated versions V_i may raise flags for articles which are not flagged by V_{Base}, we split our analysis between *overall* articles $\Omega_O(V_i)$ and *shared* articles $\Omega_S(V_i)$. $\Omega_O(V_i)$ refers to all the flags raised by V_i and $\Omega_S(V_i)$ refers to those articles which are flagged by both V_i and V_{Base}. Indeed, the articles in $\Omega_S(V_i)$ are particularly important as they have already proven to have good quality and impact. This leads us to set $\Omega_S(V_i)$ for all V_i to cover at least 95% of those raised by V_{Base}. To ensure quality, the articles in $\Omega_O(V_i)$ are continuously assessed following the approach DiD detailed in subsection 4.4. Based on Figure 3, we observe that (1) both V_{SN} and V_{TH} positively reduce $n(a)$ and $ret(a)$

and (2) that $\Omega_S(V_{TH})$ outperforms $\Omega_S(V_{SN})$ whereas $\Omega_O(V_{SN})$ surpasses $\Omega_O(V_{TH})$. Since the flags raised by V_{Base} have shown to bring strong positive impact on the *srr* reduction, it is critical to guarantee that $\Omega_S(V_i)$ are raised even faster, and thus we opted for the balanced version SizeFlags in production which combines the two extensions V_{SN} and V_{TH}, thereby bringing the best out of both of them to millions of customers. SizeFlags were rolled out to 14 countries in February 2020; since going live with this approach, we have seen the average of $n(a)$ and $ret(a)$ to drop significantly, by 17% and 40% respectfully, in the real world production environment.

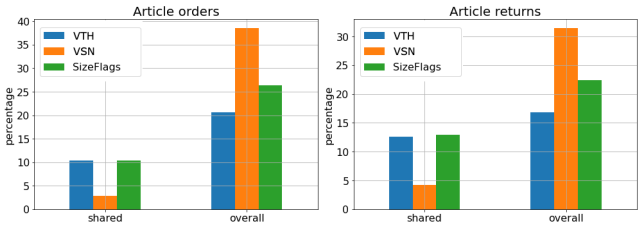

Figure 3: Impact of V_{SN}, V_{TH}, and SizeFlags compared to V_{Base}. Left: average relative reduction of $n(a)$: the number of necessary article orders before a flag is raised. Right: average relative reduction of $ret(a)$: the number of necessary article returns before a flag is raised.

5 CONCLUSION

The challenging task of reducing size and fit related returns in the context of online fashion was studied. A probabilistic Bayesian approach (SizeFlags) was introduced leveraging large-scale return data and alleviating the cold-start problem thanks to rich priors from human experts feedback and computer vision techniques. The production results demonstrated that SizeFlags effectively reduces size-related returns in online fashion. In addition, using optimized thresholds and rich priors greatly reduced the number of ordered and returned articles necessary for reducing size-related returns. Limitations of our work include the coarse definition of size and fit issues (too small and too big), and future work will explore extending the Bayesian model to a hierarchical model with a multinomial likelihood and Dirichlet prior in order to include finer-grained size and fit problems (e.g. tight on the hips, too long sleeves) and more high fidelity data such as article measurements.

6 ACKNOWLEDGEMENTS

The authors would like to thank Romain Guigourès and Yuen King Ho for the fun, positive energy, and fruitful size and fit discussions contributing to the success of this work.

REFERENCES

[1] G M. Abdulla and S. Borar. 2017. Size Recommendation System for Fashion E-commerce. In *Workshop on Machine Learning Meets Fashion, KDD.*
[2] S Ashdown. 2007. *Sizing in clothing.* Elsevier.
[3] C Barry. 2000. Happy returns: how to reduce customer returns-and their costs. *Catalog Age* (2000).
[4] Tsan-Ming Choi. 2016. *Analytical Modeling Research in Fashion Business.* Springer.
[5] Sh. Cullinane, M. Browne, E. Karlsson, and Y. Wang. 2019. *Retail Clothing Returns: A Review of Key Issues.* 301–322. https://doi.org/10.1007/978-3-030-14493-7_16
[6] M. A. Diggins, C. Chen, and J. Chen. 2016. A Review: Customer Returns in Fashion Retailing.
[7] K. Dogani, M. Tomassetti, S. De Cnudde, S. Vargas, and B. Chamberlain. [n.d.]. Learning Embeddings for Product Size Recommendations. In *SIGIR eCom'19.*
[8] H Dong, X Liang, X Shen, B Wang, H Lai, J Zhu, Z Hu, and J Yin. 2019. Towards Multi-Pose Guided Virtual Try-On Network. In *The IEEE Conference on Computer Vision.*
[9] Eddie S.J. Du, Chang Liu, and D Hutchison Wayne. 2019. Automated Fashion Size Normalization. *ArXiv* abs/1908.09980 (2019).
[10] P Guan, L Reiss, D A. Hirshberg, Er Weiss, and M J. Black. 2012. Drape: Dressing any person. *ACM Trans. on Graph* (2012).
[11] R. Guigourès, Y. King Ho, E. Koriagin, A-S Sheikh, U. Bergmann, and R. Shirvany. 2018. A hierarchical bayesian model for size recommendation in fashion. In *Proceedings of the 12th ACM Conference on Recommender Systems.*
[12] Y Ha and L Stoel. 2004. Internet apparel shopping behaviors: the influence of general innovativeness. *Journal of Retail & Distribution Management* (2004).
[13] K. Hajjar, J. Lasserre, A. Zhao, and R. Shirvany. 2020. Attention Gets You the Right Size and Fit in Fashion. In *ACM Conference on Recommender Systems. RecSys'20 Workshops (fashionXrecsys'20).*
[14] X Han, Z Wu, Z Wu, R Yu, and L S Davis. 2018. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition.*
[15] J J Heckman, H Ichimura, and P E Todd. 1997. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The review of economic studies* 64, 4 (1997), 605–654.
[16] W-L Hsiao and K Grauman. 2020. ViBE: Dressing for Diverse Body Shapes. In *Computer Vision and Pattern Recognition.*
[17] N Karessli, R Guigourès, and R Shirvany. 2019. SizeNet: Weakly Supervised Learning of Visual Size and Fit in Fashion Images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop on FFSS-USAD.*
[18] S Kedia, M Madan, and S Borar. 2019. Early Bird Catches the Worm: Predicting Returns Even Before Purchase in Fashion E-commerce. *CoRR* abs/1906.12128 (2019). arXiv:1906.12128 http://arxiv.org/abs/1906.12128
[19] Jr. C. J. Langley, J. J. Coyle, B. J. Gibson, R. A. Novack, and E. J. Bardi. 2008. Supply Chain Management: A Logistics Perspective. 8th edn. South Western College, Kentucky.
[20] J. Lasserre, A-S. Sheikh, E. Koriagin, U. Bergman, R. Vollgraf, and R. Shirvany. 2020. Meta-learning for Size and Fit Recommendation in Fashion. In *SIAM International Conference on Data Mining.*
[21] L. Lefakis, E. Koriagin, J. Lasserre, and R. Shirvany. 2020. Towards User-in-the-Loop Online Fashion Size Recommendation with Low Cognitive Load. In *ACM Conference on Recommender Systems. RecSys'20 Workshops (fashionXrecsys'20).*
[22] J Mostard and R Teunter. 2006. The newsboy problem with resalable returns: A single period model and case study. *European Journal of Operational Research* 169, 1 (2006), 81–96.
[23] E. Ofek, Z. Katona, and M. Sarvary. 2011. The impact of product returns on the strategies of multichannel retailers. *Marketing Science* 30, 1 (2011), 42–60.
[24] C Ratcliff. 2014. How fashion ecommerce retailers can reduce online returns. *Blog text, Econsultancy. Saatavissa* (2014).
[25] 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).* 801–808.
[26] V. Sembium, R. Rastogi, A. Saroop, and S. Merugu. 2017. Recommending Product Sizes to Customers. In *Proceedings of the Eleventh ACM Conference on Recommender Systems.*
[27] V. Sembium, R. Rastogi, L. Tekumalla, and A. Saroop. 2018. Bayesian Models for Product Size Recommendations. In *Proceedings of the 2018 World Wide Web Conference.*
[28] A-S. Sheikh, R. Guigourès, E. Koriagin, Y. King Ho, R. Shirvany, and U. Bergmann. 2019. A deep learning system for predicting size and fit in fashion e-commerce. In *Proceedings of the 13th ACM Conference on Recommender Systems.*
[29] G Stalk. 2006. Customer returns top $10 billion in 2005: most Canadian retailers fail to capitalize on this key customer relationship. *Canada NewsWire* (2006).
[30] A Vecchi, F Peng, and M Al-Sayegh. 2015. Looking for the perfect fit? online fashion retail - opportunities and challenges.
[31] R. Velazquez and S. M. Chankov. 2019. Environmental Impact of Last Mile Deliveries and Returns in Fashion E-Commerce: A Cross-Case Analysis of Six Retailers. In *IEEM'19.* 1099–1103.
[32] G. Walsh, M. Möhring, C. Koot, and M. Schaarschmidt. [n.d.]. Preventive Product Returns Management Systems: Review and Model 2014. *ECIS 2014 Proceedings - 22nd European Conference on Information Systems.*
[33] N Weidner. 2010. *Vanity sizing, body image, and purchase behavior: A closer look at the effects of inaccurate garment labeling.* Ph.D. Dissertation. Eastern Michigan University.
[34] S Young. 2014. Improving Library User Experience with A/B Testing: Principles and Process. *Weave: Journal of Library User Experience* (2014).
[35] Y Yu and H-S Kim. 2019. Online retailers' return policy and prefactual thinking: An exploratory study of USA and China e-commerce markets. *Journal of Fashion Marketing and Management* 23, 4 (2019), 504–518.
[36] Y. Zhang and O. Juhlin. 2015. Using crowd sourcing to solve the fitting problem in online fashion sales. *Global Fashion Management Conference* (2015).