Towards User-in-the-Loop Online Fashion Size Recommendation with Low Cognitive Load

LEONIDAS LEFAKIS, Zalando SE
EVGENII KORIAGIN, Zalando SE
JULIA LASSERRE, Zalando SE
REZA SHIRVANY, Zalando SE

One of the major challenges facing e-commerce fashion platform is that of recommending to customers the right size and fit for fashion apparel. In this work we study this topic in depth and demonstrate its various complexities focusing in particular on the challenging cold-start problem that arises when no order history is available for a specific customer. We demonstrate the multifaceted value of data obtained by involving the customer in the loop and show how it allows for an effective cold-start recommender system. We highlight our findings via detailed experiments performed on hundreds of thousands of customers and items in real world e-commerce scenarios. In addition, results and discussions are provided investigating the trade-off between the recommender's effectiveness and the customer's experience with the goal of introducing accurate solutions with low user cognitive load.

Additional Key Words and Phrases: Size and Fit Recommendation, Fashion e-Commerce, Cold-Start Recommendation, User-Centric Study, Real-World Implementations, Cognitive Load

ACM Reference Format:
Leonidas Lefakis, Evgenii Koriagin, Julia Lasserre, and Reza Shirvany. 2020. Towards User-in-the-Loop Online Fashion Size Recommendation with Low Cognitive Load. In *Proceedings of RecSys '20 Workshops: ACM Conference on Recommender Systems (RecSys '20 – fashionXrecsys '20).* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Finding fashion apparel online with the right size and fit is challenging for many customers. It is actually one of the major factors impacting not only customers purchasing decisions, but also customers satisfaction with e-commerce fashion platforms. The underlying difficulties mean that either customers remain reluctant to engage in the purchasing process, in particular with regards to new articles and brands they are not familiar with, or they purchase articles in multiple neighboring sizes to try them out and return the ones that are not fitting. Compounding the issue, customer preferences towards perceived article size and fit for their body remain highly personal and subjective which in turn influences the definition of the right size for each customer.

In order to achieve these goals, major fashion platforms are experimenting with providing size and fit advice to steer customers' behavior using a variety of approaches. One of the simplest methods used is that of size tables and aggregated article measurements [1] provided per brand and article category. This approach requires customers to find what fits them best themselves and relies on various body measurements, such as "bust", "waist", "hip", also typically measured by the customers themselves. However, these charts rarely help a customer select the best size [2]. Another

Manuscript submitted to ACM

emerging approach is that of directly addressing customers, for example via questionnaires or dialogue windows [3] to provide a size advice based on customers' explicit data. Such approaches have been recently adopted by major e-commerce platforms [4–6], and require customers to explicitly provide personal information, such as age, weight, height, tummy shape, hips form, body type, favorite brand, fit preferences (such as slim vs. normal), customers usual sizes, and so on, used to provide size advice for the customers. In a similar vein of asking customers for explicit body data, computer vision and 3D approaches [7–10] have also shown promising results in providing virtual-fit advice. Such solutions require customers to submit, typically high definition images or videos of their bodies, in predefined poses and tight fitting clothes. Such strict requirements are necessary to allow the currently available technology to infer relatively accurate body measurements, size and shape. In contrast to these approaches are the recent work that do not require any explicit data from customers to provide size advice but rather exploit customers rich order history in order to infer the appropriate size advice [11–17] on future orders.

Each of these various approaches typically rely on disparate assumptions making each appropriate for different customer segments and experiences. In particular it is obvious that the data used by each of these approaches is very different in nature and each require a different level and type of engagement from the side of the customer. Certainly a comprehensive comparison of all these approaches would be a high-value to the community- this remains out of the scope of this paper and constitutes a great future research direction. Here we aim to investigate the size recommendation problem in the so called cold-start scenario in which there is little to no order history for each customer, and thus, customers are part of the solution by providing explicit information through some sort of questionnaire. We also aim to create a solution that comes with a low cognitive load for the customers in a way that we burden the customers as little as possible while providing quality size advice.

The contributions of this work are 4-fold: 1. We analyze the sizing characteristics of apparel from hundreds of brands available in fashion e-commerce context and formulate major challenges faced in building large-scale reliable size recommender systems with or without order history; 2. We demonstrate the multifaceted value of customer metadata in effectively tackling the cold-start problem by leveraging said data to build an effective cold-start recommender comparable with those state-of-the-art solutions that have privilege access to customers order data; 3. We leverage the trade-off between the size recommender effectiveness and the required customer data and propose, for the first time to our knowledge, an accurate large-scale size recommender with low customer cognitive load which has been rolled out in six European countries covering various size system conventions. 4. With experimental evidence we furthermore demonstrate how current state-of-the-art size recommendations benefit from our findings from the cold-start problem even in a hot-start setting.

2 COMPLEXITY OF THE SIZE AND FIT PROBLEM AT SCALE

In order to highlight the scale of the size recommendation problem, we analyzed fashion articles and orders in the category of Female Upper Garments which encompasses a large variety of different fashion apparel, from dresses to denim jackets, and is strongly representative of the complexity of fashion in general and of the many obstacles that arise in personalized size recommendations in particular. In Figure 1 we present a bar plot of the number of distinct apparel sizes, in the female upper garment category, from around 2000 brands available on a large-scale e-commerce fashion platform during the 2015-2019 time period. In this plot we see that when we aggregate the list of all possible sizes for all brands (composed of all the different size systems such as numeric 38-39-40, ... standard S,M,L, ..., fractions 41 1/3, 42.5, ... confection sizes 36-38, 40-42, ... country conventions EU, FR, IT, UK ...), we reach the upper bound of 17k sizes. We also see that the scale of the size recommendation and size selection problem has grown continuously

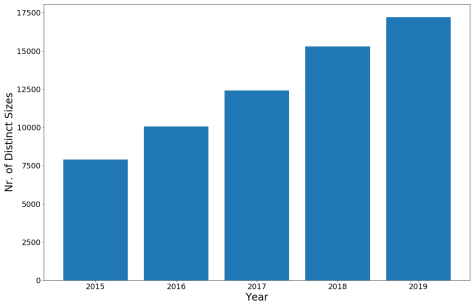

Fig. 1. Bar plot of the number of distinct apparel sizes, in women upper garment category, from hundreds of brands with different size systems and country sizing conventions during 2015 to 2019 time period. The scale of the size recommendation and selection problem continues to grow rapidly in recent years.

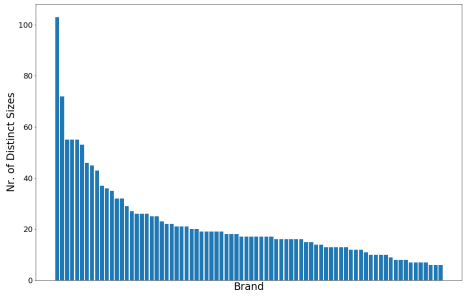

Fig. 2. Bar plot of the number of distinct sizes per brand, in women upper garment category, for 80 major brands in e-commerce fashion. Each vertical bar represents one distinct brand.

and rapidly with the number of distinct sizes more than doubling within this category over last 4 years. Diving deeper, Figure 2 represents the bar plot of the number of distinct sizes per brand for 80 popular brands. Here each vertical bar represents the number of sizes offered by one distinct brand. Brands create these distinct sizes due to multiple underlying (and often undisclosed) business and product optimization rationals [2, 18]. We can see that, already in this category, some brands offer upward of 30 different distinct sizes, leaving customers to face a challenging decision with regards to which exact size to select when shopping within those brands, and most notably what this means for them when shopping those brands with much fewer sizes to select from.

As fashion e-commerce is increasingly growing, assisting customers in buying the right size presents a huge opportunity for research in intelligent size and fit recommendation systems which can directly contribute to increasing customer satisfaction, reducing environmental footprint, and helping business profitability.

3 PRIOR WORK

Although customer-centric product recommendation is a well-studied field (see [19–23]), size recommendation is still in its infancy with only a few approaches addressing parts of this problem during the past few years [7–17, 23–26]. A large family of these approaches depend on historical data of customer orders and returns either using statistical [11–13] or deep-learning [16, 17, 27] methods, often concentrating on finding suitable embeddings to represent customers and orders [15, 24, 25]. Such approaches have the advantage of not asking customers for explicit data, thus involve a low cognitive effort from the customer. However, such solutions invariably suffer from the cold-start problem which affect thousands of existing and new customers visiting the shopping platforms everyday for which no prior order history is available in hundreds of brands and tens of fashion categories. The cold start problem has been widely studied in the context of user item recommendations [26]. Prior work has typically focused either on the article side [17], or on entering a dialogue with the customers [3–9]. The former approaches, i.e. exploiting attributes of articles, allow to alleviate the cold-start problem by using content-based filtering [23] and come with a low cognitive load by design. However, in the context of personalized size recommendations, article data does not bring sufficient information to tackle the problem. The latter approaches come with the advantage of allowing customers to become an direct and

integral part of the recommendation system but on the other side either require customers to share a considerable amount of sensitive personal data (such as age, weight, height, tummy shape, hips form, body type, favorite brand, fit preferences, etc.) [4–6] or require multiple high resolution images, videos, and in some cases 3D scans, of customers bodies in tight clothing and canonical poses [7–9]. Such imagery or 3D data is often used to create virtual-fitting solutions using recent 3D human body estimation and reconstruction approaches [28, 29]. As such, these approaches come with a high level of cognitive load for customers by demanding strong engagement and willingness from them to share images and scans of their bodies with fashion platforms.

A comprehensive comparison of these diverse approaches would be of high value to the community, but this remains out of the scope of this paper and constitutes a great future research direction. Here, in particular we focus on the the size recommendation problem in the so called cold-start scenario in which there is little to no order history for each customer. The cold-start problem in personalized size recommendation is a new emerging field and the underlying importance and necessity of requiring a diverse set of personal and body data for providing these recommendations requires deeper discussion and investigation in the fashion recommendation systems. Here we aim to investigate the size recommendation problem in this cold-start scenario, by involving customers in the loop through some sort of questionnaire and to create a solution that comes with a low cognitive load for the customers in a way that we burden the customers as little as possible while providing quality size advice.

4 SIZE RECOMMENDATION WITHOUT ORDER HISTORY

We model the problem of cold-start size recommendation as a categorical classification task, where each size is a possible class. The idea is to directly involve customers in the process by leveraging those for which we have both customer data and purchase data to learn a mapping from customer data to ordered sizes, thus allowing us to predict appropriate sizes for any new customer.

Customer Data: We use a comprehensive set of customer data gleaned from questionnaires presented to customers as part of a specialized online fashion styling service wherein customers are paired with professional stylists who then curate personalized outfits for them. These questionnaires cover a wide variety of fashion related areas, and for this study we extract from the customers' answers the subset of information related to size and fit and use it to build a feature representation for each customer. This subset consists of 20 attributes for each customer falling into three categories as can be seen in Table 1. A total of 450k questionnaires are available, each from a distinct customer which self-identified as female. The questionnaire data is projected onto an input space by calculating a vector representation for each customer. Of the 20 size-related questions on the questionnaire, 7 result in categorical variables and are one-hot encoded while the remaining 13 result in numerical variables and are normalized by mean and standard deviation.

Table 1. Features in the Questionnaire Data.

Type	Features
Overall	weight, height, age, gender
Upper body	top size, shirt collar size, shirt fit, prop. belly, top fit, prop. shoulder-waist, bust number, bust cup size, prop. shoulder-hip, blazer size
Lower body	pants size, jeans length, jeans width, prop. waist, pants waist-height , shoe size

Order data: The order data used in this work is composed of roughly 7.4 million orders placed on an e-fashion platform in the female upper body garment category. This dataset is anonymized and is not public due to various customer privacy challenges and proprietary reasons. We split these orders into training and test sets based on order timestamps. Of these 7.4 million orders, the oldest 5.6 million comprise the training set while the most recent 1.8 million form the test set. All ordered articles are associated with a numerical size in [34, 36, 38, 40, 42, 44, 46]. Both in training and testing, the target variable is the size bought by the customer. There are cases where different target values correspond to the same input vector as customers will at times buy different sizes. Nonetheless allowing the classifier to handle such ambiguity was found to be the best strategy, as opposed for example to using the median or mean size in training. Using the output of the Hot-Start recommender [13] (presented below) as a target value also proved sub-optimal.

Classifier: We experimented with a variety of multi-class classifiers and found Gradient Boosted Trees to perform best in practice. This choice comes with the added benefit of having an easily interpretable classifier, which as we shall show in the following can prove very useful. The hyper-parameters were tuned using a grid-search and cross-validation (splitting the Training Data into Train and Validation sets). In particular we found that performance saturated at 500 trees, and did not observe any over-fitting effect when growing the ensemble beyond this point. Each tree in the ensemble has a depth of 3, while the trees themselves are built, sequentially, using a learning rate of 0.01 and a sub-sampling rate of 0.5. We note that performance on the validation set was largely robust to the latter two hyper-parameters.

Hot-Start recommender baseline: The size recommender system introduced in [13] has shown to be robust and effective in a large-scale fashion e-commerce context; we thus employ it as a Hot-Start recommender baseline (built on order history data). It follows a hierarchical Bayesian approach that models jointly the probability of a size and return status (kept, too small, too big) given a customer and an article. This approach enjoys the advantages of Bayesian modeling, and in contrast with [11, 12] which have to predict the fit (good fit, too big, too small) for all possible sizes one by one, it can directly predict the probability of any size given a customer-article pair, conditionally on those articles being kept (good fit). It naturally fails in the case of new customers or customers with scarce order history. As one might expect however, the order data falls into the long tail problem, and as such, this shortcoming of current Hot-Start recommenders is quite pronounced in the context of online fashion where a large percentage of customers are in the cold-start category with none-to-scarce order history as shown in Figure 3.

Brand Size Offsets: Most brands suffer a high variance between their nominal and actual sizes caused by multiple design and business related factors, and as such the knowledge of customers' "usual" size alone is often insufficient information for providing both intra- and inter- brand size recommendations. We make use of customers return trends to gain a better understanding of various brands behaviour with respect to size and fit. Customers often tend to order, for a certain brand, their "usual" size and one size up or down, either within the same order or in a later order if they returned the first-ordered size. Therefore, using data regarding kept and returned articles, one can readily define an article offset as the difference in the sizes from the (reordered) kept articles and the ordered (but not kept) articles. An offset for a brand is then defined as the weighted average of all the article offsets in that brand. We weigh the contributions of each article by the number of sales so that the highest selling articles contribute the most to the final offset of the brand, and calculate a weighted mean (μ) and a weighted standard deviation (σ) to fit a Gaussian. We show the range of brand offsets in Figure 4, where μ and σ are calculated for 80 popular brands using 10k distinct women upper-garment purchases per brand. We highlight the importance of exploiting brand offsets in section 5 where we present our results on cold-start recommendation.

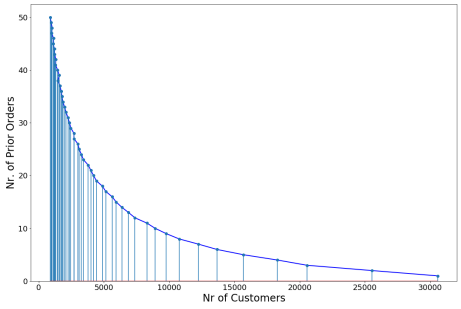

Fig. 3. Plot of the number of customers versus the number of prior orders. The long tail nature of the problem is evident where the vast majority of customers are cold-start customers with little to no order history.

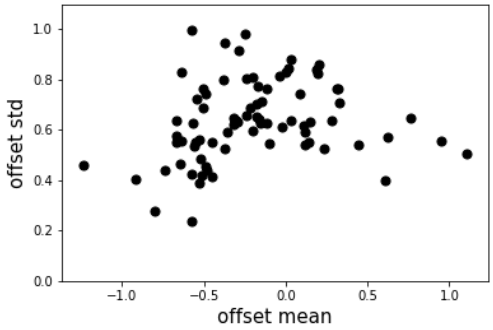

Fig. 4. Brand Size Offsets for the 80 most popular brands on the e-fashion platform. The standard deviation is plotted against the mean.

Recommender predictions: The trained classifier is combined with the brand offsets presented above resulting in our Cold-Start recommender. To obtain the final prediction of the Cold-Start recommender, the predicted class c of the trained ensemble is combined with the brand offsets by adding the brand offset mean μ_{brand} to the ensemble prediction and the final size recommendation is obtained by rounding $c + \mu_{brand}$ to the closest size. We note that we use brand offsets for the baseline Hot-Start recommender too, as they were found to be advantageous.

5 EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we evaluate the recommender systems based on the accuracy metric, defined as the percentage of times the recommender correctly predicts the size bought by the customer on the orders in the test set.

5.1 Hot-Start and Cold-Start performances

In Figure 5 we present a comparative study of the baseline Hot-Start recommender and the proposed Cold-Start recommender. We plot the accuracy of the models against the number of prior orders of each customer in the training data (as we employ temporal split of the data into training and test set, we can also speak of training and test time). As can be seen, for low number of orders regime (< 10) the Cold-Start recommender (in blue) clearly outperforms the Hot-Start one (in yellow), even in cases when a customer has a substantial amount of prior orders (> 10, < 20). We re-iterate that the Cold-Start recommender does not use any knowledge of prior orders. The Hot-Start recommender only starts outperforming the Cold-Start recommender after about 20 prior orders. This is due to the hard limitations of most current Hot-Start recommender approaches where they need a minimum set of orders per each sub-level category to perform, as has been duly noted in [13, 15], effectively restricting Hot-Start solutions to loyal customers with rich order history. Overall the performance is 58% accuracy for the Cold-Start recommender and 54% for the Hot-Start recommender (see Cold-Start (All Data) and Hot-Start (Baseline [13]) in Table 3).

In Figure 6 we present the confusion matrix of the Cold-Start recommender. We note that, even when wrong, the predictions are seldom off by more than a size. Furthermore the classifier struggles most with the more popular sizes (e.g. 40, 42). Figure 7 shows the equivalent confusion matrix for the Hot-Start recommender. The same observations

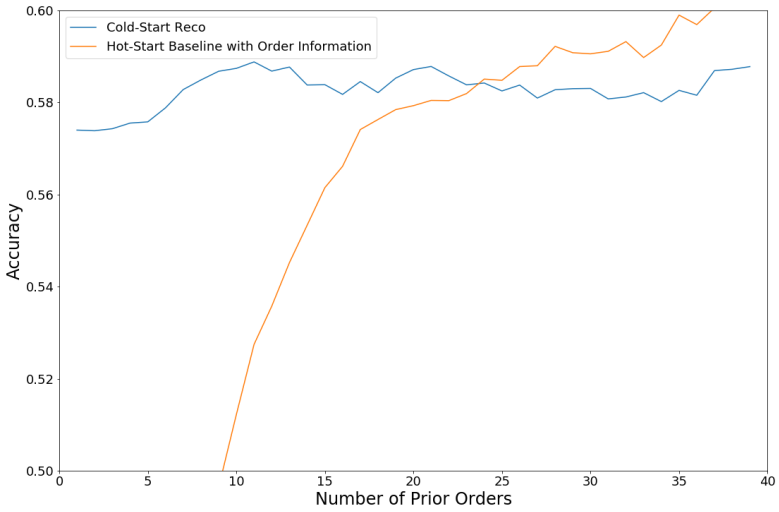

Fig. 5. Comparison of the Cold-Start recommender system presented here to the Hot-Start system presented in [13]. On the x-axis are the number of prior orders (specifically the number of prior orders of each customer in the training set). On the y-axis is the accuracy defined as the percentage of times the recommender correctly predicts the size bought by the customer on the orders in the test set.

as for the Cold-Start recommender can be made, highlighting the inherent ambiguity and complexity of the right recommendation for popular sizes.

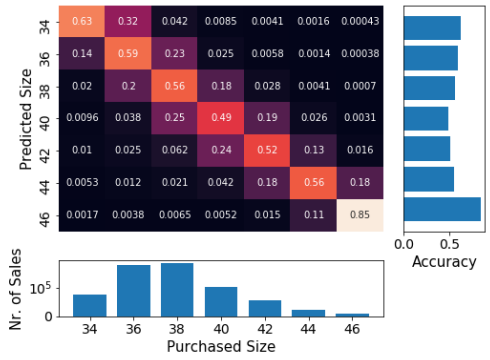

Fig. 6. Cold-Start recommender's confusion matrix, in black. The subplot below shows the distribution of sizes in sales, and the one on the right hand side shows the accuracy of the Cold-Start recommender per size.

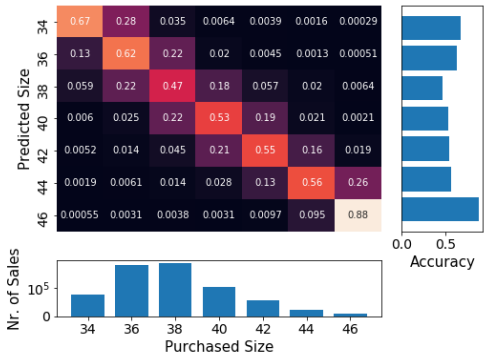

Fig. 7. Hot-Start recommender's confusion matrix, in black. The subplot below shows the distribution of sizes in sales, and the one on the right hand side shows the accuracy of the Hot-Start recommender per size.

5.2 Impact of Brand Size Offsets

As noted, the proposed Cold-Start recommender makes great use of the brand offsets. To highlight the contribution of these offsets in the performance of the recommender, we present in Figure 8 the accuracy of a Cold-Start recommender based only on the ensemble method which does not exploit the brand offsets (Cold-Start Reco Without Brand Offsets), and the accuracy of the proposed Cold-Start recommender which adds the brand offset's mean to the ensemble output. We plot these accuracies relative to a lower threshold on the standard deviation, whereby for a given threshold θ we only include those brands which have a standard deviation above θ to calculate the accuracy. As can be seen there is a clear advantage to exploiting brand offsets when making recommendations.

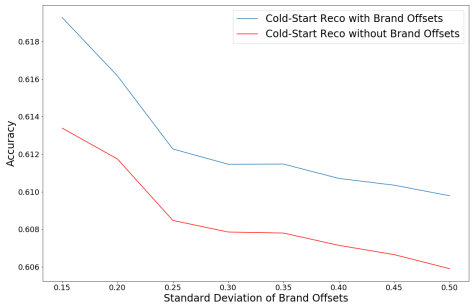

Fig. 8. Accuracy of the Cold-Start recommender depending on whether brand offsets are exploited to refine size recommendations. The results show the benefits of leveraging this information in all cases, irrespective of the brand offset's standard deviation.

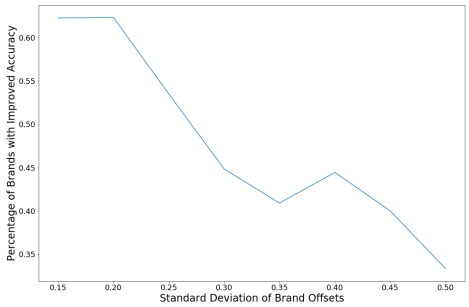

Fig. 9. Percentage of brands for which adding the brand offset's mean leads to improved accuracy plotted against a lower threshold θ on the standard deviation of the brand offsets.

In Figure 9 we present the percentage of brands for which we observe improved accuracy when adding the brand offset means relative to the standard deviation of the offsets. Again we plot against a lower threshold on the brand standard deviations. As can be seen in the case of brands with an offset standard deviation of at least 0.15 adding the brand offset means results in improved accuracy in approximately 66% of cases. As expected, the positive effect of using brand offsets diminishes as the standard deviation rises. As the standard deviation reaches 0.5 the brand offsets means lead to improved performance only in approximately 33% of cases. To address this limitation, a future work direction is to directly use the article offsets for brands suffering from such high standard deviations.

5.3 Customer coverage

In order to get a better understanding of the percentage of customers covered by various order segments[1] we plot in Figure 10 the accuracy of the recommender systems relative to these percentages. By taking all customers we achieve a customer coverage of 100% although in this case the Hot-Start Reco performs quite poorly as can be seen in the plots. On the other hand taking only those customers who have a large number of prior orders results in a Hot-Start recommender that outperforms the Cold-Start recommender but at the cost of having a low customer coverage.

[1]The orders in the test set are segmented based on the number of prior orders in the training set of the corresponding customer.

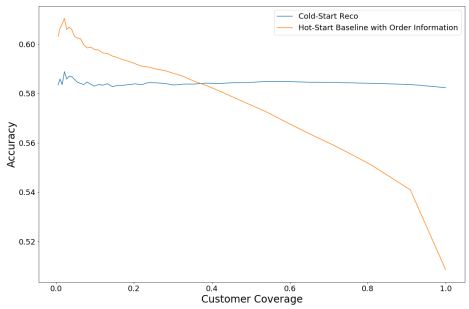

Fig. 10. Comparison of the Cold- and Hot-Start recommender systems relative to the percentage of customers covered (customer coverage).

Fig. 11. Comparison of the Cold- and Hot-Start recommender systems against the confidence level of the Hot-Start system.

5.4 Hybrid Recommendation Systems

These numbers make a strong case for the incorporation of the Cold-Start recommender system for customers with limited order history. Based on these results, a hybrid system can be envisioned with each recommender providing a recommendation for the customers it performs best on. One simple yet effective strategy would switch between recommenders according to the size of the customer's order history. A more elaborated strategy we experimented with is to use the Hot-Start recommender's confidence as a hyper parameter, switching to the Cold-Start recommender whenever the Hot-Start recommender has relatively low confidence in its predictions. Figure 11 presents the accuracy of the two recommenders plotted against the confidence of the Hot-Start recommender system. To achieve this we segment orders according to the confidence of the Hot-start recommender and plot the performance of both systems on those segments. The Hot-Start recommender outperforms its Cold-Start counterpart in those cases where it is very confident, on the contrary when it exhibits low confidence (e.g. < 0.5) the Cold-Start recommender proves to be more reliable. Note that this confidence-based hybrid approach, with an overall performance accuracy of 59%, is more effective than using the size of the order history (overall accuracy of 58%, see Hybrid (Orders) and Hybrid (Confidence) in Table 3).

5.5 Minimizing Customers' Cognitive Load

A crucial aspect of any user-in-the-loop recommendation system is the amount of cognitive load it burdens the customers with. Thus beyond the performance of the system with respect to accuracy, in practice it is of major importance to minimize the customer's cognitive load when they interact with the platform. The cold-start recommender presented in the previous section makes use of 20 explicit customer data points such as weight, height, etc. which is in reality too high. In what follows, we deep dive and investigate what performance can be obtained using a small subset of attributes towards providing a low cognitive load and critically reducing the intimate data requirements from the customers on their body shapes, etc. As mentioned in section 4, one of the added benefits of using Gradient Boosted Trees is that they result in an interpretable model, which allows us to estimate the Gini importance of each individual feature in the resulting ensemble. In turn, the attributes that come out as key are Top Size, Weight and Height.

Given the importance assigned to the Top Size attribute by the ensemble, the obvious question that arises is whether the customer provided top size would suffice to predict the size bought by the customer themselves in any future orders.

We therefore cross validated this explicit customer information with the sizes a customer actually buys on the fashion platform. We found that in fact customers only buy the size they provided in the questionnaire in roughly 57% of all orders. This observation runs counter to the intuition that customers should be good predictors of their own sizes and highlights one of the many complexities of the size recommendation problem. As customers themselves are only 57% likely to order in their provided sizes, this leaves a remaining 43% of orders where customers are unsure of what size to order and would require accurate support in the form of a size advice.

Figure 12 shows that using solely the size provided by the customer leads, as discussed above, to an under-performing recommender system (marked "Top Size" in the plot). We added a minimum information to it, in particular the brand information, which in turn enables us to use the brand offsets. As can be seen this results in a significantly better recommender system (marked "Top Size + Brand" in the plot) and highlights the importance of exploiting brand offsets when making a recommendation. As expected, the Cold-Start recommender system which has access to the full questionnaire outperforms both these systems.

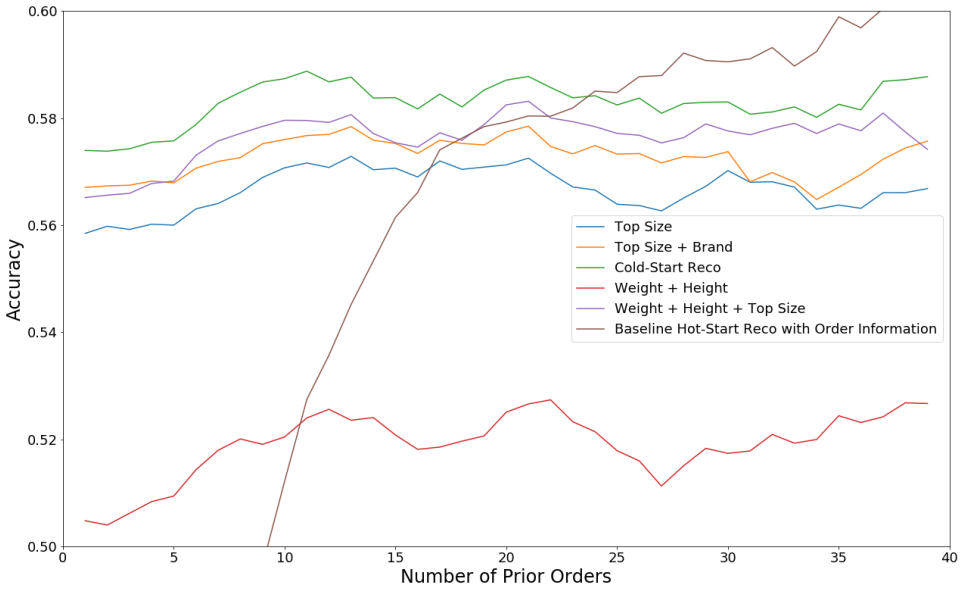

Fig. 12. Comparison of the Cold-Start and the Hot-Start systems. On the x-axis are the number of prior orders (specifically the number of prior orders of each customer in the training set)

Figure 12 also shows the performance of other flavors of the cold-start recommender system using the identified top three attributes: "Weight + Height + Top Size" in purple and "Weight + Height" in red. It is evident that asking only for Weight and Height is not enough to reliably provide a size recommendation. However, we do note that using Weight, Height, and Top Size performs closer to the full Cold-Start recommender than other flavors, and achieves an overall accuracy of 57% instead of 58% (see Cold-Start (W+H+TS) and Cold-Start (All Data) in Table 3).

While "Weight + Height + Top Size" could constitute a good trade-off between high recommender accuracy and low cognitive effort on the side of the customer, however it comes with an underwhelming customer experience which involves having to provide two intimate and privacy sensitive questions on an e-commerce shopping platform. On the other hand "Top Size + Brand" performs closely to that of "Weight + Height + Top Size". This not only highlights the importance of exploiting brand offsets when making a recommendation, but it also comes with the great advantage of not requiring intimate body data from customers. Instead customers are simply asked for their top size in one of their favourite brands. We consider this approach to be the best trade-off between accuracy and customer experience.

5.6 Performance in Production

Following the experiments shown in previous sections and given the encouraging results, we have rolled out our Cold-Start recommender to production in January 2020 on a large e-commerce fashion platform for the Adult Upper Garments category (both Men's and Women's categories). The model is currently live in six countries; Germany, Austria, Switzerland, Netherlands, Belgium, Sweden, serving approximately 3000 orders per day with an overall accuracy of 63.90%. As can be seen in Table 2 accuracy on a per country basis can vary greatly, potentially highlighting the cultural aspect of the size and fit problem which further complicates an already complex problem. This provides an exciting dimension for future work.

Table 2. Production results in different countries.

Country	Germany	Austria	Switzerland	Netherlands	Belgium	Sweden	Overall
Accuracy	58.36%	57.66%	56.56%	69.50%	68.71%	66.97%	63.90%

5.7 Leveraging Customer Data for Hot-Start Recommendation

Hot-Start recommender systems, even in the presence of rich order histories, struggle to provide highly accurate size advice. We investigated whether customer data could also be helpful in this case. The Hot-Start baseline [13] cannot readily ingest customer data so we adapted a state-of-the-art Hot-Start deep learning recommender recently proposed in [27] (denoted MetalSF), and show that our findings transfer well to MetalSF. Naturally, customer data helps in the absence of prior purchases, where the accuracy goes from 29% when using the most popular size as prediction to 56% with customer data, on par with our Cold-Start recommender. More interestingly, we plot in Figure 13 the accuracy against the number of prior orders for various input data: only order history / no customer data (in blue), order history and customer data (in yellow) and order history + Weight + Height + Top Size (in green). Customer data helps significantly up to 10 prior purchases and marginally after 20. Additionally, restricting the customer data to Weight, Height and Top Size has no impact on the performance, indicating that these variables are indeed sufficient to significantly enhance the customers' experience, even where prior purchases are available.

5.8 Summary

The overall performance of the models discussed in the study can be seen in Table 3. Cold-Reco (All Data) refers to the cold-start algorithm that uses all the customer data available and Hot-Reco (Baseline [13]) to the Bayesian Hot-Start recommender baseline. Cold-Reco (W+H+TS) refers to the cold-start algorithm using Weight Height and Top Size only, Hybrid (orders) to the Hybrid recommender based on the number of prior orders, Hybrid (Confidence) to the Hybrid recommender based on the confidence of the Hot-Start recommender. Finally, MetalSF (Original [27]) refers to MetalSF

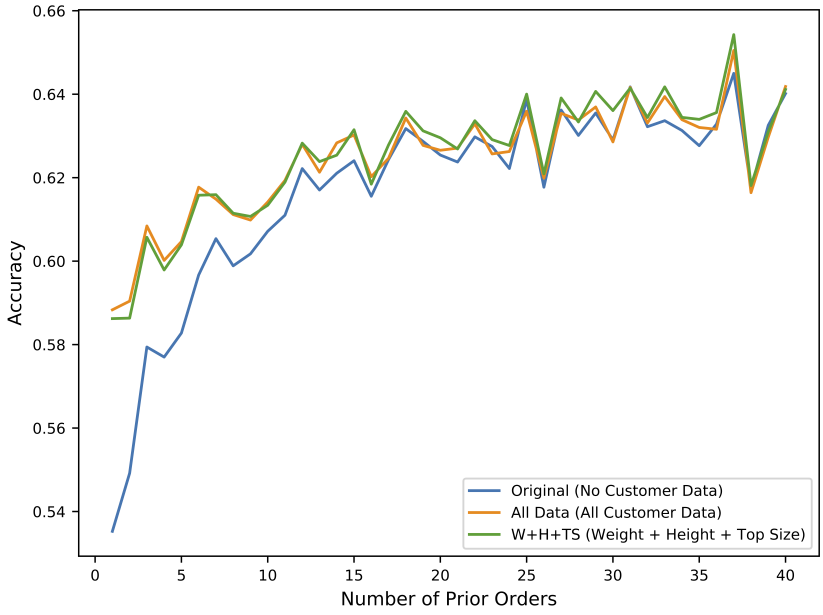

Fig. 13. Accuracy of MetalSF (Original [27]), MetalSF (All Data), and MetalSF (W+H+TS) based on the customer predicted size against the number of prior orders in the training set.

without customer data, MetalSF (All Data) to MetalSF with all customer features, and MetalSF (W+H+TS) to the use of Weight, Height, and Top Size.

Table 3. Results of various approaches studied

Reco	Cold-Reco (All Data)	Hot-Reco (Baseline [13])	Cold-Reco (W+H+TS)	Hybrid (Orders)	Hybrid (Confidence)	MetalSF (Original [27])	MetalSF (All Data)	MetalSF (W+H+TS)
Accuracy	57.89%	53.64%	57.15%	58.41%	59.27%	57.51%	61.45%	61.49%

6 CONCLUSION

We addressed a major challenge for the online fashion industry, that of user-in-the-loop personalized size recommendations with low cognitive load. With the aim of building an accurate recommender system that requires only a minimum set of explicit customer data, we further investigated 20 different customer attributes such as weight, height, top size, tummy shape, etc. for the task at hand. We experimented with different versions of the cold-start recommender system and benchmarked them against the state-of-the-art recommender systems with privileged access to rich customer order history. We presented a deep dive on the trade-off between a recommender's performance and a customer's cognitive load, and proposed a solution capable of providing accurate size advice for thousands of new and existing customers with bare minimum customer data needs. Finally we presented our results in a production environment covering six European countries and demonstrated that our approach scales up to large-scale production requirements, performs in practice at the level predicted by the experiments presented here, and to the level of industrial requirements.

REFERENCES

[1] Size charts. https://www.adidas.com.sg/help-topics-size_charts.html. Accessed: 2020-01-13.

[2] One size fits none. https://time.com/how-to-fix-vanity-sizing. Accessed: 2020-01-28.

[3] Ying Yuan and Jun-Ho Huh. Cloth size coding and size recommendation system applicable for personal size automatic extraction and cloth shopping mall. In *MUE/FutureTech 2018*, pages 725–731.

[4] THE ICONIC. https://www.theiconic.com. The customer based size recommendations are accessible on product detail pages. Accessed: 2019-11.

[5] ASOS. https://www.asos.com. The customer based size recommendations are accessible on product detail pages. Accessed: 2019-11.

[6] ABOUT YOU. https://corporate.aboutyou.de/en/. The customer based size recommendations are accessible on product detail pages. Accessed: 2019-11.

[7] Nadia Thalmann, Bart Kevelham, Pascal Volino, Mustafa Kasap, and Etienne Lyard. 3d web-based virtual try on of physically simulated clothes. *Computer-Aided Design and Applications*, 8, 01 2011.

[8] J. Surville and Thierry Moncoutie. 3d virtual try-on: The avatar at center stage. In *Proceedings of 4th International Conference on 3D Body Scanning Technologies*, 2013.

[9] Fanke Peng and Mouhannad Al-Sayegh. Personalised size recommendation for online fashion. In *Proceedings of the 6th International Conference on Mass Customization and Personalization in Central Europe*, 2014.

[10] Monika Januszkiewicz, Christopher Parker, Steven Hayes, and Simeon Gill. Online virtual fit is not yet fit for purpose: An analysis of fashion e-commerce interfaces. In *Proceedings of the 8th International Conference and Exhibition on 3D Body Scanning and Processing Technologies*, pages 210–217, 10 2017.

[11] Vivek Sembium, Rajeev Rastogi, Atul Saroop, and Srujana Merugu. Recommending product sizes to customers. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 2017.

[12] Vivek Sembium, Rajeev Rastogi, Lavanya Tekumalla, and Atul Saroop. Bayesian models for product size recommendations. In *Proceedings of the 2018 World Wide Web Conference*, 2018.

[13] Romain Guigourès, Yuen King Ho, Evgenii Koriagin, Abdul-Saboor Sheikh, Urs Bergmann, and Reza Shirvany. A hierarchical bayesian model for size recommendation in fashion. In *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, 2018.

[14] G. Mohammed Abdulla and Sumit Borar. Size recommendation system for fashion e-commerce. In *KDD Workshop on Machine Learning Meets Fashion*, 2017.

[15] Kallirroi Dogani, Matteo Tomassetti, Saúl Vargas, Benjamin Paul Chamberlain, and Sofie De Cnudde. Learning embeddings for product size recommendations. In *Proceedings of the SIGIR 2019 Workshop on eCommerce*, 2019.

[16] Abdul-Saboor Sheikh, Romain Guigourès, Evgenii Koriagin, Yuen King Ho, Reza Shirvany, Roland Vollgraf, and Urs Bergmann. A deep learning system for predicting size and fit in fashion e-commerce. In *Proceedings of the 13th ACM Conference on Recommender Systems*. ACM, 2019.

[17] Nour Karessli, Romain Guigourès, and Reza Shirvany. Sizenet: Weakly supervised learning of visual size and fit in fashion images. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019*.

[18] N Weidner. *Vanity sizing, body image, and purchase behavior: A closer look at the effects of inaccurate garment labeling*. PhD thesis, Eastern Michigan University, 2010.

[19] Yue Shi, Martha Larson, and Alan Hanjalic. Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. *ACM Computing Surveys*, 47(1):3:1–3:45.

[20] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys*, 52(1), February 2019.

[21] Xing Yi, Liangjie Hong, Erheng Zhong, Nanthan Nan Liu, and Suju Rajan. Beyond clicks: Dwell time for personalization. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 113–120, New York, NY, USA, 2014. ACM.

[22] Rose Catherine and William Cohen. Personalized recommendations using knowledge graphs: A probabilistic logic programming approach. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, pages 325–332, New York, NY, USA, 2016. ACM.

[23] Michael J Pazzani and Daniel Billsus. Content-based recommendation systems. In *The Adaptive Web*, pages 325–341. Springer, 2007.

[24] Rishabh Misra, Mengting Wan, and Julian McAuley. Decomposing fit semantics for product size recommendation in metric spaces. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, pages 422–426. ACM, 2018.

[25] Loveperteek Singh, Shreya Singh, Sagar Arora, and Sumit Borar. One embedding to do them all. *CoRR*, abs/1906.12120, 2019.

[26] Andrew I Schein, Alexandrin Popescul, Lyle H Ungar, and David M Pennock. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 253–260. ACM, 2002.

[27] Julia Lasserre, Abdul-Saboor Sheikh, Evgenii Koriagin, Urs Bergmann, Roland Vollgraf, and Reza Shirvany. Meta-learning for size and fit recommendation in fashion. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pages 55–63, 01 2020.

[28] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[29] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science. Springer International Publishing, October 2016.