Assessing Fashion Recommendations: A Multifaceted Offline Evaluation Approach

Jake Sherman
True Fit
Boston, MA
jsherman@truefit.com

Chinmay Shukla
True Fit
Boston, MA
cshukla@truefit.com

Rhonda Textor
True Fit
Boston, MA
rtextor@truefit.com

Su Zhang
True Fit
Boston, MA
szhang@truefit.com

Amy A. Winecoff
True Fit
Boston, MA
awinecoff@truefit.com

ABSTRACT

Fashion is a unique domain for developing recommender systems (RS). Personalization is critical to fashion users. As a result, highly accurate recommendations are not sufficient unless they are also specific to users. Moreover, fashion data is characterized by a large majority of new users, so a recommendation strategy that performs well only for users with prior interaction history is a poor fit to the fashion problem. Critical to addressing these issues in fashion recommendation is an evaluation strategy that: 1) includes multiple metrics that are relevant to fashion, and 2) is performed within segments of users with different interaction histories. Here, we present our multifaceted offline strategy for evaluating fashion RS. Using our proposed evaluation methodology, we compare the performance of three different algorithms, a most popular (MP) items strategy, a collaborative filtering (CF) strategy, and a content-based (CB) strategy. We demonstrate that only by considering the performance of these algorithms across multiple metrics and user segments can we determine the extent to which each algorithm is likely to fulfill fashion users' needs.

CCS CONCEPTS

• **Applied computing** → **Online shopping**; • **Computing methodologies** → *Ranking*.

KEYWORDS

fashion, recommender systems, evaluation, personalization

ACM Reference Format:
Jake Sherman, Chinmay Shukla, Rhonda Textor, Su Zhang, and Amy A. Winecoff. 2019. Assessing Fashion Recommendations: A Multifaceted Offline Evaluation Approach. In *Proceedings of Workshop on Recommender Systems in Fashion, 13th ACM Conference on Recommender Systems (recsysXfashion'19).* ACM, New York, NY, USA, 7 pages.

recsysXfashion'19, September 20, 2019, Copenhagen, Denmark

1 INTRODUCTION

Few industries touch the lives and identities of consumers as intimately as fashion. Everyone makes decisions about what to wear, and these decisions reflect not only the prevailing cultural norms [19], but also the individual identity of the wearer [3, 16]. Clothing can affect how the wearer feels and behaves as well as how others feel and behave in response to the wearer [13]. The motivations that drive fashion consumers include fashionability, individualization, self assurance, flaw minimization, and comfort, yet the importance of these motivations to fashion purchasing depends on the characteristics of the consumer [20]. That is, consumers' motivations for buying clothing are personal [21]. Consequently, any recommender system (RS) developed for the fashion domain should be evaluated on criteria that are specifically relevant to the needs of fashion consumers.

As with other domains, fashion users should receive "accurate" recommendations (i.e., recommendations that are relevant), but accuracy alone is not sufficient. Although some researchers have been urging RS developers to think beyond accuracy for more than a decade (e.g., [24]), accuracy is still the predominant focus of most RS evaluations. A survey of recent papers at the ACM RecSys Conference noted that while roughly 85% of papers used some form of offline accuracy metric, a mere 20% included a measure of diversity, novelty, or another alternative metric [12]. A dogged focus on maximizing accuracy can unintentionally degrade the end user experience. McNee, Riedl, and Konstan [14] provide an illustrative example: a travel RS that recommends solely locations that users have previously visited would perform better on most accuracy metrics than a system that recommends novel travel destinations that are more interesting to the user. Thus, non-accuracy based evaluation measures are necessary for properly evaluating RS in general, and fashion RS specifically.

Personalization is critical for good fashion recommendations. In marketing, "personalization" is operationalized as the customization of goods and/or services to meet the needs of specific consumers [8]. Different fashion consumers have different motivations for purchasing fashion items [20]. As a result, one way to evaluate the personalization of fashion recommendations is by measuring how much recommendation lists vary from user to user (e.g., the approach in [23]). In addition to list diversity, understanding the popularity bias in recommendations is important for evaluating

personalization because recommendations dominated by popular items are necessarily depersonalized.

We also want to understand how well RS perform within multiple user segments. Providing accurate and engaging recommendations is easier for users with rich interaction histories. However, within our own fashion datasets, few users have prior sales or even item views. That is, most users are new. Because personalization is so critical to fashion, approaches such as collaborative filtering (CF) that cannot provide recommendations for new users are unlikely to provide a good experience to most fashion shoppers. Consequently, we must assess how fashion RS perform with new as well as established users.

The goal of our current work is to develop a methodology for more comprehensively evaluating the extent to which fashion RS produce quality recommendations. To provide an understanding of how different types of RS perform, we evaluate three algorithms: 1) a most popular (MP) items strategy, 2) a CF-based strategy, and 3) a content-based (CB) strategy. Our evaluation method consists of multiple measures of recommendation quality performed on multiple user segments.

2 RELATED WORKS

Several approaches have been developed to address the unique demands of providing fashion recommendations. Despite the prominence of the cold-start problem inherent in fashion data, some fashion recommendation approaches have nevertheless relied on CF. Hwangbo and colleagues [11] developed a novel user-based CF approach for recommending complementary and substitute fashion items that offers interesting algorithmic ideas, but is limited in that 1) only existing products are accommodated, and 2) recommendations are not personalized to users. Rue La La, a flash sale fashion retailer, developed a latent factors CF approach for providing fashion recommendations that overcomes the cold-start problem for items by recommending product groups to users instead of individual products [9]. Although this approach *does* address the cold start problem for items, it *does not* allow them to make personalized recommendations for new users.

Other methodologies for providing fashion recommendations eschew CF in favor of models that circumvent the cold start problem by leveraging user and/or product attributes to make recommendations. De Melo, Nogueira, and Guliato [5] developed a content based (CB) fashion RS that constructs detailed clothing item attributes to build content profiles for each user and then uses k-nearest neighbors to make recommendations for new items. Although this approach overcomes the item cold-start problem, it cannot provide personalized recommendations for new users. A RS developed by Zalando [7] leverages both user and product attributes within a learning to rank (L2R) framework, allowing them to make recommendations for both new items and users. However, Zalando only measured the accuracy of recommendations and did not evaluate how the system performed within different user segments.

3 APPROACH

3.1 Evaluation

3.1.1 Data Selection. We performed separate evaluations on three different retailers. Within retailers, we trained models for women's dresses. We constructed each dataset by taking all user-product interactions that occurred in a one year period and split our data into training and test sets by allocating the first eight months to the training data and the remaining four months to the test data. Descriptive statistics for the training and test data are in Table 1 and Table 2, respectively. In both the training and test data for all retailers, the overwhelming majority of observations in the user-item matrix are unobserved (i.e., the users did not view or buy the item). Our three retailers are also similar in terms of the distribution of users across our three user segments (see Section 3.1.2).

3.1.2 User Segmentation. One of our goals was to understand how RS would perform in user segments with different product interaction histories. Many of our users have no prior interaction history. Therefore, we define "new users" as users who have no sales or views in the training data. Some users have viewed items in the training data, but never made a purchase. We consider these users "view users" (i.e., users with views in the training data, but no sales. Lastly, a minority of users have a prior purchase falling within the training data. We consider these users "sale users." We note that these user distinctions are based on the training data only. We perform our evaluations within each of these user segments as well as across all user segments to gain insight into the recommendation experience for different types of users.

3.1.3 Accuracy. Because we are primarily interested in how well RS perform at ranking items, we focus our evaluation on top-n performance [22]. To assess model accuracy, we use a modified version of normalized discounted cumulative gain (NDCG) at k. $NDCG_k$ is a normalized version of the discounted cumulative gain (DCG_k) metric, which is computed for a particular user as:

$$DCG_k = \sum_{i=1}^{k} \frac{rel_i}{log_2(i+1)} \tag{1}$$

where rel_i is the relevance label for the i^{th} item recommended to a user. $NDCG_k$ normalizes the DCG_k by dividing it by the ideal DCG_k, or the DCG_k that would be achieved by a perfect ranking. One of the limitations of typical $NDCG_k$ implementations is that if predictions are tied, the $NDCG_k$ value can be non-deterministic since the gain for tied items will be based on arbitrary ordering. To mitigate this problem, we implement the tie-aware $NDCG_k$ approach proposed in [15]. We set k to 10 as users typically see about 10 recommendations, and micro-average each user's $NDCG_k$ value together to report an aggregated value. In addition to reporting raw $NDCG_k$ values, we also report the percentage change between our $NDCG_k$ values and the $NDCG_k$ value that would result from a random ranking ($\%\Delta_r$) of the items since $NDCG_k$ will vary based on the number of items in data.

3.1.4 Personalization. To date, there is little consensus as to how best to measure recommendation diversity and personalization directly. We leverage two indirect measures that speak to diversity and personalization: an inter-user average distinct recommendations at k metric and also a relative popularity metric.

In order to measure inter-user diversity, we use the AD_k (average distinct at k) metric, which is defined as:

Table 1: Descriptive Statistics for Training Data

	Users	Products	Sales (%)	Views (%)	Unobserved (%)
Retailer 1	39,307	376	7,461(0.05%)	103,829(0.7%)	14,668,142(99.2%)
Retailer 2	42,490	865	8,276(0.02%)	143,781(0.4%)	36,601,793(99.6%)
Retailer 3	60,333	386	21,904(0.1%)	141,320(0.6%)	23,125,314(99.3%)

Table 2: Descriptive Statistics for Test Data

	New Users (%)	View Users (%)	Sale Users (%)	Products	Sales (%)	Views (%)	Unobserved (%)
Retailer 1	2,477(73.8%)	667(19.9%)	213(6.3%)	319	1,727(0.2%)	8,850(0.8%)	1,060,306(99.0%)
Retailer 2	6,048(69.0%)	1,997(22.8%)	720(8.2%)	676	5,171(0.1%)	50,443(0.9%)	5,869,526(99.1%)
Retailer 3	5,164(71.2%)	1,513(20.9%)	578(7.9%)	314	2,753(0.1%)	19,578(0.9%)	2,255,739(99.0%)

$$AD_k = \frac{1}{\frac{1}{2}(U^2 - U)} \cdot \sum_{i=1}^{U} \sum_{j=i+1}^{U} AD_{k,i,j} \tag{2}$$

where U is the total number of users, and $AD_{k,i,j}$, or the distinctness between a single pair of users, is defined as:

$$AD_{k,i,j} = |L_{k,i} \triangle L_{k,j}| = |(L_{k,i} - L_{k,j}) \cup (L_{k,j} - L_{k,i})| \tag{3}$$

where $L_{k,i}$ is the set of top-k recommended items for user i, and $L_{k,j}$ is is the set of top-k recommended items for user j. $AD_{k,i,j}$ measures the cardinality of the symmetric difference between two different users' top-k recommendations. In the case where two users' top-k recommendations are exactly the same, the value of $AD_{k,i,j}$ will be zero. When they have no items in common, the value will be $2k$. In order to avoid the $O(U^2)$ complexity associated with computing AD_k across the entire population of user pairs, we randomly sample the proportion $\frac{2}{U-1}$ of user pairs from the population of $\frac{1}{2}(U^2 - U)$ user pairs in order to create a randomly sampled set of user pairs. Then, we redefine AD_k as:

$$AD_k = \frac{2}{U-1} \cdot \sum_{i=1}^{U} \sum_{j=i+1}^{U} AD_{k,i,j} \cdot I_{i,j} \tag{4}$$

where $I_{i,j}$ is an indicator variable that takes a value of 1 when the pair of users (i, j) is in the randomly sampled set of user pairs, and a value of 0 otherwise.

In order to measure relative popularity, we use the RP_k (relative popularity at k) metric to quantify the popularity of users' top-k recommendations relative to recommending the most popular k items. RP_k is defined as:

$$RP_k = \frac{1}{U} \cdot \sum_{u=1}^{U} RP_{k,u} \tag{5}$$

where $RP_{k,u}$, or the relative popularity at k for a single user, is defined as:

$$RP_{k,u} = \frac{\sum_{i=1}^{k} Q_{u,i}}{\sum_{i=1}^{k} Q_i} \tag{6}$$

where $Q_{u,i}$ is the quantity sold of the top-i^{th} recommendation for user u, and Q_i is the quantity sold of the i^{th} most popular product across all users. In the scenario where the top-k most popular products are being recommended to all users, $Q_{u,i}$ becomes Q_i, resulting in a RP_k value of one, its upper bound.

3.2 Recommendation Algorithms

3.2.1 MP Recommendations. By definition, items are popular if they have broad appeal across a wide swath of users. Therefore, we might expect that by recommending popular items, we can achieve high levels of accuracy [4]. We determine which items are most popular based on sales. Specifically, we sum the total number of units sold for each item within our sales data on a retailer by retailer basis, and then recommend the top-k items with the highest quantity of units sold. This MP recommendation strategy serves as a baseline algorithm that gives depersonalized but broadly palatable recommendations.

3.2.2 CF Recommendations. Because some fashion RS use CF, we also include a CF-based recommendation algorithm. For our fashion items, we do not have explicit ratings of user preferences for products (e.g., a star rating of 1 to 5). Instead we must rely on implicit proxies for user preferences, in our case, product sales and views. In contrast to traditional item-based (e.g., [18]) and user-based CF (e.g., [17]), alternating least squares (ALS) is a matrix factorization CF strategy developed specifically for implicit feedback datasets [10]. ALS allows user preference to be separated from confidence in user preference, which is useful for implicit datasets since indirect measures of user preferences are inherently noisy. Because sales can be considered a stronger signal of user preference than views, we weight sales more heavily in our model. We treat views as a binary for whether or not an item was viewed since multiple views may or may not indicate increased user preference. With our trained ALS model, we make predictions for all user-item combinations, and recommend the top-k items with the highest predicted values.

3.2.3 CB Recommendations. Many fashion RS use product and/or user attributes to give recommendations, so we also include a CB approach that leverages information about users and products. This CB recommendation strategy consists of two major phases. In the

first phase, we fit an ALS CF to user sales and views and make predictions for user preferences. We then use these predictions to augment our original user-item interaction data. Specifically, if the user-item interaction was observed (i.e., was either viewed or sold), we retain the value of the original user-item interaction. If the user-item interaction was not observed, we substitute the value predicted by ALS. We then train a random forest model using the augmented outcomes as labels and information about users and products as features. For product features, we represent fashion details such as style attributes (e.g., dress shape, sleeve length) and price. For user features, we use fashion-relevant information about users such as body mass index (BMI), user age, and brand preferences. We train models separately for different retailers since the relationships between user and product features can be assumed to vary by retailer. Using the trained RF model, we recommend the top-k items with the highest predicted values.

4 RESULTS

Results across retailers are presented in Tables 3-5. To help illustrate the evaluation metrics and user segmentation, we provide depictions of results for Retailer 1 in Figures 1-3.

4.1 Accuracy

Table 3: $NDCG_{10}$ Evaluation

	Retailer 1($\%\Delta_r$)	Retailer 2($\%\Delta_r$)	Retailer 3($\%\Delta_r$)
Sale Users			
MP	0.077(340.2%)	0.025(253.8%)	0.122(703.4%)
CF	0.032(85.4%)	0.023(222.7%)	0.094(517.0%)
CB	0.046(164.4%)	0.021(194.1%)	0.108(608.5%)
View Users			
MP	0.065(269.6%)	0.031(332.6%)	0.110(623.1%)
CF	0.030(69.9%)	0.024(236.7%)	0.078(415.5%)
CB	0.033(90.2%)	0.025(258.6%)	0.115(657.2%)
New Users			
MP	0.136(681.6%)	0.025(259.6%)	0.094(519.1%)
CF	-	-	-
CB	0.025(46.0%)	0.012(68.6%)	0.089(485.8%)
Average			
MP	0.126(620.3%)	0.026(259.8%)	0.102(569.2%)
CF	-	-	-
CB	0.029(65.8%)	0.014(95.5%)	0.094(521.1%)

Note: CF cannot make predictions for new users.

$\%\Delta_r$ is % improvement over random.

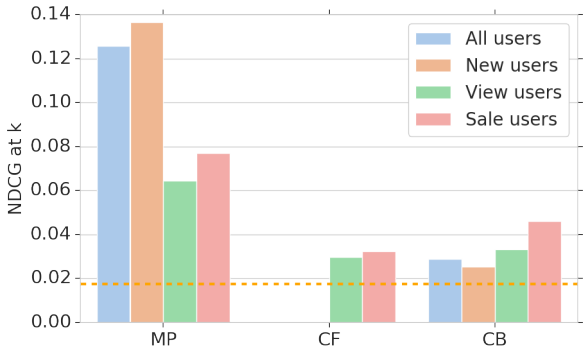

Figure 1: $NDCG_{10}$ for Retailer 1. The yellow dotted line corresponds to the $NDCG_{10}$ value for Retailer 1 that would result from a random ranking of the items.

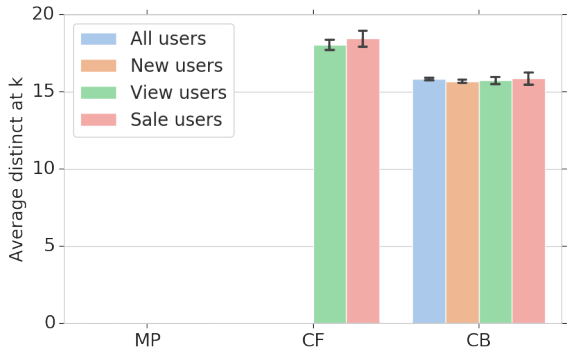

Figure 2: AD_{10} for Retailer 1. Each error bar represents the 95% confidence interval of the distribution of 1,000 bootstrap samples of $AD_{k,i,j}$ values. The MP recommendation strategy produces the same recommendations for all users, resulting in values of 0.

For $NDCG_{10}$, the MP recommendation strategy outperformed the CF and CB strategies. In general, CB recommendations outperformed CF recommendations on $NDCG_{10}$, with sale users in Retailer 2 being the only exception. Algorithm performance between user segments was dependent on retailer. For example, across all retailers, for CB recommendations, $NDCG_{10}$ was lowest for new users, but in Retailer 1, $NDCG_{10}$ for MP recommendations was higher for new than both view and sale users. Overall, although MP was generally more accurate than CB, and CB was generally more accurate than CF, a finer grained analysis by user type and retailer revealed a more complex pattern of results.

4.2 Personalization

For AD_{10}, CF provided more distinctive recommendations than CB across all three retailers for the view and sale user segments. Meanwhile, MP provided the same popular recommendations to all users, resulting in AD_{10} values of 0. Within each retailer/model combination, AD_{10} exhibited very little variance across user segments. Figure 2 shows the increased distinctiveness of CF over CB and the low within-retailer/model variance in AD_{10} across user segments for Retailer 1.

Table 4: AD_{10} Evaluation

	Retailer 1(SD)	Retailer 2(SD)	Retailer 3(SD)
Sale Users			
MP	0(0)	0(0)	0(0)
CF	18.5(3.7)	18.3(3.9)	16.7(4.1)
CB	15.9(2.8)	18.0(2.3)	12.9(3.4)
View Users			
MP	0(0)	0(0)	0(0)
CF	18.0(4.3)	18.2(4.1)	16.5(4.2)
CB	15.7(2.9)	18.0(2.1)	12.6(3.5)
New Users			
MP	0(0)	0(0)	0(0)
CF	-	-	-
CB	15.7(2.8)	18.1(2.1)	12.5(3.4)
Average			
MP	0(0)	0(0)	0(0)
CF	-	-	-
CB	15.8(2.8)	18.1(2.1)	12.4(3.4)

SD is the standard deviation between user pairs.

Table 5: RP_{10} Evaluation

	Retailer 1(SD)	Retailer 2(SD)	Retailer 3(SD)
Sale Users			
MP	1(0)	1(0)	1(0)
CF	0.38(0.06)	0.31(0.07)	0.43(0.13)
CB	0.42(0.08)	0.24(0.06)	0.55(0.13)
View Users			
MP	1(0)	1(0)	1(0)
CF	0.37(0.06)	0.31(0.07)	0.43(0.12)
CB	0.42(0.08)	0.24(0.06)	0.56(0.12)
New Users			
MP	1(0)	1(0)	1(0)
CF	-	-	-
CB	0.43(0.08)	0.24(0.06)	0.57(0.11)
Average			
MP	1(0)	1(0)	1(0)
CF	-	-	-
CB	0.43(0.08)	0.24(0.06)	0.57(0.11)

Note: CF cannot make predictions for new users.
SD is the standard deviation across users.

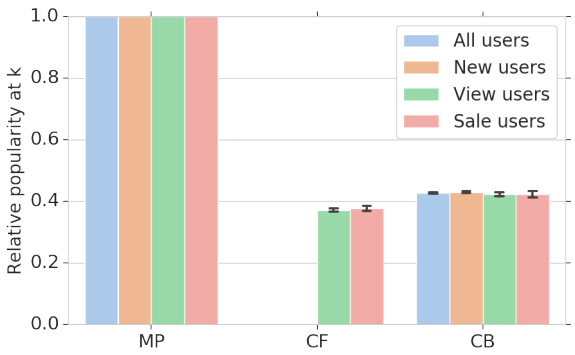

Figure 3: RP_{10} for Retailer 1. Each error bar represents the 95% confidence interval of the distribution of 1,000 bootstrap samples of $RP_{k,u}$ values. By only recommending the most popular items, the MP recommendation strategy always produces values of 1.

The MP recommendation strategy provided the most popularity-biased recommendations because by recommending the same, most-popular items to all users, MP always results in RP_{10} values of 1. CB had more popularity-biased recommendations than CF for Retailers 1 and 3, while the opposite was true for Retailer 2. Overall, recommendations for Retailer 3 were the most popularity-biased, followed by Retailer 1, and then Retailer 2. Within each retailer/model combination, RP_{10} exhibited very little variance across the four user segments.

4.3 Summary of model results

CB modeling was generally more accurate but less personalized than CF. Although CF generally provided more personalized recommendations for view and sale users compared with CB modeling, CF had much lower user-space coverage than CB modeling because CF cannot make recommendations for new users. While the MP recommendation strategy was able to provide very accurate recommendations, those recommendations were completely depersonalized, with the lowest possible AD_{10} and highest possible RP_{10}. Only by performing a holistic evaluation that includes measures of personalization and evaluations across user segments are we able to expose the shortcomings of the MP and CF recommendation strategies. Despite having lower accuracy compared with MP and lower personalization when compared with CF, CB models were able to successfully balance accuracy with personalization while making recommendations to new users.

4.4 Summary of retailer results

Given the results by model type, we suspect that the patterns of results by retailer may be driven by differences in retailer sales distributions. In general, recommendations for Retailer 3 had the highest accuracy as measured by $NDCG_{10}$, but the lowest diversity and the highest popularity bias, which may be explained by Retailer 3 having the sales distribution most dominated by popular items. As shown in Figure 4, one third of sales for Retailer 3 involve only the 3.7% of most popular items, compared with Retailers 1 and 2, where one third of sales involve the 8.7% and 12.1% of most popular items, respectively. Additionally, in most cases CB modeling was more accurate but less personalized than CF, with the exception of higher RP_{10} for CF than CB at Retailer 2, and lower $NDCG_{10}$

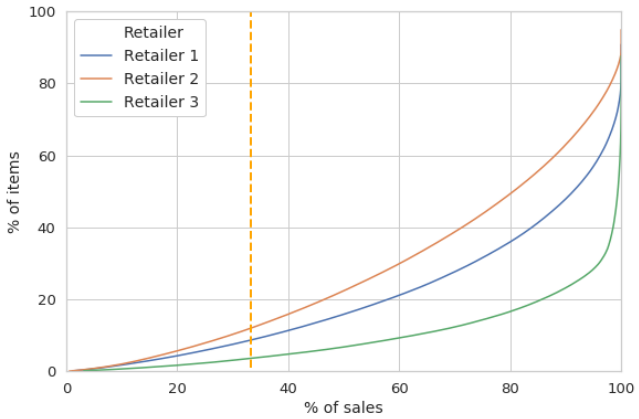

Figure 4: Sales distributions for our three retailers. Items are ordered by popularity, with the most popular items at the bottom. The set of popular items that make up a third of sales is known as the short-head, while the set of remaining items make up the long-tail [4]. The yellow dashed line provides the demarcation between the items in the short-head and long-tail.

for CB than CF for sale users at Retailer 2. This exception may be explained by Retailer 2 having the sales distribution least dominated by popular items, where the accuracy and popularity bias of CB might be directly affected by the sales distribution of the underlying retailer data.

5 DISCUSSION

Our goal was to propose an offline methodology for evaluating fashion RS. Because personalization is a critical feature of fashion, our evaluation framework includes accuracy as well as recommendation diversity and popularity bias. Moreover, because most users in our fashion datasets are new, we performed our analyses separately for users based on prior interaction history. By considering multiple metrics within multiple user segments, we gain a better understanding of how algorithm decisions are likely to influence the experience of the end users.

Although our results varied to some extent by user segment and retailer, we can still make several important conclusions. First, across all of our retailers, our data is very sparse. For comparison, the Netflix dataset and the MovieLens dataset, both of which have been used extensively for RS research [2, 6], demonstrate denser data than any of our three retailers. The overwhelming majority of users represented in the test dataset had no views or sales in the training dataset. As a result, our CF algorithm was unable to provide recommendations for over 70% of users, making CF a poor algorithm choice for fashion. In contrast, our MP algorithm was able to provide accurate recommendations for all user segments; however, because item popularity was calculated across all users, the MP algorithm provides no personalization. Our CB approach represents the best algorithm choice of the three because it provides: 1) relatively accurate recommendations, 2) an acceptable level of personalization, and 3) complete user-space coverage.

Our fashion RS evaluation approach has many advantages over more simplistic approaches; however, there are several ways in which our approach is limited. Here, we define users based on interaction history, but user groups could be defined along many axes (e.g., demographics, frequent versus infrequent shoppers). Also, our approach focused on segmenting users, not products. Content providers may also be interested in how well RS perform within specific subsets of their products (e.g., new versus classic products). Furthermore, we limited our RS comparisons to three relatively basic algorithms. Comparing different variants of these algorithms (e.g., neighborhood-based CF versus model-based CF) could provide additional nuance to our results. Future research could apply our evaluation approach to more variants of common algorithms as well as to novel algorithms specifically tailored to fashion recommendation (e.g, an algorithm focused on flaw minimization or comfort).

Here, we have proposed a more comprehensive offline evaluation. However, prior research has indicated that offline and online metrics are not always correlated [1], calling into question the utility of offline evaluation. One of reasons why offline and online metrics disagree could be that most offline evaluation methods are singularly focused on accuracy [1] and as a result, fail to capture the full range of human factors that influence users' experiences. A multifaceted evaluation approach applied to multiple user segments is more likely to promote algorithms that perform well on online metrics (e.g., click through rates, increased sales, etc.,). Nevertheless, an important next step will be performing an online evaluation to validate our offline results.

In sum, our current work demonstrates the importance of evaluating recommendations from multiple angles. By performing a multifaceted offline evaluation, we can develop a better insight into how our RS are likely to perform when encountered by real-world fashion users.

REFERENCES

[1] Joeran Beel, Marcel Genzmehr, Stefan Langer, Andreas Nürnberger, and Bela Gipp. 2013. A comparative analysis of offline and online evaluations and discussion of research paper recommender system evaluation. In *Proceedings of the international workshop on reproducibility and replication in recommender systems evaluation.* ACM, 7–14.

[2] James Bennett and Stan Lanning. 2007. The Netflix Prize. *Proceedings of KDD Cup and Workshop* (2007), 3–6. https://doi.org/10.1145/1562764.1562769

[3] Riza Casidy Mulyanegara and Yelena Tsarenko. 2009. Predicting brand preferences: an examination of the predictive power of consumer personality and values in the Australian fashion market. *Journal of Fashion Marketing and Management: An International Journal* 13, 3 (2009), 358–371.

[4] Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. 2010. Performance of recommender algorithms on top-n recommendation tasks. September (2010), 39. https://doi.org/10.1145/1864708.1864721

[5] E Viriato de Melo, E A Nogueira, and D Guliato. 2015. Content-Based Filtering Enhanced by Human Visual Attention Applied to Clothing Recommendation. In *2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI).* 644–651. https://doi.org/10.1109/ICTAI.2015.98

[6] F.Maxwell and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 4 (2015), 1–19. https://doi.org/10.1145/2827872

[7] Antonino Freno. 2017. Practical Lessons from Developing a Large-Scale Recommender System at Zalando. *Proceedings of the Eleventh ACM Conference on Recommender Systems - RecSys '17* (2017), 251–259. https://doi.org/10.1145/3109859.3109897

[8] Ronald E. Goldsmith. 1999. The personalised marketplace: Beyond the 4Ps. *Marketing Intelligence & Planning* 17, 4 (1999), 178–185. https://doi.org/10.1108/02634509910275917

[9] Stephen Harrison and Ben Wilson. 2017. Case Study : Building a Hybridized Collaborative Filtering Recommendation Engine.

[10] Yifan Hu, Chris Volinsky, and Yehuda Koren. 2008. Collaborative filtering for implicit feedback datasets. *Proceedings - IEEE International Conference on Data Mining, ICDM* December 2008 (2008), 263–272. https://doi.org/10.1109/ICDM.2008.22 arXiv:arXiv:1208.5721

[11] Hyunwoo Hwangbo, Yang Sok Kim, and Kyung Jin Cha. 2018. Recommendation system development for fashion retail e-commerce. *Electronic Commerce Research and Applications* 28 (2018), 94–101. https://doi.org/10.1016/j.elerap.2018.01.012

[12] Dietmar Jannach and Gediminas Adomavicius. 2016. Recommendations with a Purpose. (2016), 7–10. https://doi.org/10.1145/2959100.2959186

[13] Kim Johnson, Sharron J. Lennon, and Nancy Rudd. 2014. Dress, body and self: research in the social psychology of dress. *Fashion and Textiles* 1, 1 (2014), 1–24. https://doi.org/10.1186/s40691-014-0020-7

[14] Sean M. McNee, John Riedl, and Joseph A. Konstan. 2006. Being accurate is not enough. (2006), 1097. https://doi.org/10.1145/1125451.1125659

[15] Frank McSherry and Marc Najork. 2008. Computing information retrieval performance measures efficiently in the presence of tied scores. In *European conference on information retrieval*. Springer, 414–421.

[16] Riza Casidy Mulyanegara, Yelena Tsarenko, and Alastair Anderson. 2009. The Big Five and brand personality: Investigating the impact of consumer personality on preferences towards particular brand personality. *Journal of Brand Management* 16, 4 (2009), 234–247.

[17] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John T. Riedl. 1994. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. *Proceedings of the 1994 ACM conference on Computer supported cooperative work - CSCW '94* (1994), 175–186. https://doi.org/10.1145/192844.192905 arXiv:111

[18] Badrul Sarwar, George Karypis, Joseph A. Konstan, and John Riedl. 2001. Item-Based Collaborative Filtering Recommendation Algorithms Badrul. In *Tenth International World Web Conference, WWW10*. Hong Kong, 285–295.

[19] Nizar Souiden, Bouthaina M'Saad, and Frank Pons. 2011. A cross-cultural analysis of consumers' conspicuous consumption of branded fashion accessories. *Journal of International Consumer Marketing* 23, 5 (2011), 329–343.

[20] Marika Tiggemann and Catherine Lacey. 2009. Shopping for clothes: Body satisfaction, appearance investment, and functions of clothing among female shoppers. *Body Image* 6, 4 (2009), 285–291. https://doi.org/10.1016/j.bodyim.2009.07.002

[21] Kristen Vaccaro, Sunaya Shivakumar, Ziqiao Ding, Karrie Karahalios, and Ranjitha Kumar. 2016. The elements of fashion style. *Proceedings of the 29th Annual Symposium on User Interface Software and Technology - UIST '16* (2016), 777–785. https://doi.org/10.1145/2984511.2984573

[22] Daniel Valcarce, Alejandro Bellogín, Javier Parapar, and Pablo Castells. 2018. On the robustness and discriminative power of information retrieval metrics for top-N recommendation. (2018), 260–268. https://doi.org/10.1145/3240323.3240347

[23] Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. 2010. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences* 107, 10 (2010), 4511–4515.

[24] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. June (2005), 22. https://doi.org/10.1145/1060745.1060754