Size Recommendation System for Fashion E-commerce

G. Mohammed Abdulla
Myntra Designs, India
ghani.abdulla@myntra.com

Sumit Borar
Myntra Designs, India
sumit.borar@myntra.com

ABSTRACT
Understanding user size preference in addition to style preference is a critical aspect of fashion e-commerce domain. Unlike offline, in online fashion shopping, customers don’t have the luxury of trying a product and have to rely on the product image and the size charts to select a product that fits well. As a result of this gap, online shopping yields a large percentage of returns due to size and fit. Also, explicit elicitation of a users body shape or measurements does not scale well. In this paper, we propose a size recommendation system to automatically pre-select consumer’s size based on past purchase and content data without explicitly asking for users measurements. We use skip gram based word2vec model on our purchase data to learn the latent representation of all our products and users in a common size and fit space, thereby enabling a similarity notion among different products and user-products. Gradient boosting classification model is further employed on both the learnt latent features and observable features (like users estimated chest size, products fit etc.) to predict the preferred product size for a user. The effectiveness of the proposed algorithm is validated through extensive experiments on real world data. Further we derive distinct users’ body shapes and glean insights from their return behavior on our platform.

Keywords
Size Prediction, Word2vec, Personalisation

1. INTRODUCTION
Online shopping platforms provide consumers with convenience of shopping at home. Fashion, and especially apparel, is the fastest growing category in online shopping [1]. There are multiple factors like trust, logic, fit and emotions that affect user’s choice in shopping apparel through online portals [2]. In these, Clothing fit has been shown to be the most important element for consumers in determining their overall satisfaction with garments [3]. In case of unsatisfied size and fit, consumers often return the apparel that they have purchased [4].

In order to ease online shopping, most online e-commerce platforms provide free returns on all its products. Online e-commerce platforms typically observe almost double the return rates as compared to traditional offline fashion stores [5]. During the return process in our platform, we capture the reason for the return from the customer. Based on this data, we can attribute large percentage of the returns to size and fit mis-matches. Returned products incur significant operational cost on the platform and blocks inventory for sale. Furthermore, this results in poor experience for the customers and decreases their confidence for future purchases on the platform.

In comparison to traditional offline setting, in online experience there is no physical product to inspect and try. Consumer’s purchase decision rests purely on images, description and sizing charts provided with the product. Size charts require customers to remember their body measurements and compare them with product dimensions. Moreover fashion industry lacks standardization in terms of sizing [6] and the attributes associated with fashion products are highly subjective. Each category of apparel usually have similar size representations - S, M, L, XL etc. across brands, however they represent different physical measurements. For example a Calvin Klein T-shirt of size M has chest measurement of 40 inches while a Tommy Hilfiger T-shirt of size M has a chest measurement of 38 inches. Also even for a same

Figure 1: Size Recommendation displayed on Myntra Mobile App.
brand, different product lines and various fits (Slim, Regular etc.) makes choosing size a tricky process. In addition, actual products measurements and the size charts provided by brands have lot of variance. During quality check we have measured samples for each product and observed that 30% of the apparel inventory has ±1 inch error when compared with their size chart. All these factors lead to a cumbersome process for selecting correct size using a size chart, often resulting in a negative experience for the customers and hampers adoption of new consumers. There is substantial research on personalized product recommendation for users in fashion e-commerce. Recommendation based on users past interactions, taste and affinities have been studied extensively [7; 8; 9; 10]. All these recommendation systems try to model user style preferences and not size preferences. In clothing and textiles research various methods to find fit and size preference are proposed based on 3D modelling of body shapes [11; 12; 13]. These methods rely heavily on inferring body shapes from database of manually curated body shape metrics [14] or extracting body shapes from images [15]. There have also been attempts to model user size preferences in industry by the likes of True Fit and Fit Analytics. However, their approaches require users to provide body measurements explicitly via surveys or questionnaires. At Myntra, we have tens of millions of customers, and a catalog comprising of over million products across several fashion categories with more than 300k products available at any point in time. We wish to solve the problem of size and fit preference by leveraging past purchases of our diverse customer base.

In this paper, we describe a system to recommend size for a user and a product, employing a Gradient Boosted Classifier (GBC). Each product has multiple sizes available and each combination of size and product is referred as a SKU. The input vector for GBC is a blend of SKU and user vector. Further, the SKU vector is formed by concatenating observable features of SKU like size, brand, fit etc. along with latent features learnt from sales data using word2vec [16]. The user vector is computed by aggregating SKU vectors of past purchases by that user. The GBC estimates the probability of fit for a given user and a SKU. From experiments we show that model built with both observable and latent features performs better than models build solely on either observable or latent features. In our experiments we split data based on time and predict sizes for future purchases. Further we cluster these user vectors to identify prominent body types observable on our platform and derive insights into correlation between body types and return rates. Figure 1 illustrates the size recommendation displayed on Myntra mobile App.

Section 2 describes the latent features generation model and gradient boosted model to find fit probabilities. We present the experimental results in Section 3. Section 4 details the analysis of latent features and presents correlation between body shapes and return rates. In Section 5, we present our conclusions and insights learned.

2. METHODOLOGY

Our approach is to model size recommendation as a classification problem where model estimates fit probabilities for all the sizes of a given product and a user. The overall system of size recommendation is illustrated in Figure 2.

The system comprises of three stages: a skip gram based word2vec model to learn latent feature vectors, a aggregate function to compute observable feature vectors and a gradient boosted classifier to output fit probability. We discuss the three stages in detail and later illustrate three different recommendation approaches.

2.1 Observable feature vector

All the features which are based on product catalogue data are referred to as observable features data. For every SKU, we have physical measurements data like chest size, shoulder size, etc. Also, we have other product attributes like material, occasion, colour etc., which can have an impact on apparel fit. The physical measurements data of the product are continuous values and are used directly, where as other categorical product attributes are used as one hot encoded values. The concatenated vector of continuous and categorical values forms the observable feature vector for SKUs. For computing the observable user vector, we aggregate the observable SKU vectors of products purchased by a user. Also, we add few more features like mean and standard deviation of physical measurements of the SKUs purchased. The mean and standard deviation of the physical measurements captures the distribution of various sizes purchased by the user.

2.2 Latent feature vector

The latent feature vectors are generated using skip gram based word2vec model. We use user’s non returned purchase data and product content data from a single category to train category specific model. The input and output for the word2vec are pairs of SKUs purchased by the same user. After training the network, the activation of hidden layer for every SKU is the latent features vector. User purchase data can be considered as a sparse matrix $W$ where each row in the matrix represents a unique user and the columns correspond to the SKU purchased by that user. The basic idea is that SKUs purchased by a user are similar in size and fit. We exploit this information to learn a joint probability function for sequences of SKUs purchased.
by users.

We replace each SKU in the purchase matrix \( W \) by a string which is a concatenation of product attributes like brand, size, fit etc. The obvious choice here is to replace each SKU with combination of brand and size. However, further analysis showed that even in same brand and size the variance of product dimension was high due to various product lines for the same brand. In order to capture these variations, each SKU was represented by a combination of brand, size, fit and usage attributes. For example a SKU with brand ‘Roadster’, size ‘M’, fit ‘Slim’ and usage ‘Casual’ would be replaced by a word “Roadster-Casual-Slim-M”.

Inspired by the skip gram based language model, we consider each row in \( W \) as a document and attributes of a SKUs as a word in that document. We train word2vec model to arrive at representation for SKU. Formally, \( \cdot \)th row in \( W \) gives all the purchases sorted by date for user \( i \) and let this sequence be \( w_{i1}, w_{i2}, w_{i3}, \ldots, w_{in} \). Here each \( w_{ij} \in C \), where \( C \) is set of all SKUs in the catalog. The objective of the skip gram language model is to maximize the log probability

\[
\frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{q=1}^{m} \log p(w_{ij} | w_{i,j+k})
\]

where \( q \) is the hyperparameter denoting length of the purchase window. Larger \( q \) results in SKUs spanning over wide range of purchases to be considered as having same size and fit. The formulation of \( p(w_{ij} | w_{i,j+k}) \) is given using softmax function:

\[
p(w_j | w_{j+k}) = \frac{e^{w_{ij},v_j+k}}{\sum_{h \in W} e^{w_{ij},v_h}}
\]

where \( u \) and \( v \) are the input and output one hot encoded vector representation of \( w \). After we have trained the neural network, we compute the activation of the hidden layer for each SKU \( c_p \in C \) and form a latent feature vector representation \( f_p \).

We can compute the vector representation for a user \( u \), by aggregating over his purchased SKUs vectors, i.e, \( i^{th} \) user’s purchased SKUs are represented by \( f_1, f_2, f_3, \ldots, f_n \), then the user vector of equal dimension as that of SKU vector can be computed as:

\[
\frac{1}{n} \sum_{p=1}^{n} f_p
\]

### 2.3 Recommendation as Classification

The size recommendation is formulated as a binary classification in which the task is to classify if a given SKU (from all the SKUs of the product) will fit the user. Gradient boosted classifier [17] is used to predict fit probabilities for all the SKUs of the product and a user. The SKU with highest fit probability is the model’s recommendation for the user.

To train the GBC model we require SKUs that fit users and also SKUs that doesn’t fit the users. In our platform, users can return and exchange purchased products. When a product is returned, we explicitly gather reason for return from the user. Some of the reasons are "Returned due to size". "Not looking good on me" etc. Exchanges occur when user needs a different size of the same product. The positive samples for training are the SKUs that are retained by the users after the purchase. The negative samples are the SKUs returned due to size issues and the SKUs exchanged.

The training vector is a concatenation of vector user and SKU vector. Both user vector and SKU vector are formed by concatenating corresponding latent feature vector and observable feature vector. Latent SKU vectors are averaged before concatenation to transform variable size purchase history to fixed-width vectors suitable for model input.

### 3. EXPERIMENTS AND RESULTS

#### 3.1 Dataset

For our experiments, we used sales data from our platform between Jan 2015 and Feb 2017. The data was split at category level (eg: Men Shirts, Men Tshirts, Women Dresses etc) and a model is trained for each category. We present experimental results in detail for the category - Men Shirts. For Men Shirts, the dataset contained about 300K users.

We observed that some users tend to buy for more than one person with the same account. Product purchases in such cases span across wide range of sizes. To identify such users, we calculate mean and standard deviation of dimensions of all the purchased SKUs within a category for each user. A category level coherency threshold is calculated empirically. Users with standard deviation below category coherency threshold are considered as single persona users and included in the experiments.

As the objective of the model is to predict future purchases-training, validation and test data were split using time of purchase. The training data comprised of sales data between Jan 2015 and Oct 2016, validation data between Nov 2016 and Dec 2016 and test data between Jan 2017 and Feb 2017. After the data split, the training data had 1.4M purchases,

![Figure 3: Classification model architecture showing user vector concatenated with SKU vector. Both user and SKU vector are a combination of corresponding observable feature vector and latent feature vector. Latent SKU vectors are averaged before concatenation to transform variable size purchase history to fixed-width vectors suitable for model input.](image-url)
Table 1: Dataset statistics for training data

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>No of unique users</td>
<td>363,767</td>
</tr>
<tr>
<td>No of purchases</td>
<td>1,109,880</td>
</tr>
<tr>
<td>No of returns</td>
<td>182,134</td>
</tr>
<tr>
<td>No of exchanges</td>
<td>102,291</td>
</tr>
<tr>
<td>Average orders per user</td>
<td>3.84</td>
</tr>
</tbody>
</table>

validation data had 0.4M purchases and test data had 0.45M purchases. Table 1 shows various statistics of training data.

3.2 Experiments

We experimented with three different Gradient Boosted Classifier (GBC) models each having different sets of features that transform concatenated user and SKU vectors to find likelihood of fit for a given user and SKU.

3.2.1 Observable Features Model

For this model, a GBC is trained only on the observable SKU and user vectors. Based on cross-validation, we used GBC with 40 trees, maximum depth of tree as 6 and learning rate as 0.1.

3.2.2 Latent Features Model

Relying purely on latent features, this model’s training vector comprised of latent user and SKU vector. From cross validation the model parameters was fixed at 40 trees, maximum depth as 6 and learning rate as 0.1.

3.2.3 Combined Features Model

Finally, we integrated the observable features and latent features to form a combined model. In this combined model, the latent features are responsible for predicting fit likelihood using the product attributes like fit, occasion, brand etc while observable features are responsible in predicting fit likelihood using physical product measurements.

3.3 Size Prediction Results

For the evaluation, we used Precision, Accuracy and Area Under the Curve in ROC as performance metrics. Precision in our case is defined as follows:

\[
\text{precision} = \frac{\sum_{i=1}^{N} \left( y_i - \hat{y}_i \right)^2}{N}
\]  

where \( y_i \) is the ground truth and \( \hat{y}_i \in \{0,1\} \) is the model prediction and \( N \) is number of purchases in the test data set where \( y_i \) is equal to 1.

The accuracy of the model is computed as the ratio of all correct predictions by the model to all predictions by the model. Table 2 compares accuracy and precision values for the three models. For the model using only observable feature vectors, we found that feature importance assigned to the sparse categorical features of observable feature vector were low i.e., these features were not extensively used by the classifier model in making prediction. The latent features computed using the neural network which modelled co-occurrence of SKUs resulted in dense feature vectors which were assigned better feature importance resulting in improved model predictions. It is evident from these accuracy and precision values, that latent feature model performs better than observable features model whereas the combined model outperforms both individual feature models. The ROC curve in Figure 4 also shows that Area Under Curve (AUC) for combined model is larger than the individual feature models.

It was observed that fit probabilities estimated by the combined GBC model and corresponding precision scores are approximately proportional. The probabilities are the model’s confidence on the prediction. As size recommendation directly impacts user experience and revenue in our platform, we want to predict sizes with high confidence. In our live platform, we recommend sizes to users where fit probability for a given SKU is above a certain threshold. We have chosen this threshold based on the business requirements. As the threshold increases, the coverage of users for whom we can make a recommendation decreases. Table 3 shows the precision and coverage values for various probability threshold values.

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Precision%</th>
<th>Coverage%</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>79.21</td>
<td>75.72</td>
</tr>
<tr>
<td>0.6</td>
<td>82.33</td>
<td>61.45</td>
</tr>
<tr>
<td>0.7</td>
<td>83.85</td>
<td>47.43</td>
</tr>
<tr>
<td>0.8</td>
<td>85.53</td>
<td>40.08</td>
</tr>
<tr>
<td>0.9</td>
<td>89.53</td>
<td>32.85</td>
</tr>
</tbody>
</table>

Figure 4: Comparison of classification performance using ROC for 3 GBC models using different features. It can be seen that latent features model(red) performs better than observable features model(orange) and combined model(green) model is the best among the three models.

Table 2: Comparing Precision and Accuracy for 3 GBC models.

<table>
<thead>
<tr>
<th>Features</th>
<th>Precision%</th>
<th>Accuracy%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observable Features</td>
<td>65.45</td>
<td>68.27</td>
</tr>
<tr>
<td>Latent Features</td>
<td>72.21</td>
<td>78.16</td>
</tr>
<tr>
<td>Combined Features</td>
<td>75.52</td>
<td>81.28</td>
</tr>
</tbody>
</table>
Similarly, models for other categories like Men T-shirts, Men Jeans, Women Tops etc were trained. Table 4 shows the precision scores for various categories and it is perceivable that our approach can be generalized to different categories.

### 4. ANALYSIS

A central tenet in our approach is the latent features computed from the past user purchase data. In this section, we discuss the insights we derived from analysis of these latent feature vectors on SKU and User level.

#### 4.1 SKU Vector analysis

Majority of Brands in our catalog have 6 different sizes available and hence we clustered the latent SKU vectors into 6 clusters. Table 5 shows the clustering results for four prominent brands Nike, Puma, Tommy and Roadster. The values dimensions columns are the product measurements in inches. In these brands, Puma has the tendency to run a size larger than the rest of the brands. Because word representations computed using word2vec model captures notion of similar sizes of SKUs across different product lines and even across various brands it can be seen that Puma L comes closer to Nike XL. Similarly the notion of brands running a size larger is also captured by the latent vector. It can be seen in the table that sizes in Puma are similar to one size larger is also captured by the latent vector. It can be seen in the table that sizes in Puma are similar to one size larger than the rest of the brands. Because word representations computed using word2vec model captures notion of similar sizes of SKUs across different product lines and even across various brands it can be seen that Puma L comes closer to Nike XL. Similarly the notion of brands running a size larger is also captured by the latent vector. It can be seen in the table that sizes in Puma are similar to one size higher in other brands.

#### 4.2 User Vector Analysis

In order to explore male users sizing patterns, we clustered user vectors using K-means clustering algorithm. Number of clusters to be formed, \( k \) was empirically derived by plotting intra cluster variance across different values of \( k \) and observing the elbow point. We found 12 to be an optimal value for \( k \).

From the corresponding body measurements of cluster centers, body shape metric was derived for each cluster center. These body shapes metrics can be considered as the prominent body shapes of males shopping on our platform. With the help fashion experts, we further grouped these body shapes into 5 different body types namely petite, normal, muscular, boxy, oval. Using a body visualizer tool [18] we visualized all the 5 different body types. Figure 7 shows the visualization of these 5 body types.

Table 6 shows population distribution, overall return rates, top-wear and bottom-wear return rates for users with these body types. The return rates are calculated only on the returns due to size and fit issue. We can observe that body types labelled as normal and muscular have less return rate compared to other body types. This pattern is more pronounced with top-wear return rates. The reason for such a trend can be asserted back to the fact that brands and manufactures tend to design fashion apparel for the perfect body types. Also, most brands do not provide waist measurements for top-wear, however this is a significant factor in choosing correct top-wear fit. Thus, for consumers with boxy and oval body shapes, it becomes more challenging to find the right fit. Hence, the return in case of boxy and oval body shapes are higher than normal and muscular shapes. Figure 5 shows comparison of top-wear and bottom-wear return rates. It can also be observed from the figure that the return rates of top-wear vary significantly across body shapes compared to bottom-wear. This further suggests that top-wear fit is much more intricate, compared to bottom-wear.

In next step of our analysis we selected 10 random brands in our platform with similar views by users, price range and revenue contribution. In the 10 brands chosen, 5 brands were International brands while remaining 5 were Local Brands. Figure 6 shows the return rates of different body types for both Local and International brands. It can be seen that International brands have higher return rates compared to Local Brands. Similarly, these relations hold true for top-wear and bottom-wear return rates.
Table 6: Population and returns metrics for different body shapes with body type labels.

<table>
<thead>
<tr>
<th>Body Type</th>
<th>Population</th>
<th>Returns%</th>
<th>Top-wear Returns%</th>
<th>Bottom-wear Returns%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Petite</td>
<td>19.37</td>
<td>10.26</td>
<td>12.24</td>
<td>12.45</td>
</tr>
<tr>
<td>Normal</td>
<td>32.32</td>
<td>9.97</td>
<td>8.51</td>
<td>11.99</td>
</tr>
<tr>
<td>Muscular</td>
<td>5.51</td>
<td>10.51</td>
<td>9.58</td>
<td>10.39</td>
</tr>
<tr>
<td>Boxy</td>
<td>28</td>
<td>12.87</td>
<td>13.84</td>
<td>12.38</td>
</tr>
<tr>
<td>Oval</td>
<td>14.86</td>
<td>14.41</td>
<td>18.73</td>
<td>13.26</td>
</tr>
</tbody>
</table>

Figure 6: Comparison of Brand level return rates for Local and International Brands across different body types. Local brands have lesser return rates across body types and fit well for normal and muscular body types.

Local brands. This trend is amplified for bulkier body types. This supports the theory that regional spread of body types vary from the western body types and local brands are more attuned to it.

5. CONCLUSIONS

In this paper, we have described a size recommendation system that guides consumers to correct apparel size and eliminates onerous task of dealing with size charts during online shopping process.

We generate latent vectors for each product SKU utilizing word2vec model. Users are also represented in same latent size space using inherent notion of similarity in purchased sizes. Latent feature vectors capture similarity in sizes across SKUs and helps in overcoming sparsity in training vectors. We demonstrated that the proposed classifier with combination of latent features and observable features has better classification accuracy as compared to other variations. We also show that this approach can be generalized for other categories with similar results.

We clustered customers into 5 body types based on their derived measurements and compared return rates for top-wear and bottom-wear apparel. Return rates show that top-wear sizing is much more intricate and customers with petite, boxy and oval shapes have higher difficulty in finding the right fit. We also compared return rates for local and international brands for these body types. Our analysis show that local brands are more attuned to bulkier body types prevalent regionally. This information can be further utilized by brands to create targeted apparel sizes for the local population.

In our current setup, we only consider users who have purchased at least two products from the category before recommending size for that category. For future work, we would like to expand our work to utilize a combined latent vector generated over all categories, which would enable us to make size recommendation for new categories. Providing size recommendations for new categories would increase our coverage significantly and increase user’s confidence in shopping across new categories.

6. REFERENCES


Figure 7: Figure Showing body types of Male Users. (from left) Petite, Normal, Muscular, Boxy and Oval.


