Predicting Clothing Fit for Users

Yuhuan Chen, University of California, San Diego **Lengdifeng Liu**, University of California, San Diego

Siyu Ma, University of California, San Diego **Qiuyi Yang**, University of California, San Diego

1 DATASET & PREP

1.1 Dataset

The dataset, comprising 82,789 Mod- Cloth and 192,543 *RentTheRunway* entries, totals 275,332 records with 10 attributes each. It includes measurements of clothing fit, customer stature, and clothing metrics, alongside customer reviews. Details are in Tables 1 to 4.

Table 1: Dataset ModCloth Item Attributes

Attribute of Item	Description
Size	Size of the clothes
Quality	Average ratings of the quality
Category	New, dresses, outerwear, etc
Fit	Small, Fit, Large

Table 2: Dataset ModCloth User Attributes

Attribute of User	Description
Waist	Customers' waist
Cup Size	aa, b, c, etc
Bra Size	Customers' bra numeric size
Length	just right, slightly long, etc
Height	Customers' height in ft and in
Shoe Size	Customers' shoe size

Table 3: Dataset RentTheRunway Item Attributes

Attribute of Item	Description
Size	Size of the clothes
Category	gown, maxi, mini, etc
Rating	Customers' average rating
Fit	Small, Fit, Large

Table 4: Dataset RentTheRunway User Attributes

Attribute of User	Description
Bust Size	34b, 34c, 32b, etc
Weight	Customers' weight in lbs
Height	Customers' height in ft and in
Age	Customers' age
Rented	The purpose for renting clothes
Body Type	pear, athletic, petite, etc

1.2 Data Preprocess

Data preprocessing was critical for enhancing result quality. We initially converted non-numeric features to numerical codes. For example, the 'Cup size' attribute in the ModCloth dataset, expressed in letters (aa, a, b, c, etc.), was assigned numeric values from 0 to 11 based on their ordinal relationship. In the RentTheRunway dataset, the word-based 'Category' feature was transformed using One-hot encoding, due to its non-ordinal nature. Additionally, to achieve continuity, discrete values in features like 'Rating' (originally 2, 4, 6, 8, 10) were halved, remapping them to 1-5. We addressed prevalent missing values by substituting means for numeric gaps and most frequent values for non-numeric data, avoiding data loss from simple deletion. Finally, normalization was applied to non-One-hot encoded features, scaling them between 0 and 1, thus reducing bias, accelerating training, and enhancing model efficiency.

1.3 Exploratory Data Analysis

We initially undertook a visualization of the data distribution, taking the ModCloth dataset as an example, with the results illustrated in Figures 1-6. The graphical representation reveals that most features (Bra Size, Shoes, Length, Height, Cups, Fit) exhibit a normal distribution, indicating that using the mean value to address missing data maximally preserves the original data distribution.

Conversely, the data distribution also exposed the presence of anomalous outliers. For instance, in the Shoes Distribution, there was a sample with a value of 38, markedly deviating from the majority. Special handling was required for such samples. We adjusted the shoe size value of this particular sample to the mean, ensuring that the sample distribution remained unbiased. The final mean value, adjusted from 9.03 to 8.32, appeared quite rational as per Figure 3, aligning more closely with the central tendency of the distribution.

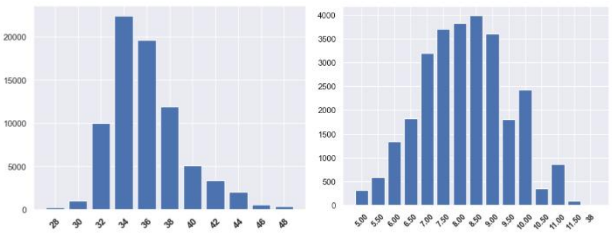

Figure 1&2: Bra Size & Shoes Distribution

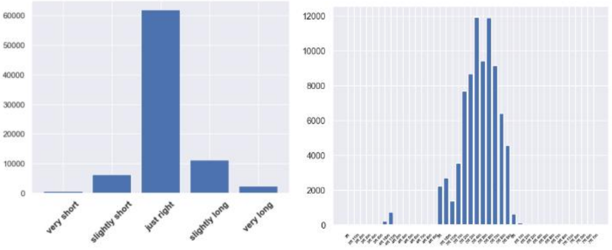

Figure 3&4: Length & Heighy Distribution

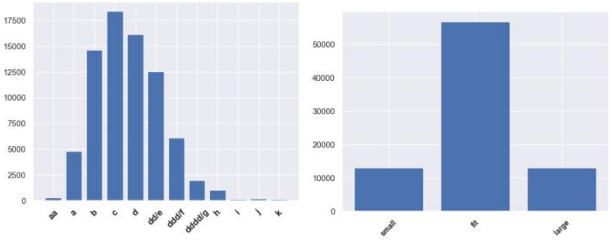

Figure 5&6: Cups & Fit Distribution

In the ModCloth dataset, we observed generally low correlations between features. However, significant correlations were noted between Size and Cup Size (0.35), Size and Bra Size (0.39), and Cup Size and Bra Size (0.45), reflecting the logical connection between body measurements and clothing size. Additionally, Fit and Quality showed a moderate correlation (0.2), indicating a relationship between how well clothes fit and their perceived quality.

In the RentTheRunway dataset, there were notable correlations between Bust Size and Weight (0.23), Bust Size and Size (0.13), and Weight and Size (0.3), again highlighting the impact of body measurements on clothing size selection. The correlation between Fit and Rating (0.25) suggests that better-fitting clothes tend to receive higher ratings, emphasizing the importance of fit in customer satisfaction.

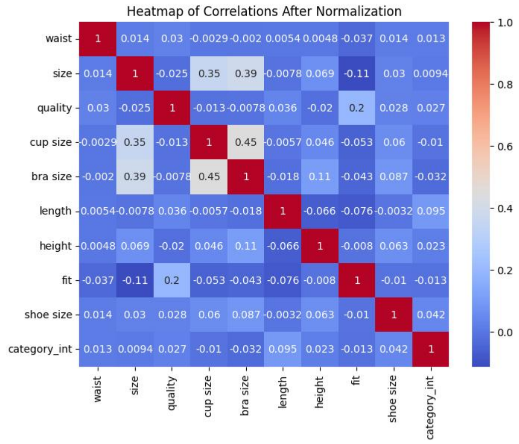

Figure 7: Heatmap of ModCloth

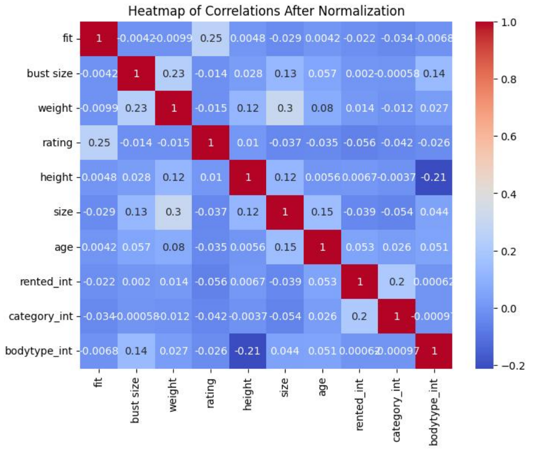

Figure 8: Heatmap of RentTheRunway

Consequently, following this kind of exploratory analysis of the data, we believe that a comprehensive extraction of various features pertaining to users and garments in the dataset can more effectively uncover the underlying connections. Particularly, there must be more intricate relationships both among the users' own metrics and between the users' feedback and the product details. It is essential to construct models capable of adeptly mining information about both users and products. Utilizing these models, we aim to extract comprehensive information from the dataset.

2 PREDICTIVE TASK

Our analysis suggests that predicting the fit of clothing for online shoppers is a valuable task, as it can reduce time and costs associated with finding well-fitting garments online. We treated 'fit' as the positive class and other sizes (like 'small', 'large') as the negative class, framing this as a binary classification problem. The dataset was split into training, validation, and test sets in an 8:1:1 ratio, with model training on the training set, validation, and parameter tuning on the validation set, and final evaluation on the test set.

Initially, our prediction model used algorithms based on Popularity, Similarity, and Bayes Personalized Ranking, focusing only on the 'Fit' feature as a baseline. However, recognizing the importance of other user and garment attributes in predicting 'fit', we incorporated these into a deep learning model. This model was trained using 'fit' as the label, enabling better feature extraction.

We evaluated the model by visualizing the distribution of both raw and processed data, giving an intuitive understanding of the model's effectiveness. Additionally, we assessed the model's performance by applying the derived embeddings to classification tasks with various classifiers, comparing their accuracy against the baseline models to highlight our model's improved efficacy.

3 MODEL

3.1 Baseline

3.1.1 Baseline: Similarity-Based Recommendation

In this model, we recommend items based on the similarity between users. We use Jaccard similarity as a feature for classification. Jaccard similarity is a measure of similarity between two sets, calculated by the ratio of their intersection to their union. We compute the Jaccard similarity between each pair of users and use it as a feature for classification. For a target user, we calculate their similarity with other users and select the most similar user to recommend items that the similar user liked. This model provides personalized recommendations but may face challenges with new users or cold start situations.

$$r(u,i) = \max_{j \in I_u \setminus \{i\}} (r_{u,j} \mathrm{sim}(i,j))$$

To optimize the model's performance, we use a validation set for parameter optimization. Specifically, we iteratively adjust the threshold parameter and select the best-performing threshold on the validation set as the final classification threshold. We then evaluate the classification accuracy on the test set using the best threshold.

3.1.2 Baseline: Popularity-Based Recommendation

In this model, we recommend items based on item popularity. For a target user, we select the most popular items and recommend them.

We optimize the model's parameters using logistic regression on a validation set and evaluate the classification accuracy on a test set. This model is simple and effective, suitable for new users and cold start situations, but lacks personalization.

3.1.3 Baseline: Bayesian Personalized Ranking (BPR) based recommendation

The Bayesian Personalized Ranking (BPR) model is a method for personalized recommendations based on implicit user feedback, prioritizing pairwise item rankings instead of global popularity or item similarity. It uses matrix factorization to infer latent preferences from user-item interactions, aiming to rank preferred items higher than others. The BPR model is refined through stochastic gradient descent, with hyperparameters adjusted to enhance performance and mitigate overfitting. Its strengths include capturing individual preferences and providing personalized item rankings without needing to calculate user similarities, though it can be computationally demanding and has difficulty with new items.

We evaluate the BPR model's performance on the test set, focusing on metrics like precision, recall, and the area under the accuracy, which provide insights into the model's ability to rank items accurately for each user.

3.2 Deep Learning Model

We chose MLP and ResNet – 18 to solve this problem. The model features four embedding layers: user embedding, cup size embedding, item embedding, and category embedding. These are used to convert the categorical features of users and items into dense vector representations. Our constructed MLP model processes user and item paths through three linear transformation blocks, each consisting of a linear layer and a ReLU activation function. Finally, by combining the features of users and items, the model generates predictions using four additional linear transformation blocks. The design of these layers aims to progressively extract and transform features in the user and item data, thereby preparing for the final predictions.

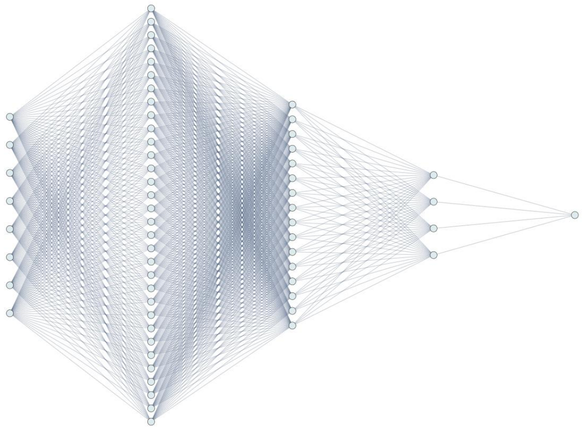

Figure 9: Structure of MLP

To prevent the issue of vanishing and exploding gradients becoming more severe with the increase in the number of network layers, leading to a decline in performance, ResNet addresses this issue through the use of residual blocks. For this task, we employ an embedding method to convert features into a matrix form. After processing the matrix data, we use ResNet-18 followed by two final fully connected layers to transform the output into the required binary classification task format. Additionally, dropout is utilized to prevent the ResNet-18 network from overfitting.

4 LITERATURE

Predicting user-product compatibility is essential in e-commerce. Several studies have explored this, such as one using Bayesian methods on a shoe dataset for size recommendations[1] and another using the CAL10K dataset for music preference predictions with a content-based similarity metric[2]. Modern techniques often use multimodal approaches, combining, for instance[3], computer vision with product images for richer information extraction.

Our datasets, ModCloth and RentTheRunway, are sourced from the paper[4]. This study attempted to build a model to address the issue of label imbalance and effectively predict the fit of clothing for users. In their work, they used a latent factor formulation to construct latent factors for both users and garments, followed by classification using Logistic Regression, K-LV, and K-LV. The prediction task in this paper is similar to ours, focusing on the fit between users and clothing.

Our work, drawing from datasets used in a study on product size recommendation, adopts a binary classification approach to predict clothing fit, contrasting with the original study's multi-class task and reliance on user and product IDs. We incorporate all available data features for improved model performance, unlike the previous study which used latent factors based solely on IDs.

The task in the paper cannot predict for users who have not previously purchased. Our method extends the prediction capability to new users by requiring personal measurements rather than past purchase data, though this demands more user input and faces challenges with data completeness and distribution, potentially impacting the effectiveness of deep model embeddings.

5 RESULT & CONCLUSION

5.1 Results

Table 5: results conclusion

Models	Accuracy on ModCloth	Accuracy on RentThe_ Runway
Similarity	0.531	0.563
Popularity	0.688	0.739
Bayes Personalized Ranking	0.527	0.579
Deep Learning 1	0.703	0.772
Deep Learning 2	0.715	0.781
DL1+LR	0.705	0.773
DL1+SVM	0.711	0.780
DL2+LR	0.707	0.784
DL 2+ SVM	0.719	0.789

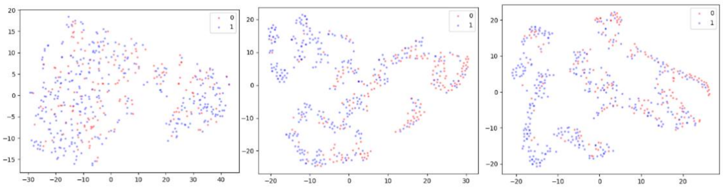

Figure 10: Basline & DL model1 & model2 Distribution of Modcloth

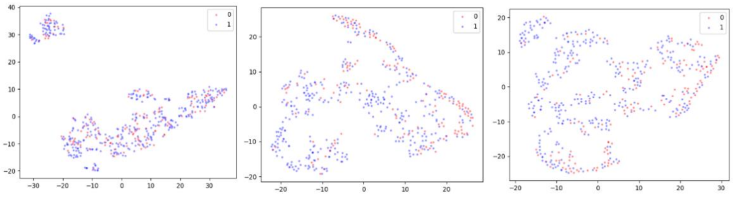

Figure 11: Baseline & DL model1 & model 2 Distribution of Renttherunway

The primary objective of this study is to delve into the potential of deep learning methods across various machine learning tasks. It is particularly noteworthy that we aimed to analyze the distinct effects of deep learning in improving baseline performance and handling complex data distributions. The experimental results vividly demonstrate the capability of deep learning models to significantly surpass baseline performance in various tasks, highlighting their excellence in tackling complex problems and enhancing predictive accuracy. Furthermore, in-depth comparative analyses with traditional machine learning methods such as SVM and logistic regression reveal their notable superiority in most cases, although the selection of the most suitable method should be contingent on the specific task and dataset characteristics.

In addition, through meticulous analysis of data distribution graphs, we observed significant overlap among data points with different labels in the original data, making it challenging to separate them effectively. However, the deep learning models we proposed have proven to be highly effective in improving data distribution, resulting in embeddings that exhibit a stronger separability in high-dimensional space. Lastly, we conducted further investigations into the embeddings generated by various models on different datasets, providing valuable insights for model selection and performance optimization in future research and applications.

5.2 Parameters

For ModCloth dataset, Similarity-Based Recommendation' threshold parameter used was 0.001. Popularity-Based Recommendation parameters were as follows: Coefficients: 1.258, Intercept: 0.780.BPR-Based Recommendatio' learning rate of 0.01, regularization parameters for user and item factors set at 0.1, and a latent factor dimension of 20. Deep Learning Model' learning rate of 0.005, optimizer set at Adam and epoch of 20.

For Renttherunway dataset, Similarity-Based Recommendation' threshold the threshold parameter used was 0.001.Popularity-Based Recommendation parameters were: Coefficients: 0.284, Intercept: 1.039. The BPR-Based Recommendation' learning rate was set slightly lower at 0.005, reflecting a more cautious approach to convergence. The regularization parameters for user and item factors were increased to 0.2 to prevent overfitting. Deep Learning Model' learning rate of 0.005, optimizer set at Adam and epoch of 20.

5.3 Conclusion

In conclusion, this study highlights the significant impact of recommendation models on prediction accuracy, with deep learning models, in particular, showcasing their potential to enhance recommendation system performance. The results emphasize the importance of carefully selecting the appropriate recommendation approach, especially when dealing with diverse datasets. These findings contribute to the ongoing advancement of recommendation systems and provide a foundation for future research in this field.

REFERENCE

[1] Sembium, V., Rastogi, R., Tekumalla, L. and Saroop, A., 2018, April. Bayesian models for product size recommendations. In Proceedings of the 2018 world wide web conference (pp. 679-687).

[2] McFee, B., Barrington, L. and Lanckriet, G., 2012. Learning content similarity for music recommendation. IEEE transactions on audio, speech, and language processing, 20(8), pp.2207-2218.

[3] Zhou, W., Mok, P.Y., Zhou, Y., Zhou, Y., Shen, J., Qu, Q. and Chau, K.P., 2019. Fashion recommendations through cross-media information retrieval. Journal of Visual Communication and Image Representation, 61, pp.112-120.

[4] Misra, R., Wan, M. and McAuley, J., 2018, September. Decomposing fit semantics for product size recommendation in metric spaces. In Proceedings of the 12th ACM Conference on Recommender Systems (pp. 422-426).