Fitting Model so Clothes Fit Models

An Empirical Study into Clothing Fitting Data from Rent the Runway

Zhouchonghao Wu

Guoyi Li

1. Introduction and Data Analysis

The goal of this project is to provide a general approach to clothing fit prediction, given some users' unwillingness to disclose information such as weight, bust size, age, body typed, etc. for privacy concerns.

The dataset used contains 192462 transactions.

Each user has the following file:

- User id – Unique ID of Each user
- Item id – Unique ID of Rented Item
- Fit – Suitability of clothes
- Rating – User rate on the item
- Review text – Text of the review
- Optional – Bust Size, weight, rating, review_text, body type, Category, Height, size, age, etc.

First, we measured the suitability of the overall user's clothes:

- `'fit': 141995,`
- `'small': 25776,`
- `'large': 24691`

Then, we measured how many users have completely filled in their own information. By understanding their reviews, comments, body size, bust size and clothing suitability to give the evaluation.

There is a total of 146,381 users with complete information. In other words, 46,081 users do not want to reveal her privacy. At present, we want to obtain effective information through analysis.

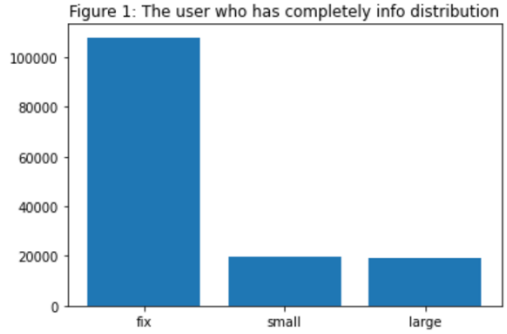

Figure 1:

Through screening, we find that when we have complete user information, we have：

1. 'fit': 108005,
2. 'small':19665,
3. 'large': 19074

Before we conduct a more in-depth analysis, we want to understand where the information is that users do not want to disclose.

Here is the answer we found:

1. User who shares their body type: 177825
2. User who shares their Bust Size: 174065
3. User who shares their age: 191503
4. User who shares their weight: 162505
5. User who shares their size: 191785

Through the above data, we quickly discovered that bust size, body type and weight account for a large proportion. In other words, that information is that users do not want to share.

#percent of total user who do not want to share:

body type: 9.6%

Bust size: 7.6%

Weight: 15.6%

age: 0.4%,

size: 0.35%

We would like to predict the fit of a product and size for a given user under all circumstances, including when the user has not shared any or some of their body measurement information.

Therefore, we want to compare the data set of users sharing their Body type, weight and Bust size information with users who do not share their private information. Detect whether these privacies affect the user experience.

First, there are a total of 45384 users who do not want to share their weight, bust size or body type. ‘fit': 33746, 'large': 5569, 'small': 6069. The distribution of fit/not fit mirrors that of the overall distribution.

Figure2: Data without Weight,bust size and body type

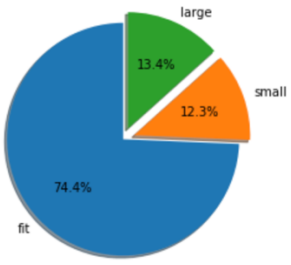

Figure3: Data with Weight,bust size and body type

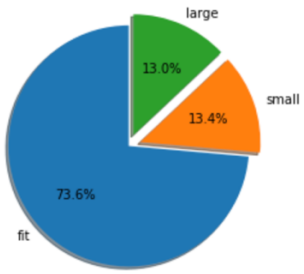

Figure 2 and 3: Despite the lack of information, the probability of getting the right clothes is high.

There are 146,381 users showing their information, ‘fit': 108005, 'small': 19665, 'large': 19074. Through Figure 3, we are surprised to find that in comparison with Figure 2, there is only a small difference between them. This is likely due to the fact that users have the option to choose their own size, and they often know their own fitting needs from their past experience. However, it is understandable that natural variations between clothes of different sizes, brands, and material can render a single person’s past experience less effective, and only deciding from their own information, they are unable to account for these natural variations and sometimes fail to achieve a perfect fit.

We believe we could extract information from the textual review and the numerical ratings. Therefore, we want to know how to analyze the corresponding rating through the text review of previous users. This way we can better understand the comments of new users and further improve the service.

For further analysis, our participants analyzed the most common words and proportions of text review.

Here is our result:

Word Count: 8688115
Distinct Words: 43440

```
[(498858, 'the'),
 (379743, 'i'),
 (318368, 'a'),
 (289663, 'and'),
 (270307, 'it'),
 (225301, 'was'),
 (195923, 'dress'),
 (178681, 'to'),
 (152039, 'this'),
 (117042, 'but')]
```

The Most frequency words shown on the left. Notice that most of them are trivial with regard to extracting information from the textual reviews, so some sort of filtering needs to be performed prior to using the raw text.

2. Prediction Task

In this report we attempt to predict how well a piece of clothing "fits." We consider it a binary variable. We consider the following features in our user database: users display their private information, for example: body type, bust size, weight, age, etc. and the review text.

The focus of this prediction task is to predict whether this product is suitable for the user (fit or not fit) given a database of reviews. In this review, we consider different sizes of a single item distinct items. We require at least 3 past transactions or 2 of the body measurements filled out if we were to generate a prediction for a given user.

More than 75% of the data in this dataset show fit. Therefore, we do not rely much on classification accuracy or classification errors; instead, we use balanced error rate (BER) to evaluate the results. The validity of the model's predictions can be checked via the labels provided by the dataset.

Our baseline model would predict for 'fit' if size is one of (4, 8, 12) and 'not fit' otherwise. These are the most common models and about half of the users fit in them.

We believe there are two approaches to this problem:

1. Using the body size information, we can create a simple mapping from body size measurements to clothing sizes with a perceptron. This is sufficient because clothing sizes are generally built to suit a certain "standard" body build, and we simply need to fine tune our model to predict our mapping.
2. By analyzing the texts, we can grasp how previous users have viewed the item and use the transaction records to predict fitting. Bag-of-words approach to text processing can map words into space. This model was chosen mainly due to the increased computation cost of other potential models. Due to our extraordinarily huge transaction database, and limited computational resources at hand, we have to make do with bag-of-words approach.

In an ideal world with complete body measurements, approach 1 would work well; however, since for our predictive task, we assume to be missing much data, we need to combine approach 1 AND 2 to make the predictions more robust and less dependent on optional user-provided information.

We also wish to clarify that the primary purpose of using a perceptron in this model is to enable non-linear decision boundaries and avoid manually setting the weights for feature representations.

Since data processing takes up significant time, we only perform rudimentary data preprocessing before run-time. For each transaction, the transaction data are first encoded. Bust size is encoded with evenly spaced floats between 0-1. Body type and clothing type are one-hot encoded. Age (in years), height (in inches), and weight (in pounds) is normalized by the corresponding max absolute value and mean value of the training set.

Note: We assume that for a fit prediction to be generated, a user must at least have a transaction history or parts of the privacy information filled in.

3. Model and Prediction

We propose a three-part architecture for our prediction task.

With input of a user (current user) and a piece of garment of a certain size, we process them in the following way:

a. Standardized Information Processing

Part (a) processes standardized information (i.e. bust size, etc.). For any given user, we compare the information they provided with any previous renter of the specific garment. If we could not find at least 20 past transactions from the specific garment in this size, we would consider transaction records of users with other garments of the same size, type, and rented for the same occasion. At the max, 150 transaction records are considered.

Any missing feature are zero-filled. We then derive a confidence vector $\vec{C}$.

$$\vec{C} \coloneqq \frac{1}{N}\sum_{k} f\,\overrightarrow{d_2^k}\cdot\overrightarrow{m^k}$$

Where f is a scalar encoding of 'fit' (-1 for fit and 1 otherwise) $\overrightarrow{d_2^k}$ is the Euclidean norm between the k-th user and the current user. $\overrightarrow{m^k}$ is a masking vector, where each element is 1 if both the current user and the past user provided the corresponding information and is 0 otherwise. N is the total number of transactions.

$\vec{C}$ is then passed through a multilayer perceptron with 2 fully connected layers of 32 ReLU units and a fully connected Sigmoid output unit.

ReLU is given by

$$ReLU(x) = \max(0, x)$$

b. Textual Information Processing

We created a textual embedding layer to process the 500 most common non-trivial words (defined by belonging to one of the categories: adjectives, adverbs, nouns, verbs that are not forms of "to be").

We avoid using one-hot encoding of words is to speed up computation more quickly and allow for considering more words.

For each prediction, we also provide the model with comments made by users who satisfy the following two criteria:

1. They previously rented of the specific garment in this size
2. They previously rented other items the current user fits in.

If we could not find at least 20 past transactions from the specific garments, we would consider transaction records of other garments of the same size, type, and rented for the same occasion. At the max, 150 transaction records are considered.

For each word in the 500-word dictionary, they are mapped to a 128-dimensional vector. All words from the included reviews are averaged.

We then attached a fully connected 256-unit ReLU layer and a Sigmoid output unit to predict fit information.

c. Summary Model

This is a multilayer neural network with 128 input units, 2 layers of 256 hidden ReLU units, and a Sigmoid output layer.

The first 64 of the input units are connected to the last hidden layer of model part (a), and the latter 64 are connected to the last hidden layer of the model part (b).

The model is trained in the following way:

First: model part (a) and model part (b) are trained on their specific inputs and labels separately.

Then: model part (a) and (b) are connected to model part (c) to be trained together. At this time, the weights of the model (a) and (b) are locked, and only model part (c) is receiving updates.

We proposed this model mainly due to the sparsity of the data we have (i.e. the number of items is astoundingly vast, and we assume no prior knowledge of the item beyond our database of transaction records). Even though either one of part (a) or part (b) can serve as standalone models for prediction, we have to assume reasonable variability between each piece of clothing and users while accommodating the privacy concerns of the users. Therefore, we combined them into part (c). Note that we will compare the performance of our model with part (a) and part (b).

We designed naïve heuristics to process the data at model runtime so that we can represent the data easily to fit our model to it. The design of the data preprocessing relies heavily on our common sense (e.g. turtlenecks would depend on very different body features than leggings, and clothing rented for a vacation might be fitted differently than those meant for work). We believe the confidence vector $\vec{C}$ provides a reasonable summary of the difference in sizes. A larger element in $\vec{C}$ means that the user's body measurements are closer to that of the other users who fit in the same garment, and therefore, we are more confident that the user fits. Otherwise, a lower confidence element means the user is less likely to fit.

We also wanted to account for the sparsity of our data with two separate models based on different information. They are essentially trained separately and were combined a single neural network that might be trained to distrust the output of either of the models and produce a more nuanced output.

We derived the model after several failed attempts. We first tried to use the simple bag-of-words model for textual understanding but found it too computationally expensive for our personal computers. We also tried using cosine similarity of the body feature vectors but found the inconsistent scaling of each feature and the unavoidable missing feature problem to terrible at prediction.

We also failed when our data terribly overfits. This was evident in the test-validation loss curve. We then increased regularization term and reduced the hidden layer size and depth. The problem was relieved.

We did not encounter scalability problems since all of our features rescaled to between 0-1.

We eventually trained our models using ADAM optimizer, with a learning rate of 5e-4, with minibatches of 1024 transactions and 15 epochs. We used a L2 regularization constant of 1e-2. We used Cross Entropy Loss as the loss function for back propagation and ADMA.

We used 6:2:2 for train, validation, and test data split. Only transactions in train are used in our prediction process.

We compared our model to its sub-models (part (a) and part (b)), as well as the ALFM model.

Compared to our model, however, the sum-models are less stable when predicting instances with fewer historical transactions and have trouble dealing with sparse data. We combine the capabilities of both and can yield a much better result. Our model, however, is more computationally expensive.

ALFM is more accurate compared our model. However, we can more easily transfer our model for more items without extra tuning. With ALFM, when adding more items/users, one would need to retune the model's hyperparameters for optimal performance.

4. Related Work

The dataset comes from "Decomposing the appropriate semantics for product size" [1]. [1] provides the suitability of the user's clothes, and the user's evaluation, rating on the product and other personal information. This provides a rich set of data, where review text information and this explicit rating can be used to build language models, such as [2] use review text information to resolve limitations, such as user privacy and incomplete information. Similar datasets also used for fit prediction include the ModCloth dataset also mentioned in [1].

[2] also proposes a way to predict clothes fit, which is considered state-of-the-art. First, they establish an aspect-aware topic model and apply it to the review text to model user preferences and product features from different aspects and estimate the user's aspect importance to the product. Then, the importance of the aspect is integrated into a novel aspect-perceived latent factor model (ALFM), which will learn the potential factors of the user and the item based on the score. In our prediction, using roughly the same idea to our model part (b), we used bag-of-words instead of aspect-aware topic model to train our data. Then we used logistic regression to predict our data.

Moreover, [3] also provide their way to prediction the product. They believe that because retailers have different sizes on different brands, the catalog sizes allocated by retailers are different, so they define two threshold parameters b_1 and b_2, which divide the continuous scale into three parts, corresponding to three Suitable categories: suitable, small and big. But they also found that they could not recommend products in the offline evaluation of the model. Therefore, they input the learned potential features of customers and sub-products into standard classifiers, such as Logistic Regression Classifier and Random Forest Classifier, to generate fitting predictions, which is quite similar to our part (a) model. We create effective functions to integrate the various attributes of users, and then make predictions to obtain more effective expectations.

Overall, for this dataset, the scant research on this problem offer similar conclusions compared to our model, despite having different specific implementations of the models. Beyond the three research papers mentioned above, we failed to find many research papers that delve into the exact problem of predicting fit of clothing from transaction data and review texts.

5. Results and Conclusions

Our baseline model received a BER of 0.285, while our prediction model got a BER of 0.249. Alternatively, if we only use model part (a) for prediction, we would get a BER of 0.256, and if we only use model part (b) for prediction, we would get a BER of 0.273.

The overall model offers marginal improvement on part (a) performing individually but substantial improvement over part (b) performing individually (2.7% and 8.7% respectively in terms of BER) and the baseline model (an improvement of 12%).

The reason part (a) performed reasonably well compared to the overall model results from the fact that the majority of the users in the dataset filled out their body measurement information.

Our overall model was able to improve on that by taking in additional information from model part (b). It is able to use the word representations to extract more detailed information from the reviews.

Model part (b), however, has its limitations. We hypothesize that this resulted from our inability to rely on context to determine meaning of the sentences. We found some comments that would appear ambiguous to our model, including one that states that

"This dress is definitely short as people have already said but it is so beautiful! I would have done better in the 6 since I am small up top."

This would be confusing because it contains "small," and "short" all in one sentence even though the dress is too big for the user. Also, many words that would have been useful for word embedding are filtered out due to rudimentary processing of words and the limitations in computing power. Some non-standard word variations, like "TIIIIIIIIGHT" or "HAWT" would be easily understood by humans but hard to account for when construct features.

We also compared our model to the ALFM mentioned in [1]. They did not provide a BER for their model. However, we are able to compare accuracy of reporting. Their model has a 76% accuracy on the test set while our model has a 72.2% accuracy on the test set. Our model did not perform as well as the ALFM model because the ALFM has two advantages:

1. It creates embedding for both users and items, which gives them model more information on how to process data regarding the users and items. Our model is agnostic in this respect.
2. It has a more complicated language model. This helps the model extract more information for textual data. Our model is much less successful in this respect.

Since our model uses gradient descent, it only has 2 manually tunable parameters: the minimum number of transactions and the maximum number of transactions allowed for each prediction. The minimum makes sure that all predictions includes a large enough set of input data while the maximum helps conserve computational power and prevents the less important reviews drowning out the more important reviews (since the embedding/confidence score are averaged).

Work Cited

[1] Decomposing fit semantics for product size recommendation in metric spaces
Rishabh Misra, Mengting Wan, Julian McAuley, *RecSys*, 2018

[2] Zhiyong Cheng, Ying Ding, Lei Zhu, and Mohan Kankanhalli. 2018. Aspect-Aware Latent Factor Model: Rating Prediction with Ratings and Reviews. In Proceedings of the 2018 World Wide Web Conference (WWW '18). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 639–648. DOI: https://doi.org/10.1145/3178876.3186145

[3] Misra, Rishabh. "Would This Clothing Fit Me?" Medium, Towards Data Science, 7 June 2020, towardsdatascience.com/would-this-clothing-fit-me-5c3792b7a83f.