An Exploration in Clothing Fit Prediction

Ruiqi Zhu, Lejia Zhang, Wenyu Zhong, Qingxin Li

1 DATASET

Our group is going to analyze Rent the Runway data [1] under the "Clothing Fit Data". The dataset is kindly provided by Professor Julian McAuley at University of California, San Diego. Our focus is to tackle specific aspects of this dataset and predict the fit of clothing items, gaining insights into customer preferences and making better clothing recommendations.

Our project goes beyond assessing how well clothes fit. We're also delving into the reasons behind rentals. By doing so, we aim to offer recommendations that are not only size-appropriate but also occasion-specific. This exploration involves a close examination of customer reviews and an understanding of the context of rentals. Our goal is to address the common challenge of fit accuracy in online shopping while enhancing our recommendations' relevance and personalization. We believe that a detailed analysis of customer feedback and rental purposes will provide us with valuable insights into consumer behavior, ultimately aiding in improving customer satisfaction in the online retail space. This project represents a step towards a more intuitive and responsive approach to online clothing rentals.

The dataset has 192,462 entries, showing a lot of different interactions. In total, there are 192,544 transactions involving 105,508 users and 5,850 different items. Every entry in this dataset tells a story about a customer renting a clothing item. It's not just basic stuff like the user ID and item ID; there's a lot more to it. In our analysis of the RentTheRunway dataset, we're looking at a rich collection of data points. The dataset includes a variety of information like fit, user_id, bust size, size, age, item_id, weight, rating, rented_for, review_text, body_type, review_summary, category, height, and review_date. This range of data gives us a comprehensive view of each customer's experience with their rented clothing.

Below are graphs for our analysis and interesting findings. To specify, size is a key factor for predicting fit, from the fit distribution graph (figure 1) we observe that about 73.6% of customers say the cloth they bought fit, about 13.4% of customers say the cloth they bought is small, and about 12.99% customers say the cloth they bought is big so that most people buy a fit cloth in their online shopping. Also, the category of clothing, ranging from formal to casual wear, informs us about the variety of styles and their relevance to specific occasions.

[1] https://cseweb.ucsd.edu/~jmcauley/datasets.html#clothing_fit

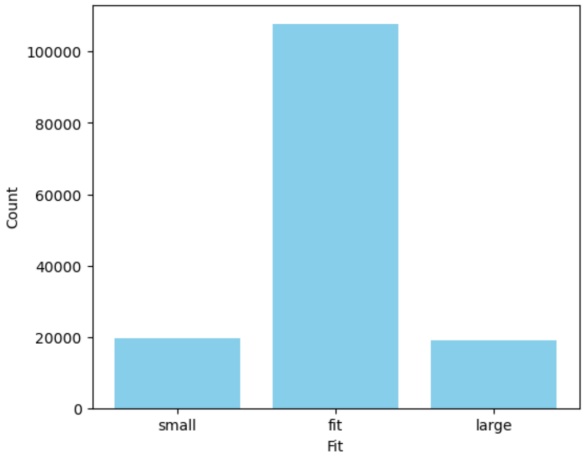

Figure 1: Fit Distribution

From the occasions and category analysis figure (figure 2), we find that on most occasions, dress is the most popular item that customers like. Also from the occasions distribution figure (figure 3), weddings, formal affairs, and parties are the top 3 occasions that customers shop online for. We have also analyzed the rating distribution (figure 4) to check how the customers would rate their orders, and most customers would like to rate a high score like 8 or 10. Review_text gives

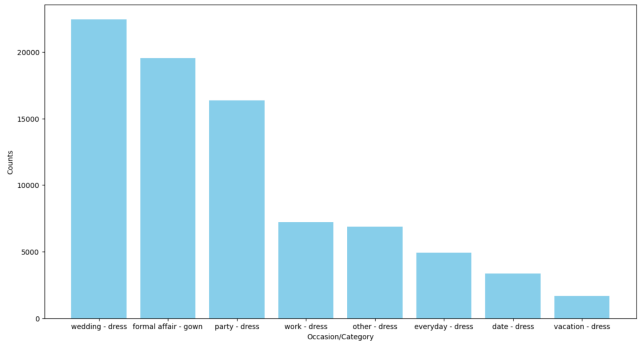

Figure 2: Occasions and Category Analysis

us certain customers' feedback in their own words, showing what they liked or didn't like about their rentals. This helps us understand their experiences. Also, we can be informed from the review_text what occasion are the clothes used for. Rented_for tells us the occasions for renting the clothes, like weddings or vacations. This helps us figure out why

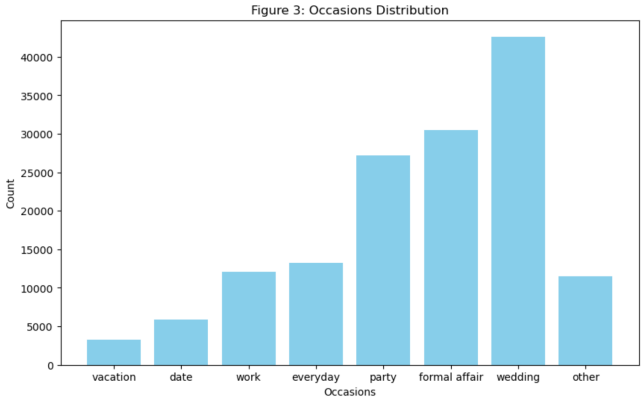

Figure 3: Occasions Distribution

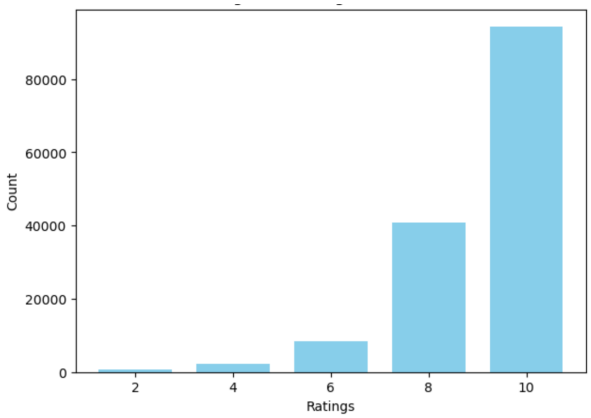

Figure 4: Rating Distribution

customers pick certain items and can guide us in making better, occasion-specific recommendations in the future.

2 PREDICTIVE TASK

We performed our predictive task on predicting the "fit" of a cloth that brings satisfactory recommendations for the users in the "*Rent the Runaway*" website based on the personal preference of the customers based on their measurements(Ex: weight, height, etc), clothing occasions(For work, vacation, etc), and the size of the clothing. Considering these factors in our predictive task can allow the users to have an improved recommendation experience in the "*Rent the Runaway*" website when buying their desired clothes.

2.1 Data Cleaning section:

To maintain data integrity, we will exclude any entries that do not contain all 15 variables. Following this data-cleaning process, we are left with 146,381 entries. This sizable dataset remains adequate for comprehensive testing and analysis.

```
all_cloth[0]
```

```
{'fit': 'fit',
 'user_id': '420272',
 'item_id': '2260466',
 'weight': '137lbs',
 'rating': '10',
 'review_text': "An adorable romper! Belt and zipper were a little hard to navigate in a full day of wear/bathroo
m use, but that's to be expected. Wish it had pockets, but other than that— absolutely perfect! I got a million
compliments.",
 'review_summary': 'So many compliments!',
 'category': 'romper',
 'height': '5\' 8"',
 'size': 14,
 'age': '28',
 'review_date': 'April 20, 2016',
 'bust_size': '34d',
 'rented_for': 'vacation',
 'body_type': 'hourglass'}
```

Figure 5: First Glance of Data

Add some description to describe features, and standardize the dataset columns, similar to the notebook example

2.2 Here's a breakdown of the features before standardized:

- `review_text`: The text of the user's review. This is unstructured text data and can be utilized for sentiment analysis or feature extraction (e.g., using TF-IDF). submission's unique ID on each page of the work.
- `review_summary`: A summary of the `review_text`. Like `review_text`, this is also unstructured text and can be processed similarly.
- `height`: User's height, typically in the format 'X' Y"'. This should be converted to a consistent numerical format, like inches or centimeters.
- `bust_size`: The user's bust size, is typically a combination of a number and a letter.

In addition to cleaning missing values, we have standardized our data by unifying measurement units. First, we addressed the `bust\_size` feature: it was divided into two separate components, `bra\_size` (numeric) and `cup\_size` (alphabetic). Using regular expressions, we extracted numeric and alphabetic parts independently. Subsequently, we transformed the height measurements: originally presented in a feet-and-inches format (e.g., '5' 8"'), they were converted into total inches. This was achieved through a custom function designed to parse and sum up the feet and inch values into a singular inch measurement. This standardization process has led to a dataset with more uniform and analytically advantageous features, proving to be particularly useful for our subsequent model construction.

2.3 Baseline Model:

(1) The model always predicts fit when looking at the clothing to every customer as a baseline, since we want to have a basic estimate for the accuracy of the dataset and need to improve based on the accuracy of the baseline model.
(2) The model predicts fit depending on the common size range of any type of cloth: if the size is between the

threshold of size 5 and size 40, the model predicts as "fit". If lower than the threshold range, the model predicts it as "small". If higher than the threshold range, the model predicts it as "large". The size of the clothing is one of the important variables to determine whether a cloth will fit the customer accurately to provide sufficient baseline recommendations.

We will first use numerical features on our prediction models. Then separately using the text features on our prediction models. Finally, we will try using several combination of both text and non-text features, measuring their performance, and then use the best combination of features using classification models such as logistic regression(One-hot encoding involved with non-text features) and SVM(Support Vector Machines) as they are efficient models that can be used to help us create predictions on the recommended clothings based on the both text and non-text features.

To assess the validity of our model's predictions, we will be using the accuracy rate in the following form: [number of correct in model prediction / total number of predictions]. Using the sklearn train_test_split function, We divide the whole dataset into 3 parts: Train set takes up 80% of data, Validation set takes up 10% of the data, and lastly Test set takes up 10% of data. We will train our models on the training set, tune the hyperparameters on the validation set before final evaluation on the test set.

3 MODEL

We proposed the following models to use: Logistic Regression, LSVM(Linear Support Vector Machine) as this predictive task is more associated with the classification problem of clothes that correctly fit each user in the Rent the Runway website. With the combination of both text and non-text features, simple regression models like linear regression might not be useful in determining the correct fit for each user as we have to consider several non-linear and non-text features like user's reviews, the preferred occasion of cloth for each user, and etc. Also, LSVM is a useful model for us as it deals with datasets that have a large number of features with text and non-text features combined, like this one from the Rent the Runway website, LSVM can still perform well. It can manage complex decision boundaries, which is helpful when dealing with diverse user preferences from the user's reviews, clothing occasions, size preference, etc.
Based on our classification models, we have selected 3 different feature representations and try to compare which ones are more effective:

- Use a mix of non-text features and simple text features like occasions, and categories of clothing such that they can be one-hot encoded and evaluate their impact and effectiveness of these features in conjunction with non-text features for predicting clothing fit.
- Use tf-idf to extract 1000 unigram features from `review_text`
- Use tf-idf to extract 1000 unigram features from `review_summary`

3.1 Accuracy comparison

Feature 1(Logistic Regression): 0.7382
Feature 2(LSVM): 0.803
Feature 3(LSVM): 0.7554
Based on the result of accuracy measures using different features and model combinations: Feature 2, which uses LSVM, has the highest accuracy at 0.803, followed by Feature 3 with an accuracy of 0.7554, and Feature 1 with an accuracy of 0.7382.

The accuracy for Feature 2 is higher than other features due to two reasons. First, the tf-idf method pulls out important words in each user's reviews more uniquely to better identify each user's preferred fit of clothes. Second, the logistic regression approach heavily relies on converting simple non-text features into numerical representations. However, the addition of these features does not substantially influence the predictive accuracy of the model. Consequently, the significance of these non-text features in improving predictive measurements is limited when utilized alongside logistic regression. Therefore, when using LSVM and Feature 2 to process our predictive modeling task, LSVM is particularly adept at handling complex decision boundaries, making it suitable for datasets with a mix of text and non-text features. Its capability of managing diverse user preferences using text brings better performance in terms of accuracy compared to Feature 1 and Feature 3, since Feature 3 is a condensed version of Feature 2, and lacks the depth and richness derived from the more comprehensive Feature 2.

3.2 Relevant Baselines

Baseline 1 accuracy: 0.7360
Baseline 2 accuracy: 0.7105
All models perform better than Baseline Model 2 because Baseline Model 2 only uses one numerical feature size to determine the "fit" of clothes to users in Rent the Runway. As we have predicted, using a combination of text and non-text features will perform better in terms of accuracy using logistic regression compared to simple linear regression with only non-text features from Feature 1. Also, Feature 2's usage of LSVM's ability to manage these intricate boundaries, especially when considering user reviews, clothing occasions, and size preferences, results in a notably higher accuracy compared to Baseline Model 2. As for Baseline Model 1, due to its simplicity and the overall fit distribution of the dataset being mostly fit, some features might not be as effective in terms of performance in the predictive task.

3.3 Optimization

We can optimize our models in many ways: For tf-idf we can use stemming/lemmatization for word pre-processing during the feature building process to enhance the effectiveness of tf-idf. Stemming involves reducing words to their root form, while lemmatization involves converting words to their base or dictionary form. This process can help in consolidating similar words and reducing feature dimensionality. Expanding the maximum number of features beyond 1000 in the tf-idf extraction process allows the model to consider a larger vocabulary, potentially incorporating more relevant terms or expressions from the user reviews. For logistic regression and LSVM we can impose a stricter balance between data, hypertuning parameters using grid-search CV using cross-validation sets.

4 LITERATURE

4.1 Existing Dataset

We are using a dataset that was provided by professor Julian McAuley at University of California, San Diego. It's located in his website *Recommender Systems and Personalization Datasets*. The dataset is under Clothing Fit Data called *Rent the Runway. Rent the Runway* is used for women to rent clothes for different occasions. The application of the dataset in our project is focused on analyzing customer preference throughout customer's review and size, and enhancing the precision of *fit* predictions. In our project, we also utilize a lot of machine learning methods.

4.2 Similar Dataset Has Been Studied

Regarding the study of similar datasets in the past, the Amazon Reviews Dataset[2], as provided by Anusha Bellam on Kaggle, is a notable example. This dataset that encompasses a broad range of products including fashion items, offers a comprehensive view of customer feedback. It includes columns such as marketplace, customer_id, review_id, product_id, product_category, star_rating, review_date, helpful_votes, and etc. This data provides a multifaceted view of customers opinions, covering aspects from product rating to detailed review content.

The utilization of the Amazon Reviews Data in past research has primarily been in understanding customer feedback and preference across various product categories. Studies have applied Natural Language Processing(NLP) techniques, particularly sentiment analysis to parse the customers' review, extracting customer sentiments such as positive, negative and neutral towards products, providing overall satisfaction level. Additionally, the rating and votes have been used to gauge the popularity and perceive quality of the products. Moreover, to enhance personalized recommendation systems, researchers employ methods such as linear regression, logistic regression or classification algorithms to predict the correct size and fit for customers. By correlating historical purchase data, product sizes, and customer feedback, these models aim to provide more accurate and personalized size recommendations, reducing the likelihood of returns due to sizing issues. Also, Exploratory Data Analysis serves as a foundation step in deriving customer insights. Researchers investigate the distribution of product ratings, which can offer immediate insights into overall customer satisfaction and product quality perceptions. Additionally, EDA enables the exploration of correlations between various variables, such as the relationship between product ratings and fit accuracy.

[2]https://www.kaggle.com/datasets/anushabellam/amazon-reviews-dataset

4.3 Exploring Contemporary Techniques

In addition to the method used in our report, there are various other technological and data-driven approaches. Size Recommendation engines are specialized algorithms used primarily in e-commerce for apparel to suggest the most fitting clothing sizes for customers. The process involves collecting extensive data, including customer body measurements, past purchase and return history, and size preferences. This data is then analyzed using machine learning algorithms to identify patterns and inform the model, which is subsequently implemented to provide real-time size recommendations. These engines utilize features such as customer body measurements, historical purchase data, return information, and direct customer feedback on fit and size. On the positive side, they enhance the customer experience by providing tailored recommendations, reducing return rates due to better-fit predictions, and potentially increasing sales and customer loyalty. Conversely, the accuracy of these systems is heavily dependent on the quality and comprehensiveness of the data collected.

4.4 Comparing Current Research Findings with Personal Observations

Our method for predicting clothing fit on the "Rent the Runway" website, while sharing the data-driven and personalization focus of standard recommendation engines, differs in some approaches and potentially in outcomes. Unlike broader recommendation engines that utilize complex algorithms and a variety of data points, our approach integrates specific factors like clothing occasions and employs a unique baseline model that initially predicts 'fit' for every customer. Additionally, the use of a clear size threshold (between sizes 5 and 40) to determine fit, small, or large is a distinct feature, differing from the nuanced, algorithmically derived size-fit

predictions of standard engines. These methodological differences, including the focused approach and simplicity in size prediction, suggest that our findings and conclusions may vary from those of existing recommendation engines. The integration of context-specific criteria and a straightforward baseline accuracy assessment potentially leads to different patterns in recommendations and variations in accuracy compared to more generalized, algorithm-based recommendation systems.

5 CONCLUSION

5.1 Result

As discussed in Section 3, we determined that the model yielding the highest prediction accuracy involves using Linear Support Vector Machine (LSVM) with Feature #2, specifically the user's *review_text*. In comparison, a logistic regression model employing non-text and basic text features only marginally enhanced the accuracy over the baseline model, with an approximate increase of 0.02.

Extensive hyperparameter tuning on combination with LSVM resulted in a peak accuracy of approximately 80.3(±0.1)% on the test set. This performance significantly surpasses that of baseline models, underscoring the efficacy of our model in accurately predicting the correct clothing fit for users on Rent the Runway. The model's success is largely attributable to its reliance on textual user reviews, as opposed to merely considering the average fit response across all users.

For the optimal hyperparameters in our best-performing model, we initially implemented `TfidfVectorizer` with the following settings:

- `max_features = 2000`: Limits the feature set to the top 2000 most frequent words or tokens in the text corpus.
- `ngram_range = (1, 2)`: Includes both unigrams and bigrams.
- `stop_words = 'english'`: Excludes common English stop words from the tokenized features.

Our modeling process revealed that Feature 1, in combination with Logistic Regression, achieved only a marginal improvement. This was partly due to insufficient weighting of simple text features amenable to one-hot encoding. Additionally, we did not fully incorporate all text features, including *review_text* and *review_summary*, since we were unable to devise an effective method for one-hot encoding the entire text corpus. Conversely, Feature 2, when paired with LSVM, performed admirably, meeting our expectations. Here, the *review_text* was particularly valuable, with its significant word features identified using tf-idf to ascertain the "fit" of clothing for each user. Feature 3, despite also employing LSVM, did not achieve the same level of success as Feature 2, mainly because the *review_summary* provided considerably less textual content for tf-idf to extract relevant word features.

5.2 Summary

We conclude that the following model: Feature 2 with LSVM performs the best because simple Logistic Regression using one-hot encoding for simple text features may not best suit our dataset as we did not consider fully on the weight of each feature in the dataset. With user's reviews(review_text), it might mention more important word features that determine whether or not the clothing is "fit" for each user. The variety in word features surpass the limitations of solely relying on simplistic text features encoded in a binary manner(Logistic Regression).

In summary, our findings highlight the significance of nuanced text analysis, specifically emphasizing the importance of user reviews (review_text) in determining clothing fit. This demonstrates the inadequacy of simplistic text feature encoding methods and urges the need for more sophisticated approaches to capture the essence of user sentiments using more efficient combination of text and non-text features towards the "fit" of clothing to each user.

6 WORK CITED

Rishabh Misra, Mengting Wan, Julian McAuley. (2018) Decomposing fit semantics for product size recommendation in metric spaces. RecSys

Yuxiao Ran, Haoyu Huang, Ka Ming Chan. 2019. CSE 158 Assignment 2 Report. https://github.com/YuxiaoRan/clothing-fit-prediction/blob/master/report.pdf

Anusha Bellam. 2022. Amazon Reviews Dataset. (October 2022). https://www.kaggle.com/datasets/anushabellam/amazon-reviews-dataset

Divine inner voice. 2023. A Guide to Measuring Recommendation Engine Accuracy: Best Practices and Pitfalls. https://medium.com/@divine_inner_voice/a-guide-to-measuring-recommendation-engine-accuracy-best-practices-and-pitfalls-4a65af9a9244