Predicting Clothing Ratings and Fits For Users Across Several Machine Learning Paradigms

Introduction

In product recommendation systems, rating prediction of reviews is a procedure often valuable in the recommendation of personalized products for users. By designing a system able to accurately predict the rating that a given user might rate a given item, a website can suggest items that have high predicted ratings to users in an effort to achieve more sales. This type of system could lead to higher revenue and greater user retention, which could in turn lead to larger amounts of data that could create even more accurate models for even higher sales.

Within the subspace of clothing products, yet another lucrative prediction is a fit prediction of clothing pieces on users. If data regarding whether or not a particular item fits a particular user exists, recommender systems may be built to recommend items that the system predicts will fit the user.

This project aims to explore the process of building product recommendation systems that provide rating predictions and fit predictions. Within these models, we look to leverage both interaction data and features of users and products in the context of clothing. We will be exploring the design and implementation of several machine learning paradigms such as linear regression, logistic regression, Jaccardian similarity, and TF-IDF to create predictive systems for ratings and fits. These systems will be compared to baseline systems that may leverage simpler prediction strategies in order to quantify the magnitude of impact more complex models have on accurate prediction.

Dataset

Throughout this project, we use data from RentTheRunway[1], which is a renting based clothing platform for women. The data, in JSON format, includes features like fit feedback (can either be 'Small', 'Fit', and 'Large'), consumer measurements, review text, rating, and others. Fortunately, the dataset we acquired required no cleaning and features were ready in JSON format for us to use.

Exploratory Data Analysis

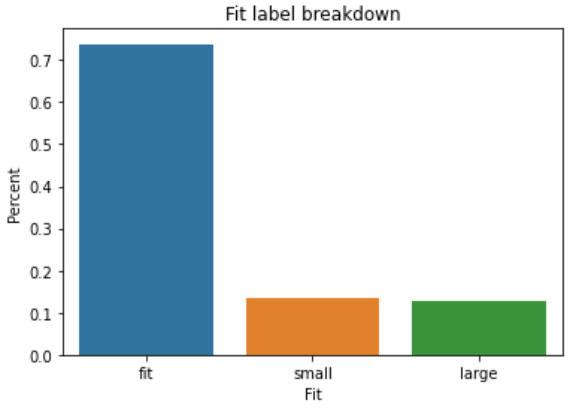

Figure 1: A breakdown of fit types. Given the label imbalance, our model will have to be cognizant of this and use hyperparameters (such as a "balanced" class weight) to ensure valid results. An imbalanced model theoretically could predict all items fits to be 'fit' as that is the most popular label, and still score an accuracy of 72%.

[1]https://www.renttherunway.com

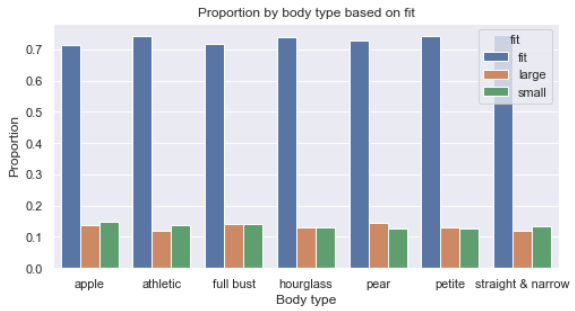

Figure 2: A breakdown of fit grouped by body type. We hypothesized that different body types could amount to varying proportions of fit. However, this figure clearly refutes that hypothesis. It seems that body type does not play a role in determining the item's fit. This result led us to not include body type as a feature in our fit prediction model.

	rating			
rented for	date	everyday	formal affair	other
fit				
fit	9.208365	9.144728	9.395493	9.317797
large	8.500535	8.308729	8.701059	8.634266
small	8.197987	7.746820	8.497739	8.458333

	rating				
rented for	party	party: cocktail	vacation	wedding	work
fit					
fit	9.280411	8.0	9.256868	9.362987	9.138252
large	8.538614	NaN	8.403587	8.709906	8.233507
small	8.396081	NaN	8.134146	8.663126	7.950027

Table 1: Average ratings grouped by 'fit' type (index) and 'rented for' (columns). Further analysis of this table reveals that 'wedding' had the highest average rating over all fit types, while 'everyday' had the lowest. Note that 'party:cocktail' was not considered when finding averages as it had missing values for 'large' and 'small' fits. After looking at the reviews of items that were bought for a wedding, it makes sense why the ratings were higher. Reviews frequently mentioned how incredible a given item looked at the wedding and the compliments it received. This demonstrates that validation from social circles plays a huge role in how a person rates an item. The more one's social circle likes a garment of clothing, the more one might like it too.

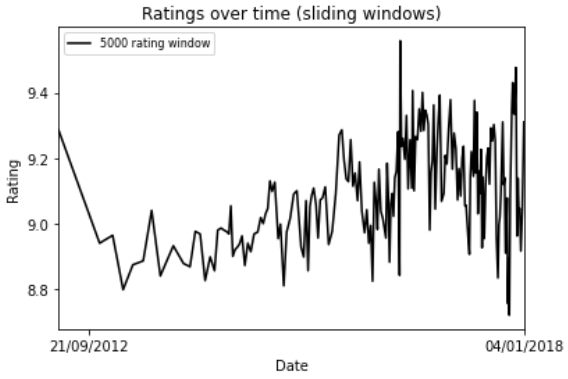

Figure 3: Ratings over time from 2012 to 2018. The average review dips around 2013 but recovers to about 9.2. Dates around 2016 to 2018 have significant ups and downs, so it is worth exploring that in a separate figure. However, figure 3 does tell us that there are no significant differences in average ratings over the 6-year time period.

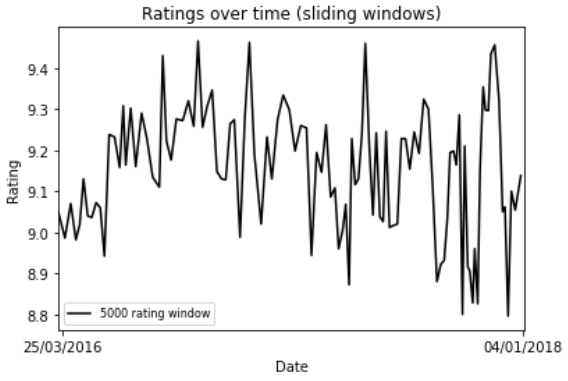

Figure 4: Ratings over time from 2016 to 2018. Although it seems like there do seem to be highs and lows month to month, the pattern is not clear. Even if a sinusoidal feature to encode the month of the year were to be implemented, we believe it would not make the model better. Instead, we are going to examine average rating differences by day and potentially encode day into our model.

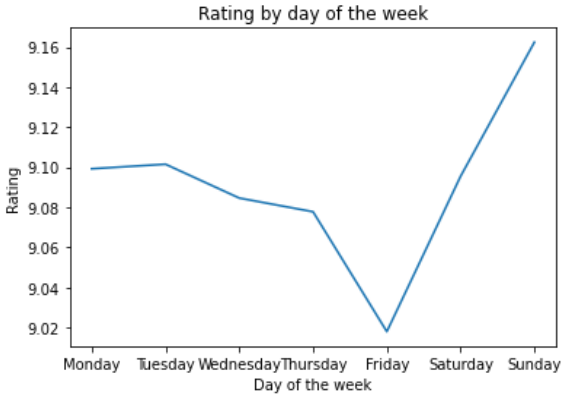

Figure 5: Average rating by day of the week. Clearly, Sunday has the highest average rating out of any day of the week. Monday to Thursday are relatively stable, with a bigger drop occurring on Friday. Saturday recovers the Friday drop and Sunday increases again. This finding makes sense, as people generally tend to be happier on the weekends, thus rating higher. As a result of this promising finding, we encoded the day of the week in our rating prediction model.

Image 1: Word cloud for all reviews' text. Note: The word cloud does not include English stop words (the, is, and, etc.) or punctuation. As expected, the most common words include 'dress', 'wore', 'wear', and 'size', which all have a neutral sentiment. Words with either a positive or negative sentiment (in the sense of fashion and/or clothing) include 'loved', 'beautiful', 'compliment', 'perfect' 'tight', and 'issue'.

Related Work

In the machine learning research space, a few studies have been done on product recommendations in the clothing set. In a deep learning study from 2019, Sheikh, Guigores, and others explore using interaction data and features to build a deep-learning-based recommendation system for clothing fits and sizes in A Deep Learning System for Predicting Size and Fit in Fashion E-Commerce. The system created boasts significantly higher performance than baseline systems used, and shows evidence that combining interaction and feature data can lead to performant recommendation systems. Another study, Decomposing Fit Semantics for Product Size Recommendation in Metric Spaces, which was performed by McAuley and others in 2018, also explores the space of clothing size and fit recommendation, using the same dataset that this project will be used for building and testing the models. This project explores a more mathematical approach using several features of users and clothing such as review text to carry out the predictions, in a pattern the paper refers to as "latent metric learning". The work serves as another data point for how user and product features can be used to build effective models.

Our project aims to design a simpler machine learning framework than these two studies for prediction, in an effort to understand how effective the strategies taught in class are in carrying out practical prediction with real-world datasets. As such, we will not be leveraging the same systems used in these two studies. Instead, we will be using their promising results as evidence that the design of using interaction data and feature data can increase performance for predictive systems.

Methodology

Rating Prediction

In this project, we aimed to create a performant predictor using both feature data and interaction data. To reach such a performant predictor, we started with a baseline predictor and worked our way up towards more complex predictors that leveraged more features or more developed machine learning models. The baseline predictor for ratings was a predictor that simply predicted the global average for every input. The next predictor used the number of exclamation marks in the review text to predict the rating. The third predictor utilized 4 features from the user and item data, which were the user's weight, height, and age; and the item's size. If any of the first 3 terms are missing from the review, the value corresponding to the attribute would be substituted for the prediction with the global average user weight, height, and/or age, respectively. The fourth feature model developed combines the prior 2 models by including the exclamation mark count of each review be the 5th feature used for prediction, while the other 4 features and their implementation from the aforementioned feature model are carried over into this one unchanged.

Interaction Data

For predicting ratings, we also wanted to leverage interaction data to build a proposed model. We explored 2 different models in the project in order to predict ratings - one that uses Jaccard similarities between users and items, and the second one that uses user and item biases.

Our proposed model for ratings combines the 5-feature model for leveraging feature data with the user-and-item-bias model for leveraging interaction data. The interaction data model is used as the main model since it produced the lower MSE out of the other feature data models, but when both the user and item bias terms were zero (i.e. the user and item were not seen before in the training data), we use the feature model to predict rating based on the user and item features.

Fit Prediction

Another type of model we wanted to consider was one that would be able to predict the fits of different items on users. In this design, we examined feature models and their performance on fit prediction.

Each of these models leverages the logistic regression machine learning approach with the "small" fit being labeled the value 0, the "fit" fit labeled as 1, and the "large" fit labeled as 2. Each model used this same encoding, but with different feature implementations. The baseline, similar to the rating prediction models, guesses the same value for every single input - the most common fit seen in the training data. Additionally, the second, third, and fourth models use the same feature implementations as the second, third, and fourth models used in the rating prediction. What we did differently in this experiment though is leverage a text-analysis model to do fit prediction. We went with both a bag-of-words model and a TF-IDF model used on the review text of reviews to build a logistic regression for prediction. The relative accuracies of both models are described in the results section.

Results

Our results comparing the MSEs and the accuracies of the various models used are presented in Figure 6 below. When comparing the different rating prediction models we employed, we observe that the baseline model of

returning the global average rating for every input has the worst MSE at ~2.01. The feature model using the user's weight, user's height, user's age, and item size as the four bias terms had the second worst MSE at ~2.006, which came to our surprise. The feature model using only the exclamation mark count in the review text of each review had an MSE of ~1.944, which is a noticeable improvement over the aforementioned models. However, out of all the rating prediction models, the one with the best MSE was the feature model which included the exclamation mark count of the review text as the fifth bias term, along with the four previously stated terms remaining unchanged. The MSE of this model was ~1.942, which while being an improvement over the exclamation mark count model, isn't as large of a difference as we had expected.

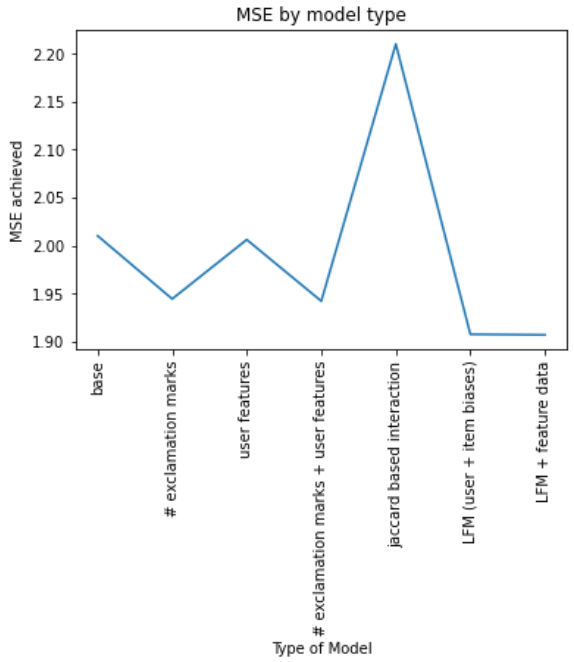

Figure 6: MSEs for rating prediction by model type. The Jaccard Similarity approach had the highest MSE, while the LFM approaches had the lowest MSEs.

The two interaction models that we explored had a stark contrast between their MSE values. The model using Jaccard similarity between the users and items gave an MSE of ~2.21, which was significantly worse than we had initially anticipated. Meanwhile, the user-and-item bias model performed significantly better with a validation MSE of ~1.907, and with a combination of interaction data and feature data, maintains an MSE of ~1.907.

Looking at the results of the accuracy of the various models in regard to fit prediction, determining the efficacy of certain models over others becomes a more nuanced process. Although the MSE of the exclamation mark count only model was somewhat worse than the model factoring in exclamation mark count only as the fifth bias term in addition to the four others stated earlier in this section, the opposite is true when comparing their accuracy rates for fit prediction. The former had an accuracy rate of ~45.05%, while the latter had an accuracy rate of only ~27.4%. The model factoring in only the four bias terms had an accuracy rate of ~25.55%. However, all of these models developed still had a greater accuracy rate than the baseline, which was only ~13.11% accurate.

Using the bag of words and TF-IDF approaches both resulted in far greater accuracy rates than any of the aforementioned approaches. The bag of words model resulted in an accuracy rate of ~75.31% while the TF-IDF approach was ~80.28% accurate, the highest of any of the models we explored in this project.

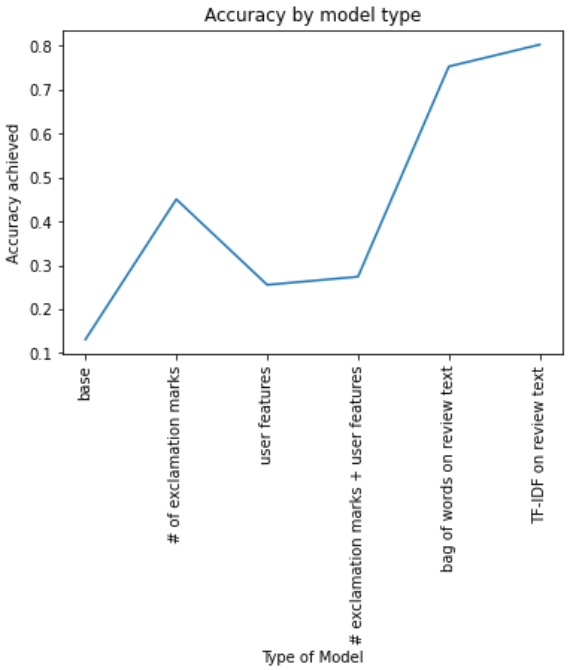

Figure 7: Accuracies for fit prediction by model type. The TF-IDF model on review text performed the best, with an accuracy of 0.802.

Existing Literature Results

Both existing papers on clothing prediction (*Sheikh* and *Misra*) used different evaluation metrics for their models. *Misra* used average AUC as their metric, but the results followed our model in performance in terms of features used. They found that adding more latent variables increased the average AUC score for their models. The models in *Sheikh* focus on providing a scalable solution using a neural network. They also found that as the NN was scaled up and fed more data, it performed better.

Our models followed the same general progression. Our best-performing model for rating prediction used a latent factor model and encoded user features as well. Our fit prediction, models with more features did perform slightly better, but the best model had a single feature, TF-IDF vectorizer on review text. However, it can be argued that a TF-IDF is not a single feature but many, since a vectorized TF-IDF is just a feature matrix model of many features.

Analysis

When observing the results of the different rating prediction models, we see that the feature model of the four bias terms actually has a worse MSE than simply counting the number of exclamation marks. This could be due to the number of exclamation marks in a review text possibly being more correlated with the excitement or positivity which the user has for that particular item. This could be more predictive of a positive rating for the item than the user's height, weight, age, or item size, as some users that may have the same height and weight may have different body types or shapes, and as such, may have different opinions about a piece of clothing of a particular size.

We also noticed in our fit prediction models that machine learning algorithms performed on the review text had much better performance than those done on the user and item features. This may be related to what was mentioned above that properties of a review's text has a much stronger association with a user's sentiment towards an item than by just using the raw user's and item's features.

Lastly, we noticed that leveraging interaction data tended to lead to better performance of models compared to just using feature data. Again, this contributes to a general trend that using plain feature data is not sufficient to produce a performant model, and more data regarding either a review's text or a user's interactions with other items are helpful to make better predictions and recommendations.

Conclusion

We performed an analysis and comparison on different machine learning algorithms in the context of rating prediction and fit prediction for users and items. From our results for this dataset, we have concluded that interaction data provides better quality predictions than simply using plain feature data, and that a combination of interaction and feature data leads to insignificant gains in performance for the algorithms being used. Additionally, using algorithms on a particular review's text generally led to higher performance than using feature data, because analysis of a review text's sentiment can show strong signs of whether a user might rate a particular item highly or whether or not a particular item might fit a user.

The comparisons and analyses performed in this project strictly leverage algorithms taught in this class, but future research should be done that considers different models more primed toward text analysis, regression, and classification problems.

References

A Deep Learning System for Predicting Size and Fit in Fashion E-Commerce. In Thirteenth ACM Conference on Recommender Systems (RecSys '19), September 16–20, 2019, Copenhagen, Denmark. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3298689.3347006
Abdul-Saboor Sheikh, Romain Guigoures, Evgenii Koriagin, Yuen King Ho, Reza Shirvany, Roland Vollgraf, and Urs Bergmann. 2019.

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