Clothing Rating and Fit Prediction Analysis | CSE 258 (MGTA 461) | Assignment 2 Final Report

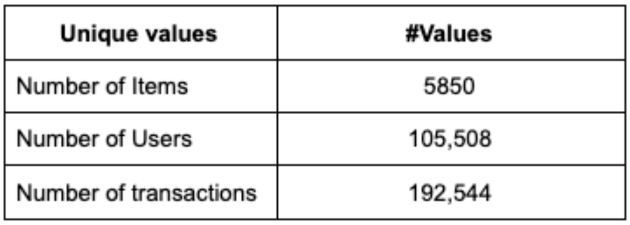
Agarwal Mrinal (A59007629) | Gupta Chirag (A59006557) |
Khanna Akanksha(A59010687) | Verma Rohan (A59013544)

1. Introduction: In this report we aim to work with the clothing fit data - Rent the Runway data set[1]. The aspects covered in this report starts with the exploratory data analysis followed by data cleaning. Next, the report elaborates the assumptions to be taken based on the identified issues and gaps with the data to avoid hindrance while working on the later parts that covers the modeling aspect. Finally, two tasks have been explained and analyzed which includes a.) Rating prediction using various technique, b.) Clothing Fit Prediction (Fit/Not Fit) based on user features. Each aspect reveals really important information for an organization to compete in this fast growing and competitive industry.

2. Exploratory Data Analysis: We are exploring the data set - Clothing Fit – Rent the Runway. The website renttherunway.com is a B2C platform where customers can select plans and rent out various outfits for different occasions. It contains clothes and accessories for different occasions ranging from party clothing to business formal clothing.
Based on the data exploration we looked at various summaries such as number of user and items, average of height, age, weight, size, and rating at various levels including body type, purpose of renting and categories.

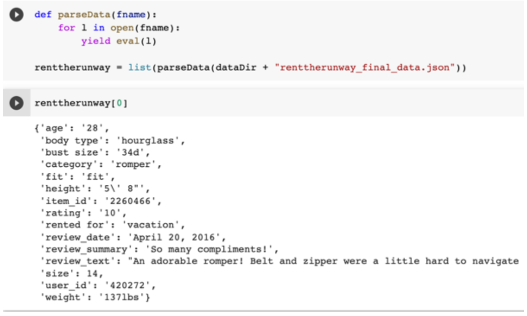
2.2 Data Summaries:

2.1 Data Description and Features Definition: The data set describes the users and the features of the user such as age, heights, review text, data of review, rating, item etc. Refer to the Figure 2.1. The dataset comprises of the following:



Unique values	#Values
Number of Items	5850
Number of Users	105,508
Number of transactions	192,544

There are columns consisting of the following features:
Fit: Fit of the product – fit, small, large
Bust Size: Bust size of the user (30A, 34D, etc.)
Category: Types of clothing item (dress, romper, etc.)
Body type: Body type of user (pear, hourglass)
Weight: Weight of user in lbs
Rented for: Purpose of renting clothing item
Age: Age of user
Review_text: Review given by user



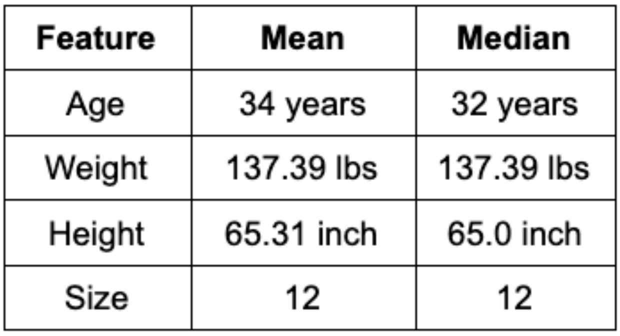
```
def parseData(fname):
    for l in open(fname):
        yield eval(l)

renttherunway = list(parseData(dataDir + "renttherunway_final_data.json"))
```

```
renttherunway[0]
```

```
{'age': '28',
 'body type': 'hourglass',
 'bust size': '34d',
 'category': 'romper',
 'fit': 'fit',
 'height': '5\' 8"',
 'item_id': '2260466',
 'rating': '10',
 'rented for': 'vacation',
 'review_date': 'April 20, 2016',
 'review_summary': 'So many compliments!',
 'review_text': "An adorable romper! Belt and zipper were a little hard to navigate
 'size': 14,
 'user_id': '420272',
 'weight': '137lbs'}
```

Figure 2.1 The figure shows the raw data and elements present in the first entry.



Feature	Mean	Median
Age	34 years	32 years
Weight	137.39 lbs	137.39 lbs
Height	65.31 inch	65.0 inch
Size	12	12

Some of the interesting plots have been shown in the report. The following plot shows the body type to average size for the data.

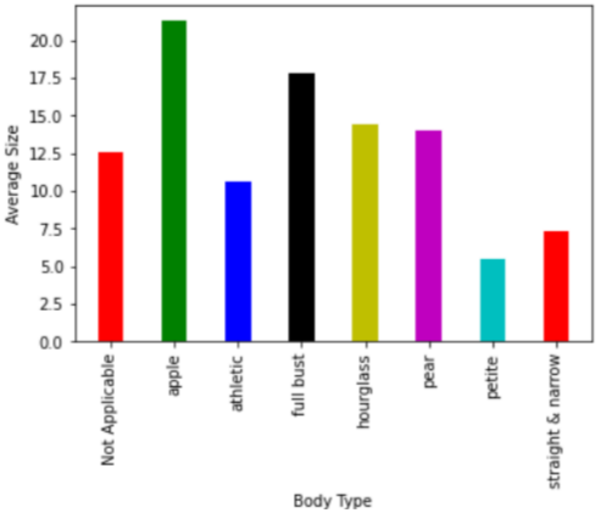


Figure 2.1 The bar plot shows the average dress of the size available based on various body types.

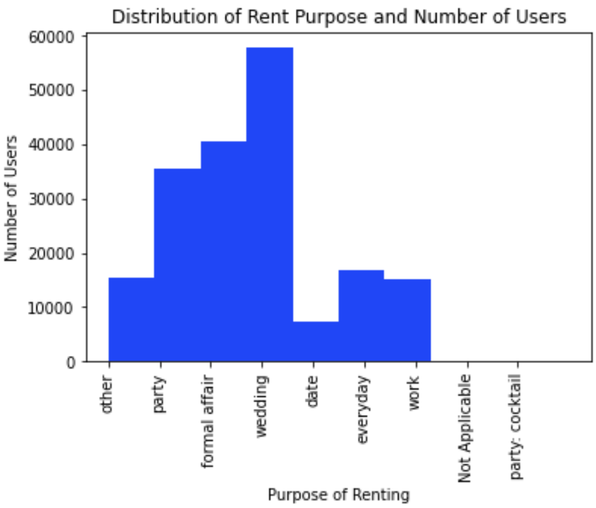


Figure 2.2 The bar plot shows the number of users based on Purpose of Renting

The figure above captures an interesting insight where we see that most number of users rented a clothing for the purpose of wedding followed by formal affair.

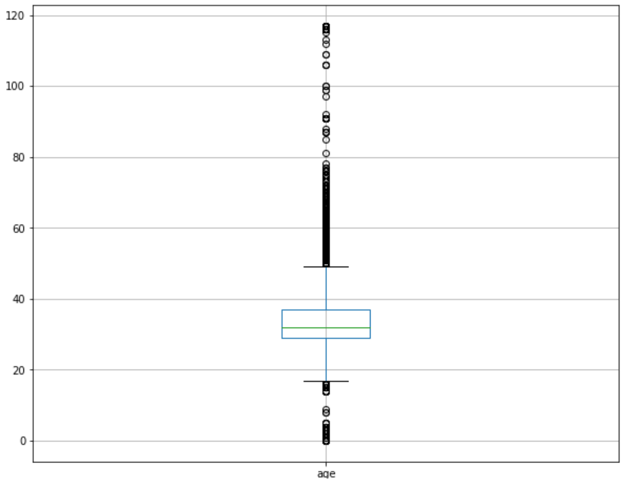


Figure 2.3 The bar plot shows the number of users based on Purpose of Renting

To analyze any sort of outliers in the data, we made boxplots for various features. The figure above shows an example of age where we see age outlier values of as high as 120 years for some users.

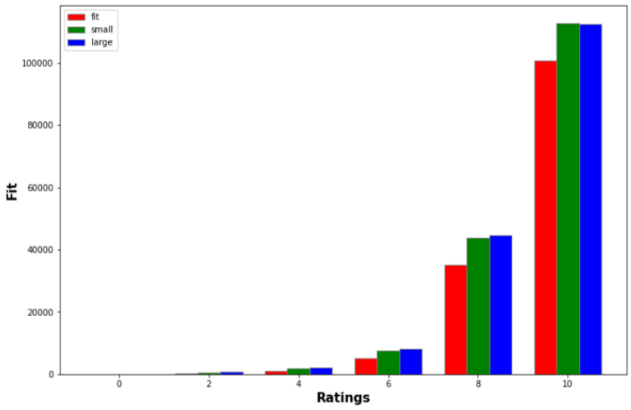


Figure 2.3 The bar plot shows the distribution of data based on Ratings.

The above figure shows an interesting aspect where we can clearly see that the data was skewed and majority of the users have given higher ratings.

2.3 Data Feature Adjustments

The data cleaning steps included:

- Removal of NULL values in Bust Size, Weight, Body Type, Height, Age, Rented For.
- **Bust Size**: Replace NaN values with “Not Applicable”.
- **Weight**: Split the entries, save the numerical part as integer type. Replace the NaN values with the mean weight.
- **Body Type**: Replace NaN values with “Not Applicable”.
- **Height**: Convert the height from feet and inches to inches, save the data as an integer. Replace the NaN values with the mean height.
- **Age**: Set the maximum value of age at 90 and replace all the values above with the mean age.
- **Rented For**: Replace NaN values with “Not Applicable”.
- **Rating**: Convert the data to numeric.

2.4 Final Data: Touch upon the clean data set that we will be using for the model now.

3. Rating Prediction: Rating prediction may be very helpful to the organization (Rent the Runway company). Knowing the dependable drivers to the rating can help the organization work on specific areas to provide better service to its customer base. In this world of competition an important aspect to stand out is to have good relationship with the customers.

In this report we attempt to explore different technique and models to predict rating and then compare the results to understand the best models for predicting the rating given by the user. The models and different approaches are described as follows:

3.0 Baseline Model

To have a better understanding of the accuracies of our models that we have build, we created baseline model using average ratings as the baseline. For cold start, we are using global average ratings. In cases where the user information is unavailable, we decided to use item average ratings. In places where item value is not available, user average rating was used.

3.1 Linear Regression Models (OLS): The linear regression method – ordinary least square was one of the approaches used for prediction of rating. Based on the knowledge from classroom and literatures referred online, various model features were tried for predicting the ratings. The main metric for evaluating the model was based on the Mean Squared Error (MSE) value.

The data was split into test and train based on random generator in the ratio of 70% train and 30% test.

3.1.1 Modeling with Numeric Features: Modeling with feature variables with numerical values were initially used. Model with and without intercept values and different features were tried to compare MSE values and plots of predicted Rating (Y Predicted) and actual Rating (Y Actual) was plotted.

After running multiple iterations of model, we could find a lowest model prediction MSE value for the order of 2.08 - 2.95 value. The actual ratings in the data were from {0,2,4,6,8,10} and the model could not produce good predictions for the lower ratings (0,2 and 4) but was fine in predicting ratings of higher order (8 and 10 mostly).

Various features including i) length of text review by user ii) age of the user iii) user weight iv) user height v) size were used in different combinations. Interaction models with interaction between different features was also tried to predict the rating to minimize the MSE value.

The following model showed the best MSE value of 2.08 relatively.

Rating ~ Intercept + age + weight + height + size + length of review

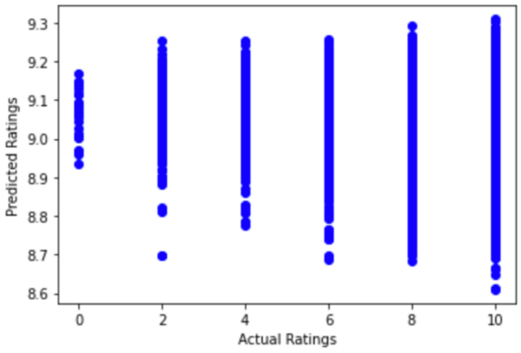


Figure 3.1.1 The figure shows the predicted vs actual rating from the Linear Regression Model with all numeric features*.

3.1.2 Modeling with Categorical Features:

Modeling using feature variables with both numerical variables (described in 3.1.1) along with categorical variable features using one-hot-encoding were also added. The features that were considered for dummy variable (one-hot-encoding) are as follows: i) Review Date (used both Year and Month), ii) Body Type, iii) Category, iv) Rented For, v) Bust Size.

The above-mentioned categorical variables were used either along with the numerical feature variables or without the numerical feature variable. Similarly, with and without intercept models were tried.

After critically analyzing possible set of model iterations, we were able to get MSE values in the range of 1.89 – 2.3 values. We saw the results improved from the models tried earlier that contained only numerical value. However, even after introducing categorical features, the rating values with the values {0,2,4} were not predicted correctly by the model.

The following model showed the best MSE 1.89 relatively.

Rating ~ Intercept + age + weight + height + size + length of review + year(categorical) + month(categorical) + body type (categorical) + rented for (categorical) + bust size (categorical)

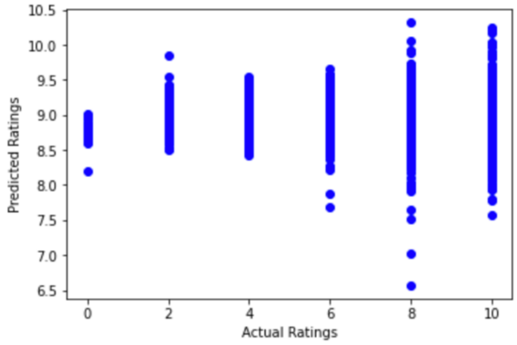


Figure 3.1.2 The figure shows the predicted vs actual rating from the Linear Regression Model with all numeric features and some categorical featured used in one-hot-encoded style*.

3.2 Linear Regression Model (Ridge Regression):

From the OLS model, the best MSE score we got was 1.89. To improve our MSE further and make better predictions we move to Ridge Regression. For this we created a sentimental analysis model based on 1000-word bag of word model and with Ridge regression to predict the rating given by the user.

3.2.1 Why are we using Ridge Regression?

Ridge regression is a supervised learning algorithm which is used in estimating the coefficients of regression models when there is a multi-correlation between the independent variables. Due to multi-correlation the usual least squares as a loss function cannot be implemented as they are unbiased. Hence Ridge regression is used as it implements L2 loss function to predict values as close as possible to actual values.
The cost function of ridge regression is shown below:

MIN(||Y-X(theta)||^2 + λ ||theta||^2)

Here, λ is the tuning parameter which is used in order to get the least amount of MSE, i.e., making a better model. We find the optimal value of λ using RidgeCV (). This helps in finding the best value of λ by iterating over an extensive list of possible λ which we create using np.linspace().

3.2.2 What is a bag-of-word model? (BoW)

A bag-of-word model is a collection of most popular words in the selected feature to apply in machine learning models. It is referred to as a bag as it is only concerned with the occurrence of the words in the document. A BoW contains the vocabulary of the words and the frequency of each word occurring in the text.

3.2.3 Implementation of Models.

For our purpose we used unigram, bigram, trigram, combination of unigram, bigram, combination of unigram, trigram, combination of bigram, trigram and combination of unigram, bigram, trigram. All the combinations are made with the similar process of removing the punctuations and putting the whole feature (in our case **review_text**) in lowercase. Based on what we learned in class, we created a function for each model through which we can extract the desired features. This is stored in a list with each input having the frequency of combination for each feature input.
Once, we have the features from a specific bag of word model, we split the data in train, test with a ratio of 70:30. We use RidgeCV() to find the optimal value of λ, and fit the Ridge regression on the train data. Once, we have the fitted model, we make predictions and calculate MSE for each model.

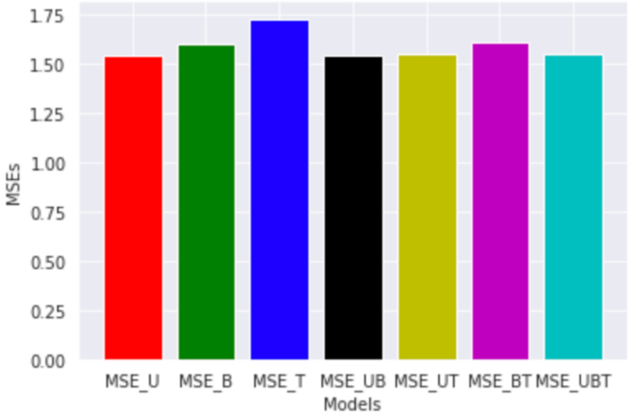


Figure 3.2.1 The figure shows the MSEs of all the models we ran using BoW and Ridge Regression.

We evaluated the above 7 models based on their MSEs, from all of these, the **combination of Unigram and Bigram gave the best performing MSE with the value of 1.54031617**.

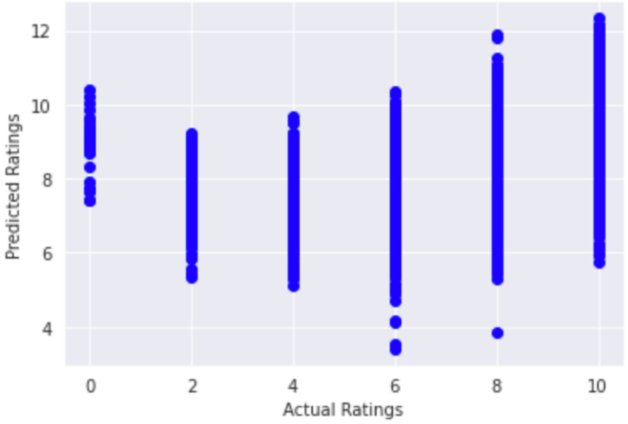


Figure 3.2.2 The figure shows the predicted vs actual rating for the model made from the combination of Unigram and Bigram.

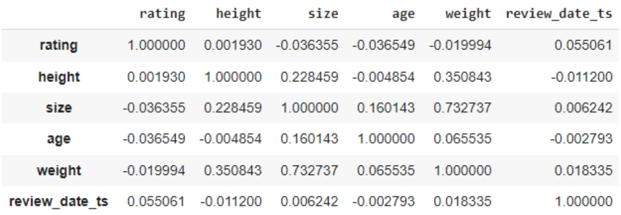
4. Clothing Fit Prediction:

All leading apparel brands and fashion e-commerce organizations emphasize highly on the personalized size and fit recommendation. Predicting the correct fit not only helps in increasing the revenue of the organization but also enhances the customer satisfaction score. Fit prediction is currently one of the biggest challenges faced by all the leading organization in fashion industry. Incorrect fit leads to the return of the product and which further decreases the profit margins. Incorrect fit accounts for over 50% of returns in e-commerce fashion industry.

Our objective is to predict whether the apparel fit to a consumer or not (fit = 1 and no fit = 0) based on the certain features. With the correct fit, fashion industry can battle the plague of losses through returns.

4.1 Co-Relation Matrix

To predict the consumer fit, we have created the co-relation matrix to check the co-relation between the attributes. Our dataset contains the records of the apparels which consumer have rented for various occasions. Each record is having the following information about the consumer like age, bust size, category, body type, fir, height, rented for, review text, review summary, rating, item id, user id. Let’s look at a co-relation matrix with few relevant features.



	rating	height	size	age	weight	review_date_ts
rating	1.000000	0.001930	-0.036355	-0.036549	-0.019994	0.055061
height	0.001930	1.000000	0.228459	-0.004854	0.350843	-0.011200
size	-0.036355	0.228459	1.000000	0.160143	0.732737	0.006242
age	-0.036549	-0.004854	0.160143	1.000000	0.065535	-0.002793
weight	-0.019994	0.350843	0.732737	0.065535	1.000000	0.018335
review_date_ts	0.055061	-0.011200	0.006242	-0.002793	0.018335	1.000000

Figure 4.1.1 The figure shows the co-relation matrix using the rating, height, size, age, weight and review datetime.

We can derive the following insights from this correlation matrix.

- Size and weight are moderately co-related.
- There are slight negative co-relations between Rating & size, rating & weight, rating & weight, review datetime & height and review datetime & age.

Same relationships can be seen in the below heatmap, where numbers 0 to 1 pertain to rating, height, size, age, weight, and review datetime.

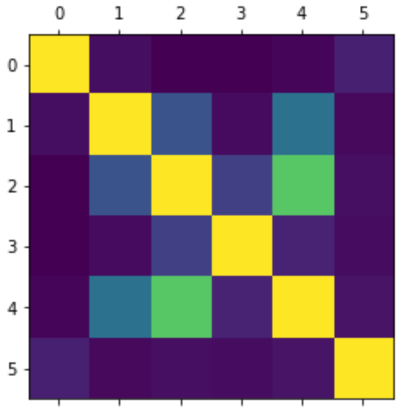


Figure 4.1.2 Heat map graph of the co-relation matrix

4.2 Classification Models

We have created a baseline model and used the various classification models to compare it with the baseline accuracy.

Following are the steps that was followed during the classification model.

- Created features and labels. Used the one-hot encoding for the categorical variables.
- Split the dataset into test and train.
- Run the model on test dataset and predict the “FIT” on the test dataset.
- Calculated the confusion matrix and Balanced accuracy score. The formula for Balanced Accuracy Score is:

Balanced Accuracy = (((TP/(TP+FN) + (TN/(TN+FP))) / 2
where TP, TN, FP and FN are True Positive, True Negative, False Positive and False Negative.

- Created the ROC curve to show the performance of a classification model at all classification thresholds. This Curve plots two parameters:
  - True Positive Rate
  - False Positive Rate

  True Positive Rate (TPR) is a synonym for recall and is therefore defined as follows:

  TPR = TP / (TP +FN)

  False Positive Rate (FPR) is defined as follows:

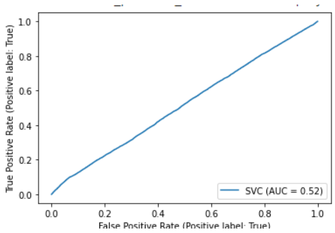
  FPR = FP / (FP + TN)

  An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. The following figure shows a typical ROC curve.
- AUC: Area Under the ROC Curve

  AUC provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example.

Below are the models that we have used to predict the "FIT".

1. Baseline Model
   - We have used the uniform strategy of Dummy Classifier. Score of the baseline model is 50%. AUC of the baseline is 0.52



2. Classification Models

   We have run the Random Forest, SVM and logistic classifier using the below feature set. For the categorical values, we have used the one-hot encoding.

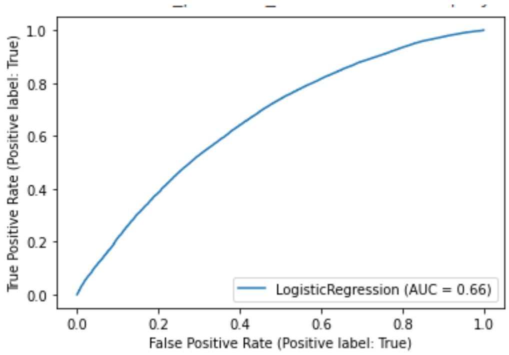
i) Features: All columns
ii) Feature: Rating, age, weight
iii) Features: Rating and Category
iv) Features: Rating and length of review text
v) Features: Rating, body type and bust size **(Proposed Model)**

Confusion Matrix:

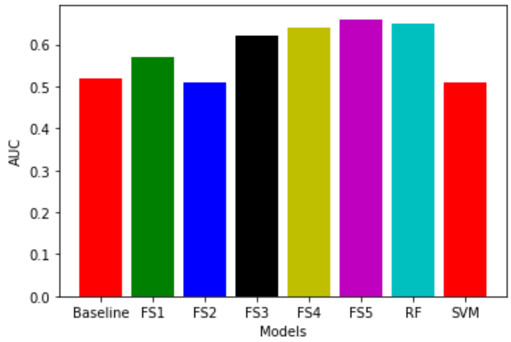
	Actual		
Predicted		Positive	Negative
	Positive	1285	13441
	Negative	1234	41174

Balanced Accuracy Rate: 0.54

ROC Curve & AUC:



We evaluated the above 8 models based on their Accuracy, confusion matrix and AUC score, from all of these, we found out the **feature set containing Rating, Body Type and Bust Size gave the best result having AUC=0.66**



5 Literature Review

[1] We reviewed a research paper that worked on the same data set for Clothing data set prediction that we are working on. The research paper mainly focused on predicting the size recommendation and fit for prediction. The paper proposed an interesting new predictive framework to tackle the predict fit problem. It captures the problem where say two users have similar comment on the size of a dress, however they marked the options in polarity i.e., one may mark oversize and other tight. The paper discusses the topic of ordinal regression to tackle such kind of issues in the data and features. The paper collected two data sets from different websites for analyzing and forming the results and concluded. This research paper is different in working from a limited number of papers already published on this topic in various aspects. The research paper focuses on capturing the fit semantics and also tries to handle label imbalance issues (example described above) using matric learning approaches.

The methodology involved the learning of fit semantics that involved the latent factor model approaches. The next was the metric learning approach that was used to label the imbalances as discussed earlier. The Data sets used were from the two websites RentTheRunway and ModCloth who deals in rented dresses for women. Laten factor models of various types were used in the experiment and one of the best was recommended.

We took inspiration from this model and tried to explore the fit prediction using approaches such as logistic regression model, random forest model and Support Vector Machine models. Further, we extended our work to also predict ratings using linear regression models. For Sentiment analysis models based on Ridge regression and Bag of Word which predicts the rating we took inspiration from a research paper written on Sentiment analysis using n-gram feature selection and combinations. [2] In this they applied Unigram, Bigram and a combination of Unigram and Bigram on Decision Tree Classifier and Kernlab Classifier. From this we learnt, the feature extraction based on Unigram, Bigram and combination of both. We further extended their understanding by adding trigrams and its combinations with unigram and bigrams.

6. Results and Conclusions

The purpose of our project was to predict the ratings given by the user and fit of the clothing based on each user.

6.1 Ratings

We implemented various models based on Linear Regression (OLS) and Ridge regression with BoW. Out of these, Ridge Regression on Unigram and Bigram combination performed the best. One of the probable reasons why Ridge Regression performed better than OLS is due to multi-correlation between the independent variables. Since Ridge uses L2 loss function we were able to solve the problem of multi-correlation. Along with that we used an optimal value of λ which was calculated using RidgeCV(). Rating prediction is important from a business point of view as it helps in understanding the needs of the user and how the company can improve itself for better customer retention.

6.2 Fit

In the context of our work, we ran the several models based on three techniques- Logistic Classifier, Random Forest and SVM. Out of these model Logistic Classification gave out the best result. We can see from Figure 3.2.2, that the Random Forest gave the result close to logistic classification.

7. Future Scope

7.1 Rating Prediction

To improve on that we can build rating prediction models based on similarity functions such as Jaccard Similarity, Cosine Similarity to improve our prediction results.

7.2 Fit Prediction

To improve on the model, we can further extend our work using gradient-boosted tree and multi-layer perceptron.

8. References

[1] Link to Data: http://deepx.ucsd.edu/public/jmcauley/renttherunway/renttherunway_final_data.json.gz

[2]https://ijarcce.com/wp-content/uploads/2016/09/IJARCCE-35.pdf
[3]https://cseweb.ucsd.edu//~jmcauley/pdfs/recsys18e.pdf
Data Set Used: **Clothing Fit Data**

Citation:

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

[2] Awachate Payal, Kshirsagar Prof. Vivek P. Improved Twitter Sentiment Analysis Using Ngram Feature Selection and Combinations