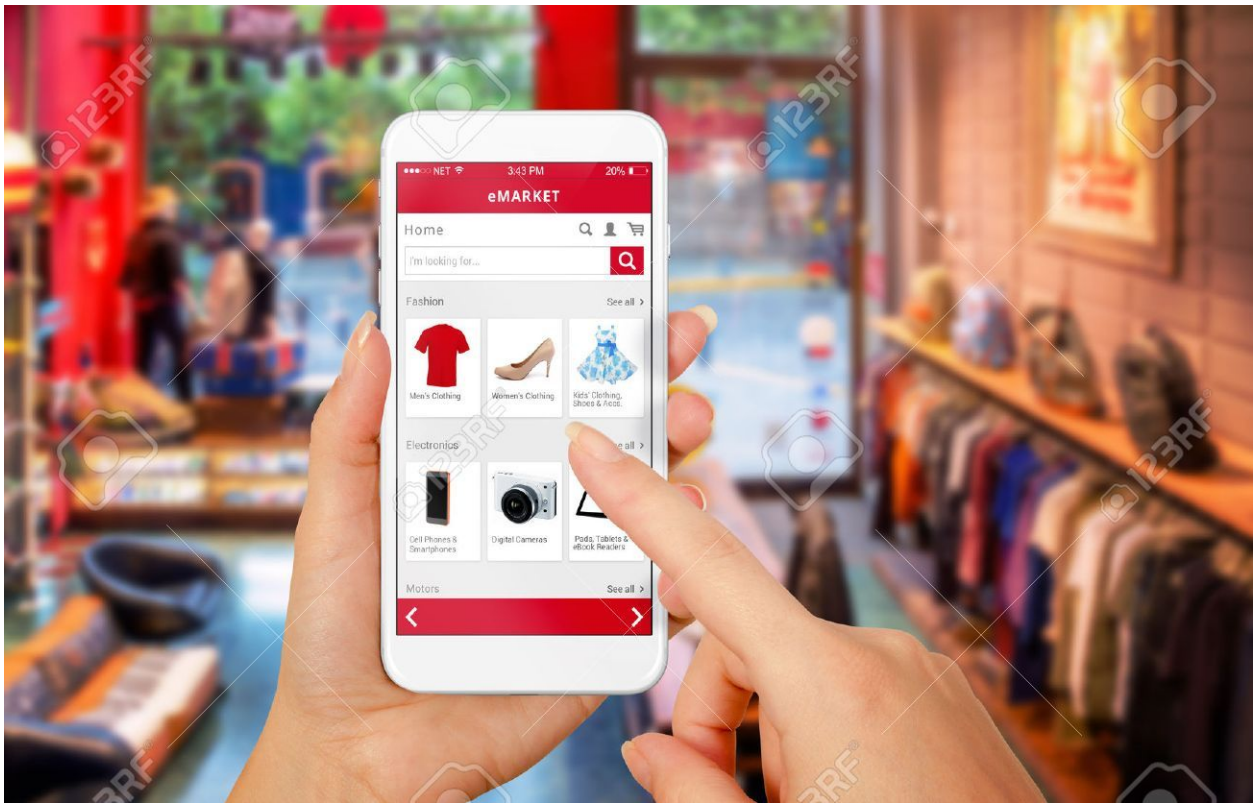


Sentiment Analysis of Women's E-commerce Clothing Review

Data Science Capstone Project 1 Final Report

Rui Cao

2019-11-13

Introduction

With the development of the network, there is an increasing number of people choose to purchase online. Given the shortage of information online, customers are always struggling with issues such as size, quantity, colors, and etc. Thus, an overview of others' reviews can help us get a quick impression of the products.

Understanding customer sentiments not only enhance the efficiency of online shopping but also give the companies an insight as to how the customers like their product. The best way to improve customer experience is by listening to them. The customer's feedbacks always comes in many different forms and in many different languages. Manually reading all customers' reviews simply wouldn't be possible. The best way to solve this issue is to build a technology model to automatically analysis the customers' feedbacks and the retailer can easily working on improving the customer's experience. The retailer also can refine sales and marketing strategies or report the important issues that might not be addressed.

Objectives

In this project, I attempt to understand the correlation of different features based on an anonymized Women's Clothing E-Commerce dataset and indicate whether the review is a positive, negative or neutral sentiment based on recommended, Rating or Positive Feedback factors.

Using Natural Processing Language(NLP) techniques to find the most popular words in the recommended or not recommended review in order to find out which features are most important to customers (color, price, size, etc). Finally, built the machine learning model and using SHAP to find the most predictable words and LIME to interpret the result.

Data Wangling

I use Women's E-Commerce Clothing Reviews on Kaggle Datasets. This dataset includes 23486 rows and 10 feature variables. Each row corresponds to a customer review and includes the features: Clothing ID, Age, Title, Review Text, Rating,

Recommended IND, Positive Feedback Count, Division Name, Department Name, and Class Name.

	Unnamed: 0	Clothing ID	Age	Title	Review Text	Rating	Recommended IND	Positive Feedback Count	Division Name	Department Name	Class Name
0	0	767	33	NaN	Absolutely wonderful - silky and sexy and comf...	4	1	0	Initmates	Intimate	Intimates
1	1	1080	34	NaN	Love this dress! it's sooo pretty. i happene...	5	1	4	General	Dresses	Dresses
2	2	1077	60	Some major design flaws	I had such high hopes for this dress and reall...	3	0	0	General	Dresses	Dresses
3	3	1049	50	My favorite buy!	I love, love, love this jumpsuit. it's fun, fl...	5	1	0	General Petite	Bottoms	Pants
4	4	847	47	Flattering shirt	This shirt is very flattering to all due to th...	5	1	6	General	Tops	Blouses

Figure 1.Dataset Overview

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23486 entries, 0 to 23485
Data columns (total 11 columns):
Unnamed: 0                  23486 non-null int64
Clothing ID                 23486 non-null int64
Age                         23486 non-null int64
Title                       19676 non-null object
Review Text                 22641 non-null object
Rating                      23486 non-null int64
Recommended IND             23486 non-null int64
Positive Feedback Count     23486 non-null int64
Division Name               23472 non-null object
Department Name             23472 non-null object
Class Name                  23472 non-null object
dtypes: int64(6), object(5)
memory usage: 2.0+ MB
```

Figure2. Dataset Information

After loading and general inspect the dataset, we find there is an ‘Unnamed’ column that needs to be removed and missing values in ‘Title’,’Review Text’,’Division Name’, ’Department Name’ and ‘Class Name’ columns. Because the ‘Review Text’ is the KEY factor in this analysis, I will fill the missing values with blank from the ‘Review Text’ column and keep the other columns as it for exploratory data analysis later.

I also add several feature variables for later EDA and NPL analysis:

Review Length - Total length of each original reviewed text

Polarity Score - The sentiment(polarity) score for each original reviewed text

Sentiment - set the sentiment to 3 class: Positive, Negative, Neutral based on Polarity Score

Filtered Review - cleaned revied text after removing punctuation, numbers, and stopwords

Clothing ID	Age	Title	Review Text	Rating	Recommended IND	Positive Feedback Count	Division Name	Department Name	Class Name	Character Count	Sentiment	Polarity score	Filtered Review
767	33	NaN	Absolutely wonderful - silky and sexy and comf...	4	1	0	Initmates	Intimate	Intimates	53	Positive	0.633333	absolutely wonderful silky sexy comfortable
1080	34	NaN	Love this dress! it's sooo pretty. i happene...	5	1	4	General	Dresses	Dresses	303	Positive	0.339583	love dress sooo pretty happened find store im ...
1077	60	Some major design flaws	I had such high hopes for this dress and reall...	3	0	0	General	Dresses	Dresses	500	Positive	0.073675	high hope dress really wanted work initially o...
1049	50	My favorite buy!	I love, love, love this jumpsuit. it's fun, fl...	5	1	0	General Petite	Bottoms	Pants	124	Positive	0.550000	love love love jumpsuit fun flirty fabulous ev...
847	47	Flattering shirt	This shirt is very flattering to all due to th...	5	1	6	General	Tops	Blouses	192	Positive	0.512891	shirt flattering due adjustable front tie perf...

Figure 3. Dataset after Data Wrangling

Exploratory Data Analysis

- ## Correlation Matrix for All Features

  The picture below shows the correlation matrix between each feature variable. Recommend IND and Rating has a strong positive relationship and Rating, Polarity Score and Recommended IND have a slightly positive relationship. some features like Age, clothing ID are unrelated to any other features. I will deep in the analysis of those independent features distribution and correlation features in this chapter.

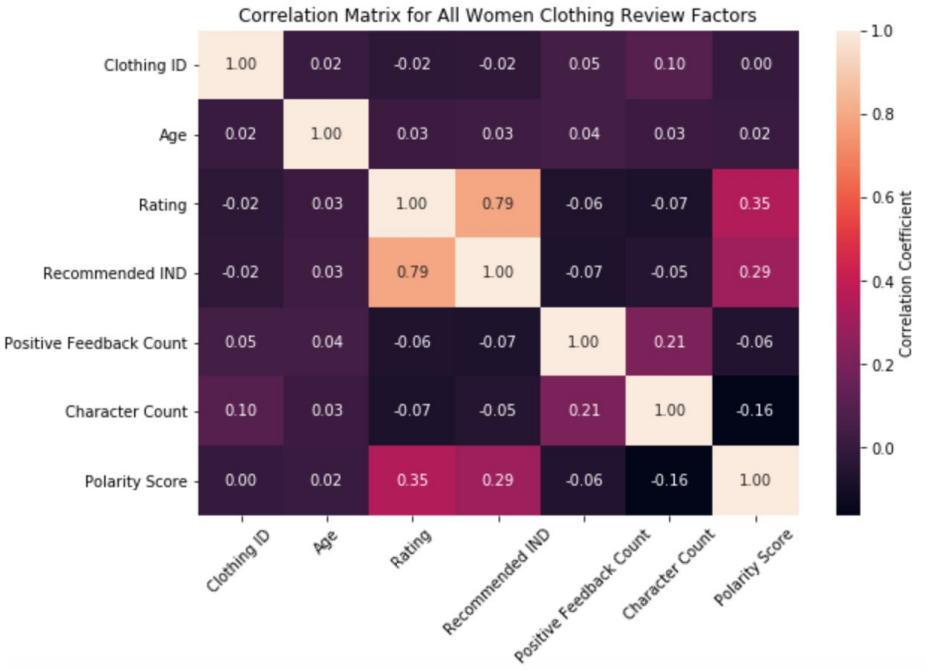


Figure 4. Correlation Matrix

- ## Customer Age Distribution

The first thing I am curious about is the customer's age range. The results really surprised me because the biggest group of reviewing customers is around 38, which I was expecting around 20-35. based on this result, we can suggest that the core market segment for this clothing brand is women between 36 and 50.

Also, there is a slight correlation between Age and Positive Feedback Count, and compared to the plot above, the age group who wrote the review most, are always giving the most positive feedback as well. As there are some high count feedback reviews, I will focus on those texts analyze later.

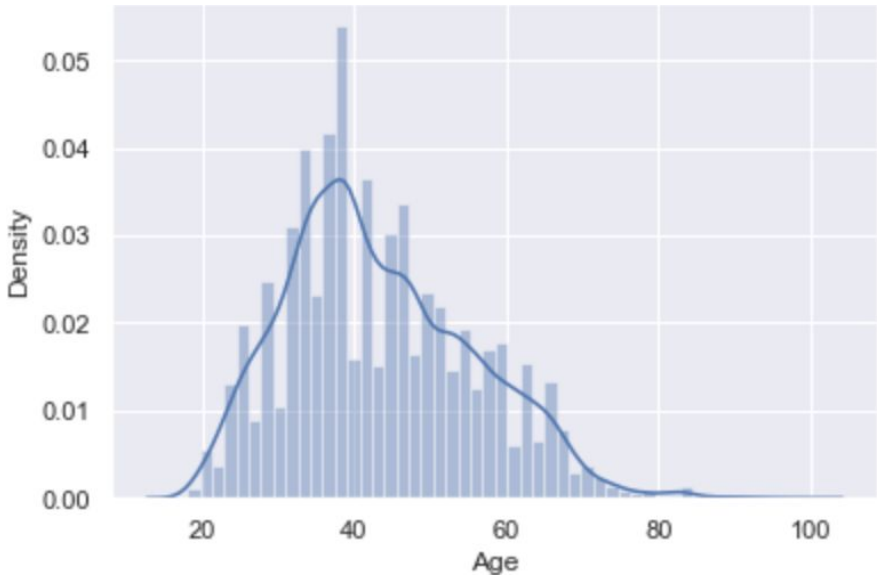


Figure 5. Age Distribution

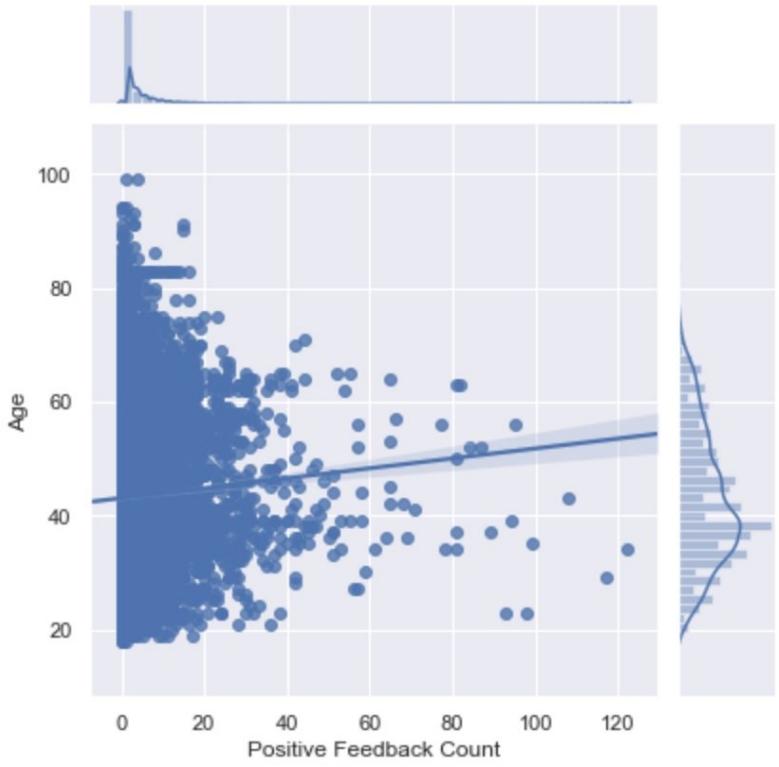


Figure 6.Customer Age vs. Positive Feedback Count Distribution

- ## Clothing ID Distribution

The total number of unique clothing ID rating in this review dataset. Due to a large number of numbers, I will only plot the top 30 clothing ID customer reviewed most. Even though we don't really know the exact item of this Clothing ID represent, however, based on my analysis, we can see that the Top 3 Clothing ID holds a large values of customer reviews, and they are in 'Dresses' and 'knits' Class, which is really useful for the company to target the market. Also, As I see in the dataset, the Top 3 items with higher reviews counts also get the high Rating and High Recommended IND number.

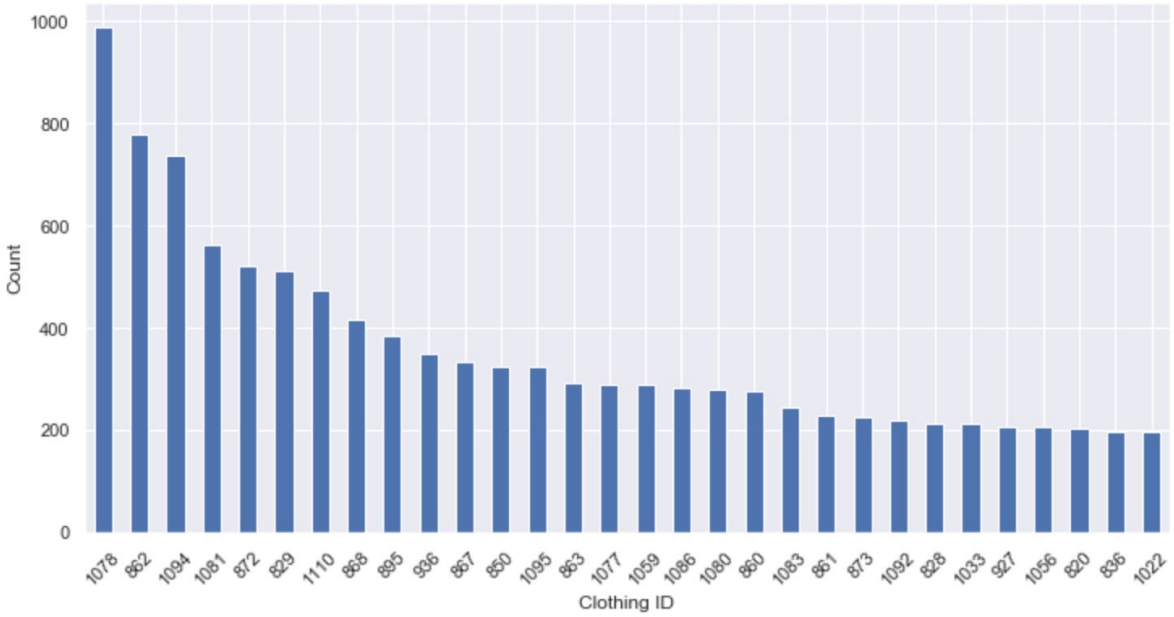


Figure 7.Top 30 Reviewed Clothing ID

	Age	Rating	Recommended IND	Positive Feedback Count	Review Length	Polarity score
mean	42.724800	4.189200	0.818400	2.862400	310.860400	0.256109
std	12.150429	1.104306	0.385592	6.773021	144.836158	0.178384
min	18.000000	1.000000	0.000000	0.000000	16.000000	-0.500000
25%	34.000000	4.000000	1.000000	0.000000	189.000000	0.149074
50%	41.000000	5.000000	1.000000	1.000000	305.000000	0.247036
75%	51.000000	5.000000	1.000000	3.000000	469.000000	0.357143
max	99.000000	5.000000	1.000000	98.000000	504.000000	1.000000

Figure 8. Dataset Describe on Items [1078, 862,1094]

- ## Customer Rating and Recommended Distribution

The majority of the reviews had been rating with a high score(3-5) on this retailer's products and most of them are extremely highly rated as a 5 out of 5. The overall customer satisfaction rate(recommended INR rate) is 82% in 20 classes, which means most people who rate 3 also recommend the item.

As we know from the overall heatmap matrix, there is a strong Correlation between Rating vs. Recommended IND, we can see more clearly in Figure 8 that Customers who rating 5 almost highly recommended, but customers who rating for 1 or 2 are recommended a small amount of the time.

Also, the Dresses and Knits which are the most popular Class also hold more negative reviews. Top, the most popular department has a less dislike ratio than the dress. I will deep to the analysis of the keywords to find which words more represents 'good' or 'bad' review.

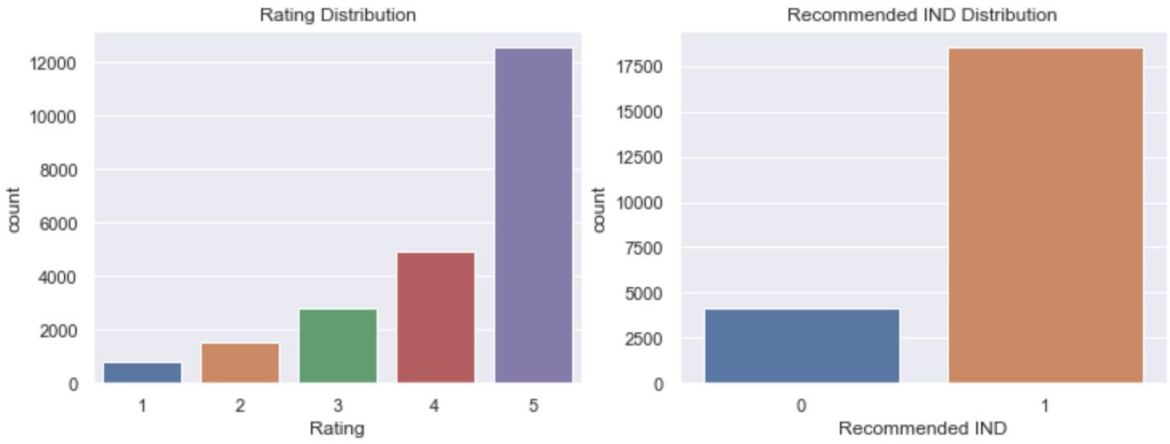


Figure 9.Rating and Recommended IND Distribution

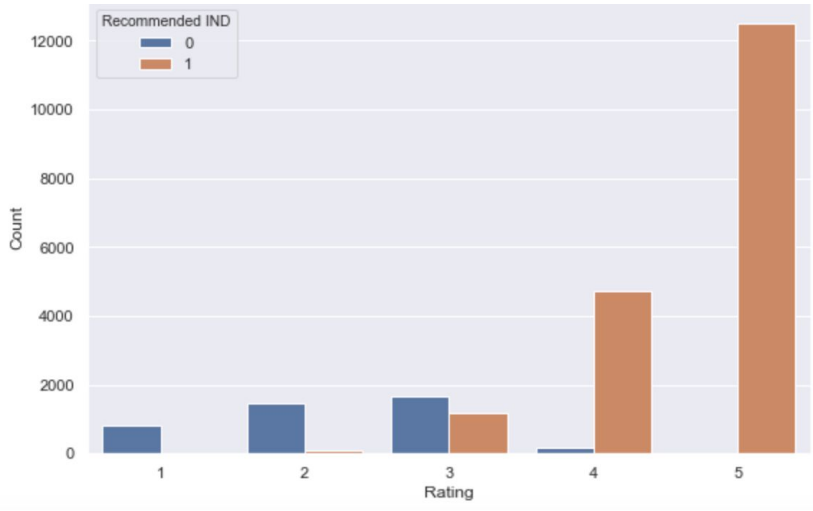


Figure 10.Numbers of Rating by Recommended IND

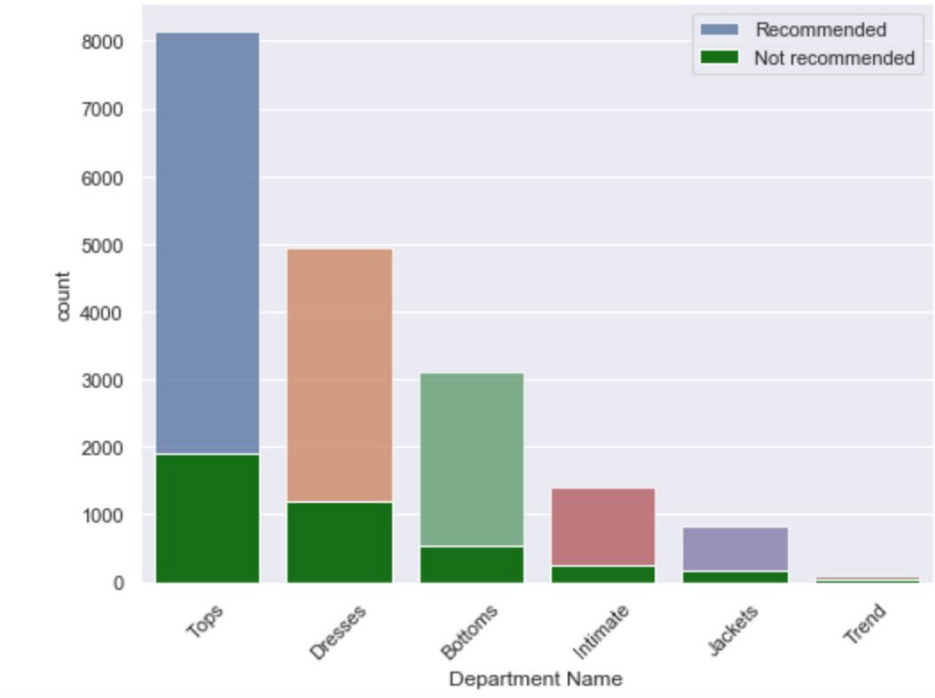


Figure 11.Recommendation Distribution in Department Name

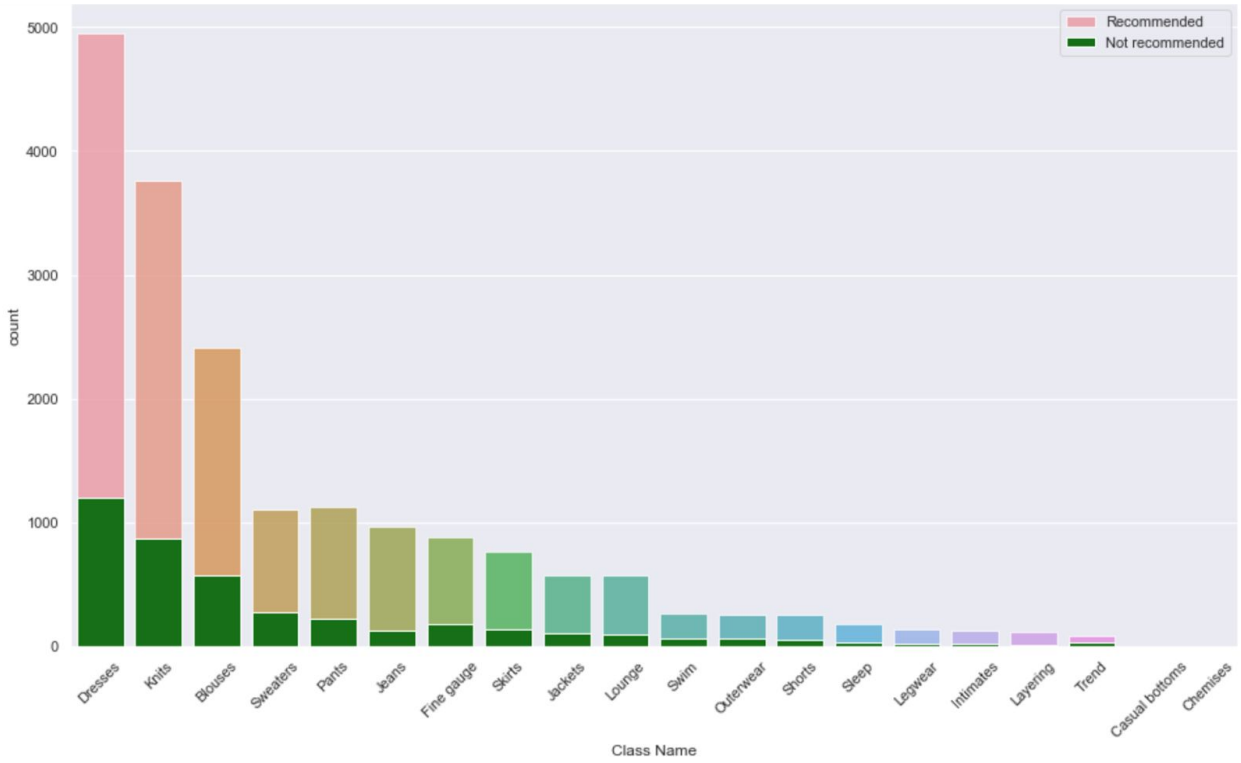


Figure 12.Recommendation Distribution in Class Name

- ## Character Count and Sentiment Analysis

  Compared to the 3 graphs below, we find that customers who give a higher rating and who recommended their product usually write a longer review compared to those who did not recommend it. I also assume that the retailer has a maximum word limit of 500, which caused the spike.

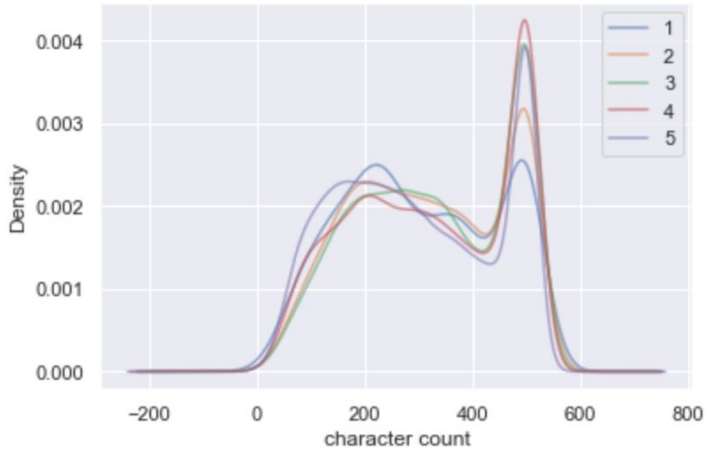


Figure13. Character Count Distribution on Rating

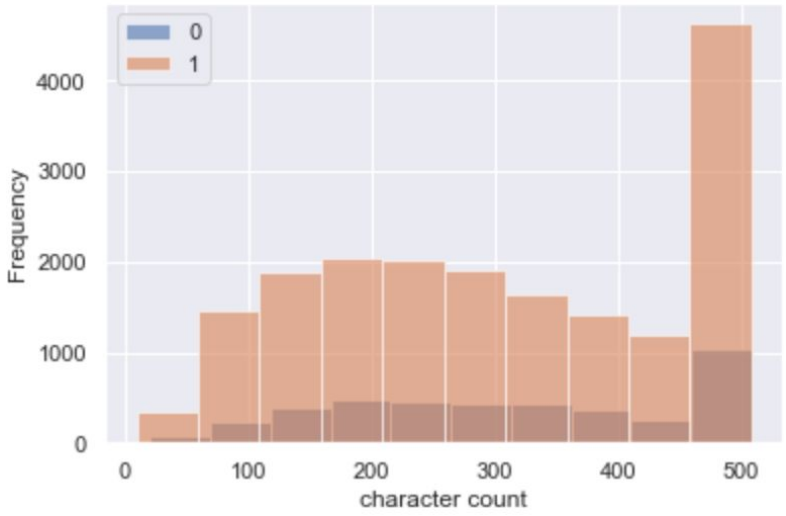


Figure 14. Character Count Distribution on Recommendation

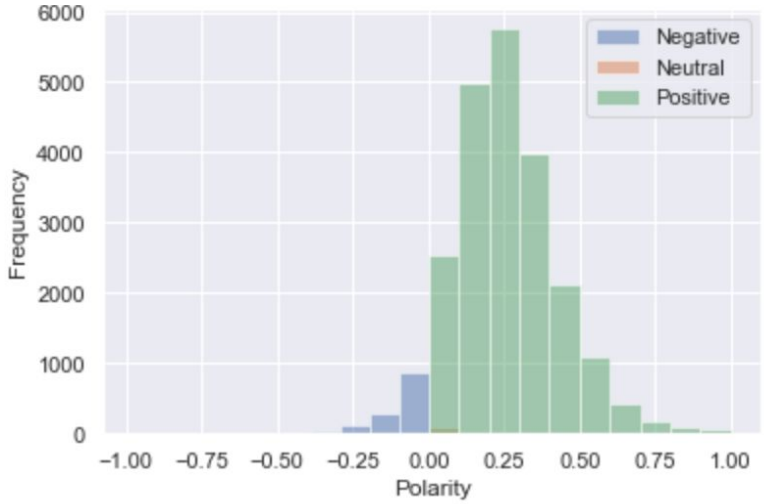


Figure 15.Sentiment Analysis

- ## Words Analysis

By checking the top 20 most common terms based on counts and average ti-idf weight, keywords are largely similar, only a slight difference in ranking. In addition to the keywords that describe the type of clothing such as ‘dress’, ‘sweater’,’ jeans’, customers use a lot of emotional words to describe their wearing experience such as ‘love’, ‘great’, ‘perfect’. The high occurrence of this review suggests that height and size is usually an important factor.

As we can find that in many comments, words like ‘fabric’, ‘comfort’, ‘soft’, ‘flattering’ shows that customers pay more attention to the quality of products. Negative words detect like ‘small’, ‘large’ can tell us that the size probably the biggest problem for the negative reviews and provide better product dimension information might help decrease the number of negative reviews.

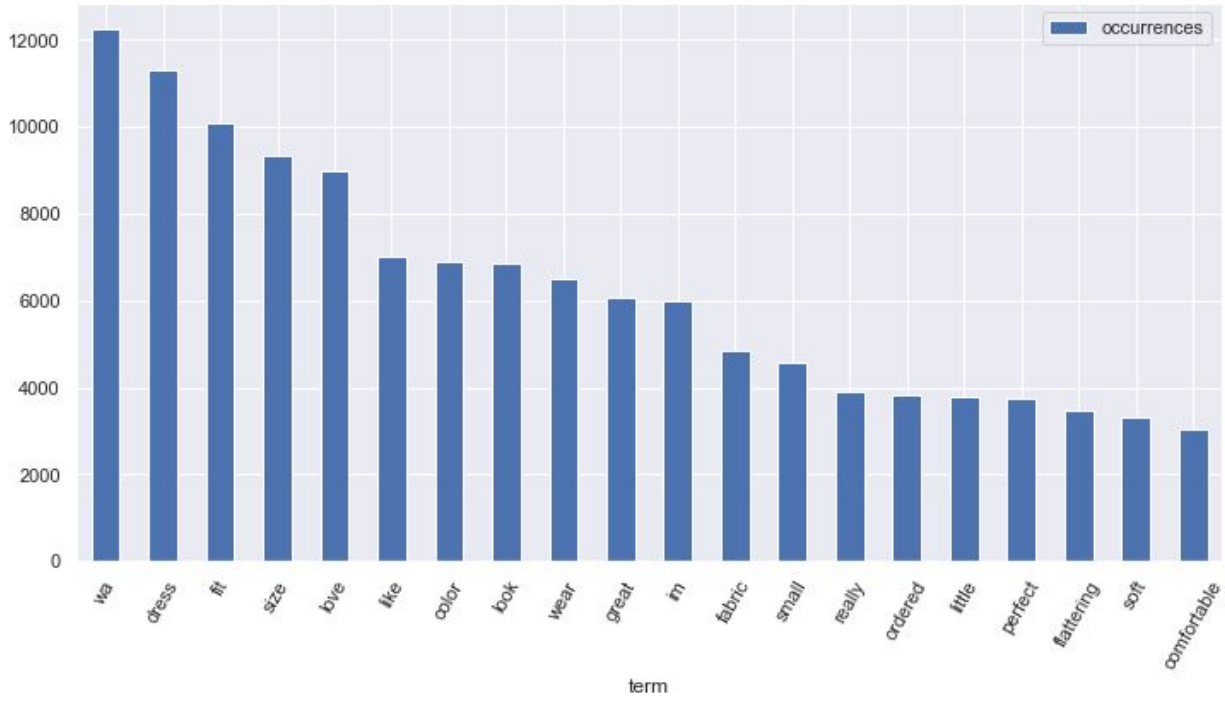


Figure 15.Top 20 Most Common Words

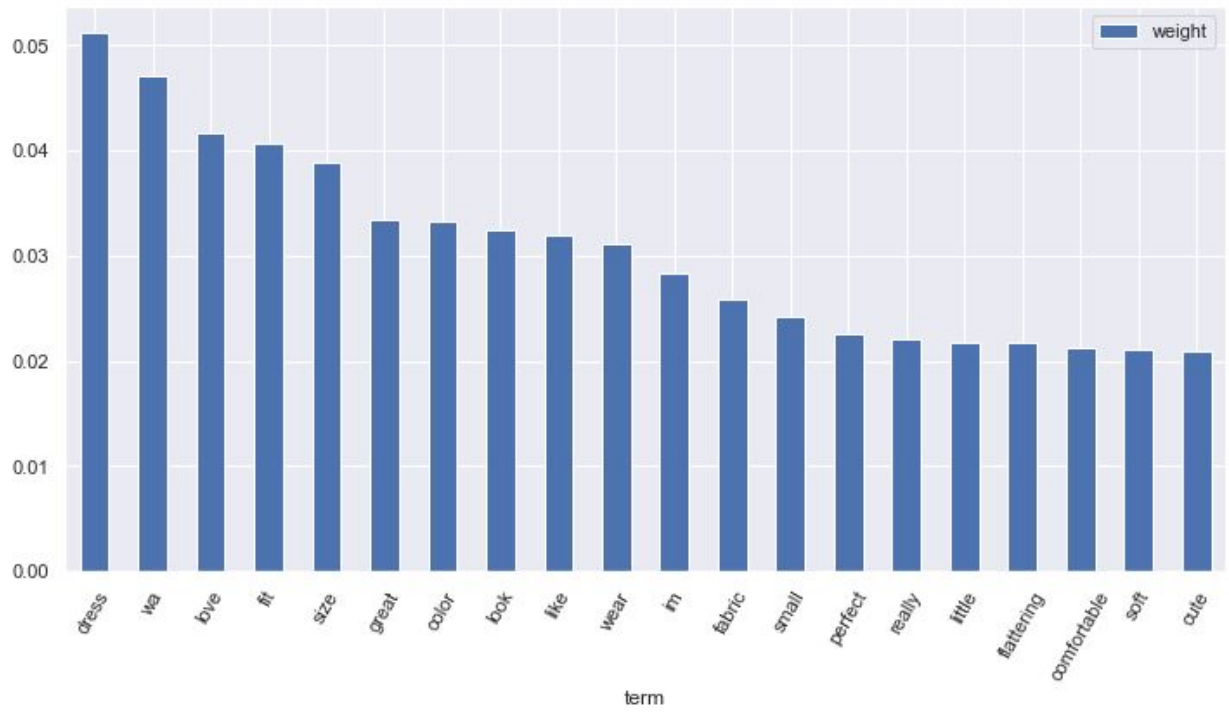


Figure 16.Top 20 Most Common Words by Average ti-idf Weight

Machine Learning Modeling and Optimization

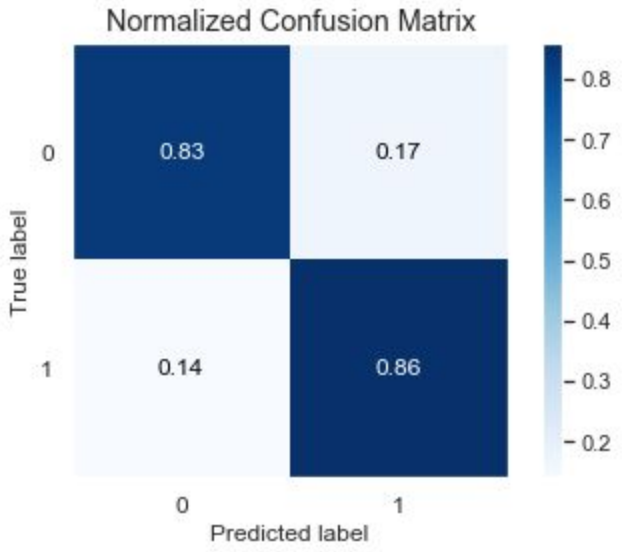
After exploring and preprocessing the review text on the dataset, the next step will use the Tf-idf weight matrix I just made to build a predictive model.

- ### Customer Sentiment Predictive Model

  In the first part of machine learning, I will focus on how well the weight matrix can predict customer recommendations. I will set the ti-idf weight matrix to X value and df['Recommended IND'] to y value.

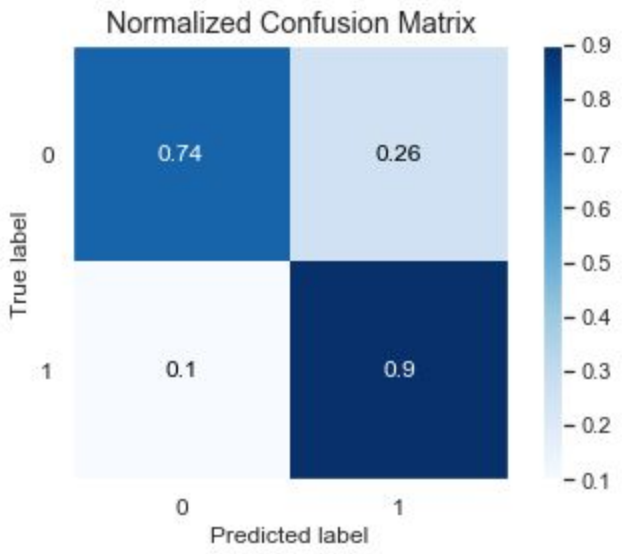
  I use 3 machine-learning algorithms: Naive Bayes, Logistic Regression and Random Forest to train-test the model separately and compare the confusion matrix, precision, recall, and f1-score to find the best solution. I also use

GridSearch on each algorithm for hyperparameter tuning.



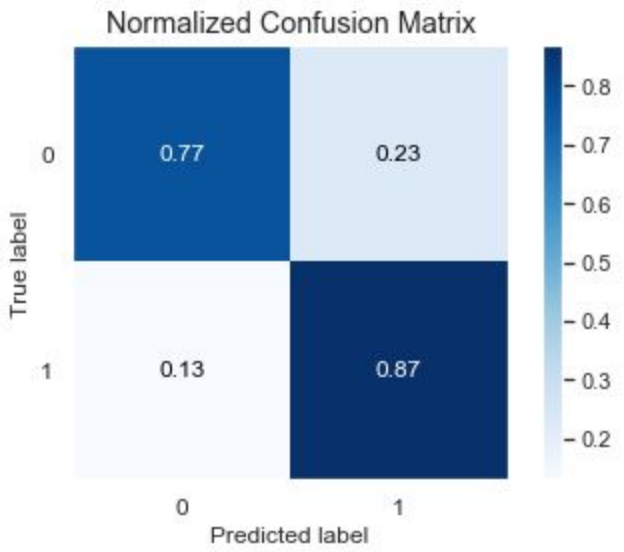
	precision	recall	f1-score	support
0	0.26	0.83	0.40	438
1	0.99	0.86	0.92	7313
accuracy			0.86	7751
macro avg	0.62	0.84	0.66	7751
weighted avg	0.95	0.86	0.89	7751

Figure 17. Naive Bayes confusion Matrix and Classification Report



	precision	recall	f1-score	support
0	0.54	0.74	0.62	1006
1	0.96	0.90	0.93	6745
accuracy			0.88	7751
macro avg	0.75	0.82	0.78	7751
weighted avg	0.90	0.88	0.89	7751

Figure18 . Logistic Regression confusion Matrix and Classification Report



	precision	recall	f1-score	support
0	0.32	0.77	0.45	572
1	0.98	0.87	0.92	7179
accuracy			0.86	7751
macro avg	0.65	0.82	0.69	7751
weighted avg	0.93	0.86	0.89	7751

Figure 19. Random Forest confusion Matrix and Classification Report

Compared to the 3 models above, the Logistic Regression Classifier has the best result in both classes. In order to deeply understand the predicted model to find out which features are most important for a model and how the predicted model makes an incorrect prediction, I will use SHAP and LIME packages to do the model interpretation.

- ## Interpreting Text Prediction with SHAP

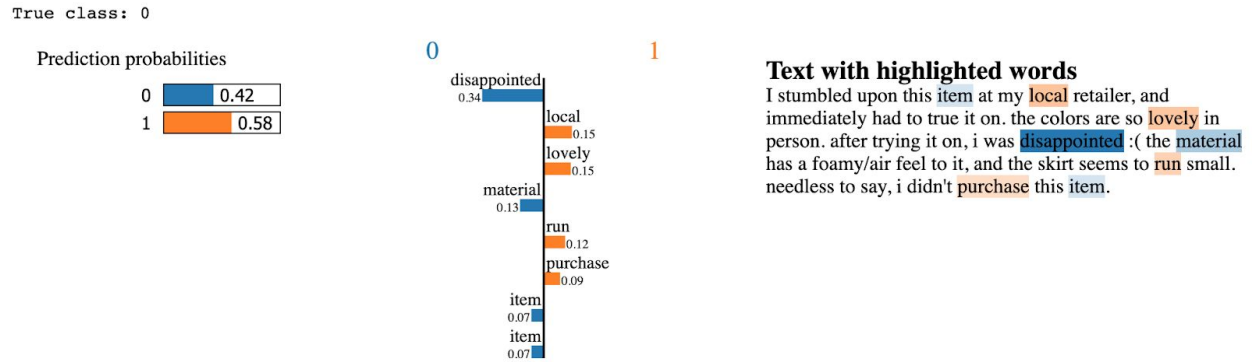
To get an overview of which features are most important to the model, we can plot the SHAP value for each feature of each sample. The following figure sorts the features by the sum of the SHAP values on all samples and uses SHAP values to show the distribution of the effect of each feature on the model output.

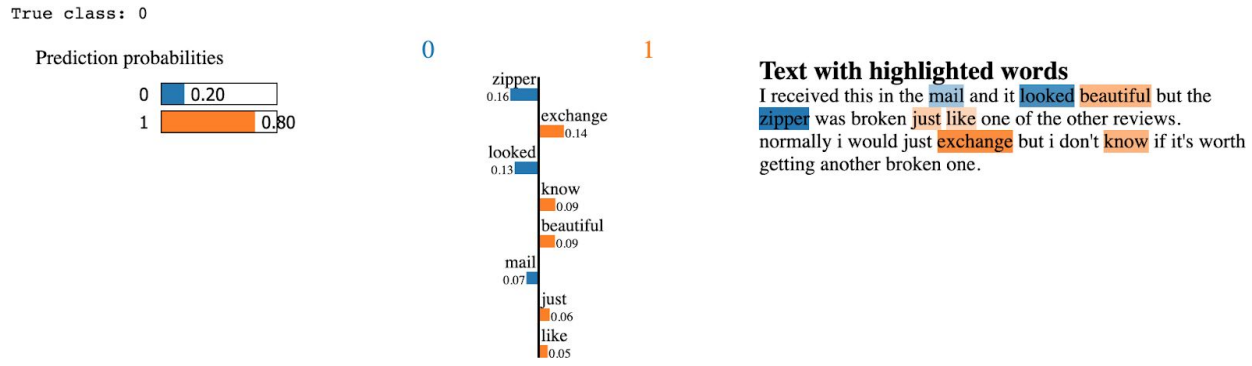
Customer review with emotional words like ‘love’,’great’,’perfect’, ‘'comfortable’, ’compliments’; quality words like ‘soft’,’ fits’ are more likely to recommend the item. for the customers who don’t recommend, they have a big concern about the ‘fabric’,’ material’,’ quality’ and ‘fit’. Also with a strong emotional word’ disappointed’.

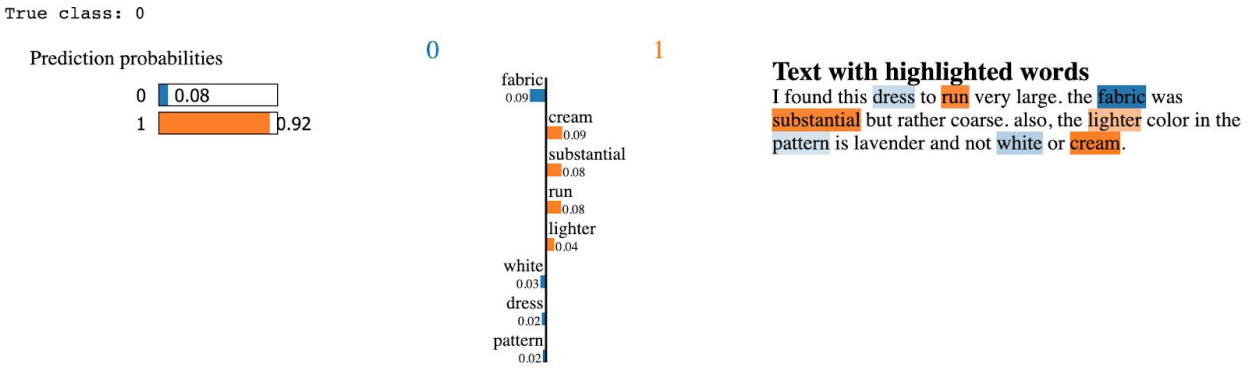
Figure 20. Most Important Features for Predicted Model

- Interpreting Text Prediction with LIME

The Logistic Regression Classifier produced a good prediction(over 90%) accuracy on recommended Items (class 1) but has low precision, recall, and f1-score on not recommended items (class 0). What features make the prediction wrong? In this chapter, I choose 15 of the text of the not recommended items which were predicted wrong and interpreting by LIME. Here are the 3 examples shown below:







Based on the several wrong labeled reviews, I think In order to improve the model performance, we need to negation into consideration. Also, this dataset has 19314 reviews in class 1 and only 4172 in class 0. This unbalanced sample may also cause the low accuracy of the not recommended review. To resample the dataset need to be considered as well to optimized the model.

Conclusion

In this sentiment analysis of women's clothing review project, I believe that I have made a good machine learning model to predict customer recommendations based on their review text. My model has over 90% precision/ recall on recommended items and 74% recall on not-recommended items which should give the company an overview about customers experience and find the key features that impact their marketing strategy.

I also interpreting the model by using LIME to get more details about which words or which factors really matters to the customers.There is also something needs to be improved such as an unbalanced sample, Take a more rigorous approach to mine the text data, such as categorizing products, controlling for spelling errors; Apply a more extensive list of algorithms (including Deep Learning) to the text data and create an ensemble of models for better prediction. The company can working on them based on my model.