Sentiment Analysis and Product Recommendation for E-Commerce Women's Clothing Business

What is sentiment Analysis?

Sentiment Analysis is a process of identifying opinions and provide a quantifiable result using AI and Natural Language Processing techniques to process raw data. In today's world customer reviews have become the place where people share their opinions. Sentiment analysis is used in e-commerce platforms to analyse gathered customer feedback.

Problem statement

Sentiment Analysis is an opportunity to understand what customer opinions/emotions are when they recommend products, by analysing online conversations. The goal is to develop a supervised machine learning model to predict user sentiment whether positive, negative, and neutral and to predict whether the product is recommended to other users.

Industry/ domain

Targeted Industry is Fashion retailers. Fashion retailers are discovering new ways to collect customer and competitor information to provide personalised experience to their own customers and attracting new customers. Some of the concerns of the customers during online shopping are size fitting, colours, quality, or material of the garment and return policy.
This can be applied to other industries such as general retailers, airlines, and hospitality.

Stakeholders

- Potential stake holders could be any fashion retailer with a platform to collect user feedback using review system or social media. Following Australian popular fashion retailers are some potential examples. Review, Witchery, Forever New, Crossroads, Sportsgirl.
- Benefits of Sentiment Analysis
  - Maintain brand reputation
  - How to maximise customer satisfaction
  - Optimise marketing campaigns
  - Fine tuning of new product launches

- Costs can be saved with Sentiment Analysis
  - Minimise Financial Loss for the brand due to Negative feedbacks
  - Reduce High Customer Churn over rate

Minimise abandoned Purchases
Reduce stock return rate

Business question

Sentiment Analysis can help businesses and services to understand their customer's experience and make necessary changes to improve the quality of their services.

Business Question:

- Understand customer's opinion whether positive, negative, or neutral.
- Identify customer product recommendation.
- Identify likes/dislikes.
- Identify Target audience for the shop.
- Identify most/least reviewed Items.
- Focus on what division/department or clothing class to be improved in the business.
- Stock management and optimization

In a business it is vital to understand all negative and positive customer feedback to improve the business or brand value.

- False positive is an incorrectly predicated negative feedback. Correct understanding of negative feedback is important to take corrective actions.
- False negative is an incorrectly predicated positive feedback. Correct understanding of positive feedback is important to continue to do what is already successful. Therefore, it is vital to minimize false negatives and positives to get an accurate understanding of the customer feedbacks to improve the business.

Data question

- What model can be used to analyse customer sentiment.
- Analyse the customer review and predict the nature of sentiment such as positive, negative, and neutral and whether the product is recommended or not?
- What are the target and recommended features?
    - Customer Review, Title, Ratings and Recommendation

Data

- Dataset downloaded from https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews
- This dataset includes 23486 rows and 10 feature variables. Each row corresponds to a customer review, and includes the variables:

    - Clothing ID: Integer Categorical variable that refers to the specific piece being reviewed.
    - Age: Positive Integer variable of the reviewer's age.
    - Title: String variable for the title of the review.
    - Review Text: String variable for the review body.
    - Rating: Positive Ordinal Integer variable for the product score granted by the customer from 1 Worst to 5 Best.
    - Recommended IND: Binary variable stating where the customer recommends the product where 1 is recommended, 0 is not recommended.
    - Positive Feedback Count: Positive Integer documenting the number of other customers who found this review positive.
    - Division Name: Categorical name of the product high level division.
    - Department Name: Categorical name of the product department name.
    - Class Name: Categorical name of the product class name.

Data is real commercial data. Dataset had some unnecessary columns (unnamed: 0) some values were missing in Title, Review Text, Division, Department and Class. This data is not available on an ongoing basis.

Data science process

Data analysis

Data Wrangling

- Removed missing data rows from Title and Review Text columns.
- Identified zero values can be zero.
- Text Pre-processing
    - Cleaned text by
        - Combined Review Text and Title.
        - Converted short from word to original form.
        - Removed double quotes, special characters, single characters, and numbers.
        - Reduced multiple spaced to one.
        - Reduced multiple newlines to one.
        - Texts were tokenized.
        - Removed stop words and punctuation.
        - Applied Lemmatization.

EDA

Display first five records of Dataset

	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	Intimates	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 - Data Frame

Dataset Information (Data Types)

```
df.info()
```

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

Figure 2 - Data Types

Null Values

```
df.isnull().sum()
```

```
Unnamed: 0                    0
Clothing ID                   0
Age                           0
Title                      3810
Review Text                 845
Rating                        0
Recommended IND               0
Positive Feedback Count       0
Division Name                14
Department Name              14
Class Name                   14
dtype: int64
```

Figure 3 - Null Values

Unique Values

```
df1.nunique()
```

```
Clothing ID                  1095
Age                            77
Title                       13983
Review Text                 19656
Rating                          5
Recommended IND                 2
Positive Feedback Count        79
Division Name                   3
Department Name                 6
Class Name                     20
dtype: int64
```

Figure 4 - Unique Values

Pearson Correlation Matrix

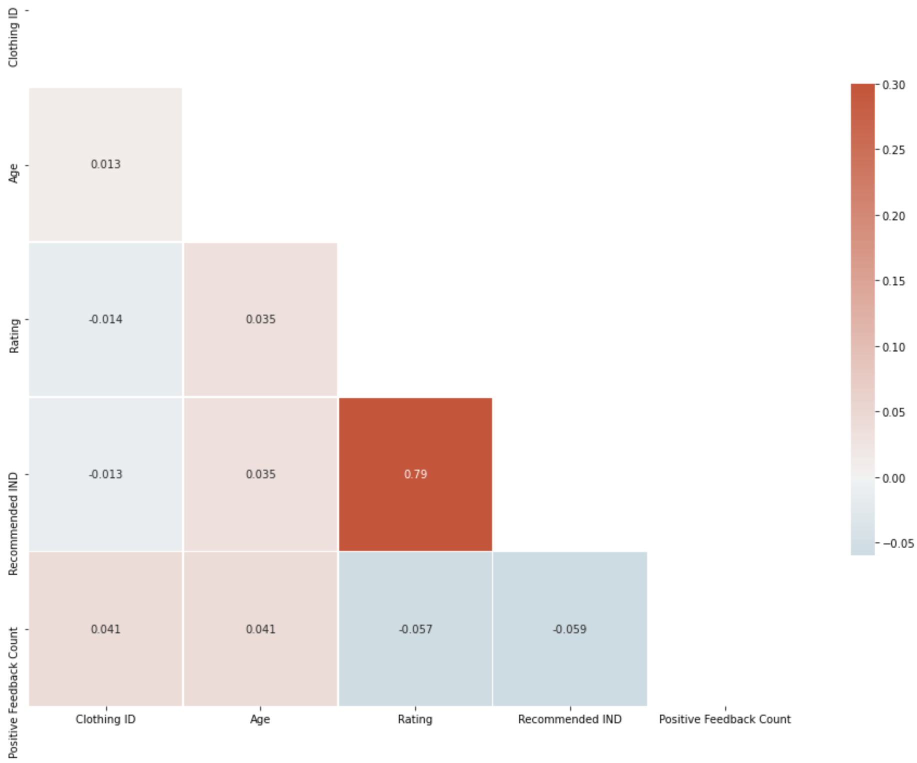

Figure 5 - Pearson Correlation Matrix

Note : Rating and Recommended IND are highly correlated.

Histogram

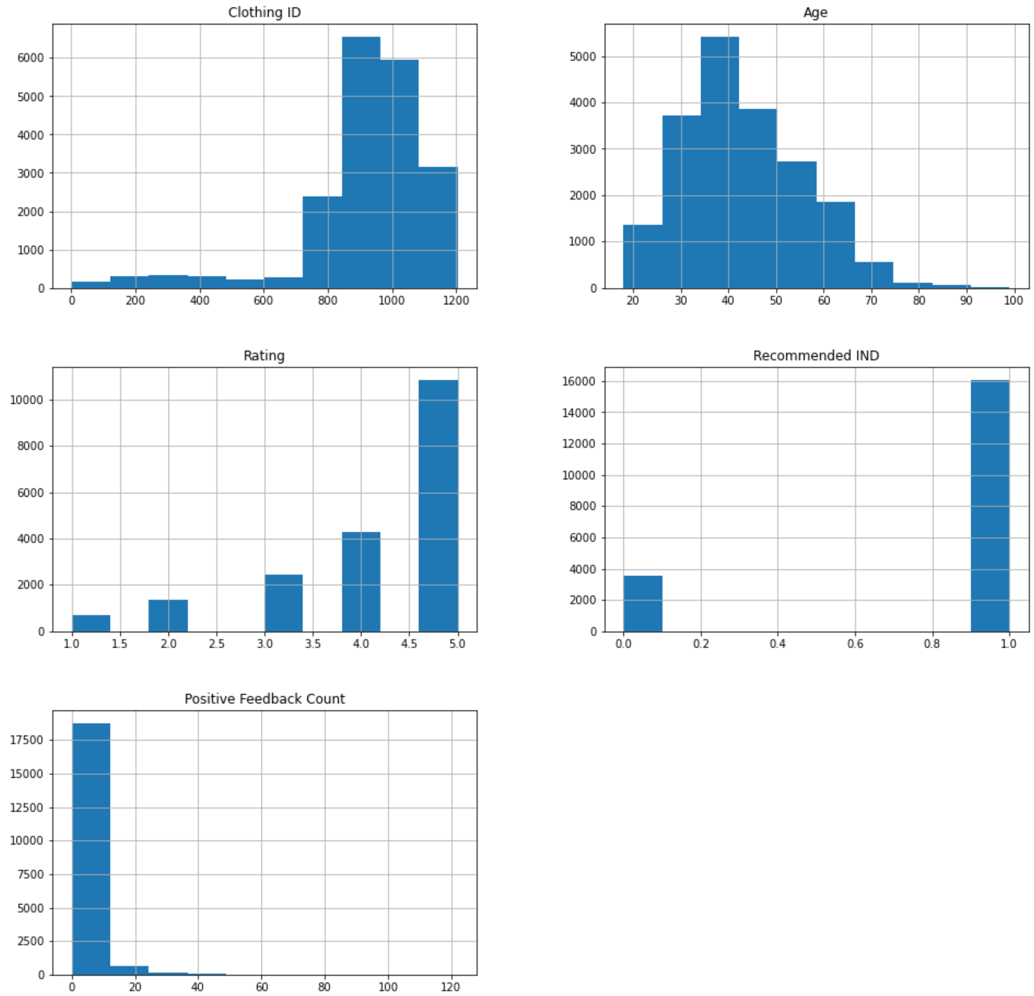

Figure 6 - Histogram

Note: clothing which has high rating (4 and 5) are mostly recommended while clothing which has low rating are unlikely to be recommended.

Rating and Product Recommendation

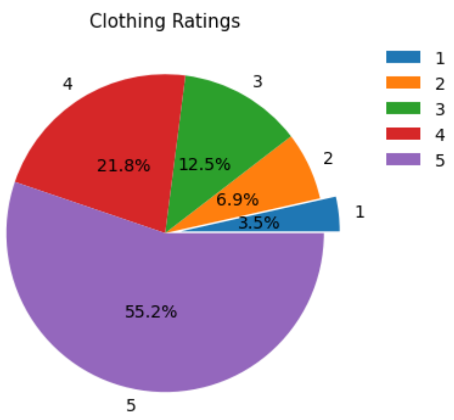

Figure 7 - Ratings

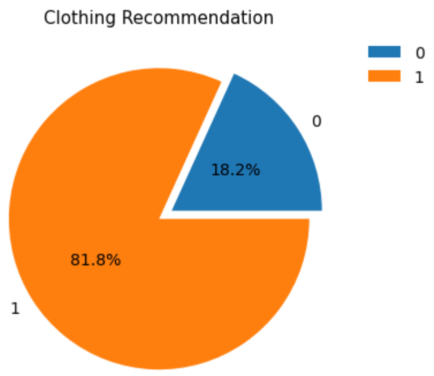

Figure 8 - Recommendation

Note : Based on above chart we can see that 65.2% has the highest rating of 5, followed by 21.8 % of rating of 4. Product recommendation percentage is 81.8%.

Rating vs Recommendation

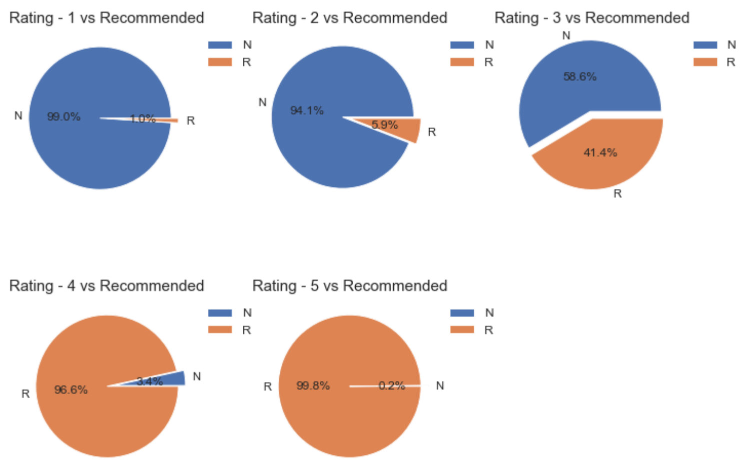

Figure 9 - Rating vs Recommendation

Note: Ratings with 1 and 2 have the least percentage of product recommendation (1% and 6%), while Rating 4 and Rating 5 have the highest percentage of product recommendation. Rating 3 has mix results of product recommendations.

Review Distribution among Division, Department and Class

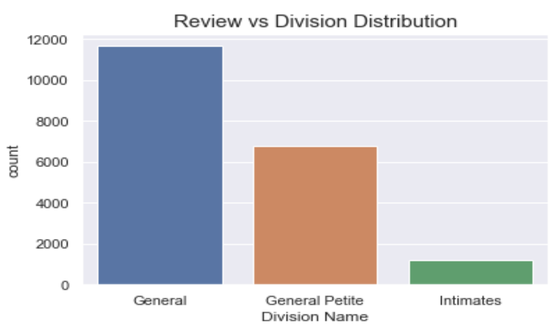

Figure 10 - Review Distribution with Division

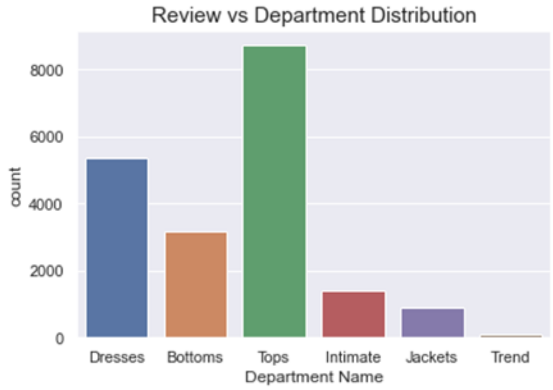

Figure 11 - Review Distribution with Department

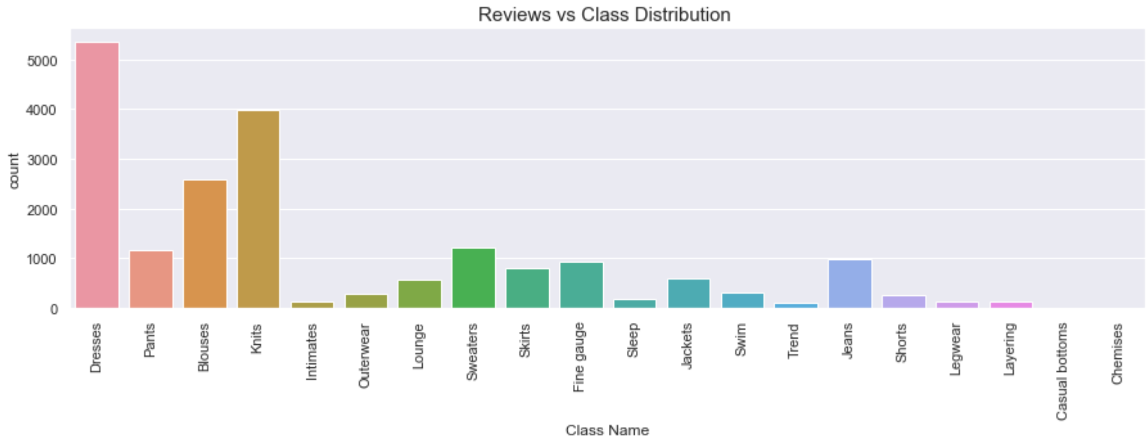

Figure 12 - Review Distribution with clothing Class

Note : The majority of the reviews(59.3%) were under General Division, followed by General Petite (34.5%) and Intimates(6.2%). Trend Department has received the least number of reviews. Dresses and knits have received most of the reviews. Casual bottoms and chemises class have received the least number of reviews.

Top 10 Recommended Items vs Division

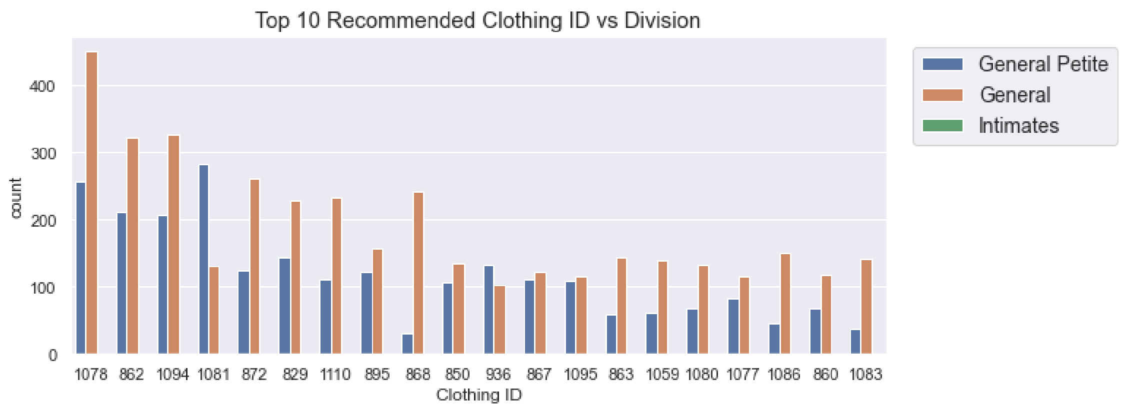

Figure 13 - Most Recommended Items vs Division

Note: Most recommended Items belong to General Petite Division followed by General Division.

Top 10 Not – Recommended Clothing Items vs Division

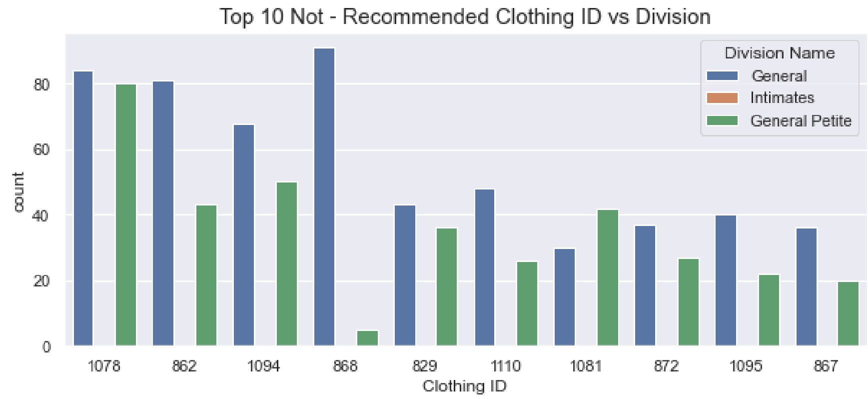

Figure 14 - Least Recommended Items vs Division

Note : Most of the not recommended Items are belong to General Division.

Rating and Recommendation vs Division

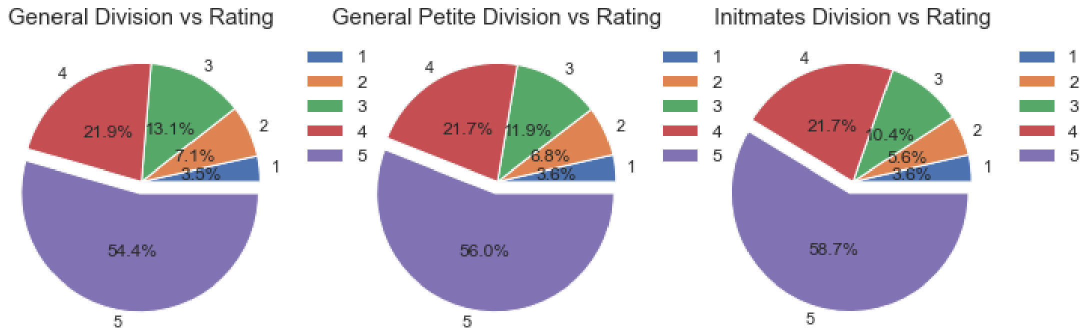

Figure 15 - Division vs Rating

Note: Every department has majority of rating 5 and 4. All Departments are performing in almost same level in average.

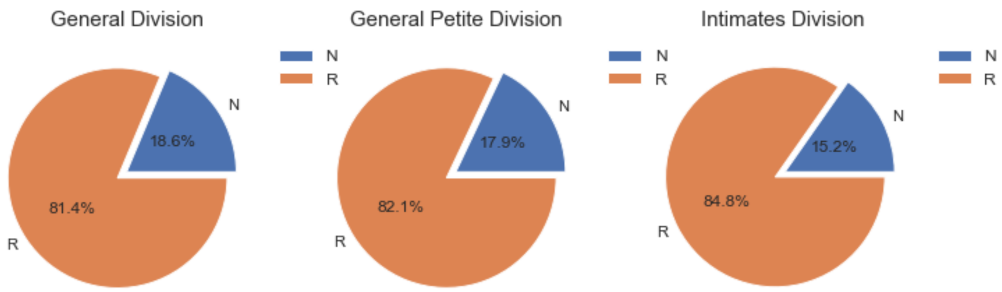

Figure 16 - Division vs Recommendation

Note : More than 10000 items have been recommended under General Division which is 81.4% of total product reviews under General Division. Intimates has the highest percentage of product recommendation which is around 1000 reviews. Overall General department has high power of generating revenue compared to another division

Department vs Rating and Recommendation

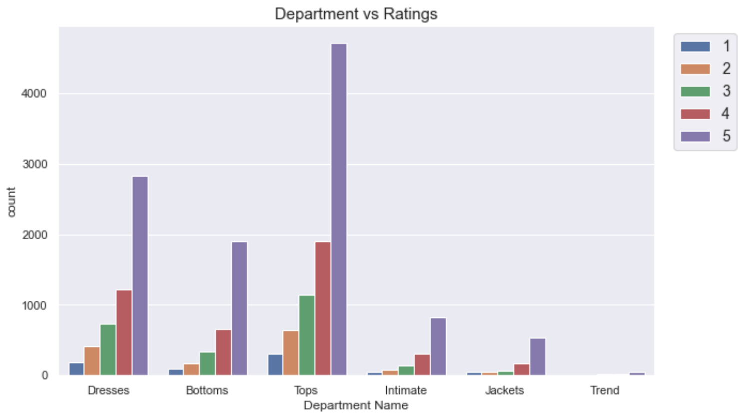

Figure 17 - Department vs Rating

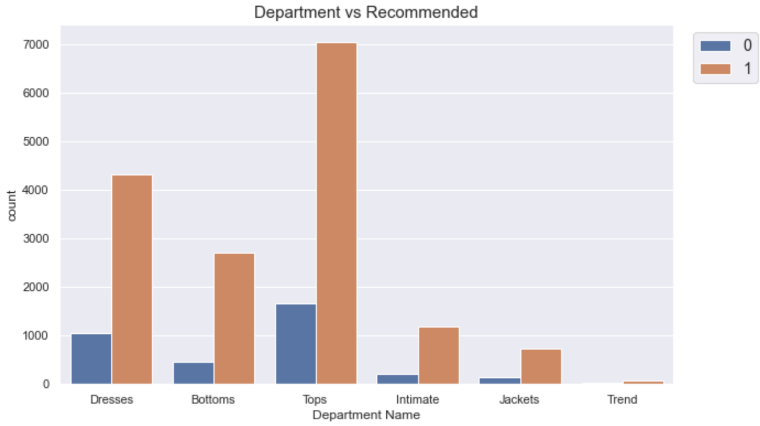

Figure 18 - Department vs Recommendation

Note: Tops department has the highest amount of product recommendation and highest rating, followed by Dress Department. Trend department has least number of reviews and recommendation.

Class vs Rating and Recommendation

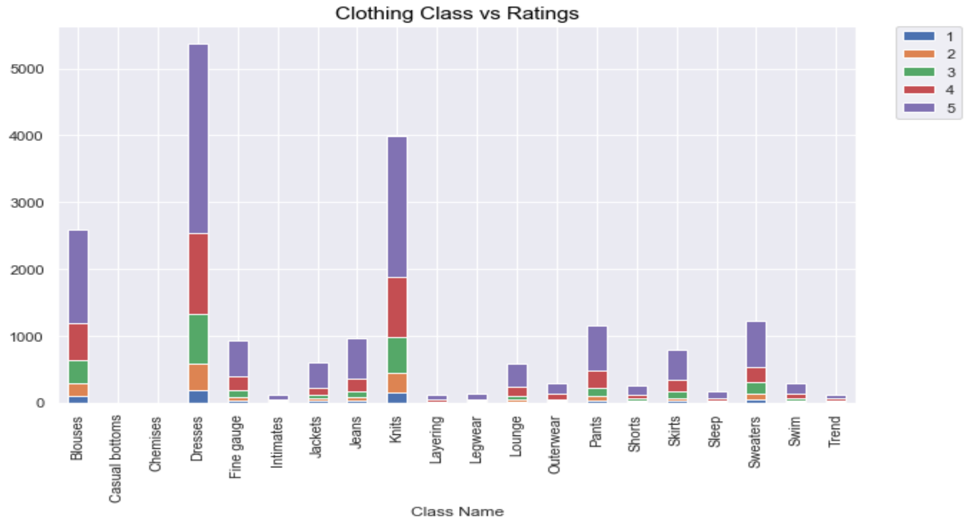

Figure 19 - Clothing Class vs Ratings

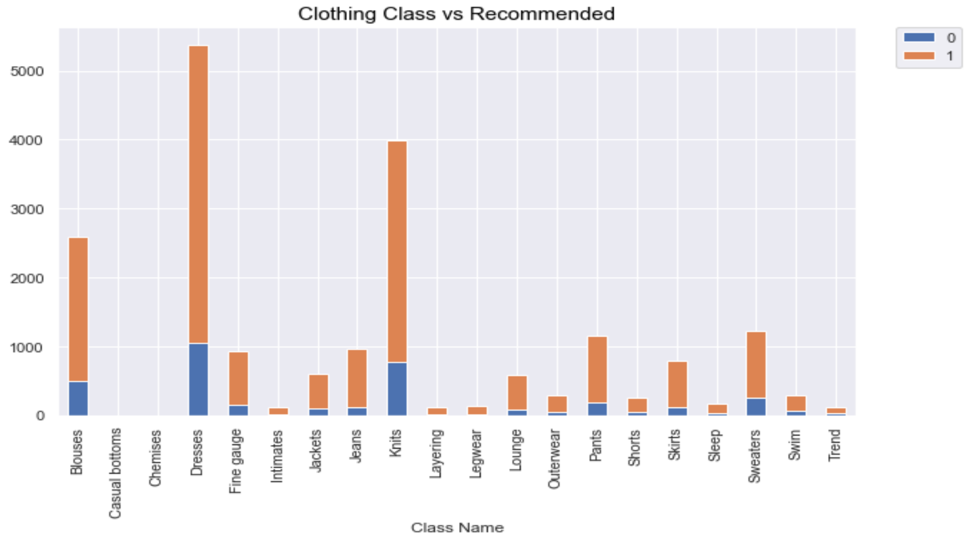

Figure 20 - Clothing Classes vs Recommendation

Note : Dresses, Knits and Blouses have the highest and lowest ratings reviews. Recommendations are higher in every class

Age

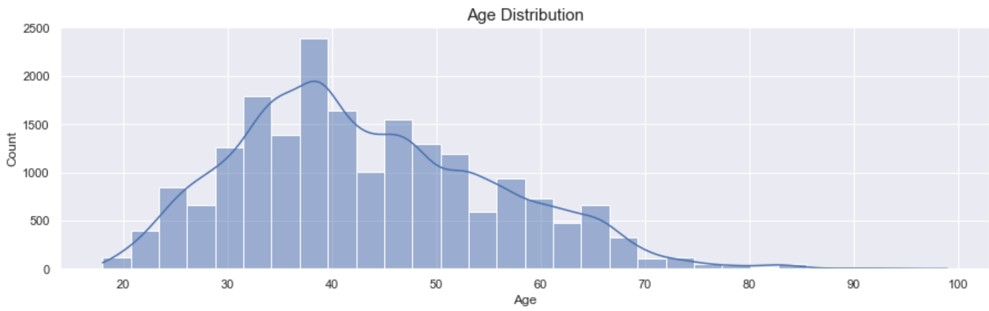

Figure 21 - Age Distribution

Note : Target audience for the reviews are age between 30 to 50

Age vs Clothing Class

Figure 22 - Age vs Clothing Class

Note : Target Review audience for Swim Class 30 to 65 (age range) while Intimate review audience vary between age 30 to 70.Casual bottoms and Chemises only popular among young customers. Dresses, Blouses, Knits, Outerwear, Sweaters and Sleep Classes target audience are range between 20 to 80.

Word Cloud

Figure 23 - Word Cloud for High Rating and Low Rating

Note: Figure 23.1 - Shows top words for the highest rating and for product recommendation. As we can see that most of words are positive such as love, comfortable, look, great and they could be related to dresses, shirts related reviews.
Figure 23.2 - Shows top words for the lowest rating and zero product recommendation. As we can see those areas to be focus on based on top words are related to dress, material, colour, size

Top 20 Trigrams

Figure 24 - Top 20 words

Note : Based on above top 20 trigrams indicates the most common words used in reviews for the shop those are mainly related to Dress class and sizes. Those top trigrams indicates positive trend towards the shop.

Modelling

Main Features Selected: Review Text, Title, Rating, Recommendation IND
Rating and Recommended IND are found that highly correlated.
Models Used:

- Logistic Regression
- Naïve Bayes
- Random Forest Classifier
- Gradient Boosting Classifier
- Stacking Classifier

As this is a supervised classification problem with two target variables. We have applied two separate classification models to predict sentiment classification and product recommendation classification with above mentioned models.

Feature Engineering

- Polarity as a new feature from TextBlob library
- Counter Vector as Features
- TF-IDF (Word level, n-Gram and Character Level) as Features
- Created New Text feature by combining Review and Title text.
- Text/NLP as features
    - Character count, Word count, Average word length and Average sentence length
    - POS based Features (Noun Count, Adjective's count, Adverbs counts and Numeral count)
- Topic Modelling as Features

Feature Selection

Observations with Different Feature Selection to predict the sentiment is Positive, Negative or Neutral.

Counter vector and TF- IDF as Features (reference: notebook: feature engineering part – 1)

	Count Vectors	WordLevel TF-IDF	N-Gram Vectors	CharLevel Vectors
Naïve Bayes	0.615751	0.637959	0.524754	0.558878
Logistic Regression	0.642184	0.635901	0.550211	0.613260
Random Forest	0.642726	0.653450	0.536887	0.638501
Gradient Boosting	0.638501	0.671867	0.499404	0.662983

Figure 25 - Feature Engineering part 1- Results

Note : Word level TF- IDF had the highest accuracy with Gradient Boosting. But the results are under baseline accuracy which is 77%.

Counter vector and TF- IDF with Combined Text (Title and Review) as Features (Reference: notebook: feature engineering part – 2)

	Count Vectors	WordLevel TF-IDF	N-Gram Vectors	CharLevel Vectors
Naïve Bayes	0.650092	0.661792	0.532878	0.584552
Logistic Regression	0.685841	0.669375	0.568844	0.636117
Random Forest	0.650200	0.680425	0.539920	0.647492
Gradient Boosting	0.665367	0.696891	0.513271	0.700899

Figure 26 - Feature Engineering part 2- Results

Note: As we can see that results were improved by combining two text columns.

Model Evaluation

Combination of TF-IDF, Text/NLP, Polarity, and combined text as Features (Reference: notebook FE)

Sentiment Classification Model Results

	WordLevel TF-IDF
Logistic Regression	0.758279
Naive Bayes	0.673892
Random Forest	0.955991
Gradient Boosting	0.783034
Stacking Classifier	0.812961

Figure 27- Sentiment Classification Results

As we can see that Random Forest has the highest accuracy. As this is an imbalanced dataset, we must investigate Recall (false negative) and Precision (false positive) to select the final model. In this business problem Recall is more important to consider than Precision. Out of all models Random Forest classifier has the highest recall value. We can conclude that sentiment analysis can predict with 95 % accuracy with Random Forest Classifier. (Model training time: 23 seconds)

Classification Report (Sentiment Classification) for Random Forest

```
 Accuracy Score for test data = 0.9559907580591924
Classification Report for test data
              precision    recall  f1-score   support

          -1       0.97      0.97      0.97      3036
           0       0.96      0.94      0.95      3036
           1       0.93      0.96      0.95      3017

    accuracy                           0.96      9089
   macro avg       0.96      0.96      0.96      9089
weighted avg       0.96      0.96      0.96      9089
```

Figure 28 - Classification Report for Sentiment Classification Model

Confusion Metrics for Random Forest Classifier

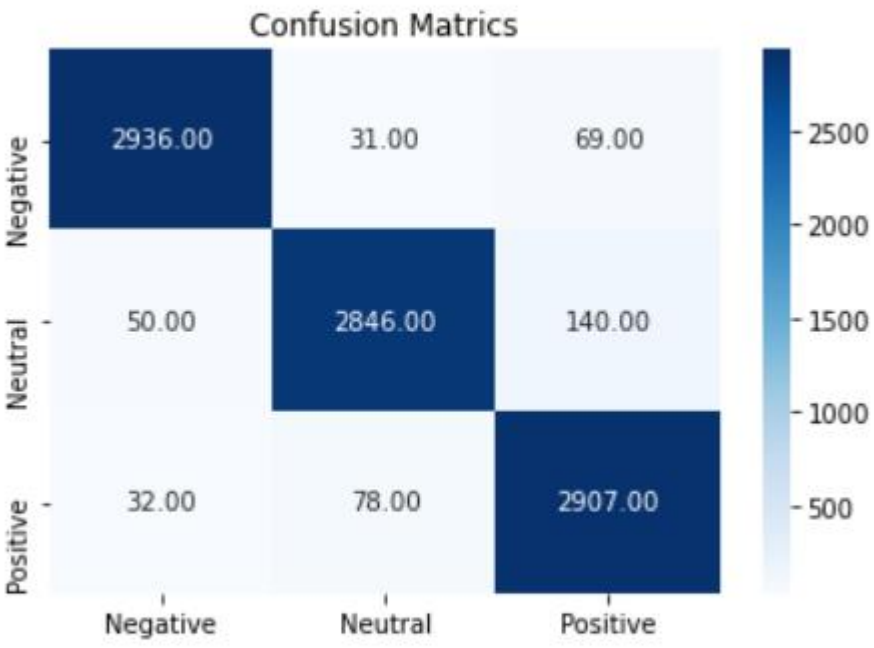

Figure 29 - Confused Metrics for Sentiment Classification

Recommended Classification Model Results

	WordLevel TF-IDF
Logistic Regression	0.904118
Naive Bayes	0.848640
Random Forest	0.941414
Gradient Boosting	0.912821
Stacking Classifier	0.822222

Figure 30 - Recommendation Classification Results

As we can see that Random Forest has the highest accuracy. As this is an imbalanced dataset, we must investigate Recall (false negative) and Precision (false positive) to select the final model. In this business problem Recall is more important to consider than Precision. Out of all models Random Forest classifier has the highest recall value. Product recommendation can predict with 94 % accuracy with Random Forest Classifier. (Model training time:15 seconds)

Classification Report (Recommendation Classification) for Random Forest

```
 Accuracy Score for test data = 0.9414141414141414
Classification Report for test data
              precision    recall  f1-score   support

           0       0.94      0.94      0.94      3217
           1       0.94      0.94      0.94      3218

    accuracy                           0.94      6435
   macro avg       0.94      0.94      0.94      6435
weighted avg       0.94      0.94      0.94      6435
```

Figure 31 - Classification Report for Recommendation Classification

Confusion Metrics for Random Forest Classifier

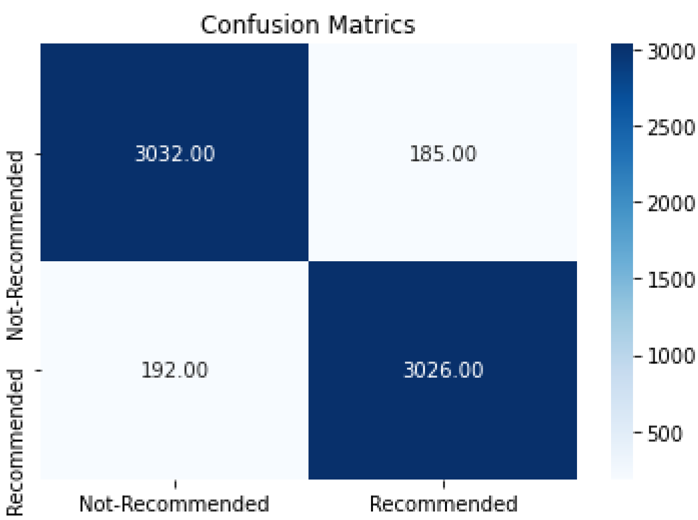

Figure 32 - Confused Metrics for Recommended Classifier

Outcomes

- Sentiment can be predicted with 95 % accuracy and Recommendation can be predicted with 94% with Random Forest Classifier.

Implementation

- What are the considerations for implementing the model in production?
    - Machine learning model needs to be trained to handle challenging aspects of the language, such as:
        - Double negatives
        - Tone of the words

Data answer

Sentiment can be predicted with 95 % accuracy and Recommendation can be predicted with 94% with Random Forest Classifier. Both Models are predicting above their baseline accuracy which is 77% and 82%.

Business answer

Model will help organization to understand customer feedback by categorising into positive, negative, and neutral and product recommendation by analysing a customer review. Target age group for the shop identified as 30 to 55. Highest rated clothing classes are Dresses, Knits, Blouses, sweaters, and pants. Trend, legwear, sleep, Layering, and Intimate class are having very less of rating 5. Intimate cloths are popular among young age groups. Negative reviews are related to issues with size, colour, fabric, quality, dresses, sweaters, and shirts. That information can be used for stock management and optimization by understanding consumer buying habits.

Response to stakeholders

Sentiment can be predicted with 95 % accuracy and Recommendation can be predicted with 94% without manual analysis of the reviews. Analysing general sentiment can help to maintain the brand reputation by setting and evaluating current KPI. Proper understanding of what values to the customer most can lead to better customer satisfaction and reduce customer churn over with a potential to attract new customers. Target audience for the shop and customer likes and dislikes can be used to optimise marketing campaigns. While having, a high volume of positive sentiment and product recommendation can lead to lower return rate. As we can see that derived information from data can be applied to make required business decisions (Data Driven Decision Making) to increase sales and revenue.

Future Development

- Apply Neural Network Model for the Analysis.
- Combine social media website reviews to retain the model with web scrapping techniques.
- Expand Model to predict the customer churn over.
- Combine the model with financial and stock data to predict revenue and stock.
- Develop a chatbot to collect customer reviews and perform chatbot sentiment analysis.

Sulakkhana Herath

End-to-end solution

The project can be divided into following four main steps,

- Model Development
- Model Training
- Model Testing
- Export trained model as joblib pickle files.
- Develop a Flask based web application using the exported models.

References

Project Notebooks:

- Capstone project analysis Part 1: EDA
- Capstone project analysis Part 2: FE_Part_1 (Reference for Accuracy enhancement with different features)
- Capstone project analysis Part 3: FE_Part_2 (Reference for Accuracy enhancement with different features)
- Capstone project analysis Part 4: FE_with_Model (Final Notebook)

Dataset:

- Womens Clothing E-Commerce Reviews.csv

Deployed Documents:

- app.py
- recommendation_model.joblib
- sentiment_model.joblib
- Procfile
- requirements.txt
- index.html
- style.css