Project Report on

Sentiment Analysis on Women E-Commerce Reviews

Submitted by

Group No. 6 [Batch: June 2020, Location: Bangalore]

Group Members

1. **Abhinav Mahoorkar.**
2. **Shivalik Ghosh.**
3. **Mohd. Tanveer Equabal.**
4. **T J John.**
5. **Swetha Dharmavaram.**
6. **Vineet Singh Munday.**

Team Leader.

Abhinav Mahoorkar

Mentor Name.

Ms. Vibha Santhanam

Industry Review.

Businesses are running customer services since a long time. The traditional approach of customer service is to contact customers through emails, posts and telephones and ask them to provide feedback on company's products and their services. These days companies especially online businesses, have ratings and reviews section on their websites for their products. It is not easy to read each and every given review online manually. Not just this, but sometimes it becomes difficult to make sense of those reviews also; for example, the reviews containing incorrect spellings or shorthand words etc. This is where Data Science comes into picture.

Using Data Science techniques, for example NLP (Natural Language Processing) the ratings and reviews from the website, can be extracted. This technique helps to retrieve user reviews and understand why bad reviews were given. For example – 'Word Clouds' are a popular way of displaying how important words are in a collection of texts and N-grams helps looking for words association. These techniques coupled with a few others help Data Scientists making sense of reviews.

Once the reviews are extracted, Data Scientists can further segregate them and do 'Sentiment Analysis'. With this information, e-commerce can efficiently maximize user satisfaction by prioritizing product updates that will have the greatest positive impact.

It's no secret that the majority of consumers today turn to ratings and reviews to help them make informed purchase decisions. 'Power Reviews' research found that 97% of shoppers consult reviews, and 89% consider them an essential resource when making a purchase decision.
Reviews are important for a variety of different product categories. But they're becoming increasingly important for apparel brands and retailers. Although apparel has lagged behind other categories in terms of ecommerce sales, a growing number of shoppers are opting to make apparel purchases online. Today, almost half of shoppers prefer to shop for clothing and shoes online. This shift in preferences will drive a projected $90 billion in ecommerce apparel sales in 2018. And by 2022, that number is expected to grow to $123 billion.
However, this momentum doesn't come without challenges. On average, 40% of apparel products purchased online are returned. This isn't surprising, as it can be challenging for a consumer to assess qualities such as size and fit without a visit to a brick-and-mortar store.
Now is the time for apparel brands and retailers to think about how they can leverage user-generated content — like ratings and reviews, questions and answers, and user-submitted images and videos — to help shoppers make more informed purchase decisions.

Literature Survey.

A few studies have been conducted both in India and abroad over a period of time regarding the impact that reviews have on an E-commerce organization. Following are a few of the studies that have been reviewed hereunder as they would facilitate a clear backing for carrying out the present study.

- (Riedl and Rene, June 2010). Purchase decisions are ultimately guided by consumers' perceptions of the product or service. Negative perceptions of the product, service, brand, or company almost always translate to a lost sale or, at a minimum, expressed lower intentions to purchase the advertised product or service.
  If the goal is to increase online spending among consumers, particularly female consumers, then consumers' perception of online shopping ultimately matters the most.

- (Yang and Lester D. (2003a)). Recent surveys point out that the gender gap has been disappearing. In addition, the numbers of male and female internet users are equal.

- (Jackson, 2001) The ever-shifting Internet population: a new look at Internet access and the digital divide, 2003). It is noted that though young women and men use the internet equally often, they use it differently, and this may influence the motivations of buying online.

- (Yang and Lester D. (2003a)). Compared with non-shoppers, Internet shoppers have been characterized as more impulsive and less risk-averse, and they were more likely to seek convenience and variety and less likely to be brand and price conscious than non-shoppers.

Project Statement.

‘Sentiment Analysis’ of the customer on the women e-commerce clothing reviews to predict whether the product will be recommended by the customer or not.

Data Dictionary.

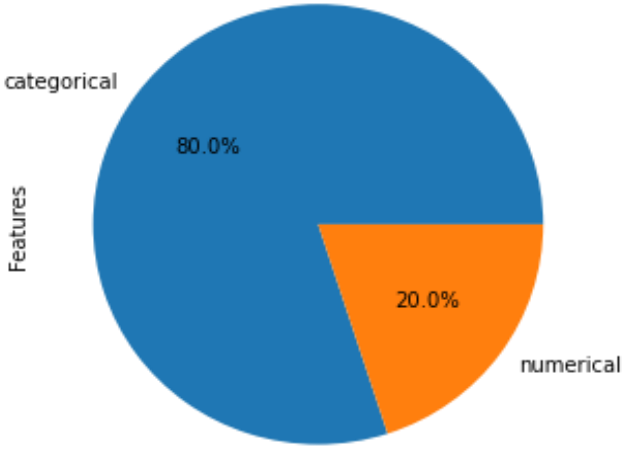
Dataset: Women E-commerce Clothing Reviews.

Data Dictionary:

1) Number of Instances: 23486
2) Number of Attributes: 10
3) Attribute information:

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

Variable Categorization:



- There are 8 Categorical Variables and 2 Numerical Variables.

Preprocessing Data Analysis.

Count of Null Values:

	Null Values Count
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

Title column has 3810 missing values which is around 16%. The other independent feature of the dataset having significant null values is ‘Review Text’.
Columns like ‘Department Name’, ‘Division Name’ and ‘Class Name’ have 14 null values each.

Treating Null Values:

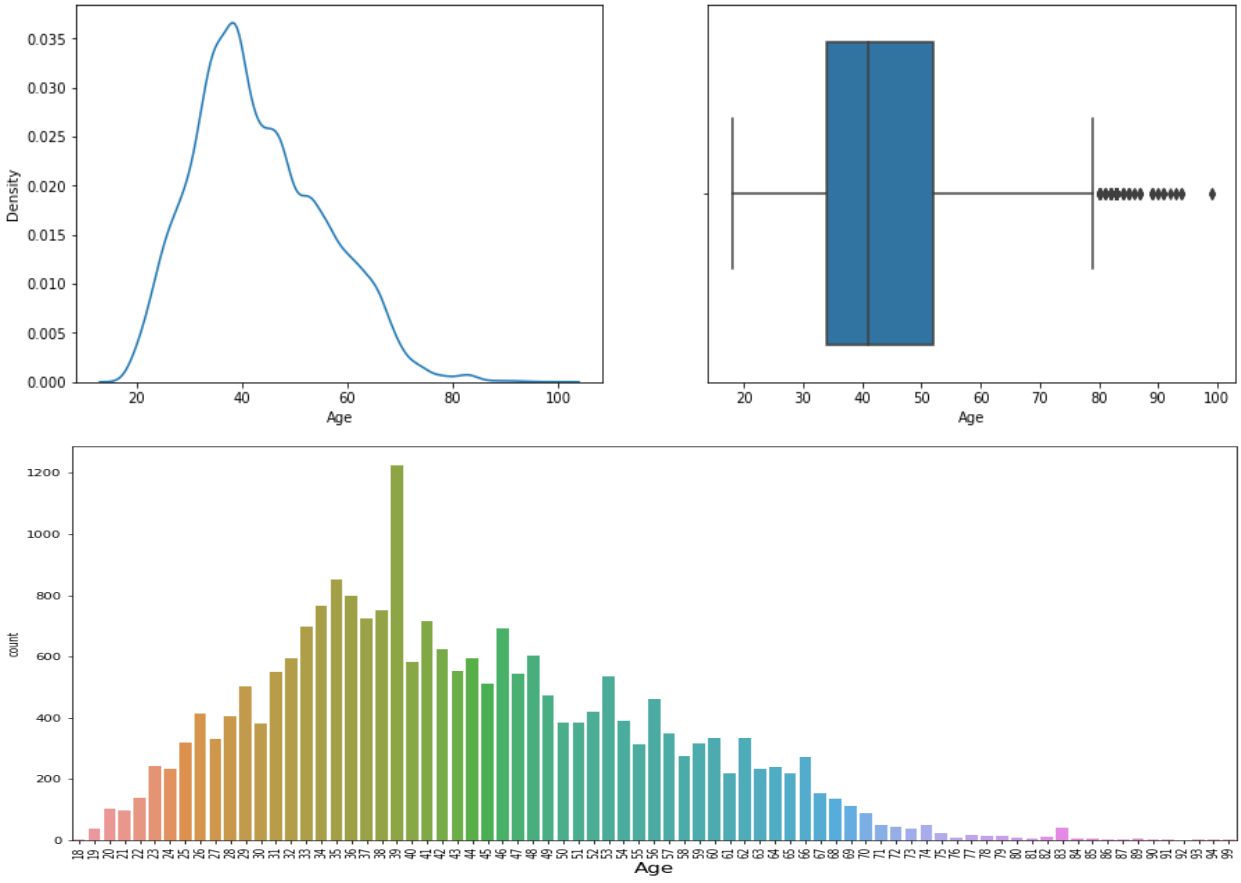
As majority of null values are present in the 'Title' and 'Review Text' columns, so we concatenated the content of these two columns and created a new column called 'Review'.
And dropped the 'Title' and 'Review Text' columns. By which the total null values in the review column became 844.
Then we dropped the rows containing null values to get the final shape of the dataset as 22629 rows and 9 columns.

Data Exploration (EDA).

1. Univariate Analysis.

Distribution of variables:

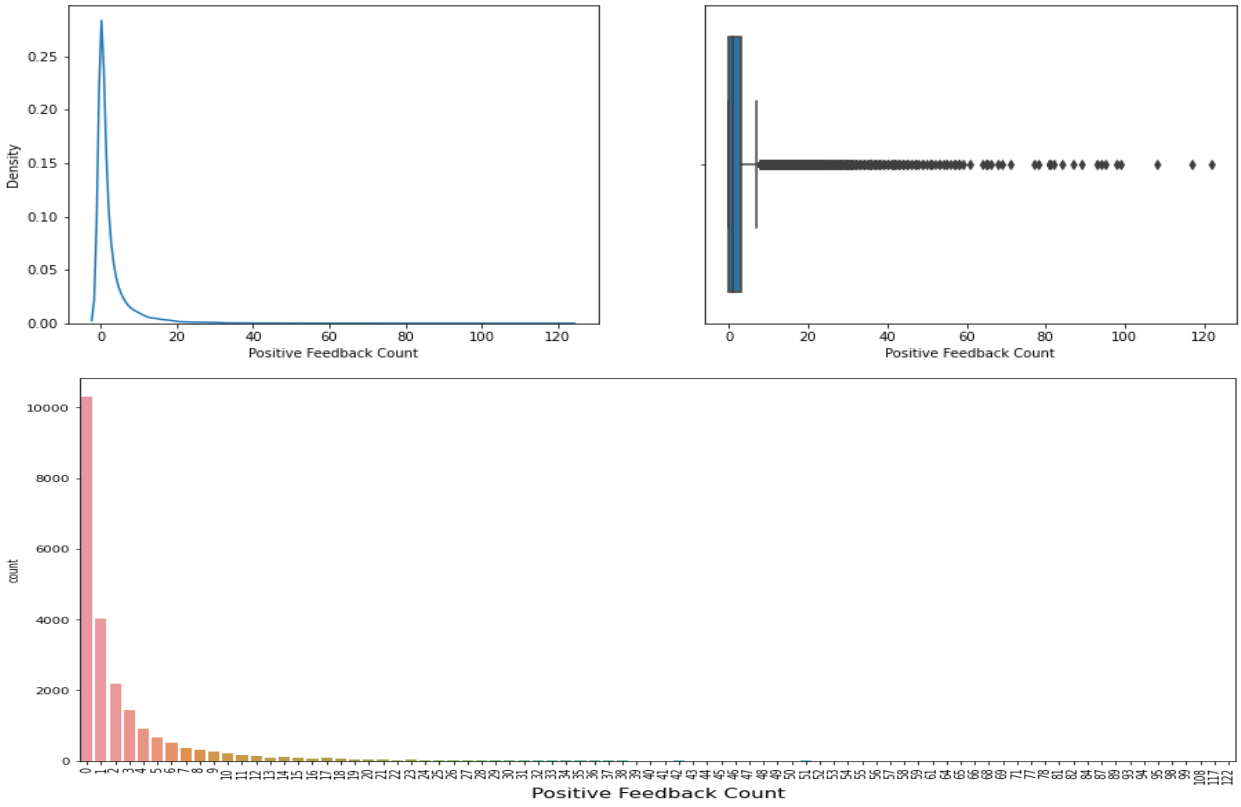
Age:



Inference:

- Age data is right skewed.
- Data has few outliers present on the higher side of the age above 80.
- Majorly, 50% of the reviews are written by the people of age between 33 and 52.
- Most reviews are written by the people of age 39 which is above 1200.
- Least reviews are written by the people of age less than 19 and greater than 83.

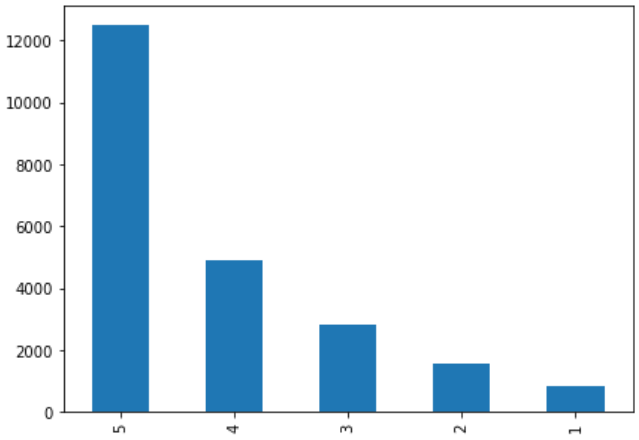
Positive Feedback Count:

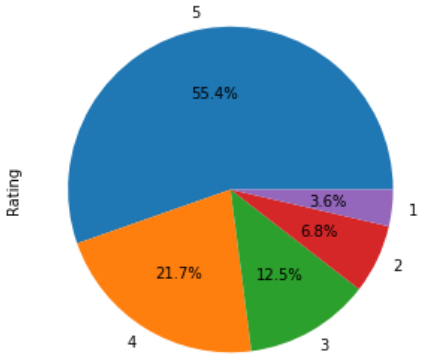


Inference:

- The Positive Feedback Count data is highly right skewed.
- Data has large number of outliers present on the higher side of the count above 8.
- Majorly, 50% of the reviews has either 0 or 1 or 2 positive feedback counts.
- Most number of reviews has 0 positive feedback count.

Rating:

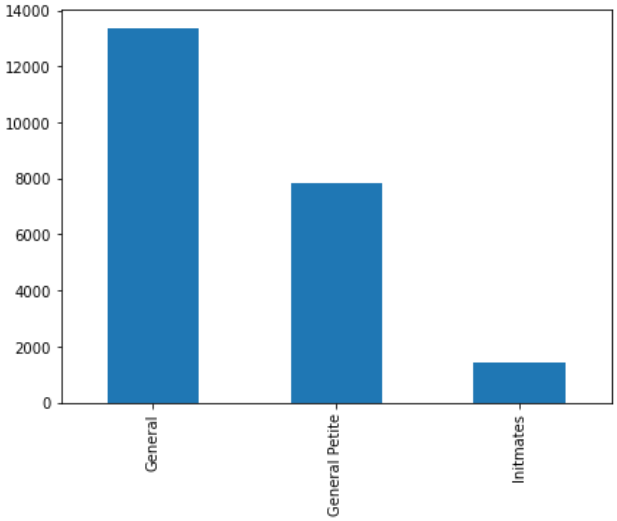


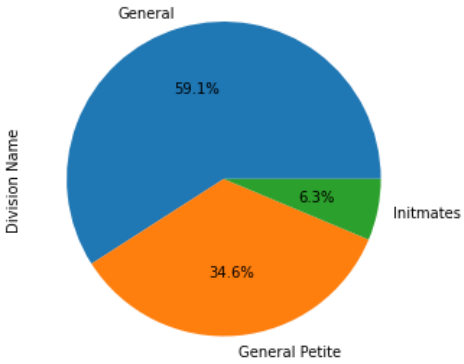


Inference :

- Most Ratings are of 5 star; which are above 12000 (55.4%).
- Least Ratings are of 1 star; which are less than 1000 (3.6%)

Division Name:

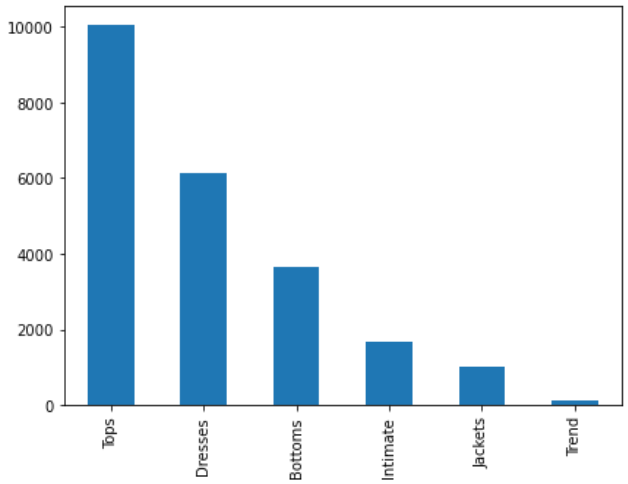


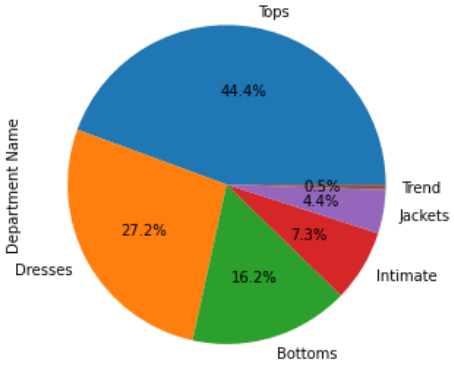


Inference:

- Most reviews are written on the General Division as compared to other Divisions; which are above 13000.
- Least reviews are written on the Intimates Division as compared to the other Divisions; which are less than 2000.
- General Petite Division has moderate number of reviews around 8000.
- Percentage of Reviews: General Division - 59.1%, General Petite Division - 34.6%, Intimates Division - 6.3%.

Department Name:

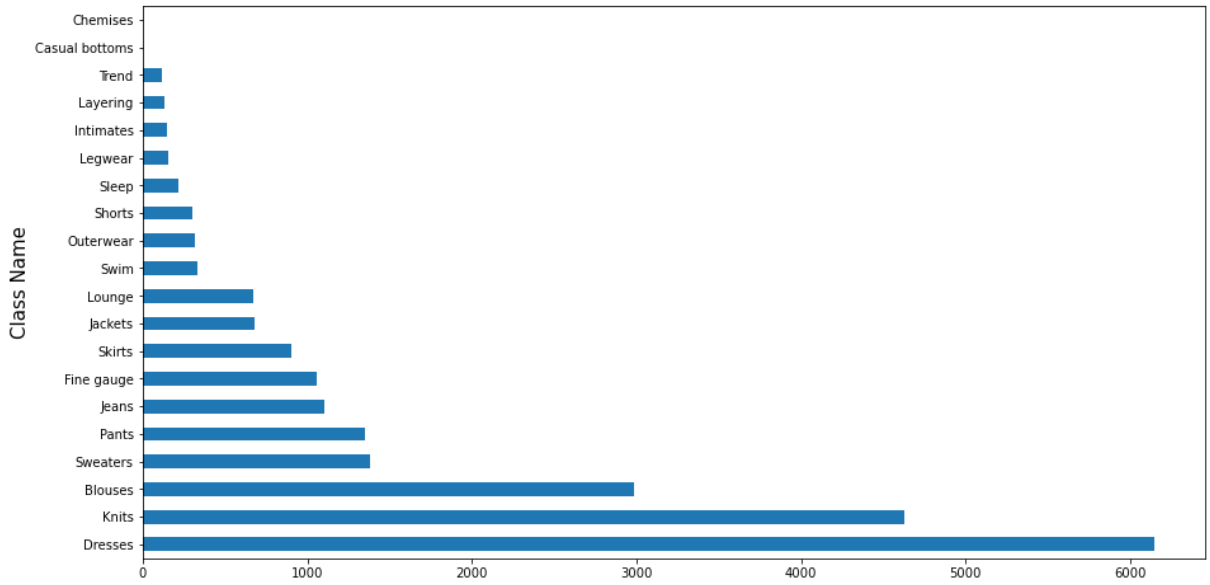




Inference:

- Most reviews are written on the Tops Department as compared to other Departments; which are around 10000 (44.4%).
- Least reviews are written on the Trend Department as compared to other Departments; which are very much less (0.5%).

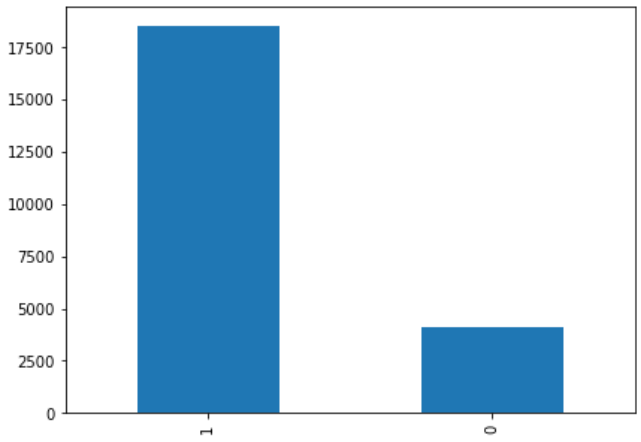
Class Name:

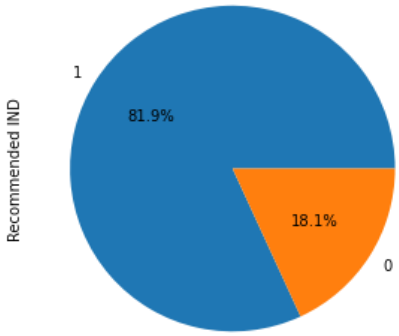


Inference:

- Most reviews are written on the Dresses Class as compared to other Classes; which are above 6000 (27.2%).
- Least reviews written on Casual bottoms and Chemises Classes as compared to other Classes; which are very much less.

Recommended IND:



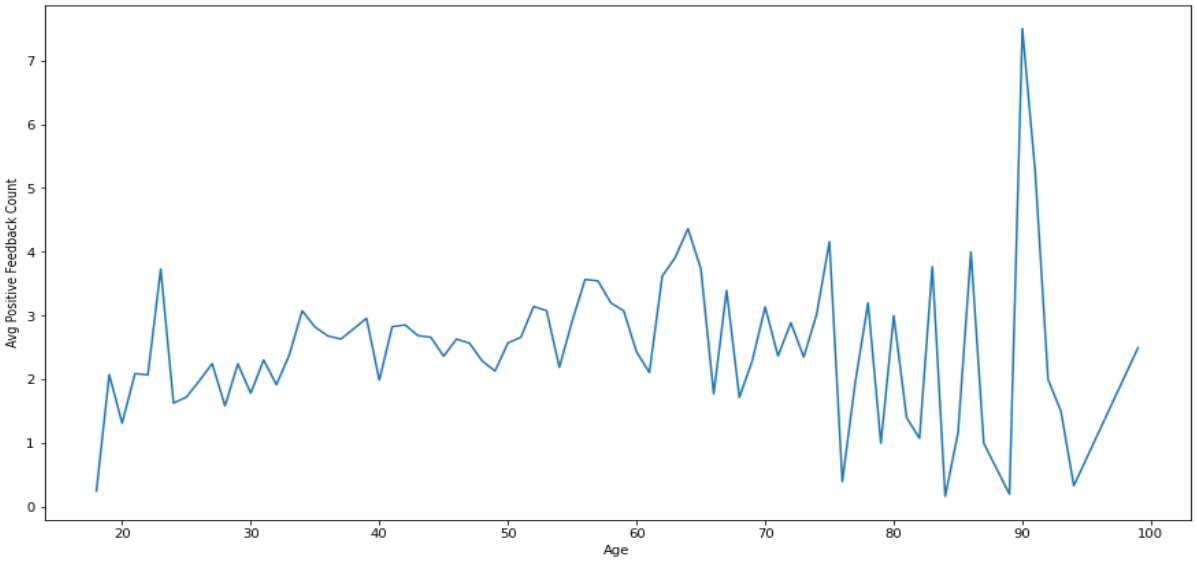


Inference:

- Women who recommended the product are more as compared to the women who did not recommended the product.
- Women who recommended the product are 81.9% and who did not are 18.1%.
- Women who recommended the product are above 17500 and who did not are less than 5000.

2. Bivariate Analysis.

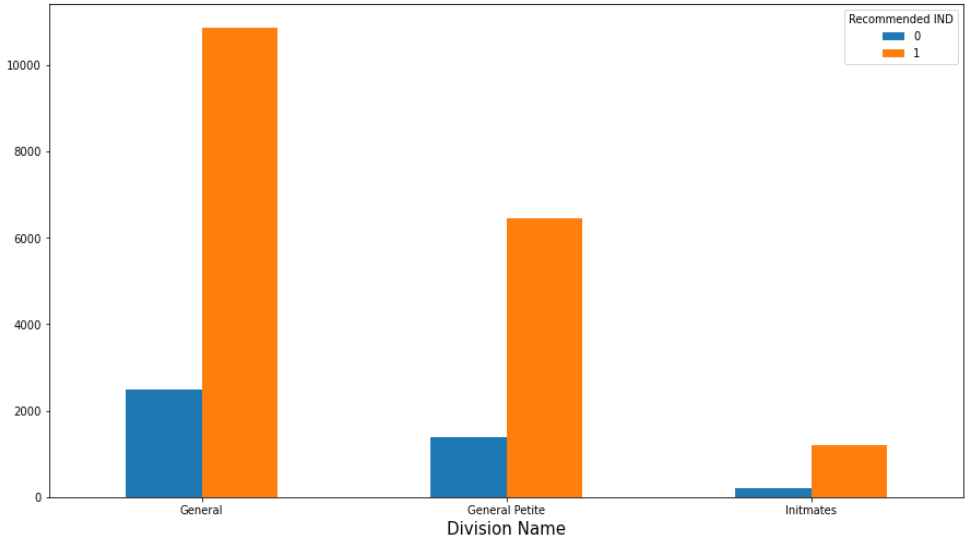
Age vs Positive Feedback Count:



Inference:

- The magnitude of fluctuation of the Average Positive Feedback Count increases as the Age increases.

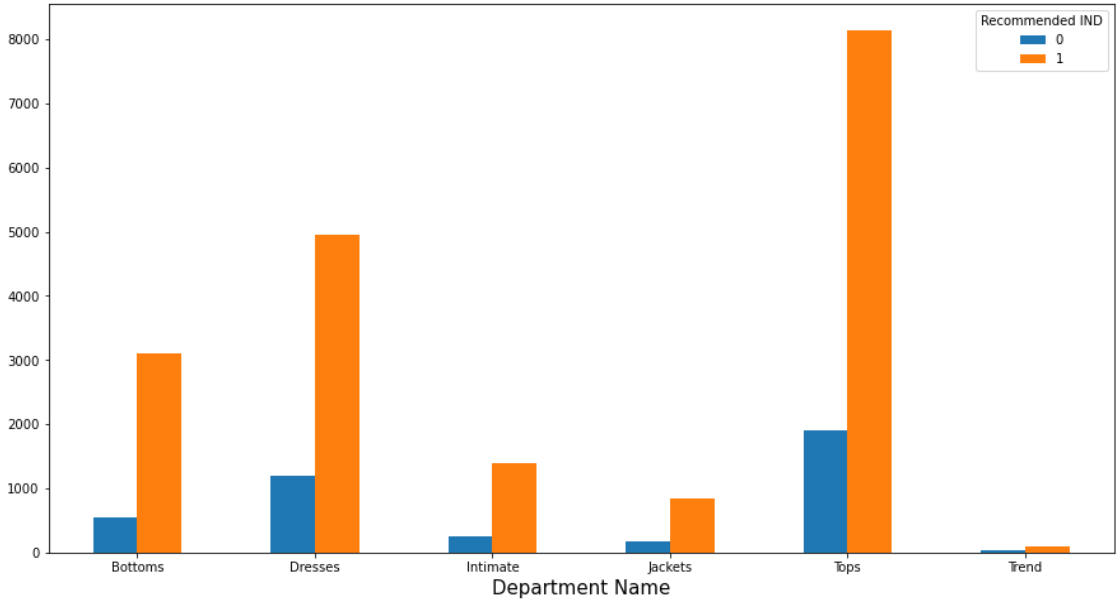
Recommended IND vs Division Name:



Inference:

- Division which is most recommended is General as compared to others and Division which is least recommended is Intimates as compared to others.
- Division which is most not recommended is also General as compared to others and Division which is least not recommended is also Intimates as compared to others.

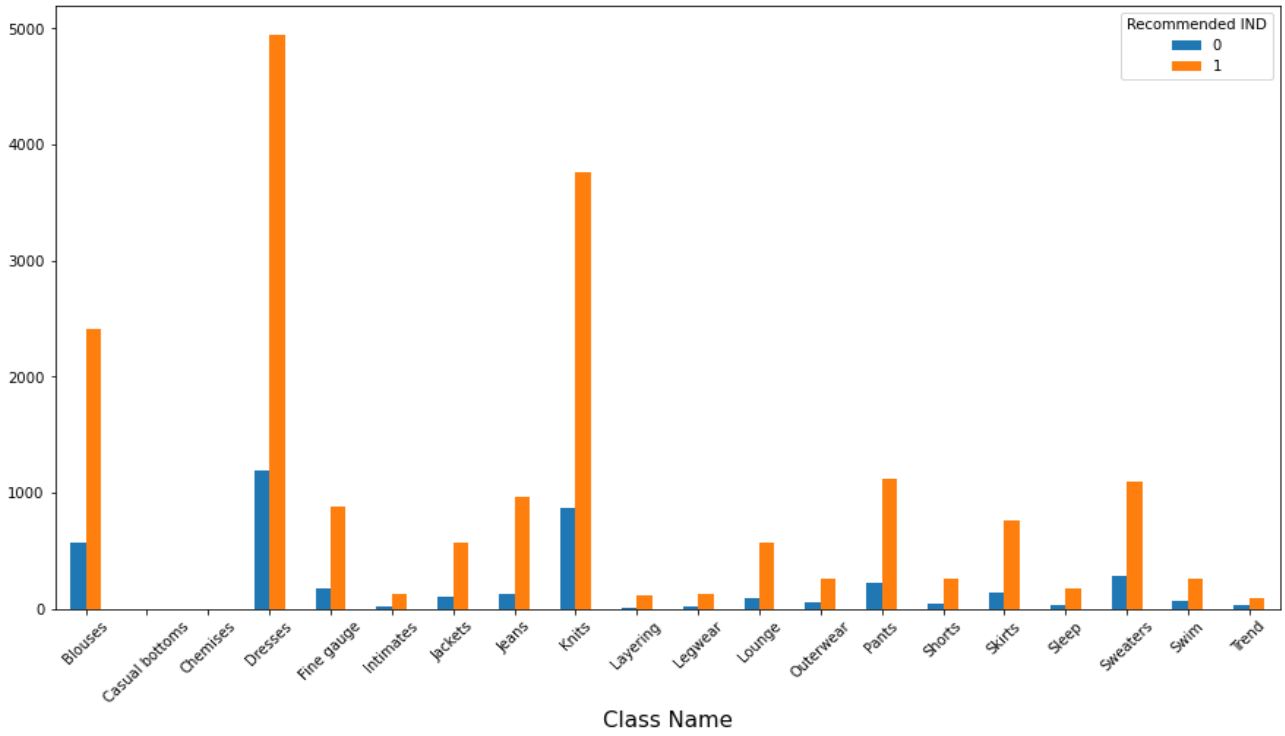
Recommended IND vs Department Name:



Inference:

- Department which is most recommended is Tops as compared to others and Department which is least recommended is Trend as compared to others.
- Department which is most not recommended is also Tops as compared to others and Department which is least not recommended is also Trend as compared to others.

Recommended IND vs Class Name:

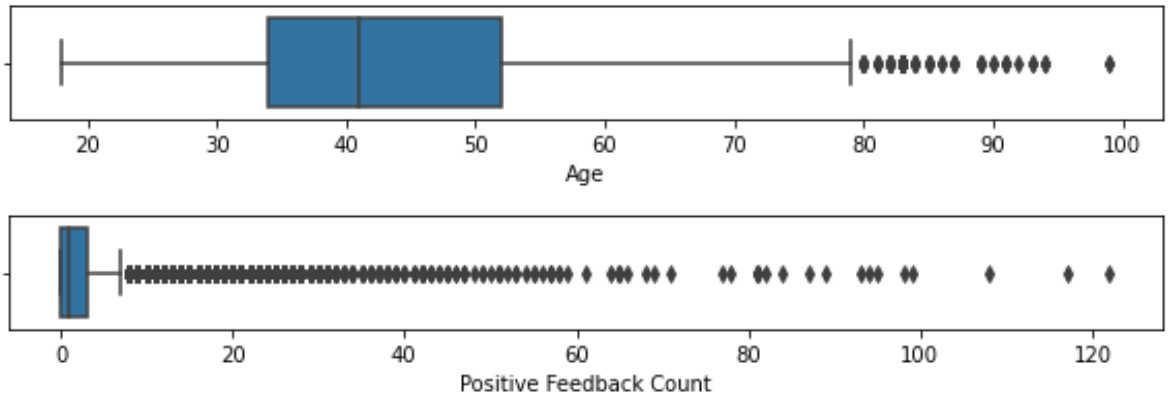


Inference:

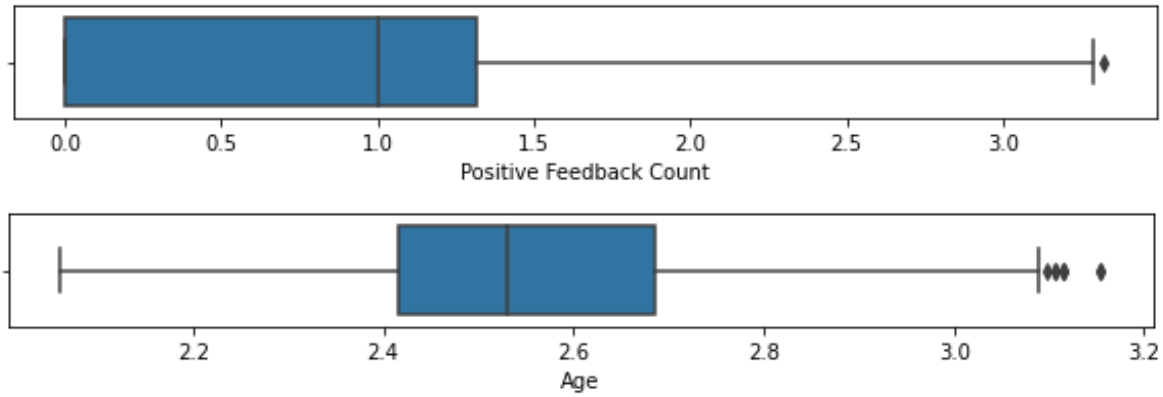
- Class which is most recommended is Dresses as compared to others and Class which are least recommended are Casual bottoms and Chemises as compared to others.
- Class which is most not recommended is also Dresses as compared to others and Class which are least not recommended are also Casual bottoms and Chemises as compared to others.

Outlier Treatment.

There are two numerical columns to check for outliers ‘Age’ and ‘Positive Feedback Count’.



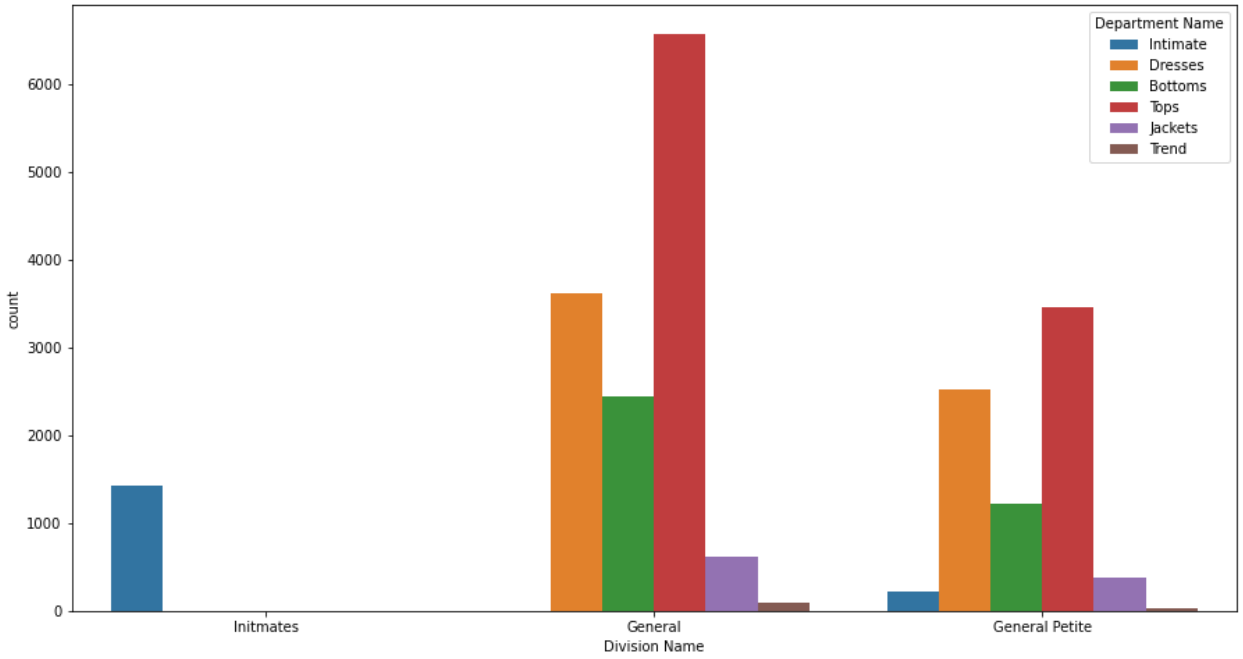
As we can see that both ‘Age’ and ‘Positive Feedback Count’ has outliers present.
So we applied ‘square root transformation’ twice to shrink the outliers into the range.

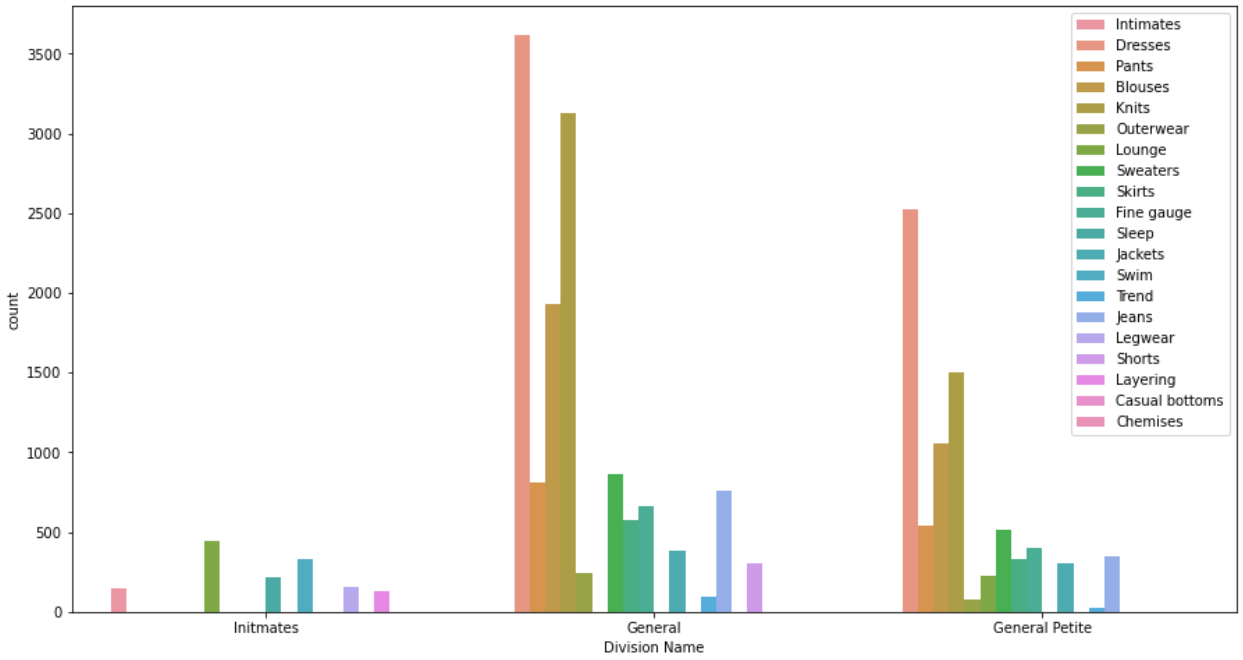


Now we can see after transformation the outliers have come within the range and the distribution of the data has become near to normal.

Feature Engineering.

Division Name:





- As we can see from the above two plots Division ‘General’ and Division ‘General Petite’ has almost same Departments as well as same Classes; we merged these two Divisions as a one ‘General’ Division.

- There are two target variables in the dataset 'Rating' and 'Recommended IND'. As we have selected 'Recommended IND' as our target variable, we need to drop 'Rating' as it is highly correlated (0.79) with the 'Recommended IND'.

- Dropping Clothing ID column as it is redundant for the analysis.

Text Analysis.

Review is the only text column present in the dataset.

When checked the top 20 most occurring words present in the reviews with and without the stop words there is lot of difference.

WordCloud

WordCloud is a visualization library which helps us to remove the stop words from the text data and display the words in a colorful way with the size of the words depends upon the number of occurrence of the words in the text data.
More the occurrence display size will be bigger; by which we can analyze the words.

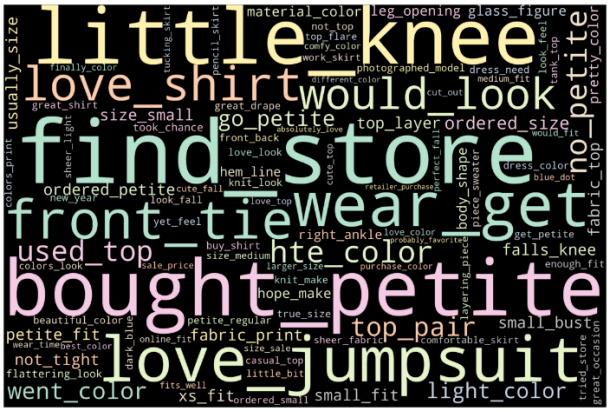
Unigrams.

These are nothing but the words or grams; the name itself say 1-word.

Recommended = 1 **Recommended = 0**

Bigrams.

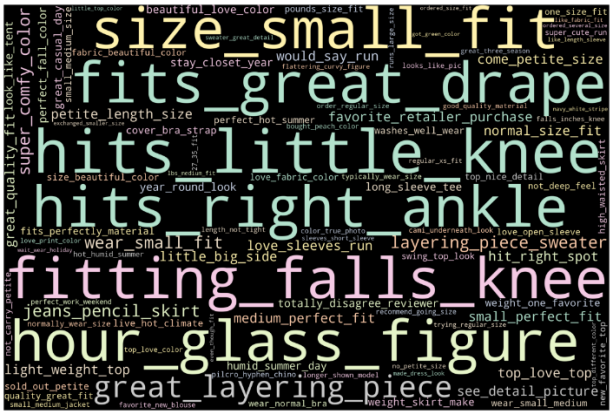
As the name says these are the combination of two words considered as a gram.



Recommended = 1 **Recommended = 0**

Trigrams.

As the name says these are the combination of three words considered as a gram.



Recommended = 1 **Recommended = 0**

Text Processing.

This part contains pre processing of text data to make it ready for the modeling.

Removal of special characters and numbers.

The first step is to remove the special characters and numbers from the text data.
Special characters are nothing but symbols like hyphen, spaces, hash etc.
For which we used ‘Regular expression’ library; we defined a function to replace these special characters and numbers with empty string (‘’).

Reducing length of the words.

In this step we cut down the unusually lengthy words.
Unusually length words are for example; ‘finallllllllly’, these kind of words.
In English language there are no words which contains letters repeating more than 2 times the consecutively.
To overcome this, we used ‘Regular expressions’ library and defined a function which replaces the more than 2 consecutive letters with empty string (‘’).

Correction of word spellings.

In this step we checked and corrected the spellings of the words present in the text data.
To check and correct the spellings we used a library called ‘Textblob’.
This Textblob checks each word from the text data and if there is a spelling mistake it will correct it and if there is no spelling mistake it will leave the word as it is.

Lemmatization.

In this step we bring down the words to their base form which has morphological meaning.
We used ‘Spacy’ library lemmatizer to perform this task.
Defined another function which takes each word from the text data and lemmatizes it.

Once we cleaned the text data the next step is to convert the text data into numeric form or creating vectors; so that we can use them into a model.

Count Vectorizer.

In this count vectorizer, it takes the text data and firstly remove the stop words from it.
Then it converts the text data into a vector, if you convert that vector to an array and then convert it into a dataframe, then you will get DTM (document term matrix).
DTM is a mathematical matrix that describes the frequency of terms(words) that occurred in a text data.

TF-IDF Vectorizer.

This is also a vectorizer which converts text data into a vector.
TF-IDF: Term Frequency Inverse Document Frequency.
Tfidf Vectorizer will tokenize documents, learn the vocabulary and inverse document frequency weightings, and allow to encode new document.

Word Embedding – Word2Vec.

We have also tried out word embedding using word2vec library.

This also converts text data into vectors but these vectors are different from vectors we get by doing count vectorizer and Tfidf vectorizer.
We get sparse matrix when we do Count vectorizer and Tfidf vectorizer, but when we do word2vec we get dense matrix, by which we can find out the relation between the words.
Which works on neural network.

As the above text processing part complete we concatenated the two dataframe:

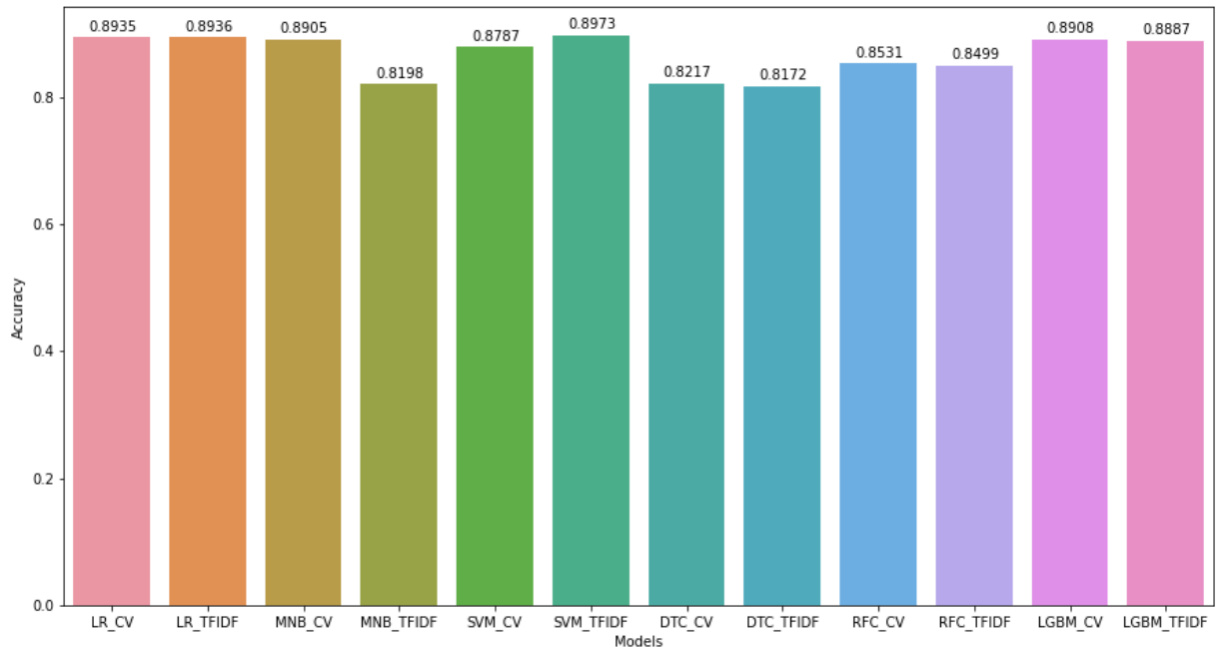
1. **Text data with count vectirizer and rest other processed columns.**
2. **Text data with Tfidf vectorizer and rest other processed columns.**

Modeling.

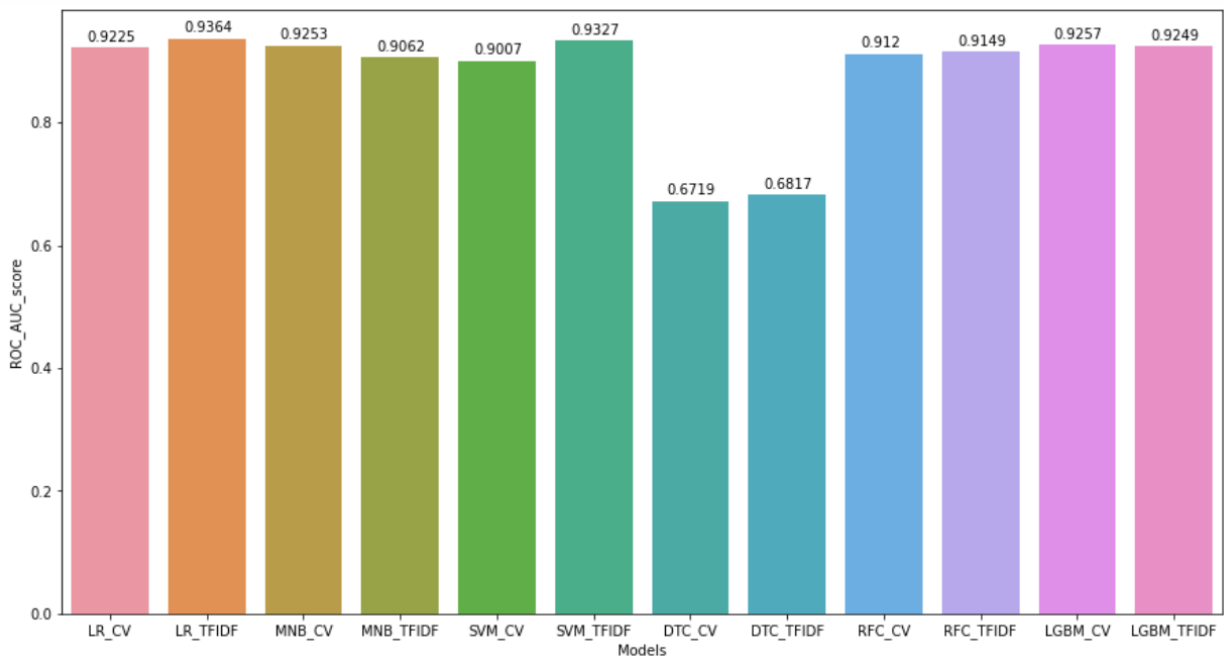
We have built 12 models of different algorithms with combination of count vectorizer or Tfidf vectorizer and compared their results.

Algorithms:
Linear Regression.
Naïve Bayes.

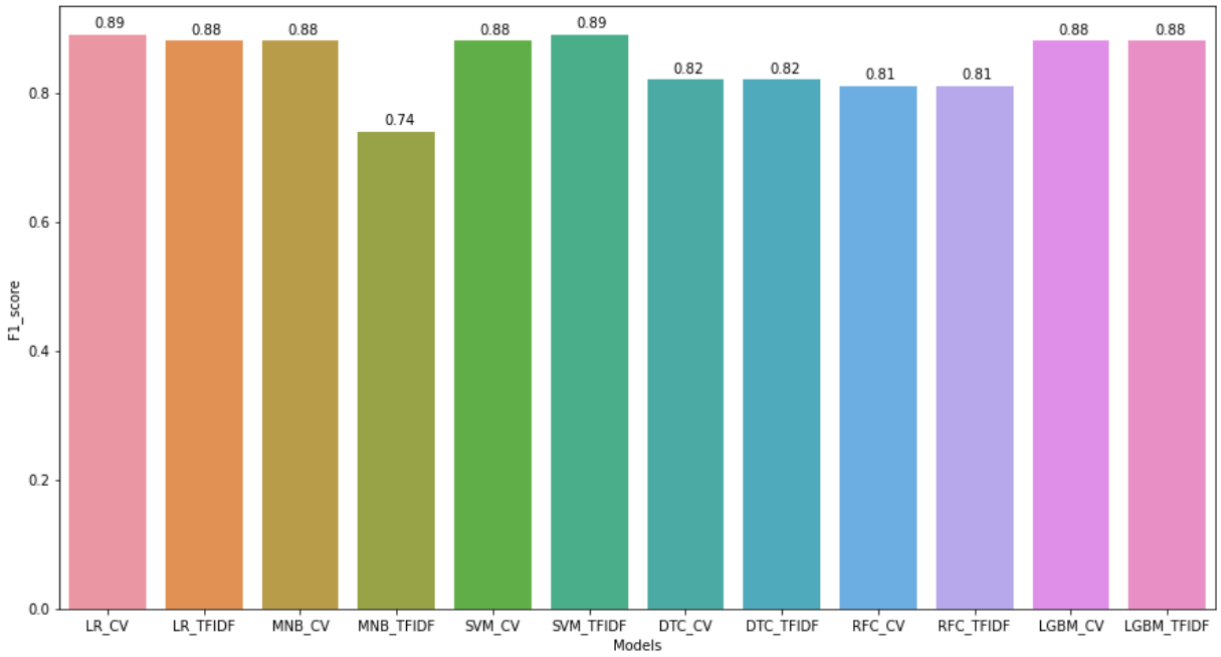
Support Vector Machine.
Decision Tree.
Random Forest.
LightGBM.



We can see SVM_TFIDF is giving the Highest accuracy.



We can see that LR_TFIDF is the with highest roc-auc score and SVM_TFIDF is giving the second highest.



We can see LR_CV and SVM_TFIDF are giving the highest f1 score.

After comparing all the metrics of the models; we can see the best results are from the SVM_TFIDF model, so we consider SVM_TFIDF model further for doing hyper-parameter tunning.

After doing hyper-parameter tunning our final model results are:
Accuracy: 0.8973
AUC score: 0.9327
F1 score: 0.89

There is no much difference between the tunned model and the base model as the dataset is huge the base model is giving the best results.
We can see class 1 is predicting well as compared to the class 0; as the dataset is imbalanced class 1 is the majority class and 0 is the minority class.

As we see the results of the final model, we can say that the model is doing good, as majority target column dependency is on text data; so better modeling can be done using 'Neural Network' algorithms.

Project Outcome:

The project outcome is economically the model we built will be helping the sellers on the platform and the platform itself understand which are the products that are selling and the ones that are not.
This project will significantly help the women e-commerce clothing sites to improve/modify or create new products which will satisfy the customer needs and as an upside increase the business for these sites.

Complexity Involved:

There is quite a bit of complexity involved since we did Natural language Processing

We learnt and used many libraries:

1. Regular expressions (Regex).

2. Spacy.

3. NLTK.

4. WordCloud.

5. Textblob.

6. Word2Vec.

The final challenge was to handle the huge amount of data which it became after text processing part with the moderately configured machine.

we tackled ‘Memory error’.

Learning:

As show above, we learnt using different libraries, handling huge dataset, tackling of memory error, little bit of flask api.
To improve the model performance and accuracy, next time we will learn the ‘Neural Network’ apply those algorithms by learning some more libraries/tools of Natural Language Processing.

Deployment.

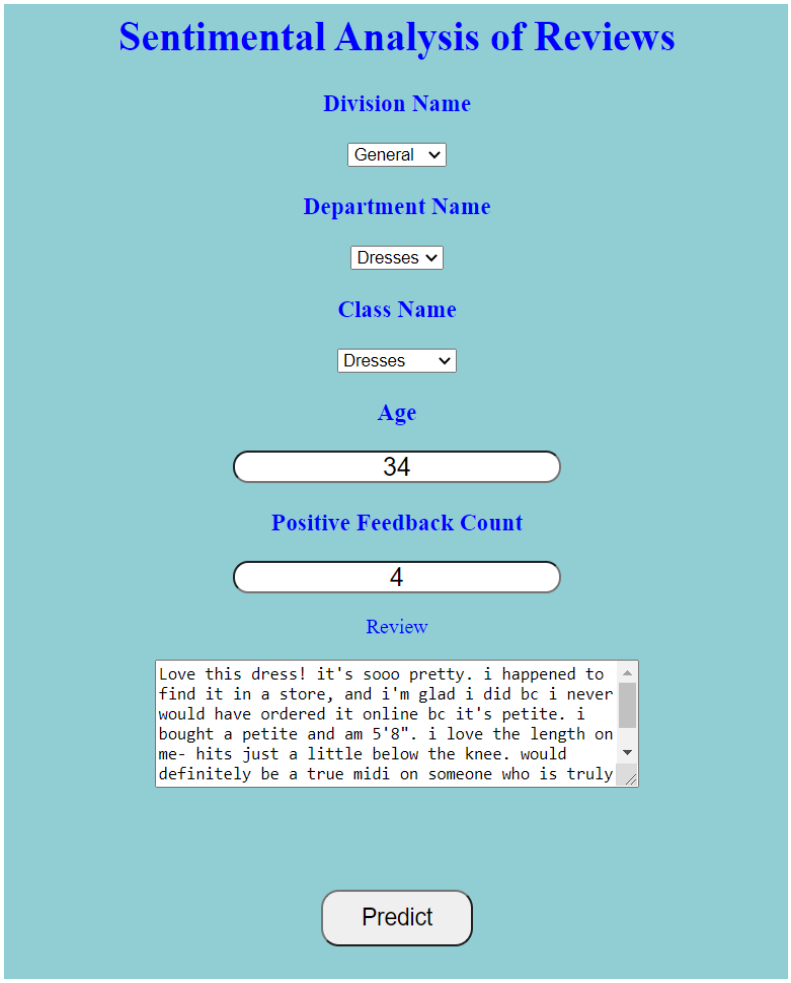
Building Flask api:

Flask is a web framework for python. It provides functionality for building web applications. Out of the numerous functionalities that Flask has to offer, we have used predominantly 3 features:

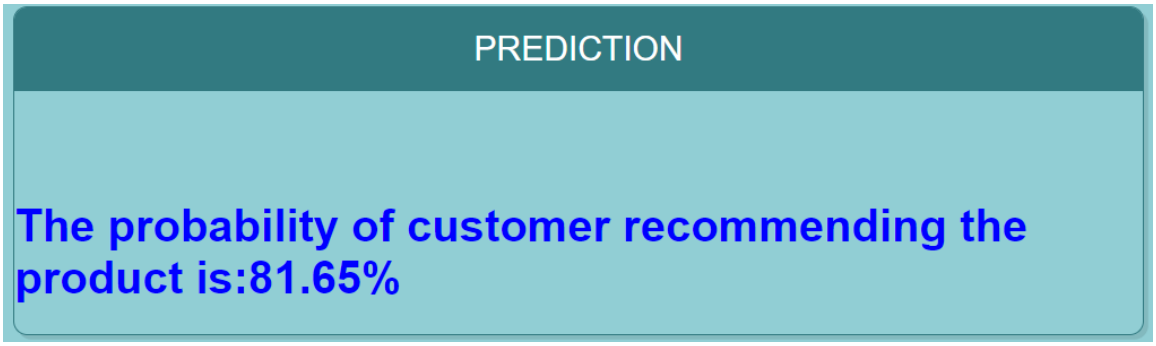
- Flask forms: To create a survey form to get the user input for different features.
- HTTP requests: Used to take the inputs given and store them in a variable and converted to a DataFrame.
- Rendering Templates: Used two templates index.html which hosts the the main form. As soon as the user hits predict, the results.html file gets rendered.

Below are a 2 snapshots of the Flask api running on local machine:

Index.html:



Result.html:


PREDICTION
The probability of customer recommending the product is:81.65%