2023, May 19th Presented to Professor Behzad Ahmadi

Data Mining

Women's Clothing Review

Prepared by Shyamala Venkatakrishnan
Qihuan He (Abby), Heejae Roh

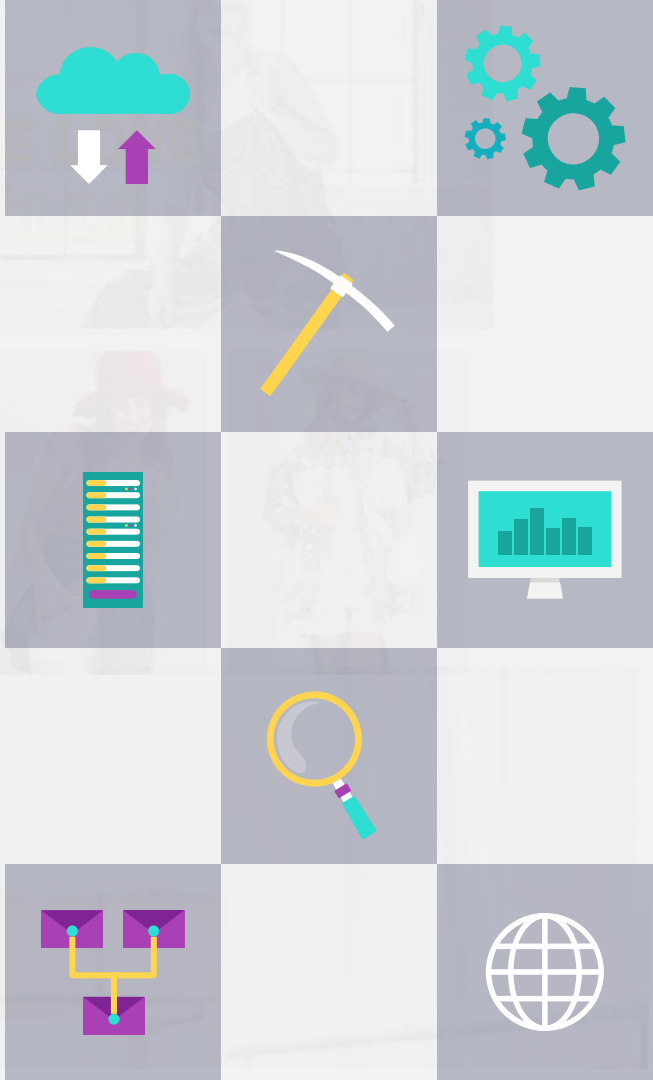
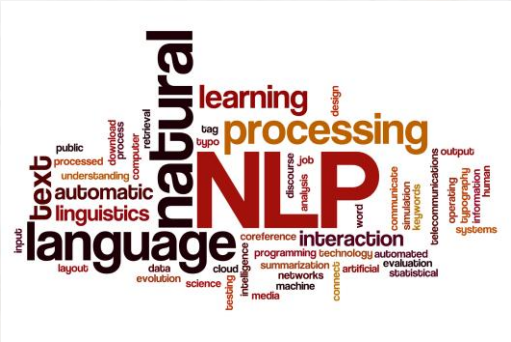
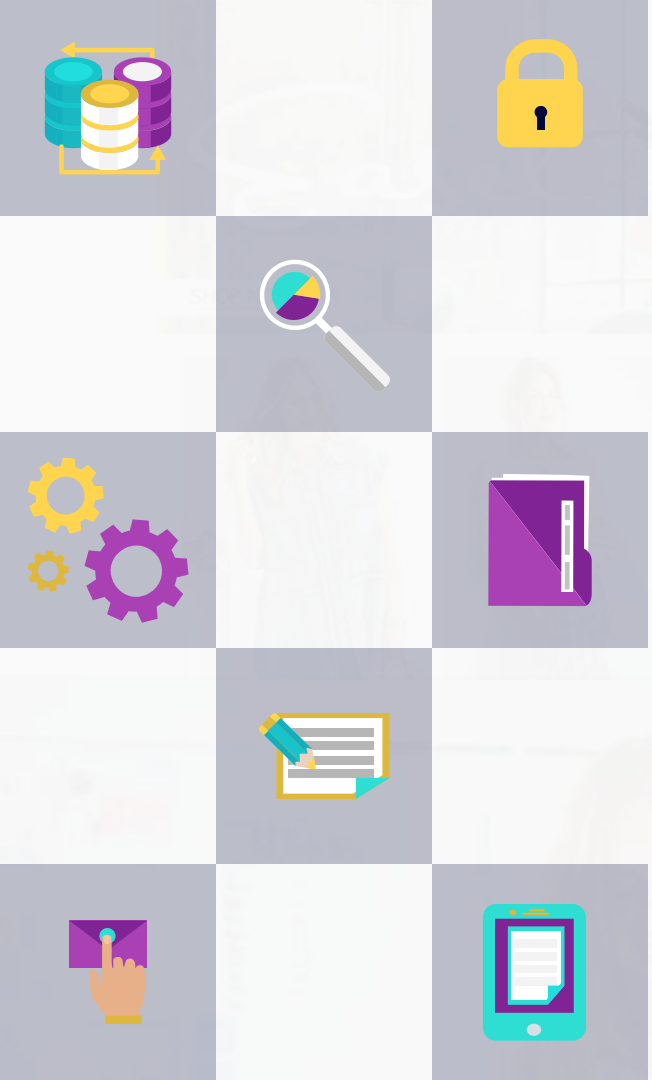


Table of contents



Primary Question

- What is the maximum **insight** we can get from **text mining**?
- **Predict** if the customer would recommend a product or not based on **review text.**



About the Data & EDA

01

Data Cleaning & Exploratory Analysis

Data Cleaning

	A	B	C	D	E	F	G	H	I	J	K
1	id	Clothing ID	Age	Title	Review Text	Rating	Recommended IND	Positive Feedback Count	Division Name	Department Name	Class Name
2	0	767	33		Absolutely wonderful - silky an	4	1	0	Initmates	Intimate	Intimates
3	1	1080	34		Love this dress! it's sooo pretty	5	1	4	General	Dresses	Dresses
4	2	1077	60	Some major design fla	I had such high hopes for this d	3	0	0	General	Dresses	Dresses
5	3	1049	50	My favorite buy!	I love, love, love this jumpsuit.	5	1	0	General Petite	Bottoms	Pants
6	4	847	47	Flattering shirt	This shirt is very flattering to all	5	1	6	General	Tops	Blouses
7	5	1080	49	Not for the very petite	I love tracy reese dresses, but th	2	0	4	General	Dresses	Dresses
8	6	858	39	Cagrcoal shimmer fun	I aded this in my basket at hte l	5	1	1	General Petite	Tops	Knits
9	7	858	39	Shimmer, surprisingly	I ordered this in carbon for stor	4	1	4	General Petite	Tops	Knits
10	8	1077	24	Flattering	I love this dress. i usually get ar	5	1	0	General	Dresses	Dresses
11	9	1077	34	Such a fun dress!	I'm 5"5' and 125 lbs. i ordered t	5	1	0	General	Dresses	Dresses
12	10	1077	53	Dress looks like it's ma	Dress runs small esp where the	3	0	14	General	Dresses	Dresses
13	11	1095	39		This dress is perfection! so pret	5	1	2	General Petite	Dresses	Dresses
14	12	1095	53	Perfect!!!	More and more i find myself re	5	1	2	General Petite	Dresses	Dresses
15	13	767	44	Runs big	Bought the black xs to go	5	1	0	Initmates	Intimate	Intimates
16	14	1077	50	Pretty party dress with	This is a nice choice for holiday	3	1	1	General	Dresses	Dresses
17	15	1065	47	Nice, but not for my b	I took these out of the package	4	1	3	General	Bottoms	Pants
23476	23474	1104	32	Much better in person	Yes, this is a great dress! i wasn	5	1	0	General Petite	Dresses	Dresses
23477	23475	1104	41	Cute dress	Cute dress but not for me. the	3	1	0	General Petite	Dresses	Dresses
23478	23476	522	27	Cheeky!	These bottoms are very cute bu	4	1	0	Initmates	Intimate	Swim
23479	23477	1094	39	Entrancing	I'm so impressed with the beau	4	1	5	General Petite	Dresses	Dresses
23480	23478	1104	32	Unflattering	I was surprised at the positive r	1	0	0	General Petite	Dresses	Dresses
23481	23479	1005	42	What a fun piece!	So i wasn't sure about ordering	5	1	0	General Petite	Bottoms	Skirts
23482	23480	862	35			5	1	0	General Petite	Tops	Knits
23483	23481	1104	34	Great dress for many	I was very happy to snag this dr	5	1	0	General Petite	Dresses	Dresses
23484	23482	862	48	Wish it was made of c	It reminds me of maternity	3	1	0	General Petite	Tops	Knits
23485	23483	1104	31	Cute, but see through	This fit well, but the top was ve	3	0	1	General Petite	Dresses	Dresses
23486	23484	1084	28	Very cute dress, perfe	I bought this dress for a weddir	3	1	2	General	Dresses	Dresses
23487	23485	1104	52	Please make more like	This dress in a lovely platinum i	5	1	22	General Petite	Dresses	Dresses

- The dataset on e-commerce women's clothing reviews from Kaggle.
- Initially, it comprised 23,486 observations and 11 variables.
- Removing ‘id’ column and missing values.

Primary and Target variable

	A	B	C	D	E	F	G	H	I	J	K
1	id	Clothing ID	Age	Title	Review Text	Rating	Recommended IND	Positive Feedback Count	Division Name	Department Name	Class Name
2	0	767	33		Absolutely wonderful - silky an	4	1	0	Initmates	Intimate	Intimates
3	1	1080	34		Love this dress! it's sooo pretty	5	1	4	General	Dresses	Dresses
4	2	1077	60	Some major design fl	I had such high hopes for this d	3	0	0	General	Dresses	Dresses
5	3	1049	50	My favorite buy!	I love, love, love this jumpsuit.	5	1	0	General Petite	Bottoms	Pants
6	4	847	47	Flattering shirt	This shirt is very flattering to all	5	1	6	General	Tops	Blouses
7	5	1080	49	Not for the very petit	I love tracy reese dresses, but th	2	0	4	General	Dresses	Dresses
8	6	858	39	Cagrcoal shimmer fu	I aded this in my basket at hte l	5	1	1	General Petite	Tops	Knits
9	7	858	39	Shimmer, surprisingly	I ordered this in carbon for stor	4	1	4	General Petite	Tops	Knits
10	8	1077	24	Flattering	I love this dress. i usually get ar	5	1	0	General	Dresses	Dresses
11	9	1077	34	Such a fun dress!	I'm 5"5' and 125 lbs. i ordered t	5	1	0	General	Dresses	Dresses
12	10	1077	53	Dress looks like it's m	Dress runs small esp where the	3	0	14	General	Dresses	Dresses
13	11	1095	39		This dress is perfection! so pret	5	1	2	General Petite	Dresses	Dresses
14	12	1095	53	Perfect!!!	More and more i find myself re	5	1	2	General Petite	Dresses	Dresses
15	13	767	44	Runs big	Bought the black xs to go	5	1	0	Initmates	Intimate	Intimates
16	14	1077	50	Pretty party dress wit	This is a nice choice for holiday	3	1	1	General	Dresses	Dresses
17	15	1065	47	Nice, but not for my	I took these out of the package	4	1	3	General	Bottoms	Pants
23476	23474	1104	32	Much better in perso	Yes, this is a great dress! i wasn	5	1	0	General Petite	Dresses	Dresses
23477	23475	1104	41	Cute dress	Cute dress but not for me. the	3	1	0	General Petite	Dresses	Dresses
23478	23476	522	27	Cheeky!	These bottoms are very cute bu	4	1	0	Initmates	Intimate	Swim
23479	23477	1094	39	Entrancing	I'm so impressed with the beau	4	1	5	General Petite	Dresses	Dresses
23480	23478	1104	32	Unflattering	I was surprised at the positive r	1	0	0	General Petite	Dresses	Dresses
23481	23479	1005	42	What a fun piece!	So i wasn't sure about ordering	5	1	0	General Petite	Bottoms	Skirts
23482	23480	862	35			5	1	0	General Petite	Tops	Knits
23483	23481	1104	34	Great dress for many	I was very happy to snag this dr	5	1	0	General Petite	Dresses	Dresses
23484	23482	862	48	Wish it was made of	It reminds me of maternity	3	1	0	General Petite	Tops	Knits
23485	23483	1104	31	Cute, but see through	This fit well, but the top was ve	3	0	1	General Petite	Dresses	Dresses
23486	23484	1084	28	Very cute dress, perfe	I bought this dress for a weddir	3	1	2	General	Dresses	Dresses
23487	23485	1104	52	Please make more lik	This dress in a lovely platinum i	5	1	22	General Petite	Dresses	Dresses

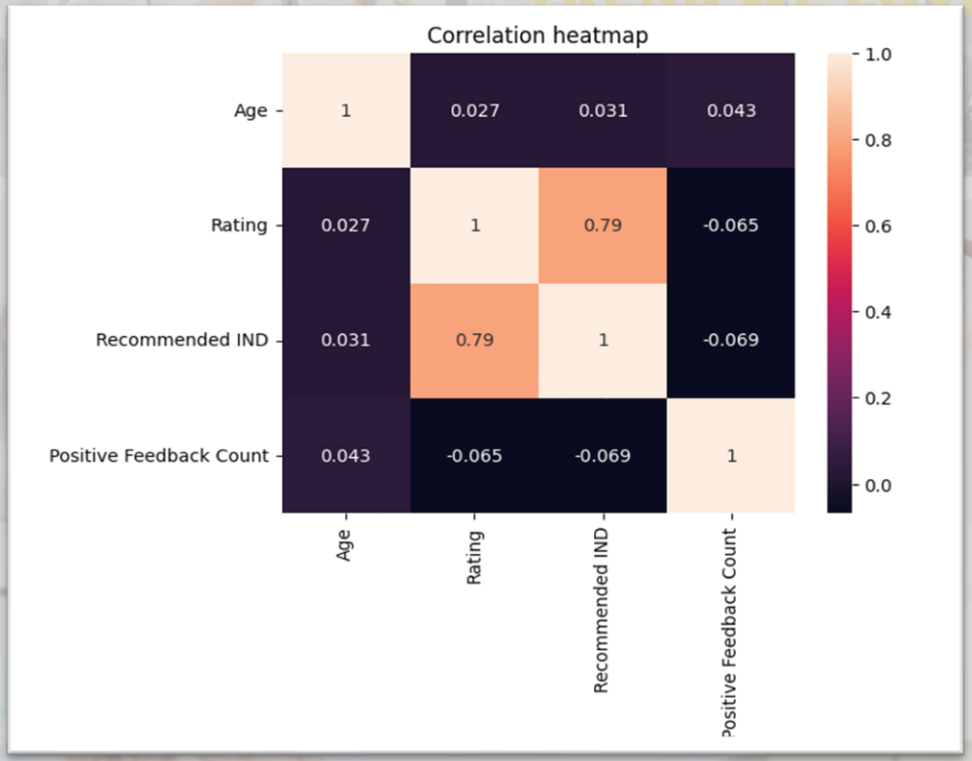
- The dataset contained 19,662 observations and 10 variables.
- Target variable is the Recommended IND (indicator) column
- Focusing on 'Review Text' for Text Analysis

Descriptive Analysis

X	mean	sd	median	trimmed	mad	min	max	range	skewed	kurtosis	se
clothing_id	921	200	936	956	158	1	1205	1204	-2.10	5.35	1.4
age*	43.3	12.3	41	42.6	11.9	18	99	81	0.52	-0.14	0.1
title*	6922	3950	7022	6908	5056	1	13983	13982	0.02	-1.14	28.2
review_text*	9829	5674	9830	9829	7284	1	19656	19655	0.00	-1.20	40.4
rating	4.18	1.11	5	4.4	0.00	1	5	4	-1.28	0.71	0.0
recommended_ind	0.82	0.39	1.0	0.9	0.00	0	1	1	-1.65	0.72	0.00
positive_feedback_count	2.65	5.83	1.0	1.38	1.48	0	122	122	6.34	67.49	0.04
division_name*	1.47	0.61	1.0	1.38	0.00	1	3	2	0.94	-0.15	0.00
department_name*	3.35	1.63	3.0	3.43	2.97	1	6	5	-0.16	-1.67	0.01
class_name*	7.92	5.22	8.0	7.53	5.93	1	20	19	0.58	-0.69	0.04

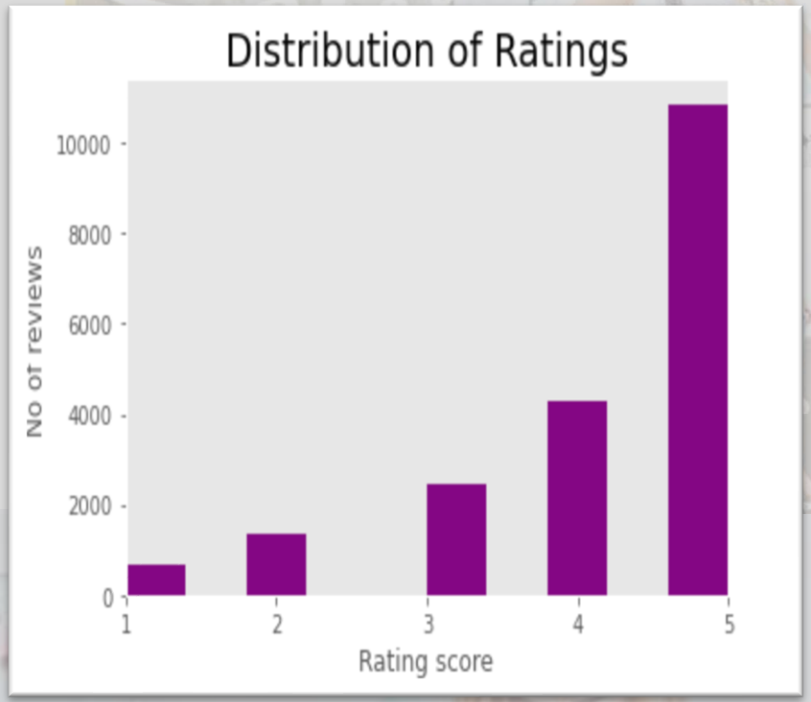
- The recommended_ind variable has a mean of 0.82, with the large proportion of values equaling 1.
- The age variable ranges from 18 to 99, with a median value of 41.

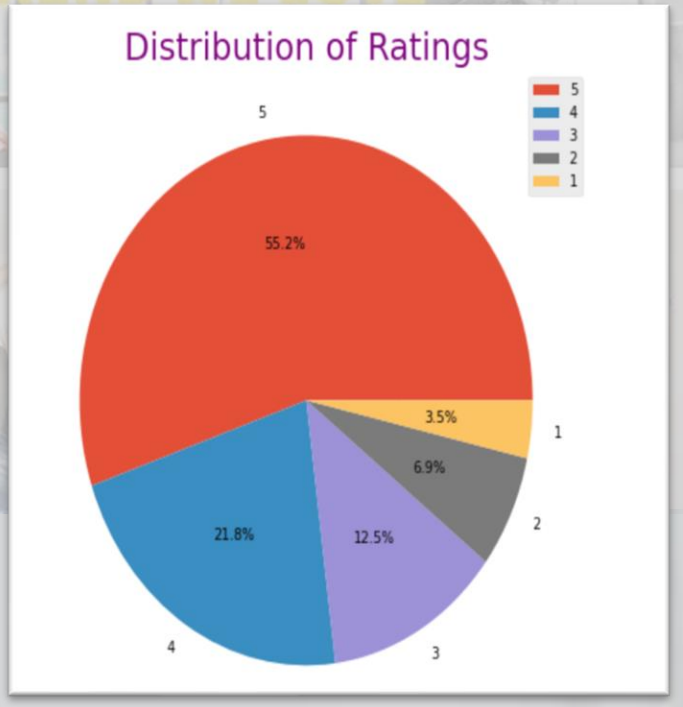
Correlation Matrix



- The recommended_ind and Rating is highly related as 0.79.
- The other variables are rarely related to the recommended_ind.

Rating, EDA1





- The majority of customers (approximately 77%) were satisfied with the women's clothing products, giving a rating of 4 or 5.
- A small percentage of customers (approximately 3.5% and 6.9%) gave ratings of 1 or 2, respectively.

Rating, EDA2

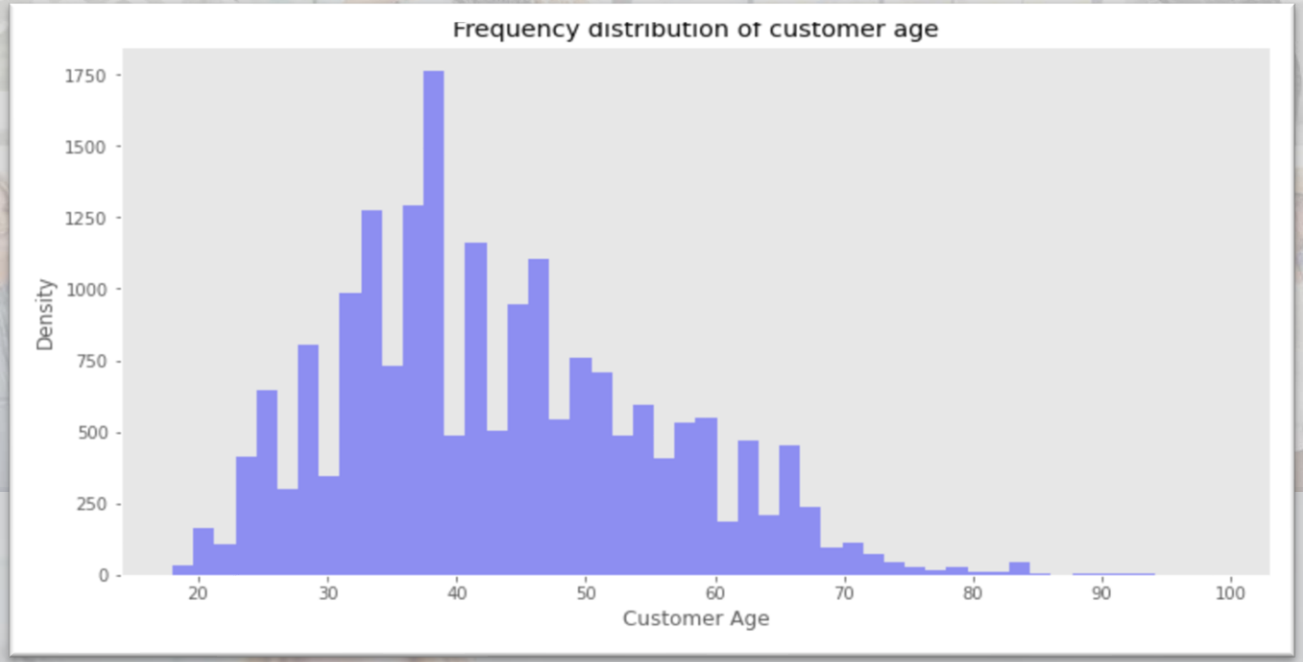
- The treemap shows the total number of reviews for each division, department, and class of clothes.
- A color scale indicates the average ratings received for each class.
- Jeans, Jackets, and Lounge products received the highest average rating of 4.3 and above.
- Knits, Dresses, Sweaters, and Blouses received an average rating of 4 to 4.1.

Data Cleaning after EDA

	A	B	C	D	E	F	G	H	I	J	K
1	id	Clothing ID	Age	Title	Review Text	Rating	Recommended IND	Positive Feedback Count	Division Name	Department Name	Class Name
2	0	767	33		Absolutely wonderful - silky an	4	1	0	Initmates	Intimate	Intimates
3	1	1080	34		Love this dress! it's sooo pretty	5	1	4	General	Dresses	Dresses
4	2	1077	60	Some major design fla	I had such high hopes for this d	3	0	0	General	Dresses	Dresses
5	3	1049	50	My favorite buy!	I love, love, love this jumpsuit.	5	1	0	General Petite	Bottoms	Pants
6	4	847	47	Flattering shirt	This shirt is very flattering to all	5	1	6	General	Tops	Blouses
7	5	1080	49	Not for the very petite	I love tracy reese dresses, but th	2	0	4	General	Dresses	Dresses
8	6	858	39	Cagrcoal shimmer fun	I aded this in my basket at hte l	5	1	1	General Petite	Tops	Knits
9	7	858	39	Shimmer, surprisingly	I ordered this in carbon for stor	4	1	4	General Petite	Tops	Knits
10	8	1077	24	Flattering	I love this dress. i usually get ar	5	1	0	General	Dresses	Dresses
11	9	1077	34	Such a fun dress!	I'm 5"5' and 125 lbs. i ordered t	5	1	0	General	Dresses	Dresses
12	10	1077	53	Dress looks like it's ma	Dress runs small esp where the	3	0	14	General	Dresses	Dresses
13	11	1095	39		This dress is perfection! so pret	5	1	2	General Petite	Dresses	Dresses
14	12	1095	53	Perfect!!!	More and more i find myself re	5	1	2	General Petite	Dresses	Dresses
15	13	767	44	Runs big	Bought the black xs to go	5	1	0	Initmates	Intimate	Intimates
16	14	1077	50	Pretty party dress with	This is a nice choice for holiday	3	1	1	General	Dresses	Dresses
17	15	1065	47	Nice, but not for my b	I took these out of the package	4	1	3	General	Bottoms	Pants
23476	23474	1104	32	Much better in person	Yes, this is a great dress! i wasn	5	1	0	General Petite	Dresses	Dresses
23477	23475	1104	41	Cute dress	Cute dress but not for me. the	3	1	0	General Petite	Dresses	Dresses
23478	23476	522	27	Cheeky!	These bottoms are very cute bu	4	1	0	Initmates	Intimate	Swim
23479	23477	1094	39	Entrancing	I'm so impressed with the beau	4	1	5	General Petite	Dresses	Dresses
23480	23478	1104	32	Unflattering	I was surprised at the positive r	1	0	0	General Petite	Dresses	Dresses
23481	23479	1005	42	What a fun piece!	So i wasn't sure about ordering	5	1	0	General Petite	Bottoms	Skirts
23482	23480	862	35			5	1	0	General Petite	Tops	Knits
23483	23481	1104	34	Great dress for many	I was very happy to snag this dr	5	1	0	General Petite	Dresses	Dresses
23484	23482	862	48	Wish it was made of c	It reminds me of maternity	3	1	0	General Petite	Tops	Knits
23485	23483	1104	31	Cute, but see through	This fit well, but the top was ve	3	0	1	General Petite	Dresses	Dresses
23486	23484	1084	28	Very cute dress, perfe	I bought this dress for a weddir	3	1	2	General	Dresses	Dresses
23487	23485	1104	52	Please make more like	This dress in a lovely platinum i	5	1	22	General Petite	Dresses	Dresses

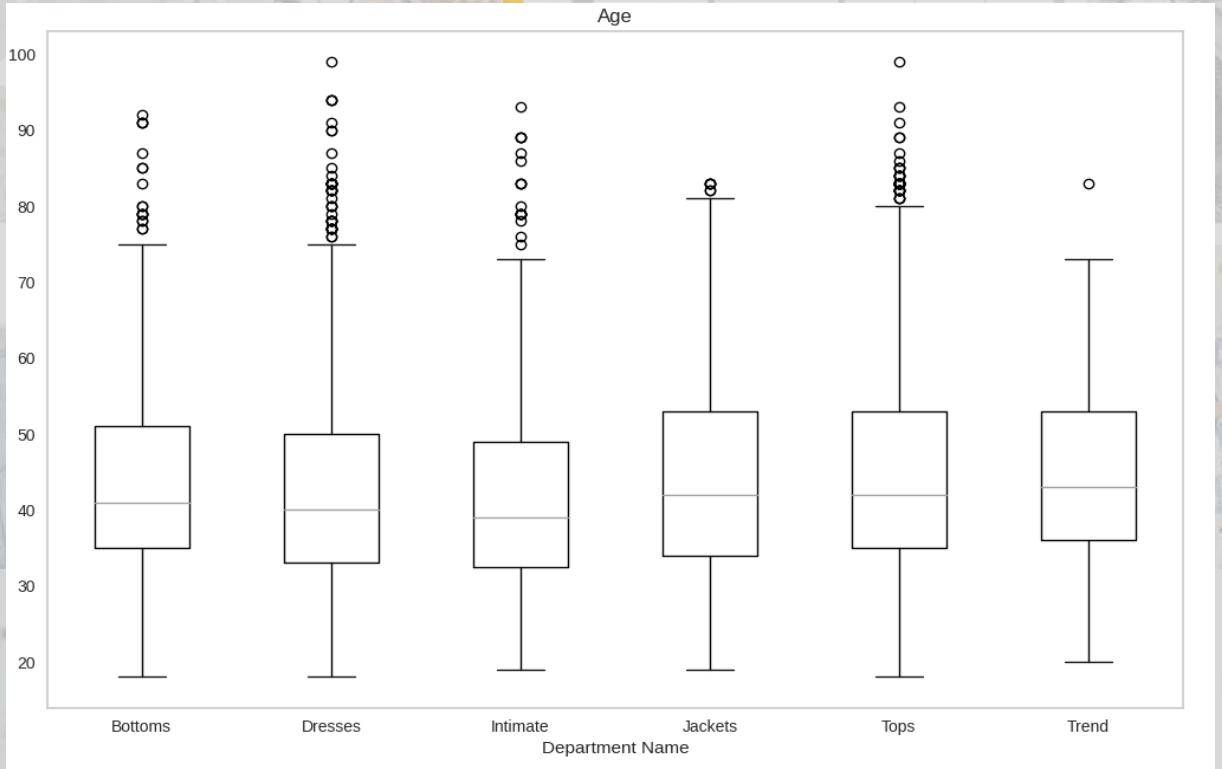
- After conducting EDA, performed data cleaning once again.
- Gaining more insight into the dataset, we can reconsidered changing approach to handling missing values or data cleaning.

Frequency Distribution of Age



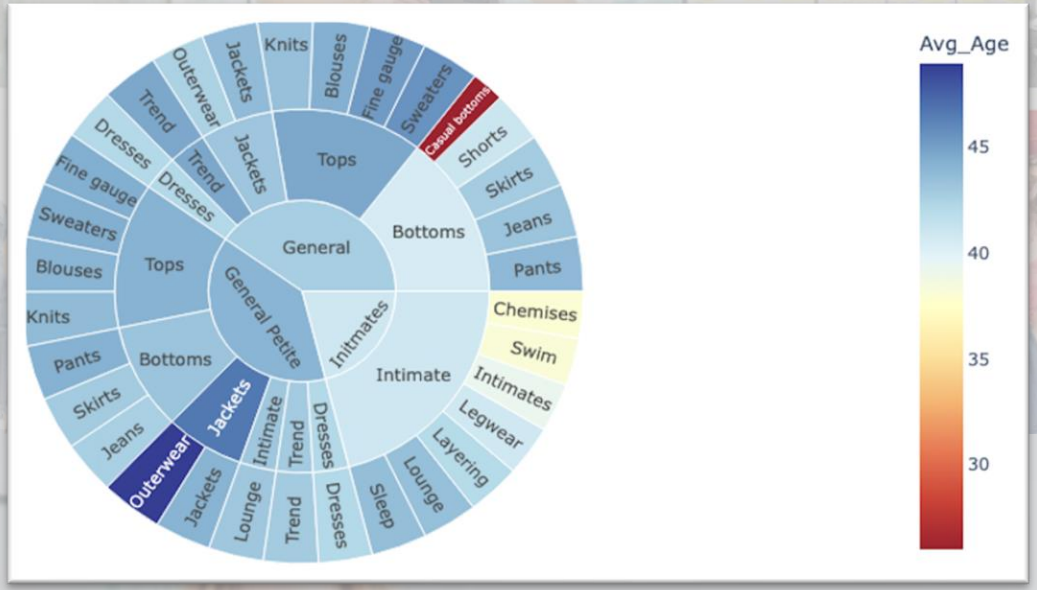
- The histogram indicates that the majority of customers fall within the 30-50 age group.
- Middle-aged customers predominantly purchase and review clothing items.
- Designers and retailers should consider producing more clothes for customers aged 30-50.

Distribution of Age



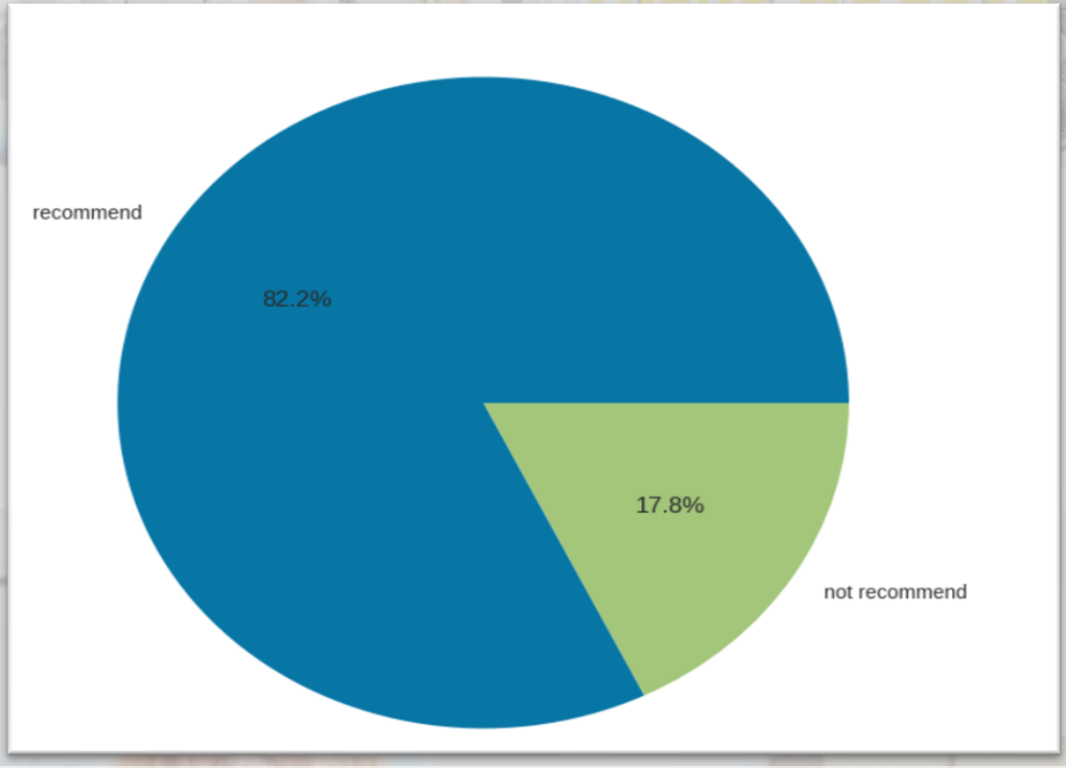
- Customers who purchased intimates tend to be slightly younger than those who provided feedback on jackets and tops.
- The age distribution for each category includes outliers aged over 75 years.

Distribution of Age by Different types of clothing items



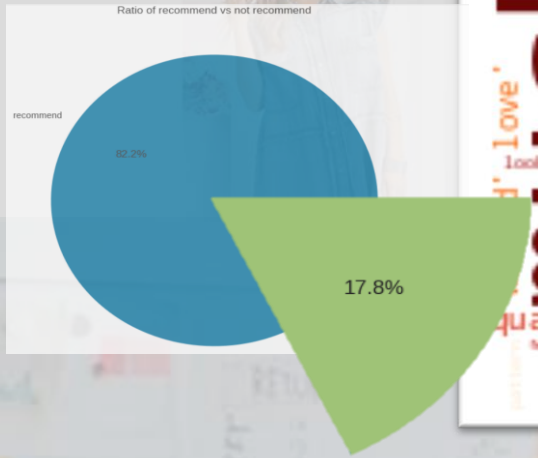
- The sunburst chart reveals the age distribution of customers purchasing various clothing items.
- Customers aged above 45 predominantly buy outerwear and jackets.
- Customers aged 35 to 40 prefer chemises, swimwear, and intimates.
- The casual bottoms category is popular among customers below the age of 30.

Recommended Indicator



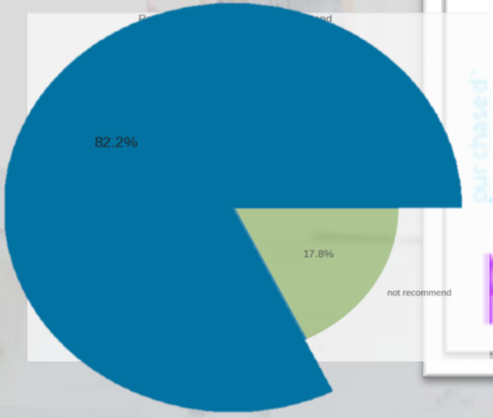
- The pie chart shows customer recommendations for the products.
- The majority of customers (82.2% of feedback) are likely to recommend the products.

Word-Cloud, Not Recommended

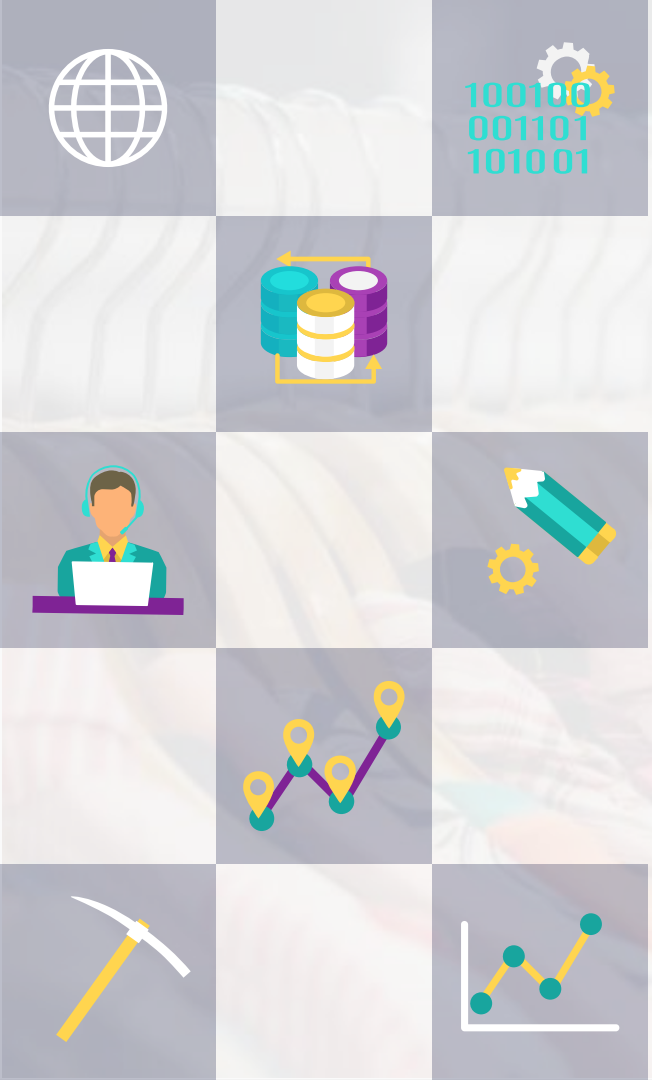


- The word cloud represents customer reviews of the products they did not recommend.
- Keywords from the negative feedback include fit, small, large, disappointed, unfortunately, and little.

Word-Cloud, Recommended



- The word cloud represents customer reviews of the products they recommended.
- Keywords from the positive feedback include perfect, love, beautiful, comfortable, soft, style, etc.



Text pre-processing

02

Data preprocessing: Data cleaning

Step 1 — Convert text to lowercase

Step 2 — Remove Punctuation

Step 3 — Use regular expression to identify non-ascii characters in text

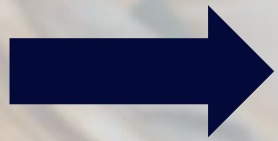
Step 4 — Remove stopwords

Step 5 — Word lemmatization to convert to root form

Data preprocessing: Data cleaning

Before

I love, love, love this jumpsuit. it's fun, flirty, and fabulous! every time i wear it, i get nothing but great compliments!



After

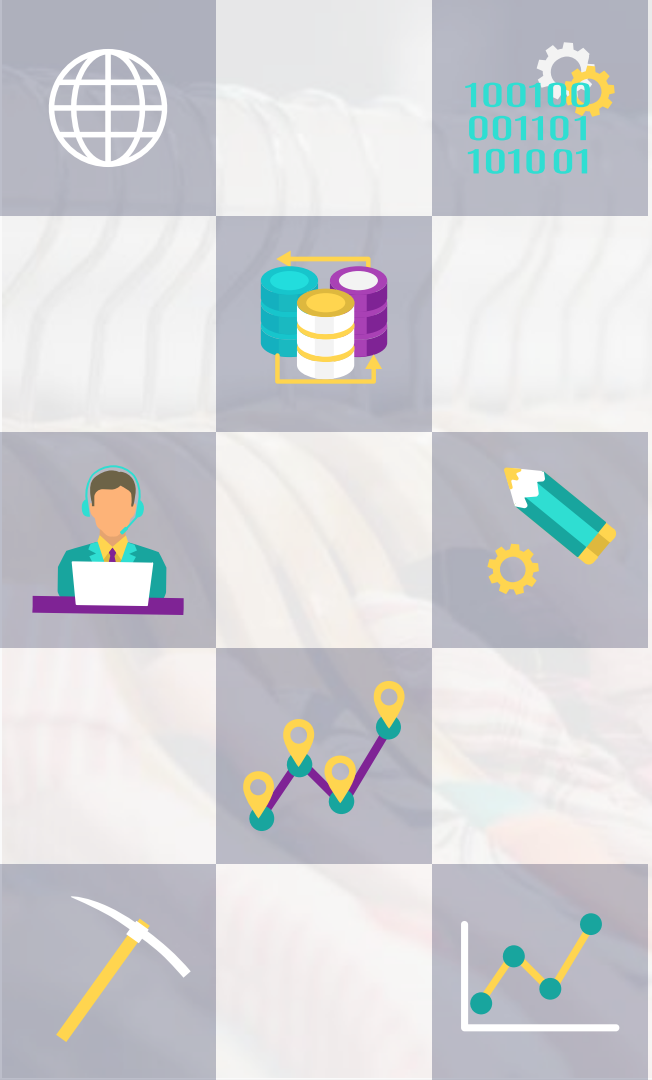
love love love jumpsuit fun flirty fabulous every time wear get nothing great compliments

Data preprocessing: Text tokenization

review_text
could zip reorder petite medium ok overall top half comfortable fit nicely bottom half...

could | zip | reorder | petite | medium | ok | overall | top

- Tokenization to divide text into smaller pieces of words



Feature Extraction

03

Feature Extraction methods used

Bag of Words:

- A straightforward and simple feature representation.
- Captures the presence and frequency of words.
- Treats all words with equal weight.
- **Applications** - Text classification tasks, sentiment analysis, or topic modeling.

Source : https://dudeperf3ct.github.io/lstm/gru/nlp/2019/01/28/Force-of-LSTM-and-GRU/

Disadvantages:

- Loss of word order and context
- Consider the two sentences: **"I love playing football and hate cricket"** and it's vice-versa **"I love playing cricket and hate football"**. Bag of Words approach will result in similar vectorized representations although both sentences carry different meanings.
- Ignores relative importance of the words across documents.

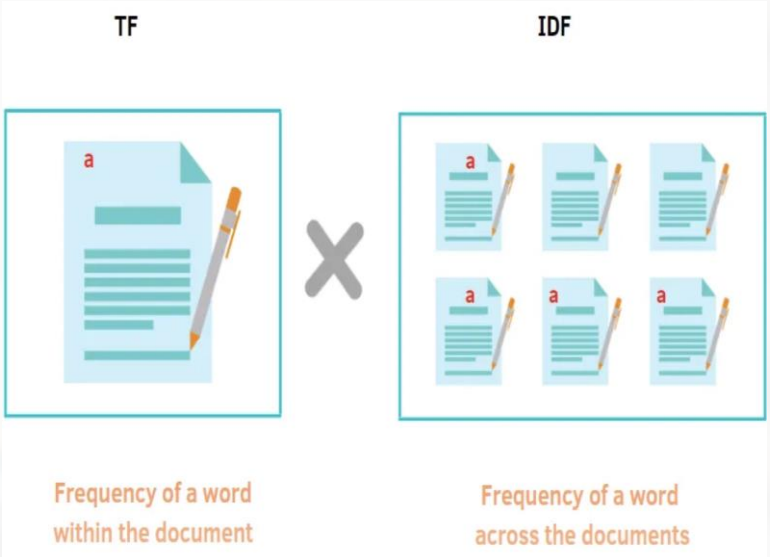
	the	red	dog	cat	eats	food
1. the red dog →	1	1	1	0	0	0
2. cat eats dog →	0	0	1	1	1	0
3. dog eats food →	0	0	1	0	1	1
4. red cat eats →	0	1	0	1	1	0

Source : https://www.kaggle.com/code/samuelcortinhas/nlp3-bag-of-words-and-similarity#2.-Bag-of-Words

Feature Extraction methods used

TF-IDF

- Takes into account not only the frequency of a word in a document but also its importance in the entire corpus.
- TF-IDF is calculated by multiplying two components:
  - **Term Frequency (TF):** Measures the frequency of a word within a document.
  - **Inverse Document Frequency (IDF):** Measures the rarity of a word across the entire corpus.
- The TF-IDF score for a word in a document is calculated as **TF * IDF.**
- Assigns higher weights to words which are common in individual documents and rarely occurring in the entire set of documents.
- Ex: **Assigns lower weights** - Common words like ‘the’, ‘them’,’and’,etc.
  Assigns higher weights - Unique words like healthcare, sports,etc.
- **Applications** - Information retrieval, document clustering, text summarization etc.,



Disadvantages

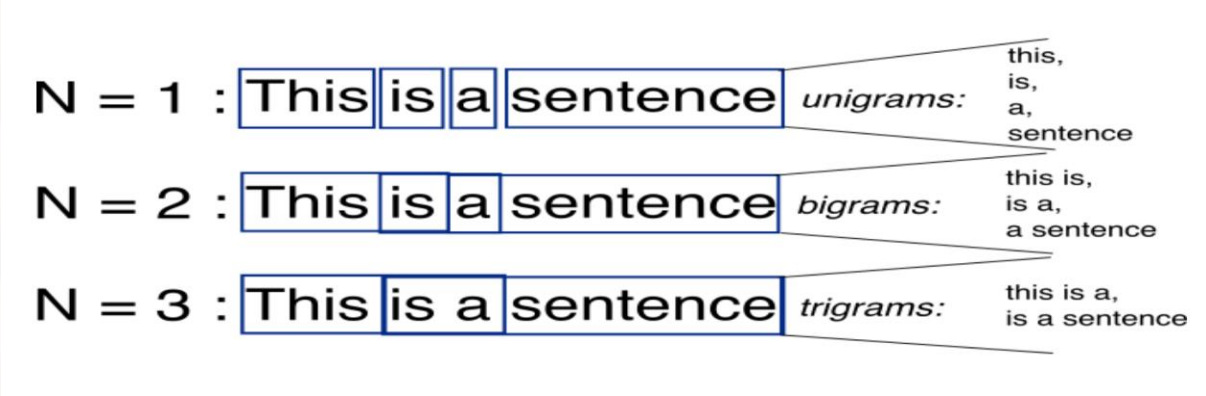
- Lack of semantic understanding
- Neglect of word order and context
- Difficulty handling synonymous terms

Source : https://betterprogramming.pub/a-friendly-guide-to-nlp-tf-idf-with-python-example-5fcb26286a33

Feature Extraction methods used

- **Parameters used for Bag of words:**

```
from sklearn.feature_extraction.text import CountVectorizer

vectorizer = CountVectorizer(ngram_range=(1, 2), max_features = 10000, stop_words='english')

bow = vectorizer.fit_transform(df_modeldata['review_text'])
```



Source : https://slidetodoc.com/lecture-2-ngram-kaiwei-chang-cs-university-of/

Feature Extraction methods used

- **Parameters used for TF-IDF**

```
from sklearn.feature_extraction.text import TfidfVectorizer

tfidf=TfidfVectorizer(ngram_range=(1, 2),max_features = 10000,stop_words='english')

tfidf_matrix=tfidf.fit_transform(df_modeldata['review_text'])
```

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
19657	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
19658	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
19659	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
19660	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
19661	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

19662 rows × 10000 columns

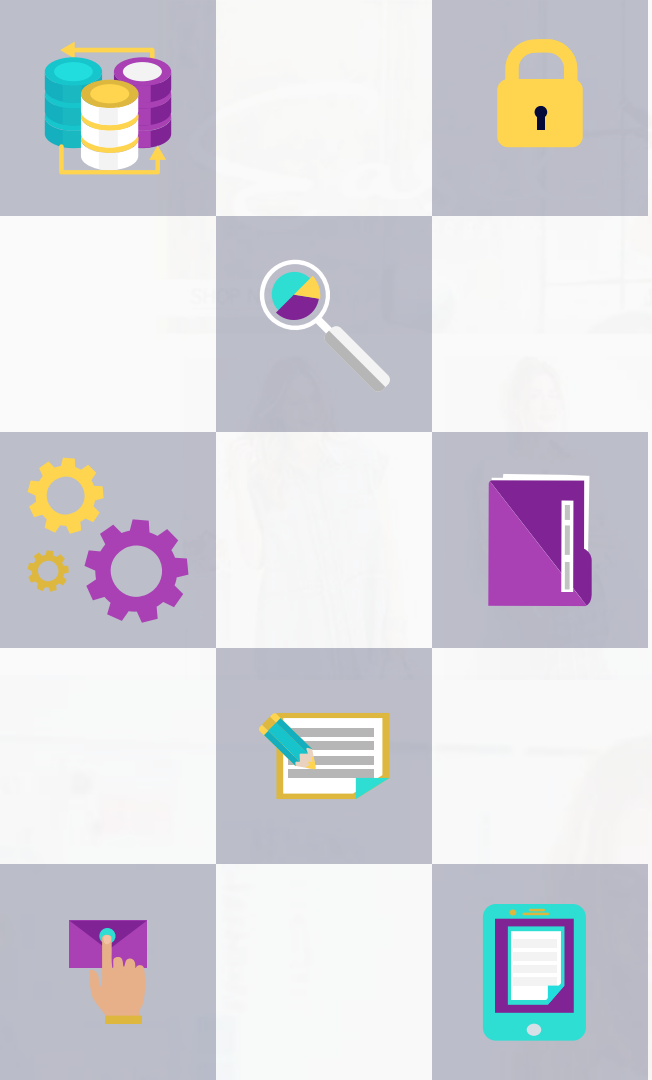
Model Building

Problem: Analyze the customer review text to predict if the customer would recommend a product or not.

Model Building using:

- Logistic Regression
- Decision Tree
- Random Forest

1. Splitting of Training and Testing data - **70% and 30%**
2.
3. Comparing the performance of the models using **F1-score**
4. Hyperparameter Optimization

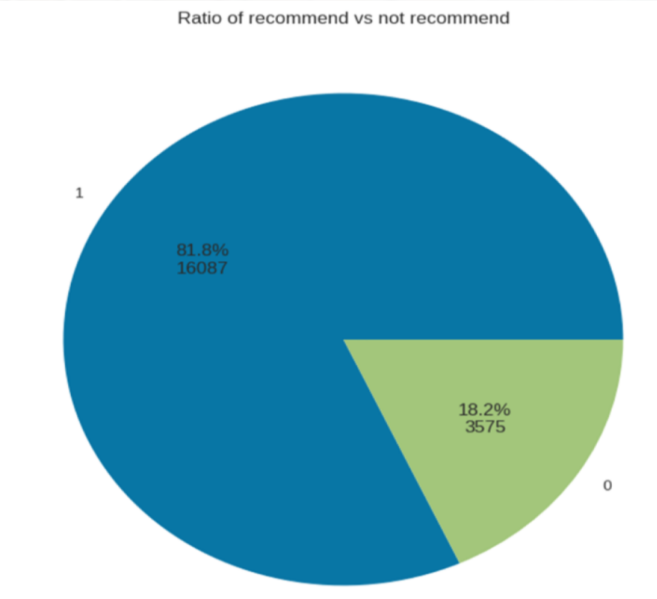


Model Building 1

04

Decision Tree, Random Forest,
Logistics Regression

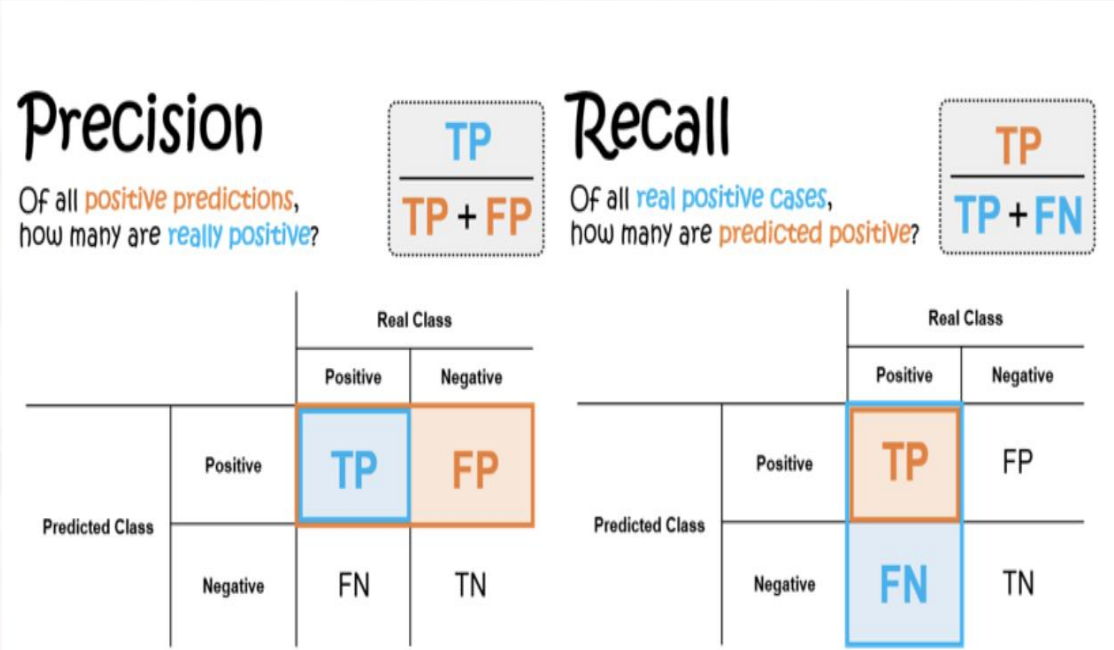
F-beta scores of the models



Accuracy metric is not compared since the target variable is imbalanced.

Metrics like Precision and Recall are compared to understand the impact of both FP and FN.

Source : https://towardsdatascience.com/precision-and-recall-made-simple-afb5e098970f



Calculating F1 score

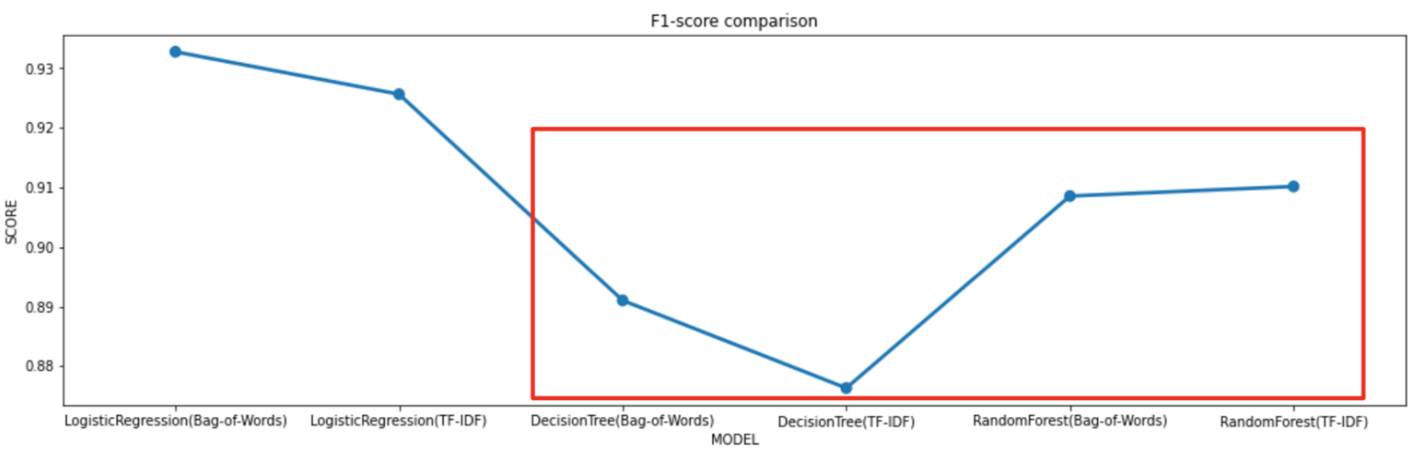
$$F_{beta} = \frac{(1+\beta^2)precision * recall}{\beta^2 * precision + recall}$$

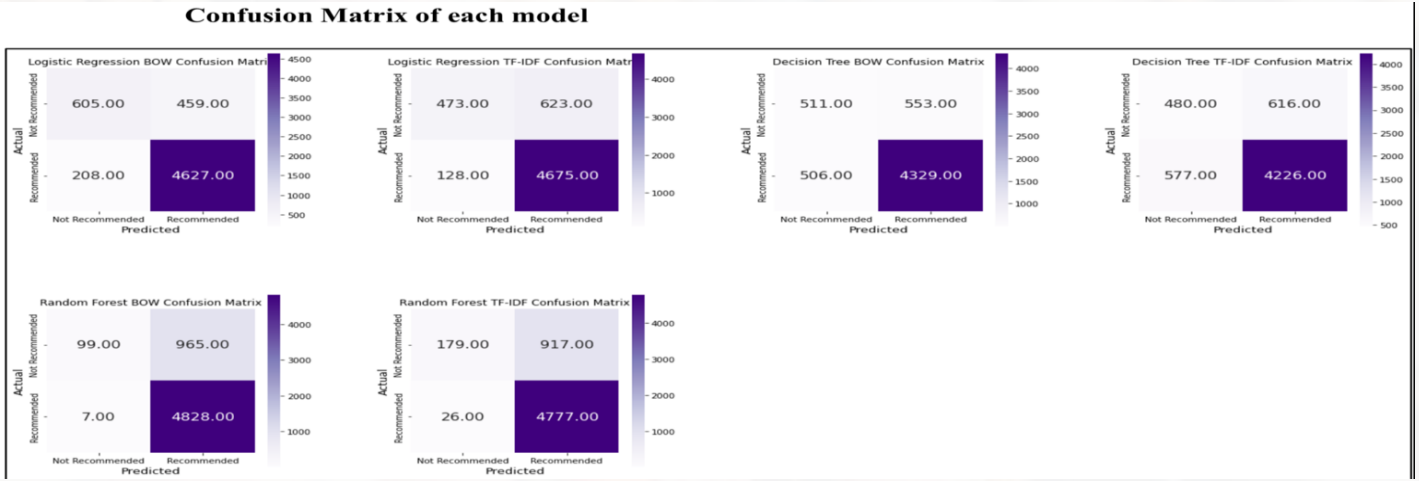
Beta = 1

$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

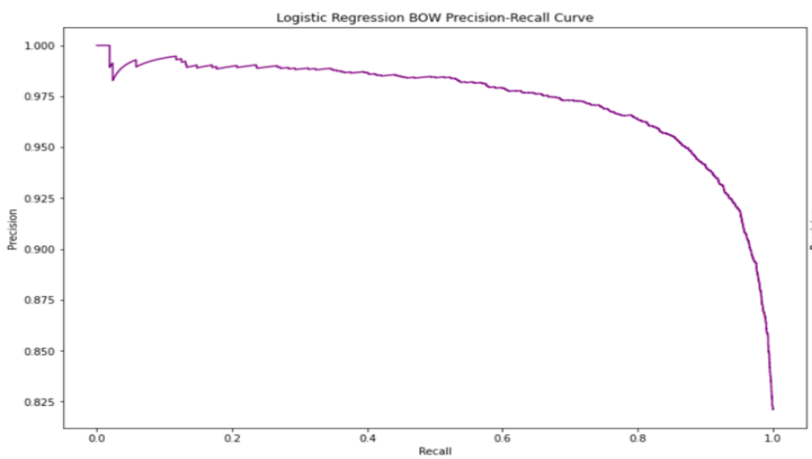
- Beta = 1, FN and FP have equal impact
- Beta > 1 (1 to 10), to place more weight on recall, where False Negatives are more crucial than FP.
- Beta < 1 (0.5), to place more weight on precision, where False positives are more crucial than FN.

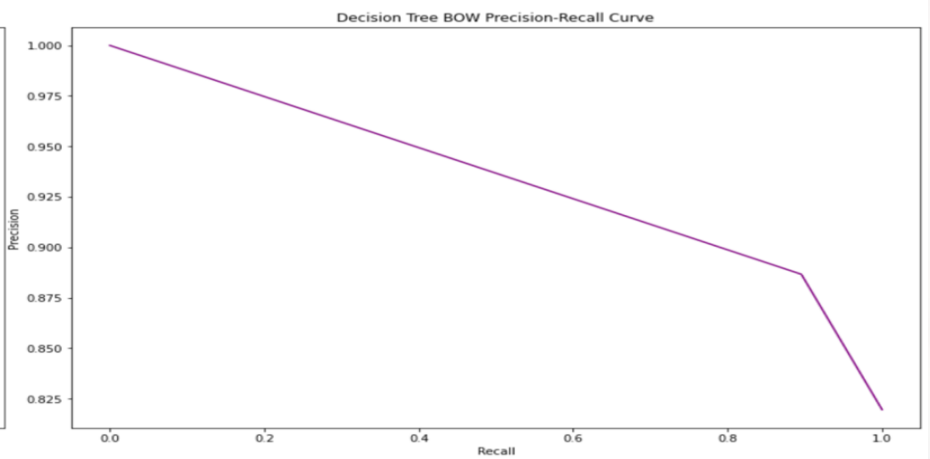
F1-score comparison

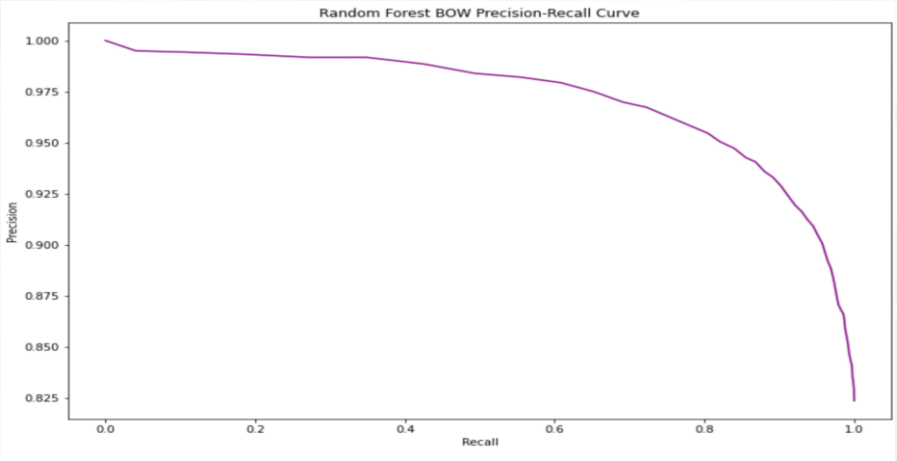




Precision Recall curves







Hyperparameter Optimization for Random Forest using RandomizedSearchCV

- Understanding different parameters of the Random Forest Classifier

```
rf =  RandomForestClassifier()
rf.get_params().keys()
```

```
dict_keys(['bootstrap', 'ccp_alpha', 'class_weight', 'criterion', 'max_depth', 'max_features', 'max_leaf_nodes', 'max
_samples', 'min_impurity_decrease', 'min_impurity_split', 'min_samples_leaf', 'min_samples_split', 'min_weight_fracti
on_leaf', 'n_estimators', 'n_jobs', 'oob_score', 'random_state', 'verbose', 'warm_start'])
```

- Defining parameter values

```
param_grid = {
    'n_estimators': [25, 50, 100, 150, 200, 400, 600, 1000, 1200, 1400],
    'max_features': ['sqrt', 'auto', 'log2', None],
    'max_depth': [3, 6, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, None],
    'max_leaf_nodes': [3, 6, 9],
    'min_samples_leaf': [1, 2, 4],
    'min_samples_split': [2, 5, 10]
}
```

Hyperparameter Optimization for Random Forest using RandomizedSearchCV

- Defining RandomizedSearchCV

```
from sklearn.model_selection import RandomizedSearchCV

random_search_rf = RandomizedSearchCV(estimator = rf, param_distributions = param_grid, n_iter = 50,
                                      cv = 3, verbose=2, random_state=42, n_jobs = -1)
```

- Finding optimal set of values for each of the hyper parameters.

```
from datetime import datetime

start_time = timer(None) # timing starts from this point for "start_time" variable
random_search_rf.fit(x_train_bow,y_train_bow)
timer(start_time) # timing ends here for "start_time" variable
```

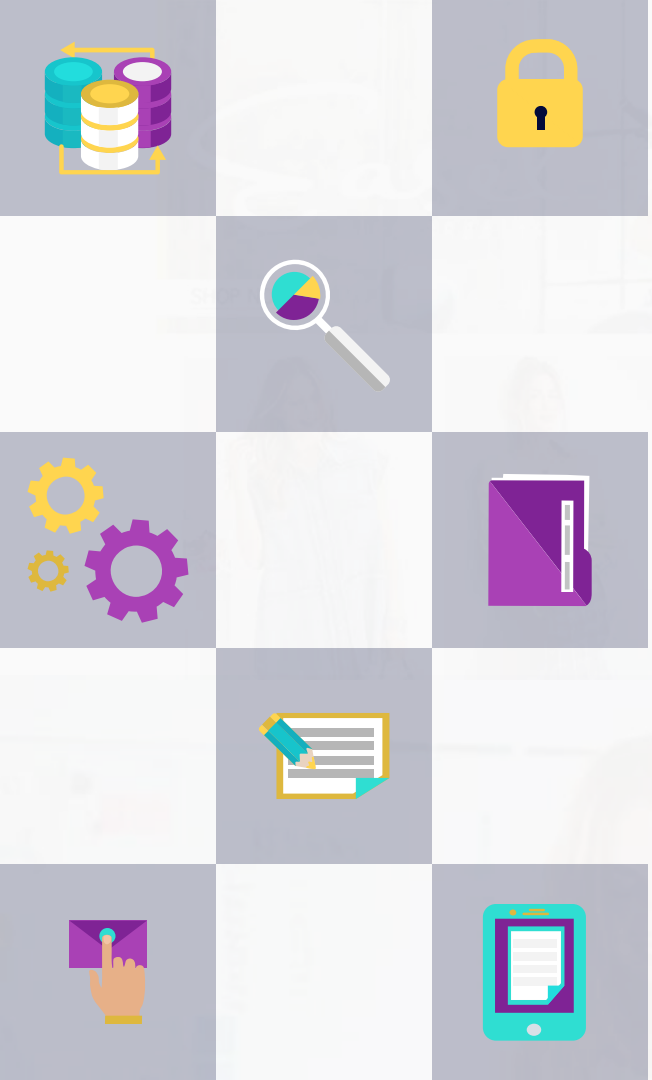
```
Fitting 3 folds for each of 50 candidates, totalling 150 fits

 Time taken: 0 hours 5 minutes and 18.27 seconds.
```

- Best set of hyperparameter values can be found by using best_estimator() function.

```
random_search_rf.best_estimator_
```

```
RandomForestClassifier(max_depth=20, max_features=None, max_leaf_nodes=9,
                       min_samples_leaf=4, min_samples_split=10,
                       n_estimators=25)
```

```
rf_hyper_tuning =  RandomForestClassifier(max_depth=20, max_features=None, max_leaf_nodes=9,
                       min_samples_leaf=4, min_samples_split=10,
                       n_estimators=25)
```

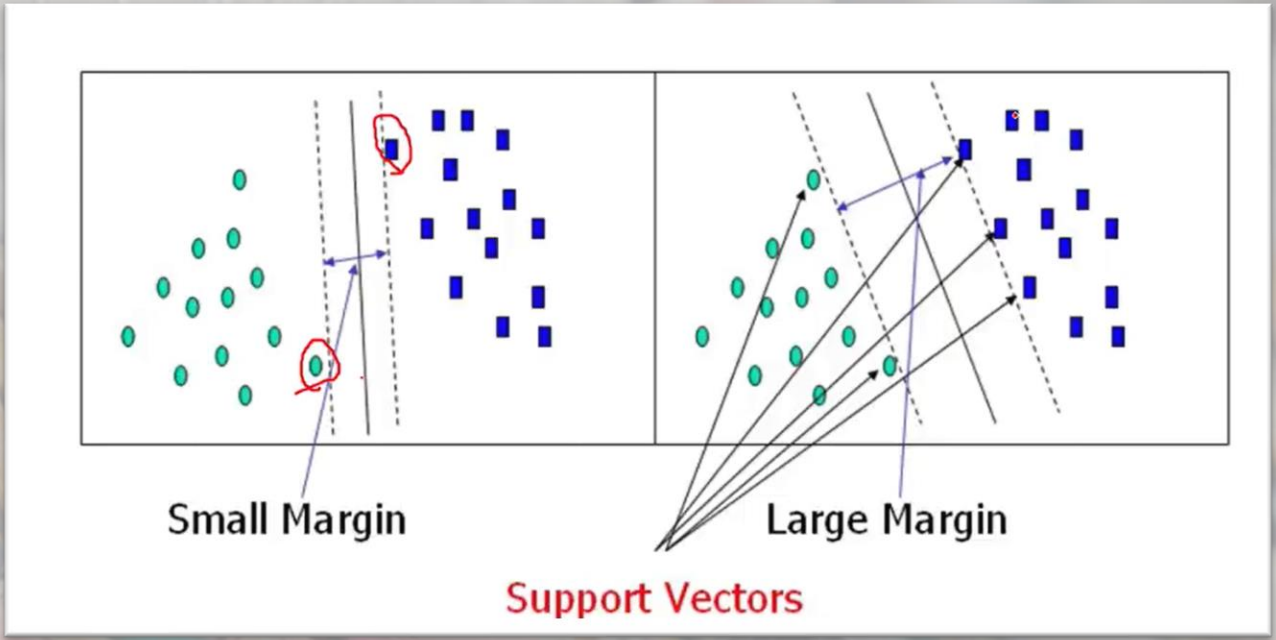


Model Building 2

05

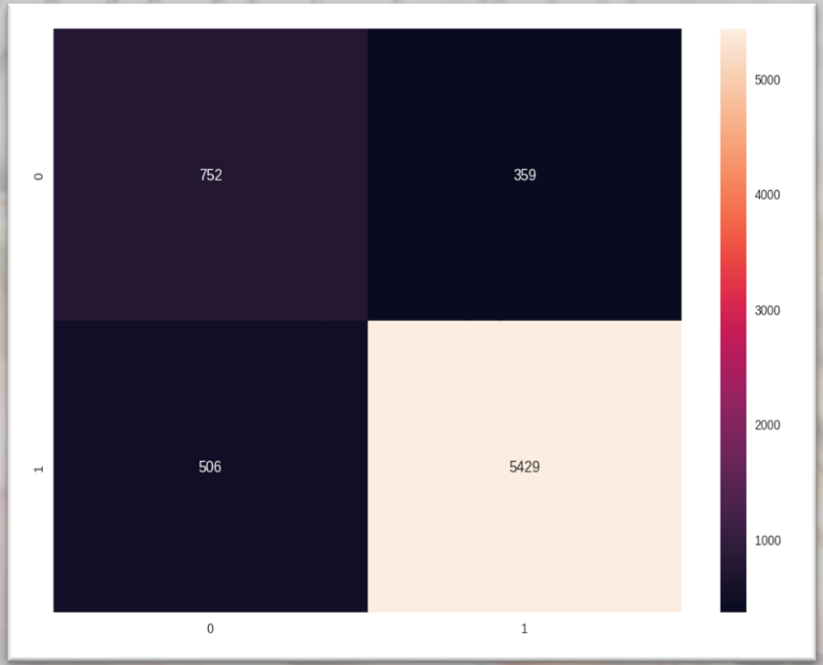
SVM, Naïve Bayes Classifier

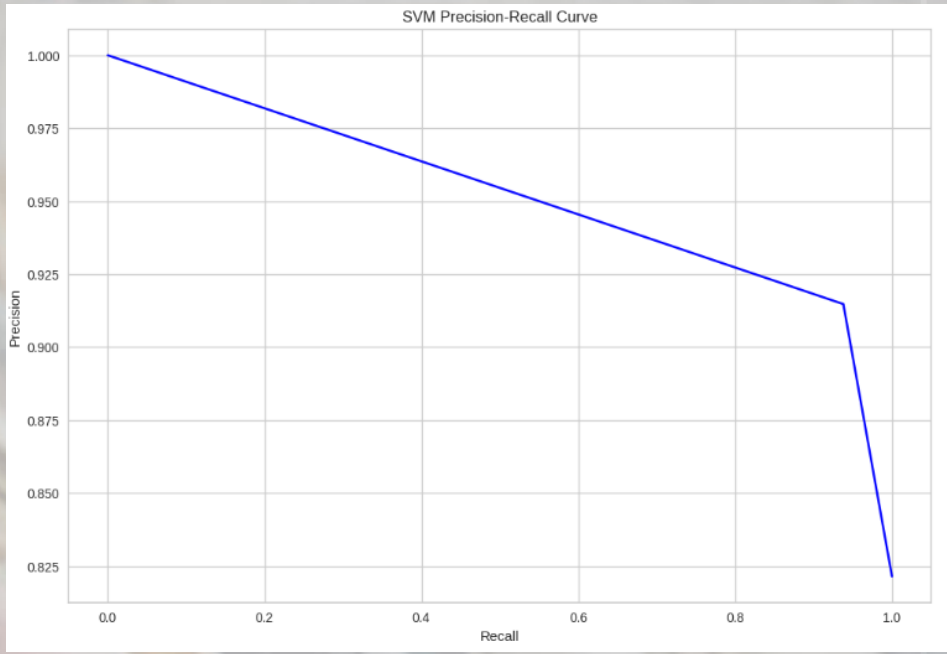
Support Vector Machine



- SVM aims to find the best decision boundary that separates the data points into different classes with the maximum margin.
- The margin is the distance between the hyperplane and the closest data points of each class.

SVM Model Performance





F1 score: 0.926

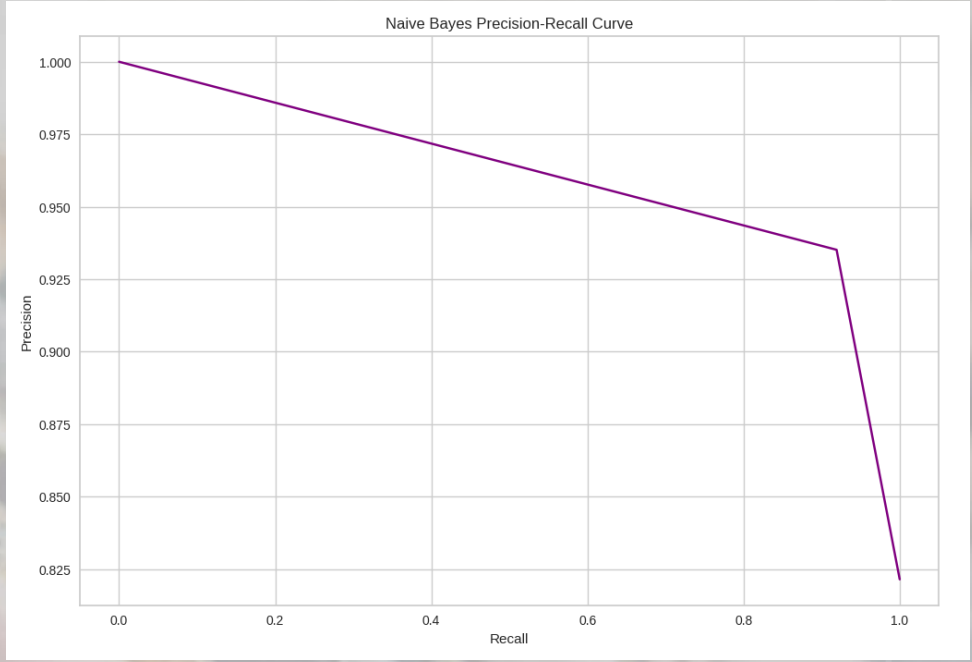
Naive Bayes Classifier

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

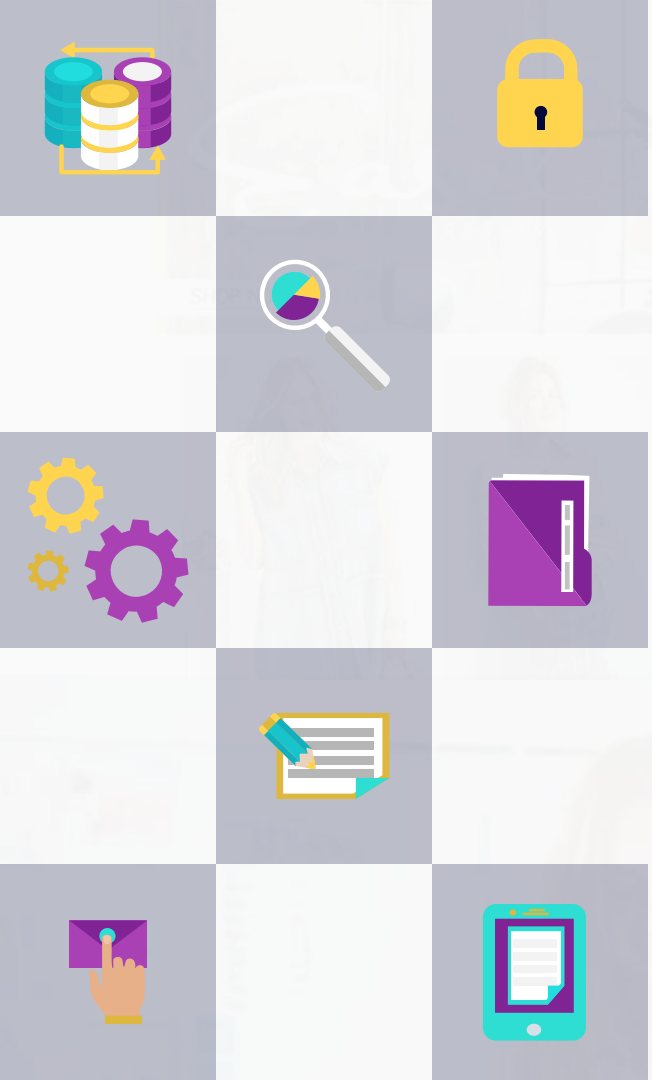
- A: class
- B: features

- The Naïve Bayes classifier is built upon Bayes' theorem, which calculates the conditional probability of a class given the observed features -P(class|features).
- Select the class with the highest probability as the predicted class label.

Gaussian Naïve Bayes Model Performance



F1 score: 0.927

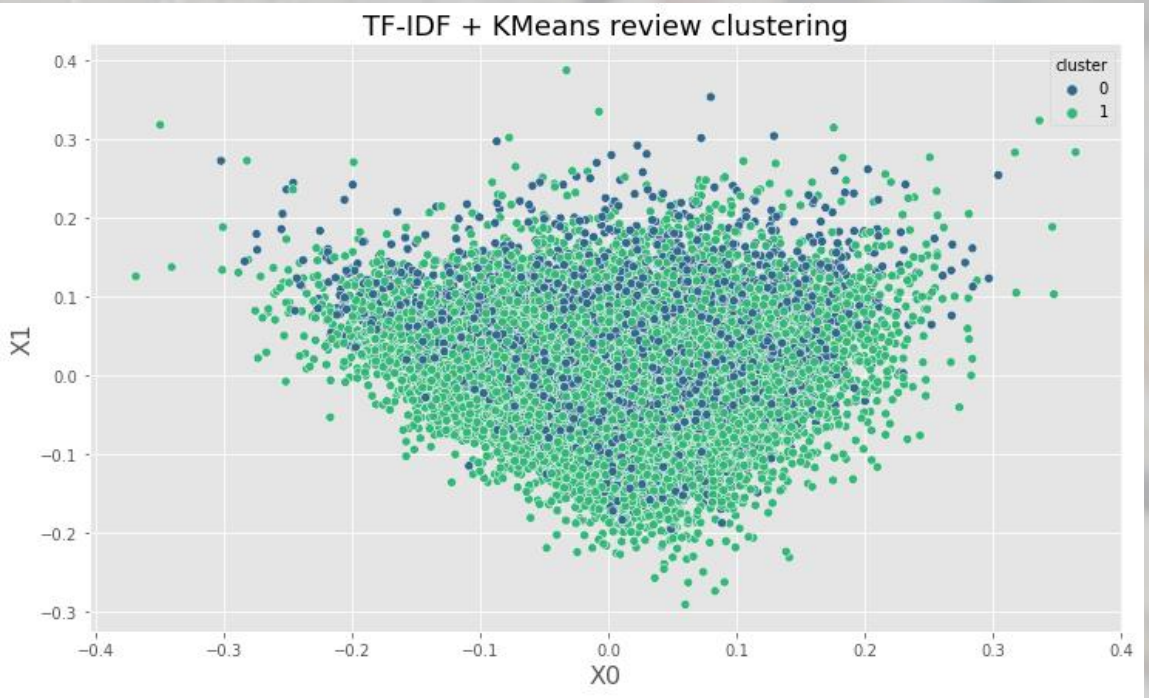


Model Building 3

06

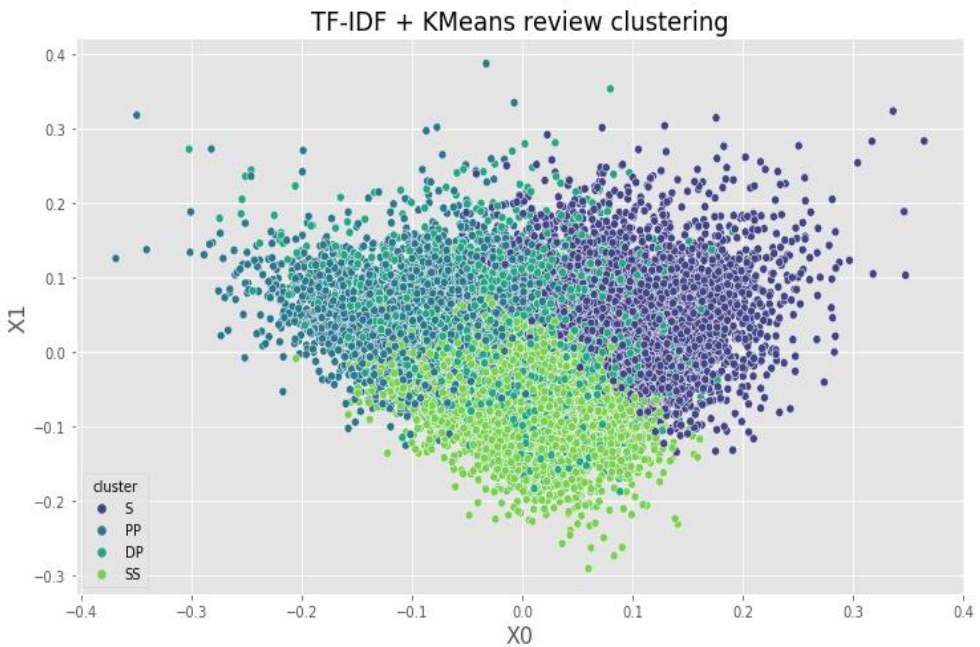
K-means Clustering, Topic Modeling

Process of K-means Clustering



- In the process, the number of clusters was determined in the order of 2, 3, 10, and 4 manually.
- There were many overlapping areas, or there were many cases where analysis was impossible.

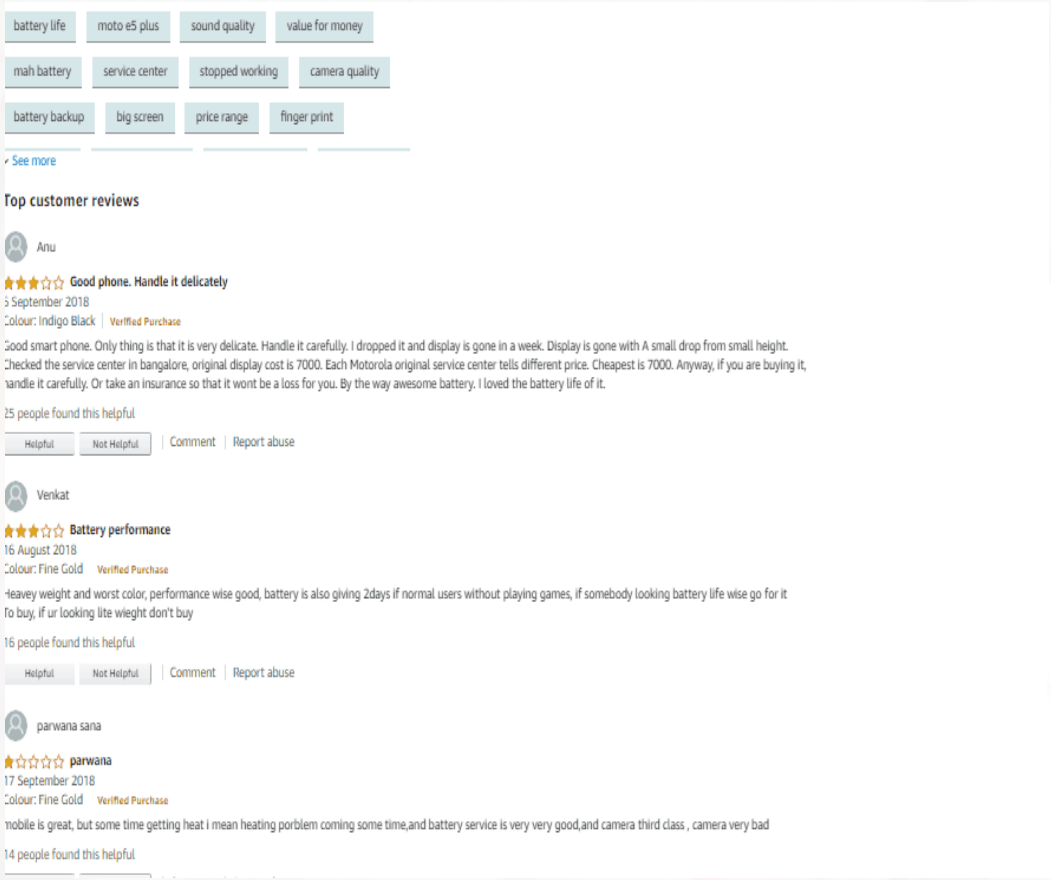
Process of K-means Clustering



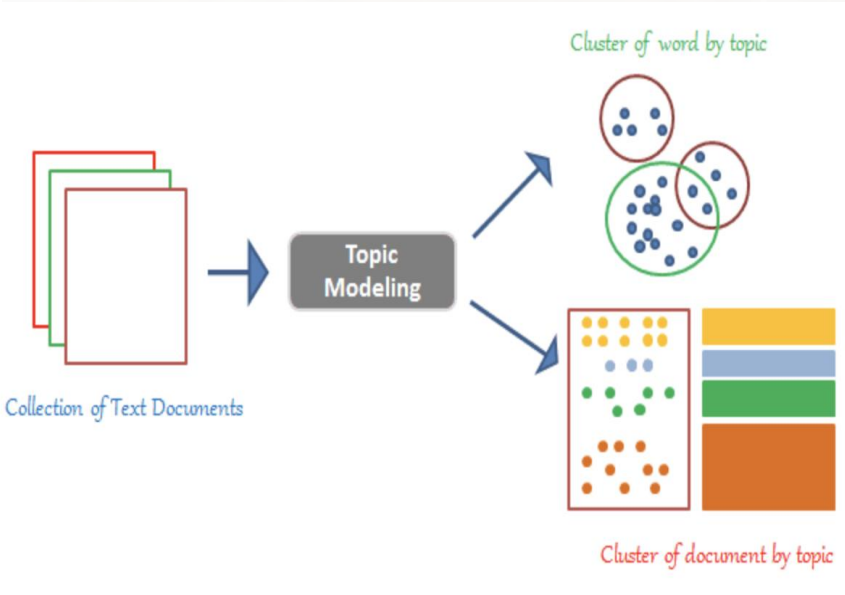
- Finally identified four distinct clusters that not perfectly but effectively separated the data.
- Each cluster was examined for 10 keywords, checking significant overlap between them.
- The number next to each text means the total number of times it appeared in the Top 10
- Named by leaving only the text that appeared once or twice.

Trimmed Top Keywords		
Cluster 0 (S)	medium(1), top(2), order(1), run(1), large(1), small(1), size(2)	Size
Cluster 1 (DP)	like(2), great(2), fabric(2), size(2), dress(1)	Dress & positive
Cluster 2 (SS)	shirt(1), sweater(1), fabric(2), like(2), top(2)	Shirt, Sweater
Cluster 3 (PP)	perfect(1), comfortable(1), pant(1), jean(1), great(2)	Pants & positive

Topic Modeling

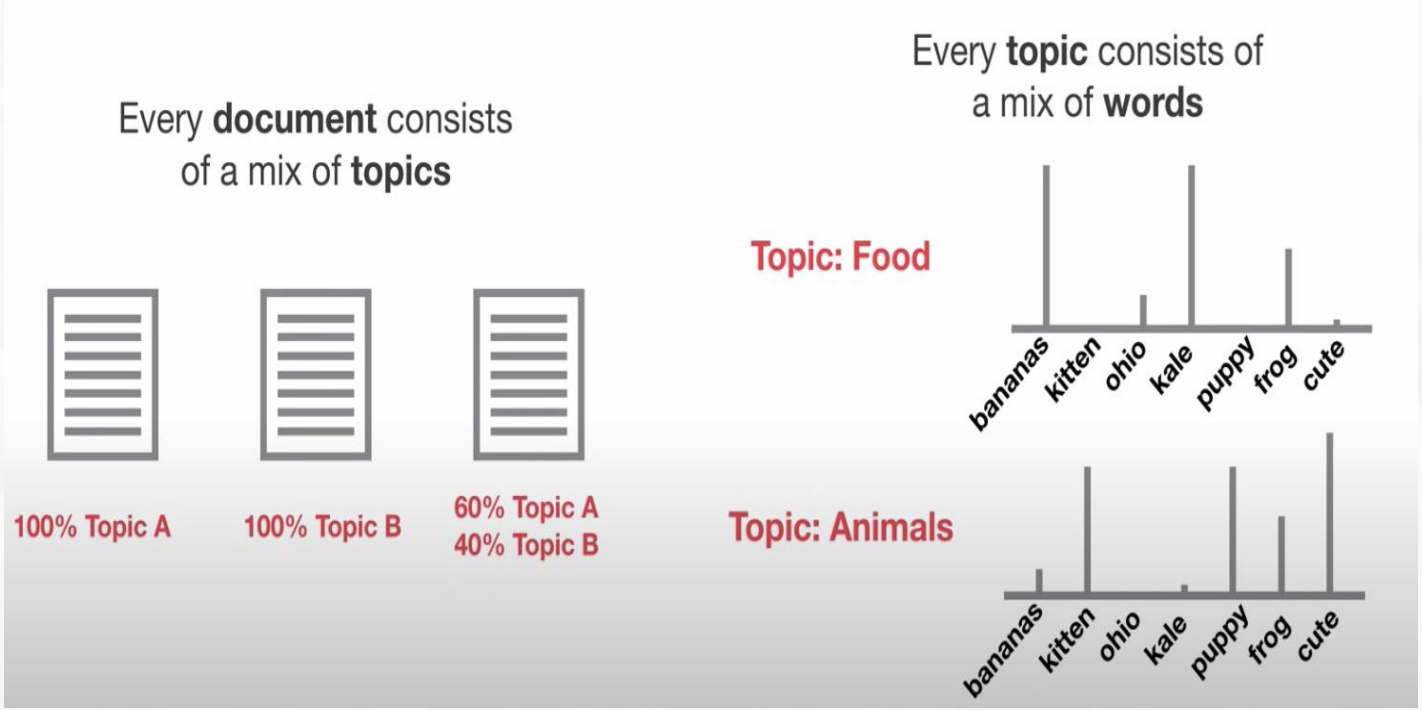


- A process to automatically identify topics present in a text object and to derive hidden patterns exhibited by a text corpus..



Topic Modeling using LDA

LDA (Latent Dirichlet Allocation), one of the most popular topic modelling algorithms is used from gensim, python library. Probabilistic approach to topic modelling. Latent - hidden, Dirichlet - type of probability distribution.



Topic Modeling using LDA

How LDA works?

- **Goal:** To find the topic mix in each document and word mix in each topic.
- **Input :** Document-term matrix, number of topics, number of iterations

Gensim LDA:

- Randomly assigns each word in the document to one of the 3 topics.
- The word and topic assignment is repeated for several times based on the below two points:
  - How often a word occurs in the topic
  - How often the topic occurs in the document.

- **Output** : Top keywords in each topic. If the topics are not interpretable, input parameters can be altered and process can be repeated.

Topic Modeling using LDA

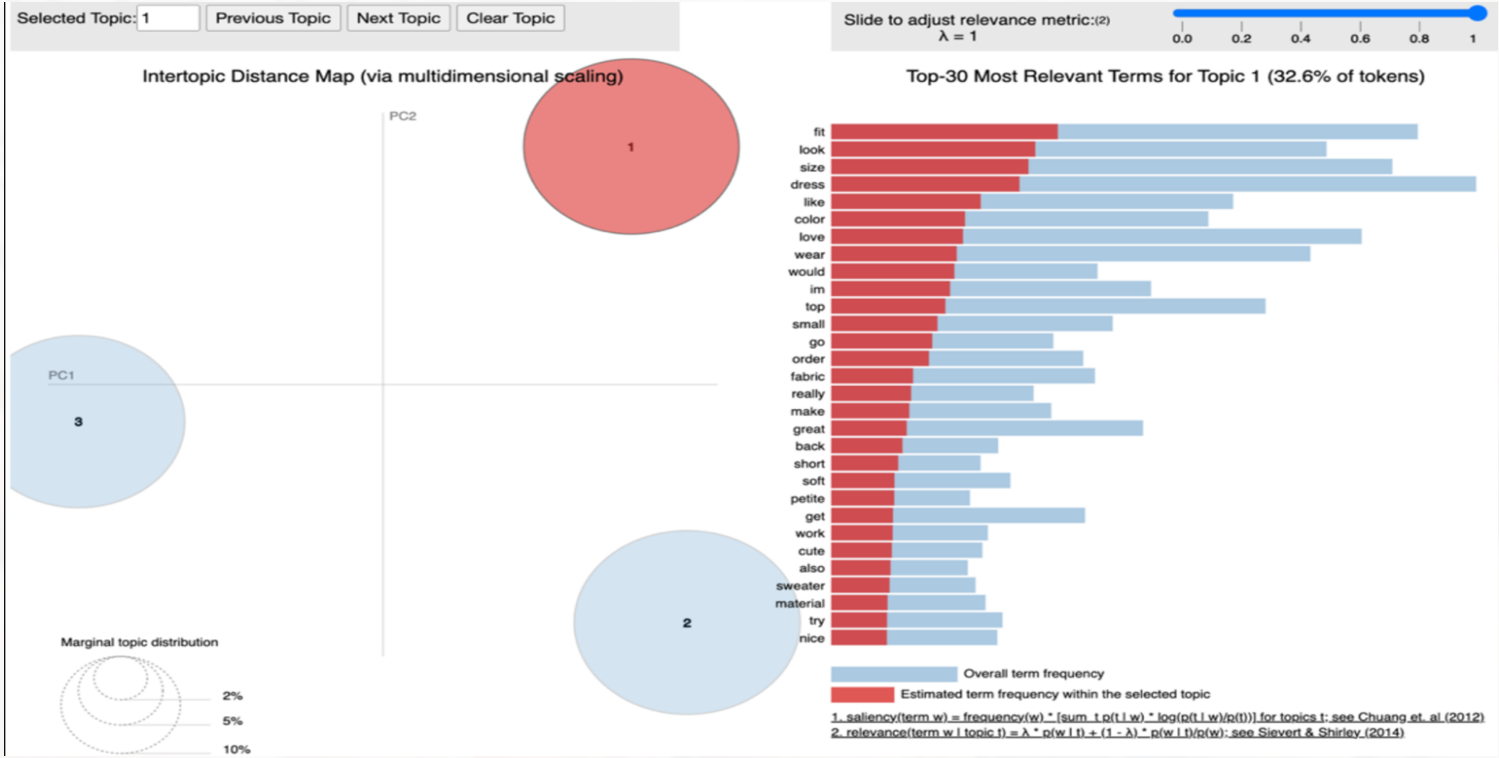
```
#num_topics = 3

%time ldamodel = lda(doc_term_matrix,num_topics=3,id2word=dictionary,passes=10)
ldamodel.print_topics()
```

```
CPU times: user 1min 6s, sys: 3.9 s, total: 1min 10s
Wall time: 1min 16s

[(0,
  '0.047*"size" + 0.028*"top" + 0.028*"dress" + 0.027*"small" + 0.025*"fit" + 0.017*"order" + 0.017*"im" + 0.017*"large" + 0.011*"medium" + 0.010*"look"'),
 (1,
  '0.023*"great" + 0.022*"color" + 0.018*"fit" + 0.018*"dress" + 0.017*"wear" + 0.015*"jean" + 0.014*"love" + 0.014*"look" + 0.013*"top" + 0.012*"comfortabl
e"'),
 (2,
  '0.027*"dress" + 0.019*"fabric" + 0.016*"fit" + 0.016*"color" + 0.016*"look" + 0.015*"skirt" + 0.014*"sweater" + 0.013*"soft" + 0.011*"material" + 0.011*"nic
e"')]
```

Topic 1	Fit and Size
Topic 2	Great-looking dress
Topic 3	Quality

Topic Modeling using LDA

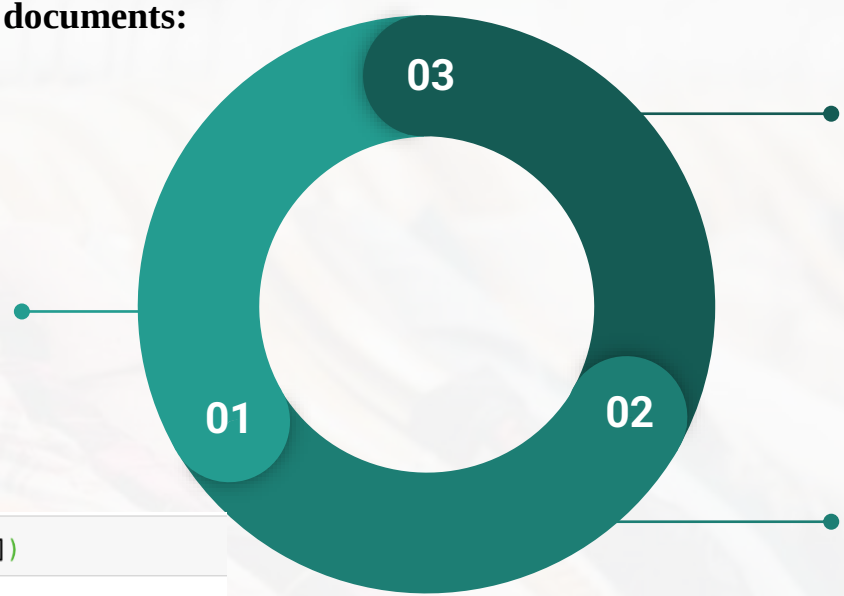
Topic clusters are visualized using **pyLDAvis - LDA Visualization library**



Topic Modeling using LDA

Topic assignment to the documents:

The **probability** that a document can be tagged to any of the **3 topics** is generated by LDA for all the documents.



For each document, **the topic with a probability higher than the threshold score** is assigned as the corresponding topic.

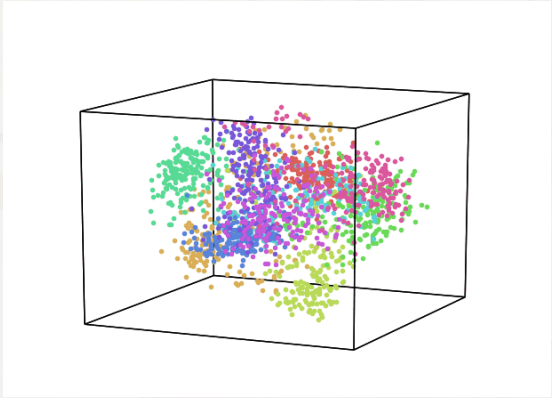
Threshold score is calculated as the average of all the topic probabilities of the reviews.

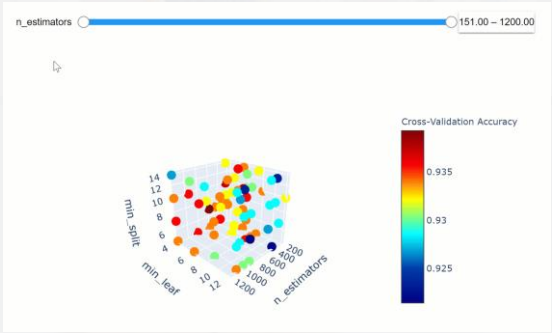
```
print(ldamodel[doc_term_matrix][0])
```

```
[(0, 0.20165072), (1, 0.2029862), (2, 0.5953631)]
```

Future scope

- Feature extraction techniques – Word2Vec, AvgWord2Vec, BERT, etc – to capture semantic and syntactic relationships between words.
- Other Hyperparameter optimization methods – Grid Search, Bayesian optimization.
- NLP techniques – Named Entity Recognition, Text summarization,etc





Conclusion

Data & EDA

Data Cleaning
Exploratory Analysis

01

Text pre-processing

Preparation for
Text Analysis

02

Feature Extraction

Bag of words
Tf-Idf

03

04

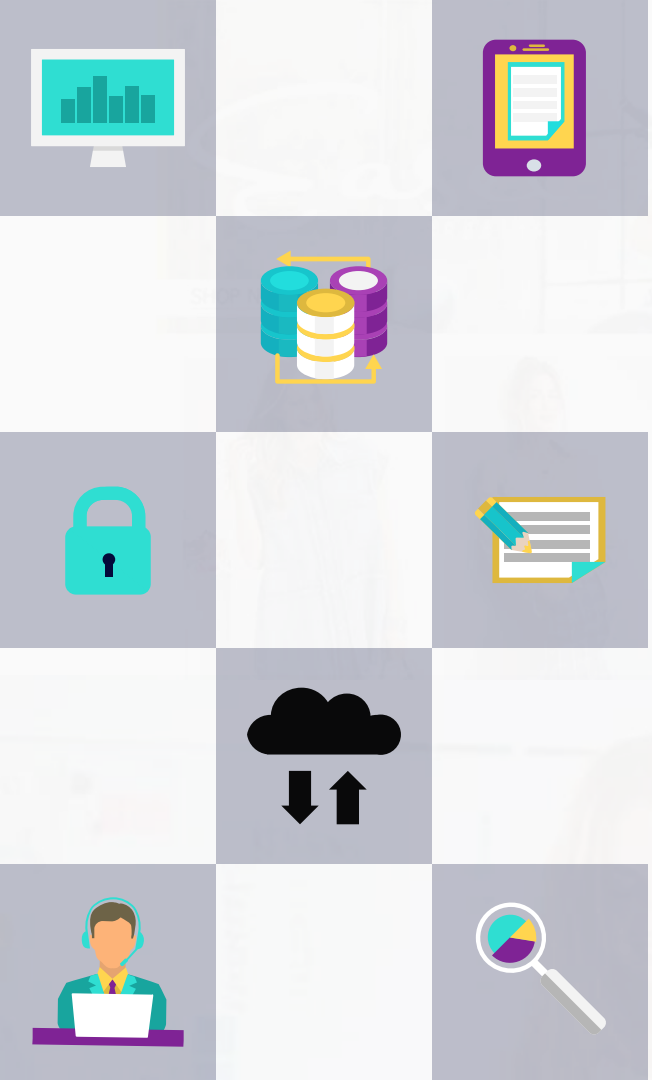
Decision Tree
Random Forest
Logistic Regression

05

SVM
Naïve Bayes Classifier

06

K-means Clustering
Topic Modeling



Thanks!

Do you have any questions?

REFERENCE

- NICAPOTATO. (2018). Women's e-commerce clothing reviews. kaggle. Retrieved from https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews?datasetId=11827&sortBy=voteCount&searchQuery=eda
- MATTHEW CONNOR. (2022). NLP: Comparative RNN & DL Models with detailed EDA. kaggle. Retrieved from https://www.kaggle.com/code/azizozmen/nlp-comparative-rnn-dl-models-with-detailed-eda
- Shankar297. (May 31, 2022). A Complete guide on feature extraction techniques. Analytics Vidhya. Retrieved from https://www.analyticsvidhya.com/blog/2022/05/a-complete-guide-on-feature-extraction-techniques/
- Purva Huilgol. (February 28, 2020). Quick introduction to bag-of-words (BoW) and TF-IDF for creating features from text. Analytics Vidhya. Retrieved from https://www.analyticsvidhya.com/blog/2020/02/quick-introduction-bag-of-words-bow-tf-idf/#:~:text=Bag%20of%20Words%20just%20creates,vectors%20are%20easy%20to%20interpret.
- Doug, Steen. (September 19, 2020). Precision-recall curves. Medium. Retrieved from https://medium.com/@douglaspsteen/precision-recall-curves-d32e5b290248
- Andrea, D'Agostino. (November 24, 2021). Text clustering with TF-IDF in python. Medium. Retrieved from https://medium.com/mlearning-ai/text-clustering-with-tf-idf-in-python-c94cd26a31e7
- Shivam5992. (August 24, 2016). Beginners guide to topic modeling in python. Analytics Vidhya. Retrieved from https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/

REFERENCE

- Aravind CR. (July 26, 2020). Topic modeling using gensim-LDA in python. Medium. Retrieved from https://medium.com/analytics-vidhya/topic-modeling-using-gensim-lda-in-python-48eaa2344920

- Gunjit, Bedi. (2018, November 9). A guide to Text Classification(NLP) using SVM and Naive Bayes with Python. Medium. Retrieved from https://medium.com/@bedigunjit/simple-guide-to-text-classification-nlp-using-svm-and-naive-bayes-with-python-421db3a72d34

- sepandhaghighi. (2018, June 4). How to get accuracy, confusion matrix of binary SVM classifier equivalent to multiclass classification?. Cross Validated. Retrieved from

- A Dash of Data. (2020). Natural language processing (Part 5): topic modeling with latent dirichlet allocation in python. YouTube. Retrieved from https://www.youtube.com/watch?v=NYkbqzTlW3w

- Alice, Zhao. (2018). Topic modeling. github. Retrieved from https://github.com/adashofdata/nlp-in-python-tutorial/blob/master/4-Topic-Modeling.ipynb

- slidesgo. (2023). Data mining project proposal. Retrieved from https://slidesgo.com/recent?premium=1&utm_source=mandrillap&utm_medium=email&utm_campaign=welcome%20premium&utm_term=explore