E-Commerce

AN ANALYSIS BASED ON WOMEN CLOTHING REVIEW DATA

PROJECT NOTES - I

(i) Problem statement, need of the project and data report
(ii) EDA
(iii) Insights from EDA

Problem statement

- This is a Women's Clothing E-Commerce dataset revolving around the reviews written by customers.

Based on recommendations provided by customers, store is looking forward to get the various products recommendations.

Need of this project

- The online marketing space is in constant shift as new technologies, services, and marketing tactics gain popularity and become the new standard. Online store owners are one of the many different segments affected by these constant evolutions. In order for these business owners to survive and thrive, they need to be able to make better decisions faster. This is where need of this project or analysis comes into picture.

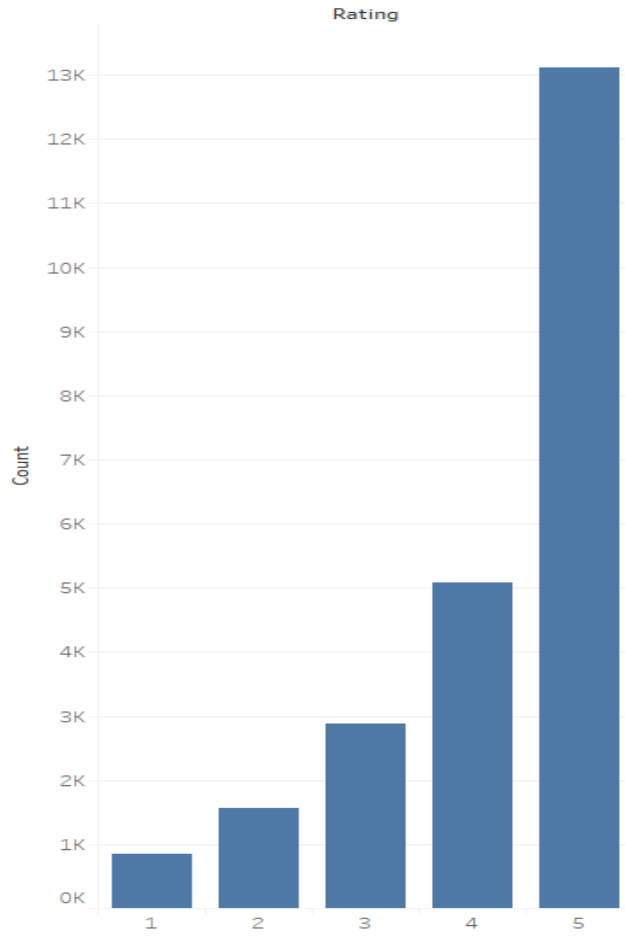
Data Report

This dataset includes 23486 rows and 10 feature variables. Each row corresponds to a customer review, and includes the variables:

Column Name	Data Description
Clothing ID	Integer Categorical variable that refers to the specific piece being reviewed.
Age	Positive Integer variable of the reviewers 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

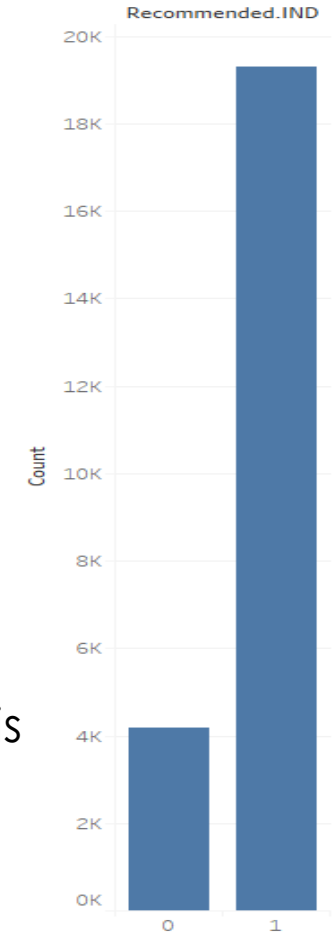
Univariate Analysis

Distribution of Rating



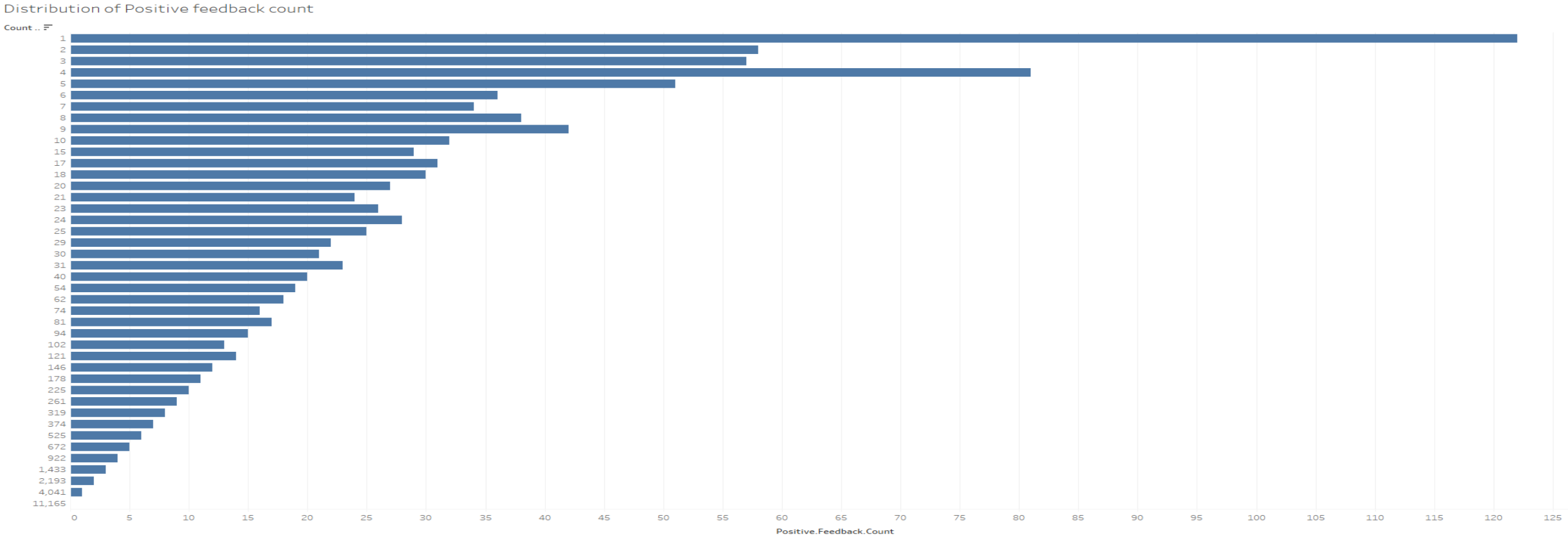
5 star rating has topped the chart with most number of count.

Distribution of Recommendation



Recommendation of products is much more higher. 19300 items have been recommended out of 23472 items.

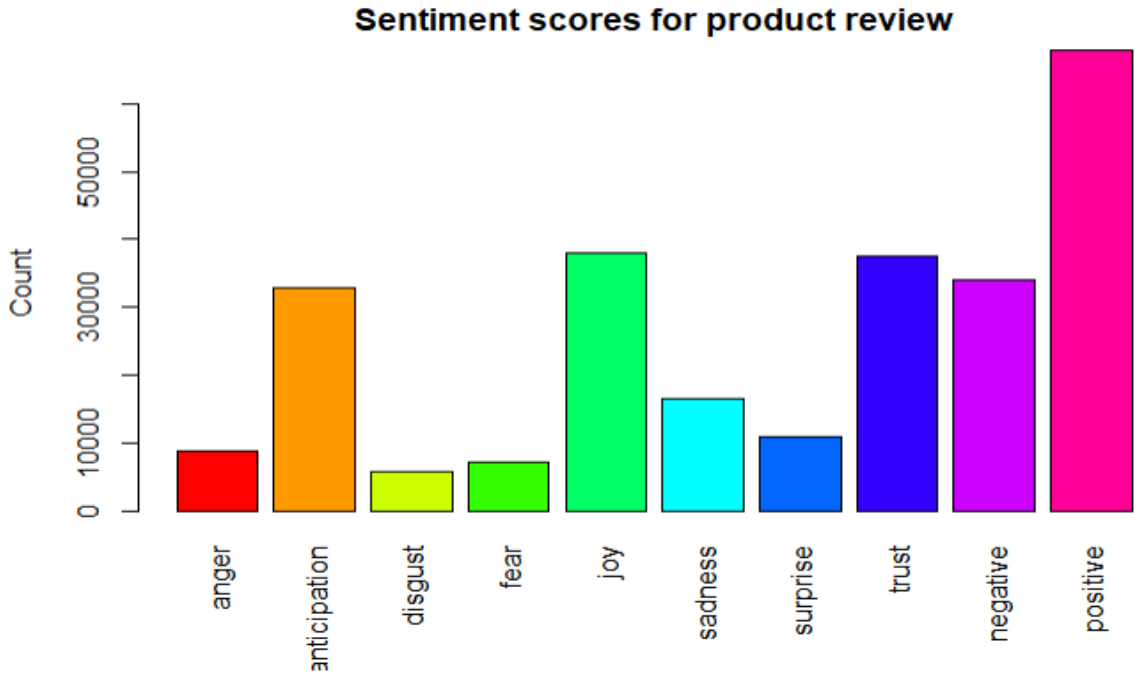
Univariate Analysis



Text and Sentiment analysis of "Review Text"

After text analysis, Wordcloud has been created and we can see that the words like 'great', 'fit', 'like', 'perfect' and 'comfortable' have been used more in reviews.

Sentiment analysis of Review comments shows that customers' sentiments are more towards Positive, Joy, Trust and anticipation. Negative and Sadness sentiments can also be observed.

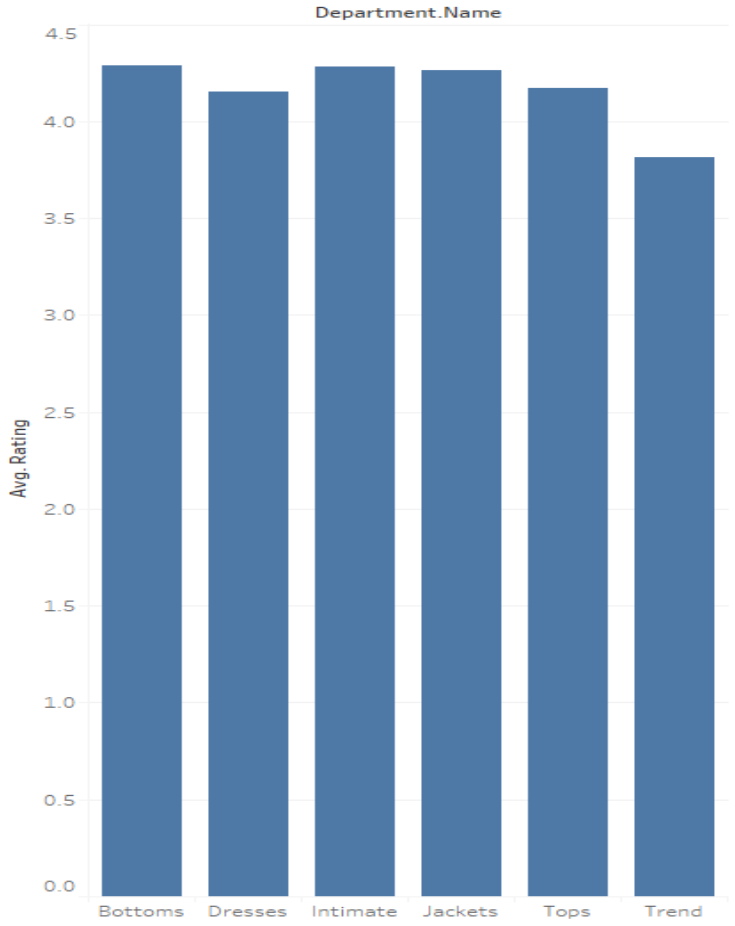


Text Analysis of “Title”

Text analysis of “Title” variable shows that love, beautiful, cute and perfect words have been used more than other words

Bivariate Analysis

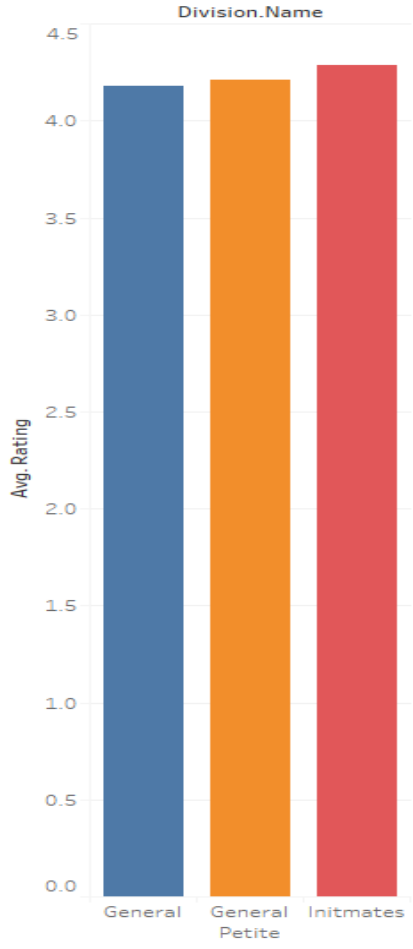
Dept vs rating



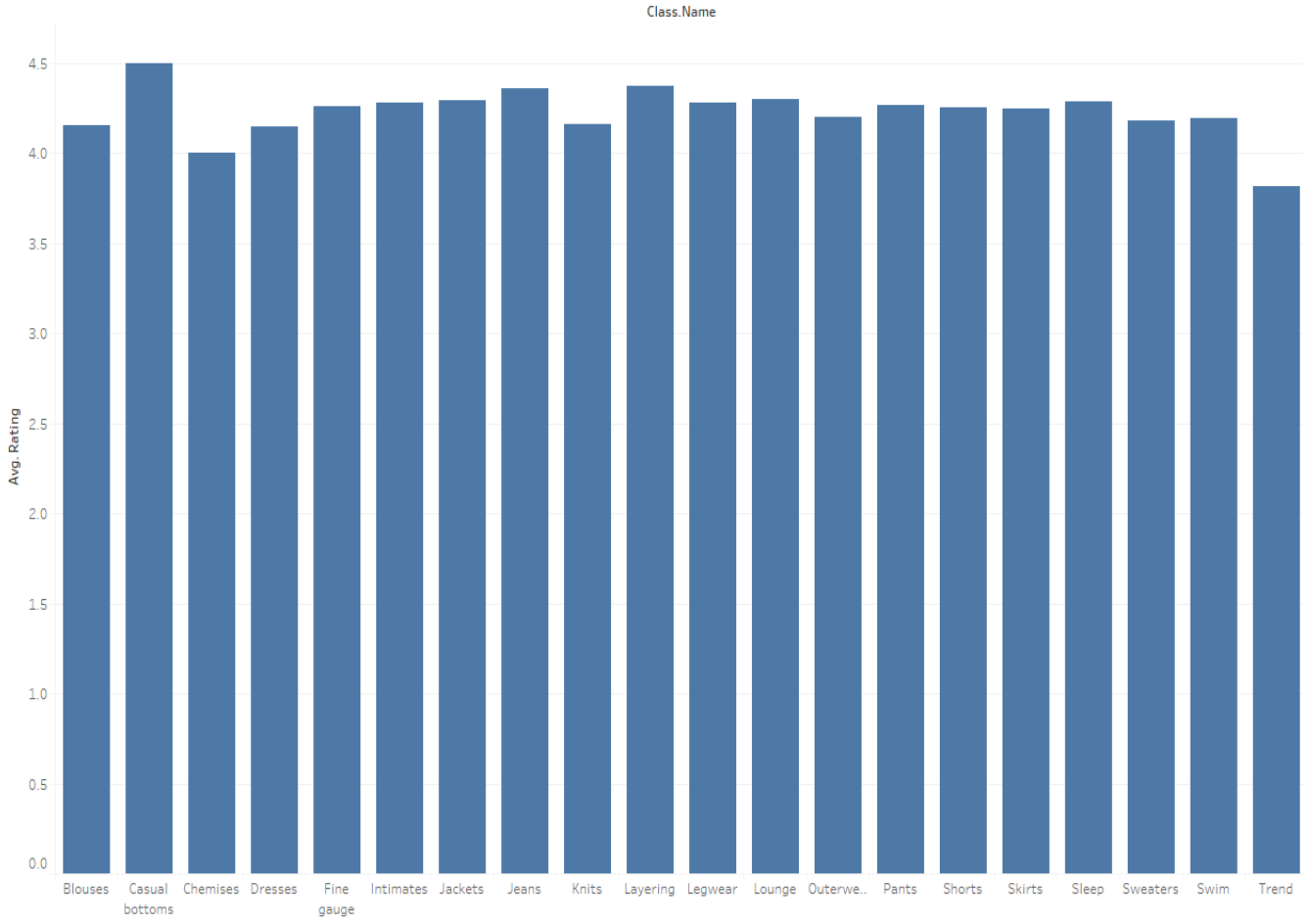
In terms of ratings, although all the departments have performed well but highest number of ratings have gone to department of bottoms and then Intimate, Jackets and Tops respectively.

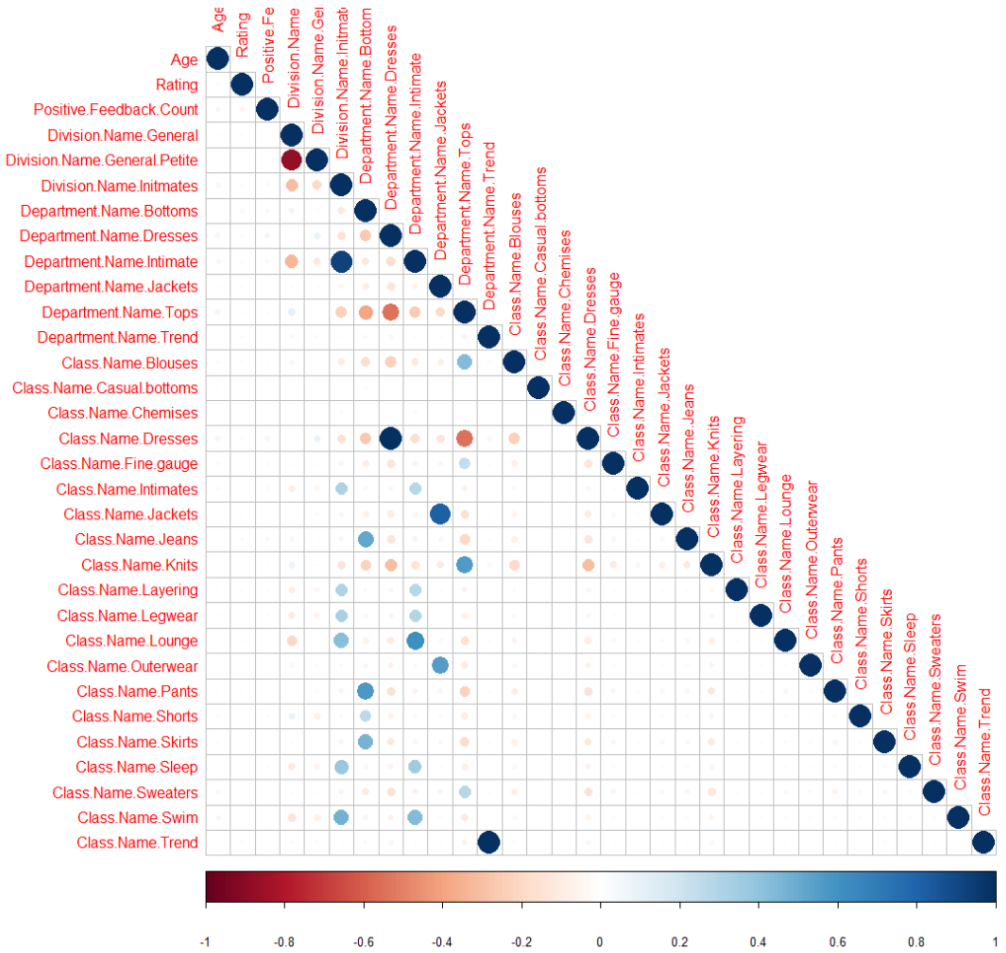
If we talk about Divisions, then division of Intimates has topped the chart.

Division vs rating



Bivariate Analysis





In terms of individual products, Casual bottoms and Layering are the most rated products.

Removal of unwanted variables/ Missing value treatment/ Addition of new variables

- Two unnecessary variables have been removed that were Id and Clothing Id. These variables do not add any value in analysis.
- There were blank values in the dataset for Division, Department and Class. Since these values were 0.05% of total data, Hence, the values have been removed from the dataset. Also, without knowing the item name, there was no benefit for doing the analysis for item's recommendation and reviews.
- Added numeric variables for Division, Department and Class name in order to get correlation among variables.

Insights from data

- Data seems unbalanced since in target variable, sample of one class is much higher than the other. [4172 samples for “Not Recommended” and 19314 samples for “Recommended”]. This may lead to overfitting.

To overcome this issue, we have three methods : undersampling, oversampling and SMOTE. Oversampling and SMOTE can be used to increase the cardinality of minority class.

- Using univariate, bivariate and text analysis, we can get some really useful insights such as:

Almost all products are high rated.

Large number of products is being recommended.

All the divisions are doing really well in terms of ratings and most ratings are for division of Intimates.

Some products have mixed reviews but majority of them are positive.

Overall text and sentiment analysis shows that majority of titles and reviews are positive, joyful and continuing the trust.

PROJECT NOTES - II

(i) Model building and Interpretation
(ii) Model tuning

Age bracketing and ‘Title’ variable’s data addition as column

- Age variable has been divided into groups and added into column to get better insight and to add some good contribution into model creation.
- After text mining, Texts of title variable have been added into column which is expected to be useful in model creation.

Treatment of imbalanced data

- Since the provided dataset is imbalanced. So, SMOTE has been applied on training dataset for better model creation.

Logistic regression classification

- Various logistic regression models have been created with different variables.

1st model :

Confusion matrix :

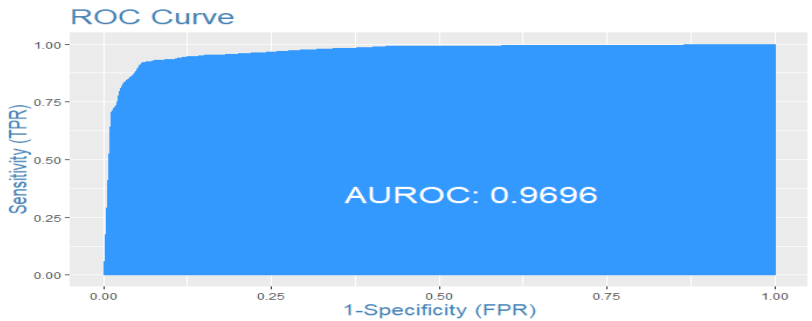
actual	prediction 0	1
0	1140	97
1	401	5404

Accuracy: 92.9%

Sensitivity: 95%

Specificity: 85%

Misclassification error: 0.07

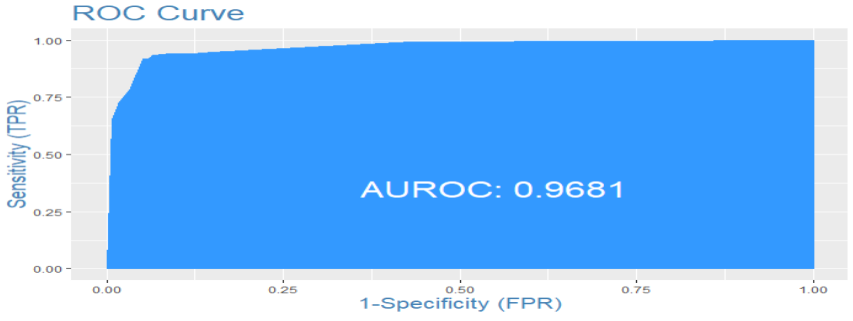


2nd model :

Confusion matrix :

actual	prediction 0	1
0	1162	75
1	454	5351

Accuracy: 92.5%
Sensitivity: 94%
Specificity: 90.9%
Misclassification error: 0.07



Logistic regression classification

3rd model :

Confusion matrix :

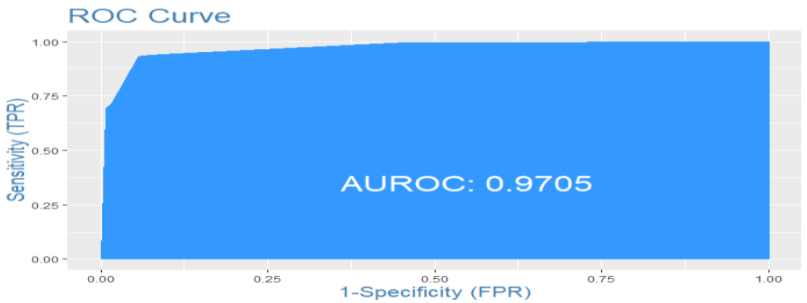
actual	prediction 0	1
0	1168	69
1	393	5412

Accuracy : 93.4%

Sensitivity : 94%

Specificity : 90.8%

Misclassification error: 0.07

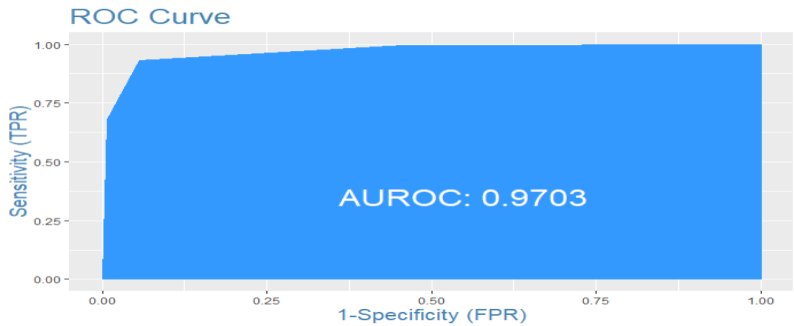


4th model :

Confusion matrix :

actual	prediction 0	1
0	1168	69
1	393	5412

Accuracy: 93.4%
Sensitivity: 93.2%
Specificity: 94.4%
Misclassification error: 0.07



Naïve Bayes classifier

1st model :

Confusion matrix :

Actual	Predicted 0	1
0	1128	109
1	3826	1979

Accuracy : 44.1%

Sensitivity : 34%

Specificity : 91.2%

2nd model :

Confusion matrix :

Actual	Predicted 0	1
0	1097	140
1	422	5383

Accuracy : 92%

Sensitivity : 92.7%

Specificity : 88.7%

Naïve Bayes classifier

3rd model :

Confusion matrix :

Actual	Predicted 0	1
0	1100	137
1	343	5462

Accuracy : 93.2%

Sensitivity : 94.1%

Specificity : 88.9%

4th model :

Confusion matrix :

Actual	Predicted 0	1
0	1168	69
1	393	5412

Accuracy : 93.4%

Sensitivity : 93.2%

Specificity : 94.4%

k - NN classifier

1st model :

Confusion matrix :

Actual	Prediction 0	1
0	1101	151
1	313	5477

Accuracy : 93.4%

Sensitivity : 94.6%

Specificity : 87.9%

2nd model (k = 5):

Confusion matrix :

Actual	Prediction 0	1
0	1017	235
1	289	5501

Accuracy : 92.6%

Sensitivity : 95%

Specificity : 81.2%

k - NN classifier

2nd model (k = 7) :

Confusion matrix :

Actual	Prediction 0	1
0	1011	241
1	271	5519

Accuracy : 92.7%

Sensitivity : 95.3%

Specificity : 80.7%

Random forest classifier

1st model :

Confusion matrix :

Actual	Prediction 0	1
0	1157	80
1	385	5420

Accuracy : 93.4%

Sensitivity : 93.4%

Specificity : 93.5%

Error rate : 0.07

2nd model :

Confusion matrix :

Actual	Prediction 0	1
0	1158	79
1	378	5427

Accuracy : 93.5%

Sensitivity : 93.5%

Specificity : 93.6%

Error rate : 0.06

Random forest classifier

3rd model :

Confusion matrix :

Actual	Prediction 0	1
0	1168	69
1	393	5412

Accuracy : 93.4%

Sensitivity : 93.2%

Specificity : 94.4%

Error rate : 0.07

Model performance and tuning

- While creating the models based on various algorithms, different combination of variables has been used to get better results.
- For tuning of random forest models, parameters has been manipulated with various values such as mtry.

Best models out of each kind of models

- Logistic regression : model 3
- Naïve bayes : model 3
- Knn model : model 2 (k=7)
- Random forest : model 2

Best model out of all models

- Although, most of the models are providing fine results but 2nd model of Random Forest classifier seems to be giving the best results after considering the performance elements such as Accuracy, Sensitivity, Specificity and error rates.
- Most important variable that comes out from this analysis is Rating.
- Interpreting the best model,

On fitting with test data, accuracy of model is 93.5% with 93.5% sensitivity and 93.6% specificity. Error rate of this model is 6% which is better than all the models.

Tuning of this model has been done with mtry, So at mtry = 2, it gave the best result.

SMOTE has been applied on training dataset of this model to make the dataset balanced then it has been tested with testing data.