REPORT – CAPSTONE PROJECT II

Muzaffer Estelik
estelik.muzaffer@gmail.com

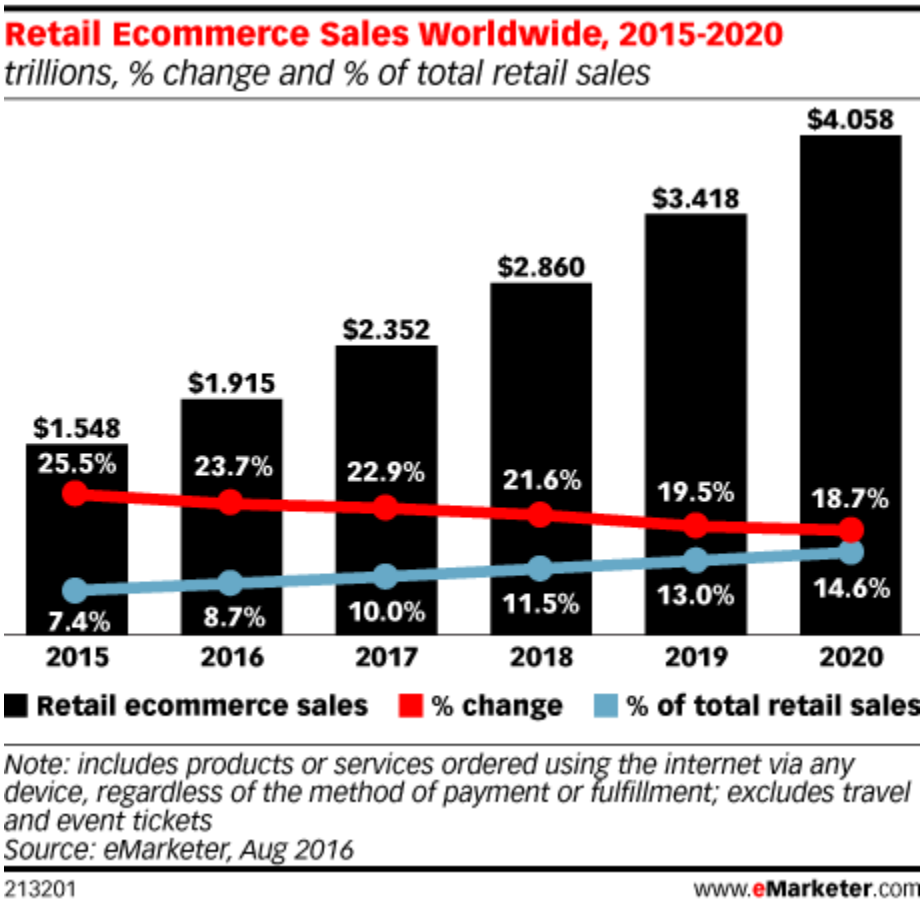
1. PROBLEM

Prediction of ratings based on women's e-commerce reviews.

a. Introduction:

E-**commerce** is the activity of buying or selling of products on online services or over the Internet[1]. Ecommerce refers to commercial transactions conducted online. This means that whenever you buy and sell something using the Internet, you're involved in e-commerce.

Needless to say, e-commerce has grown by leaps and bounds since the first time e-commerce transaction was committed in 1994. BigCommerce cites that e-commerce is growing 23% year-over-year, and according to eMarkerter, global e-commerce sales are expected to top $27 trillion in 2020 — and that's just statistics for the retail sector.



1 https://en.wikipedia.org/wiki/E-commerce

REPORT – CAPSTONE PROJECT II

Global retail e-commerce sales are projected to reach $27 trillion by 2020. It's obvious that, each day e-commerce will be more and more important for the companies.

This project aims using NLP techniques, different machine learning models and Deep Learning for predicting the rates of the products into 2 categories (good, bad).

2. DATA WRANGLING

a. Examining the Features and Samples

Features:

F.Nu	Name	Explanation	Type
1	Clothing ID	Integer Categorical variable that refers to the specific piece being reviewed	Linear
2	Age	Positive Integer variable of the reviewers age	Linear
3	Title	String variable for the title of the review	Text
4	Review Text	String variable for the review body	Text
5	Rating	Positive Ordinal Integer variable for the product score granted by the customer from 1 Worst, to 5 Best	Linear
6	Recommended IND	Binary variable stating where the customer recommends the product where 1 is recommended, 0 is not recommended	Binary
7	Positive Feedback Count	Positive Integer documenting the number of other customers who found this review positive	Linear
8	Division Name	Categorical name of the product high level division	Text
9	Department Name	Categorical name of the product department name	Text
10	Class Name	Categorical name of the product class name.	Text

b. Class Distribution:

The ratings are scored by the customer from 1 Worst, to 5 Best. I added up a new feature (Positively Rated) splitting these ratings to a binary classification adding rating scores 1,2,3 as bad (0) and 4,5 as good (1).

There are 23486 rows, each representing a review for a different clothing item.

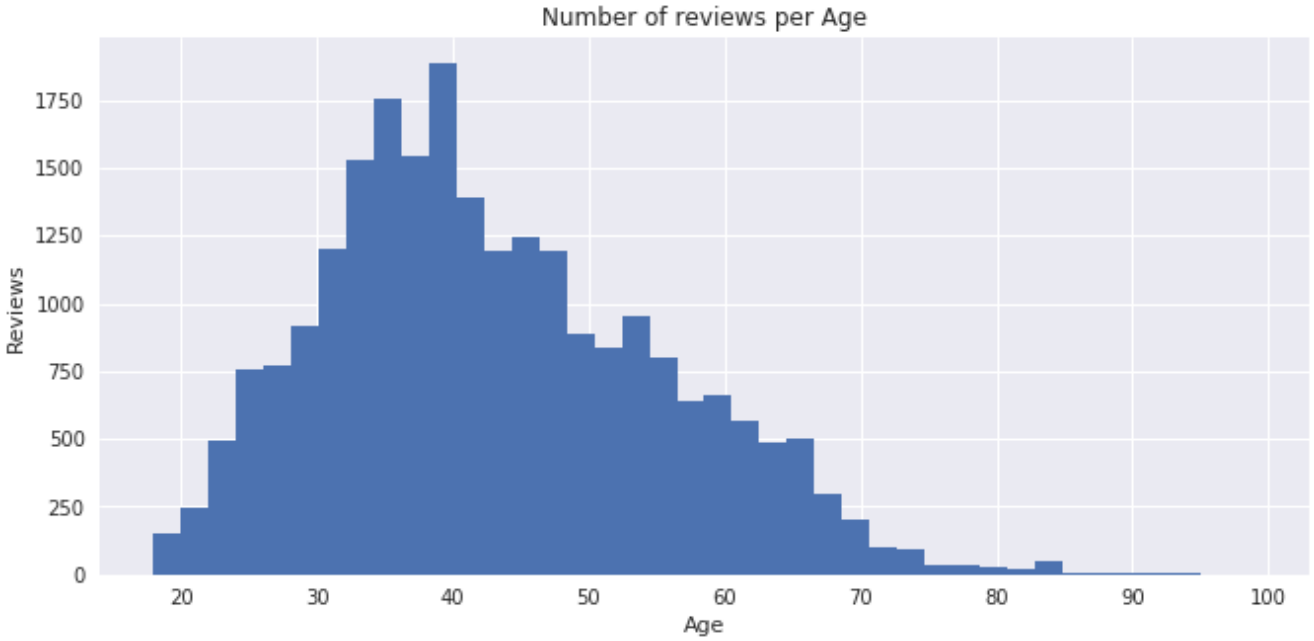
There are 10 attributes like clothing id, age, title, review text, etc. The data set is labeled with 2 different classes. Class 1 corresponds to positively rated item. Class 0 corresponds to negatively rated item.

The data set is looking like biased towards positive ratings but there are enough samples of negative ratings for ML models to make predictions.

3. EXPLORATORY DATA ANALYSIS

There are 23486 sample reviews in the data set. I created several plots to better understand the characteristics of the data

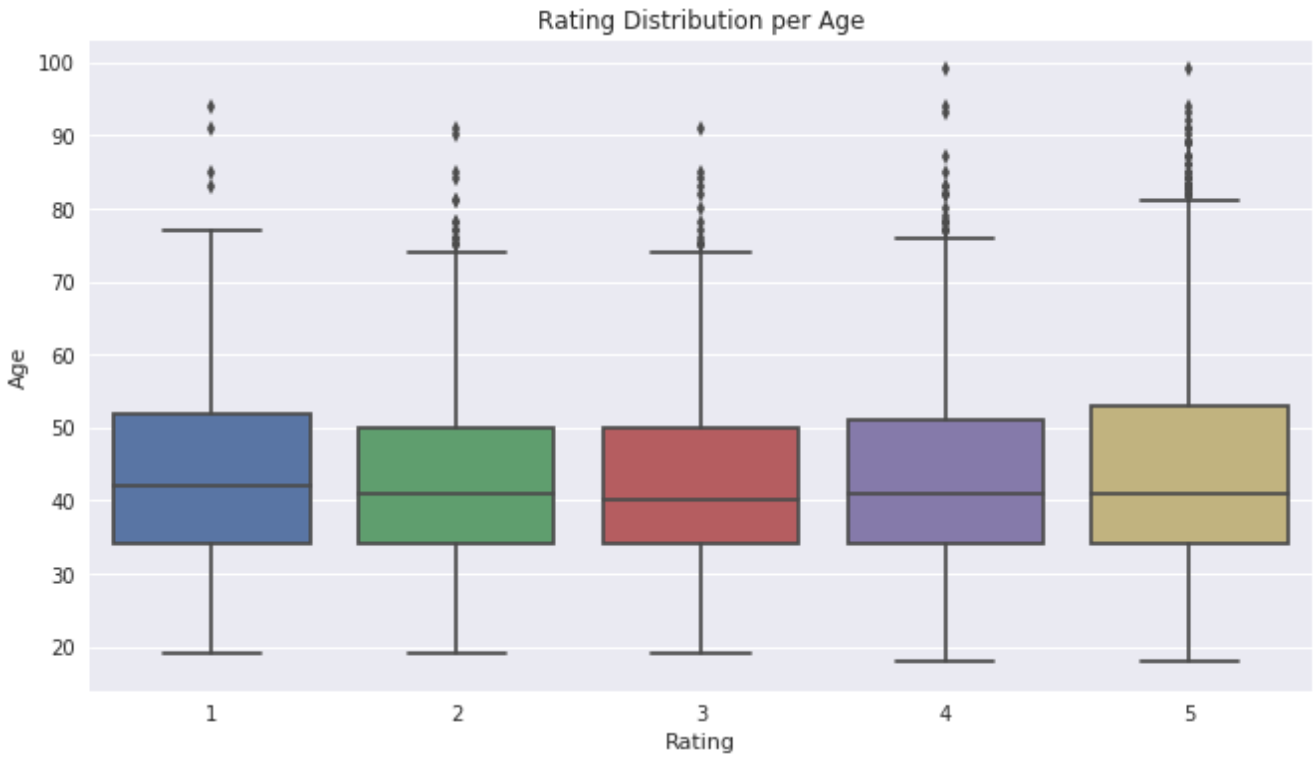
a. Number of reviews per age:



According to this plot, we can say that, 25-50 is the most reviewing age group. In other words, this age group is main target of the company.

REPORT – CAPSTONE PROJECT II

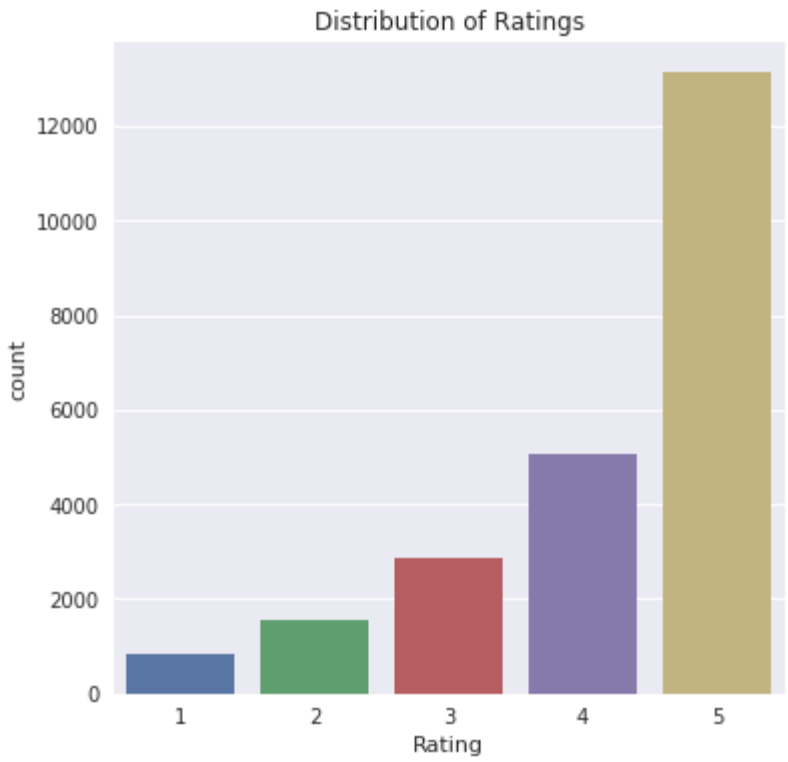
b. Distributions of ratings per age:



This plot is created to see if the happiness of the customers with the products are changing according to age group or not. It looks like, upper/lower quartiles and medians are pretty close to each other, which means age is not a factor on the ratings of the reviews.

REPORT – CAPSTONE PROJECT II

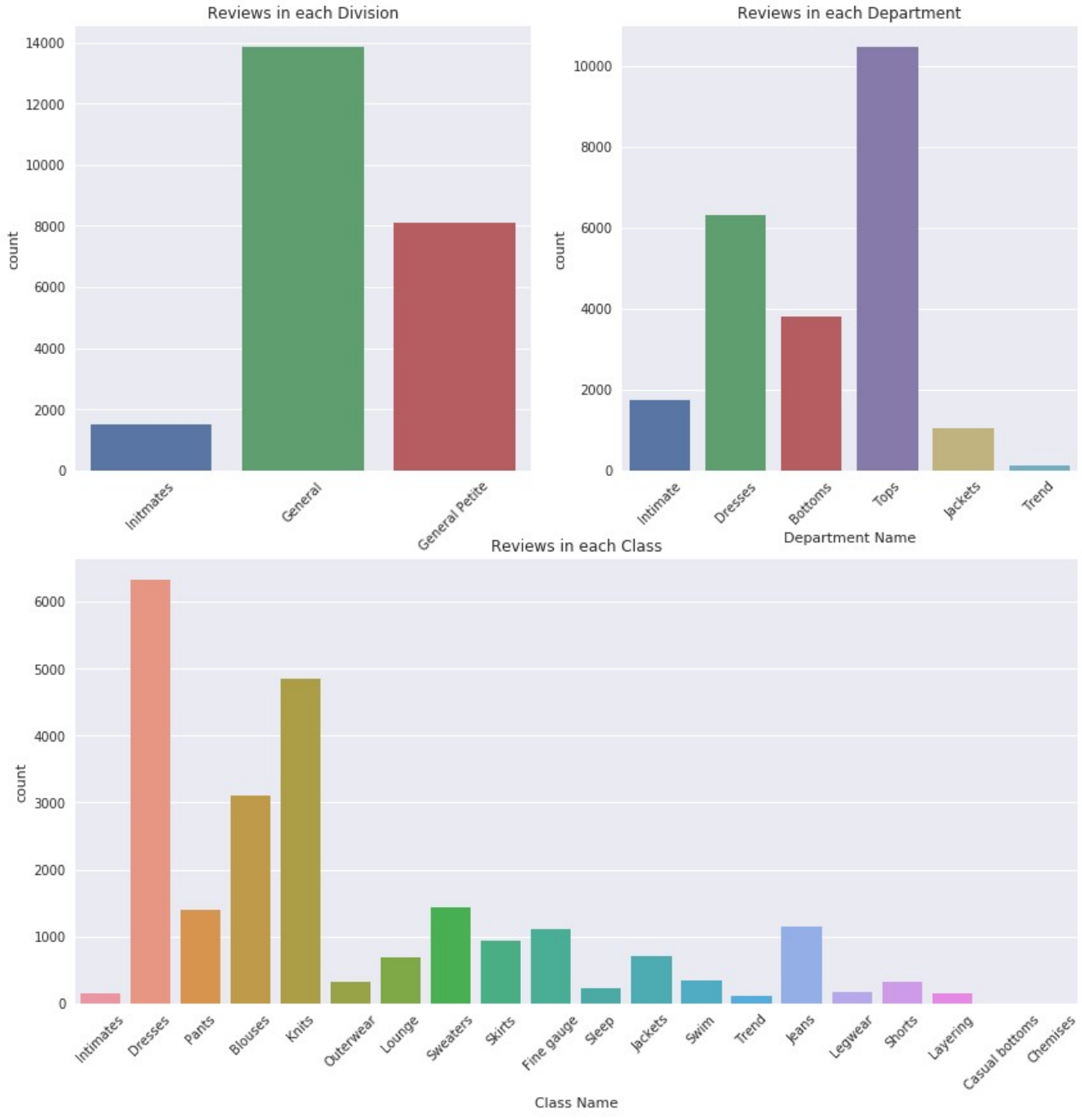
c. Distribution of ratings



Rating scores are showing a general satisfaction but of course companies need to work more to make it better. On the other hand, data looks like imbalanced but we have enough data for our ML models to make successful predictions.

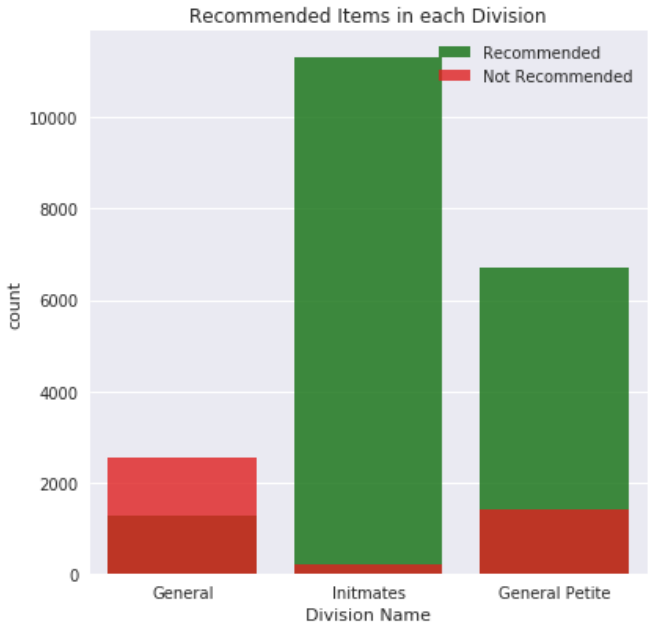
REPORT – CAPSTONE PROJECT II

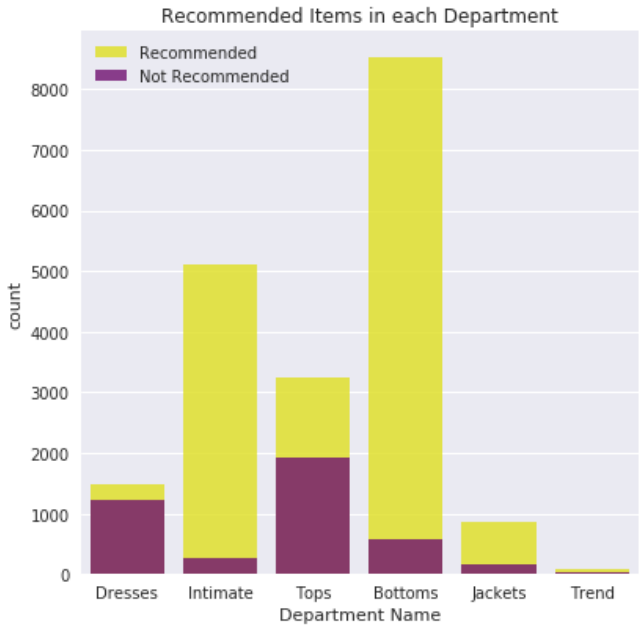
d. Distributions of sales amounts per Division/Department/Class

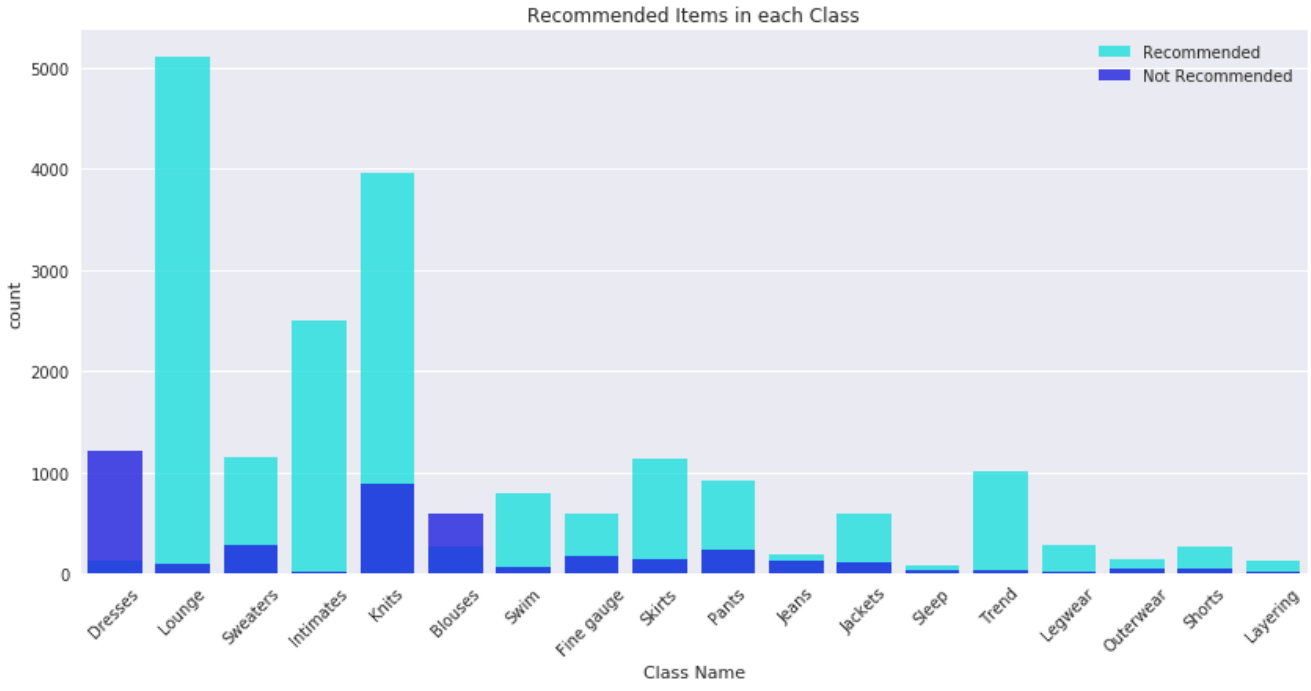


This plot is showing kind of the performance of the popularity of the department, division and classes.

Now, let's see the same plot with recommended feature added up version.



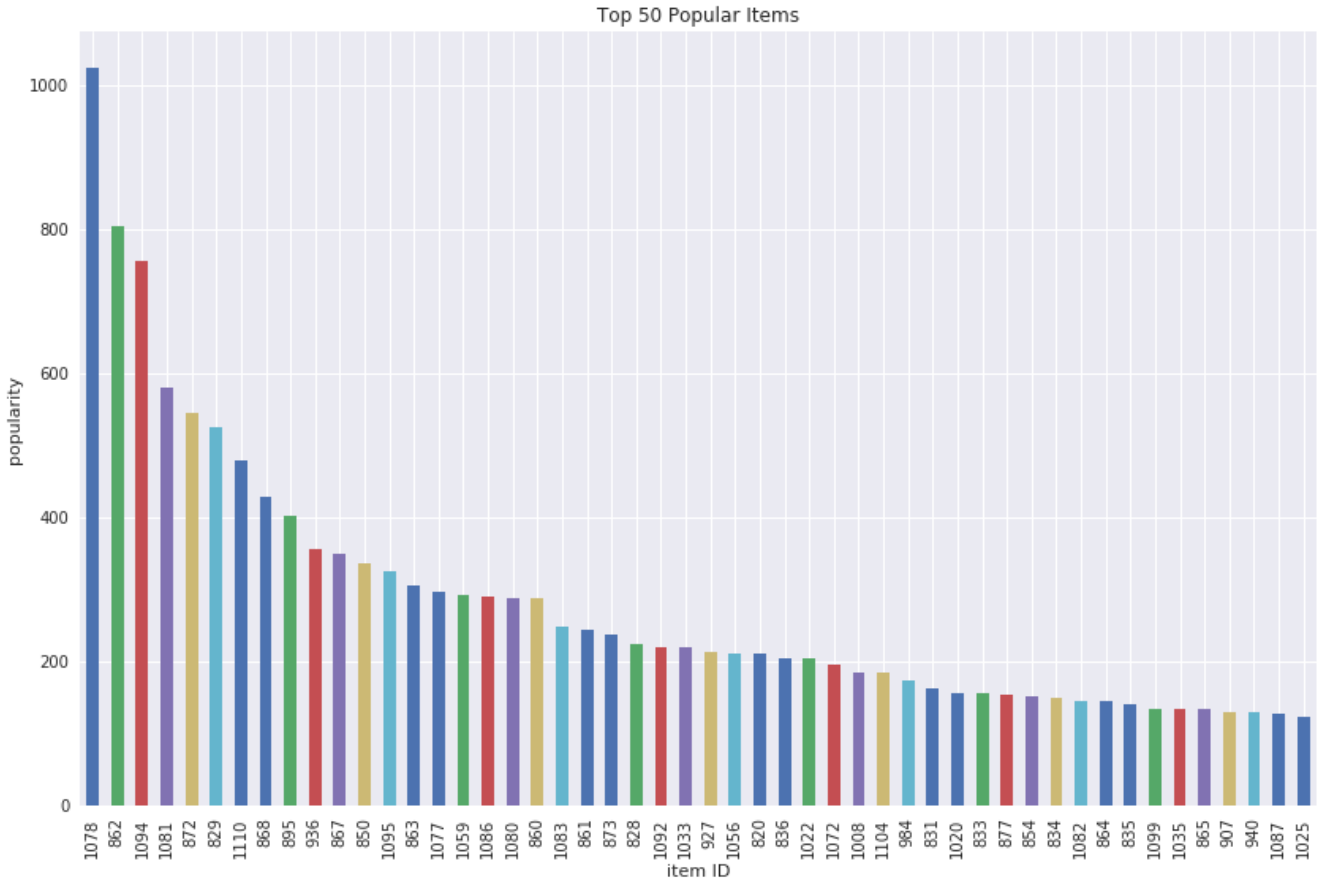




Dresses class doesn't look normal comparing the amount of reviews and good ratings. But I couldn't find any reason why it appears like bad with this plot.

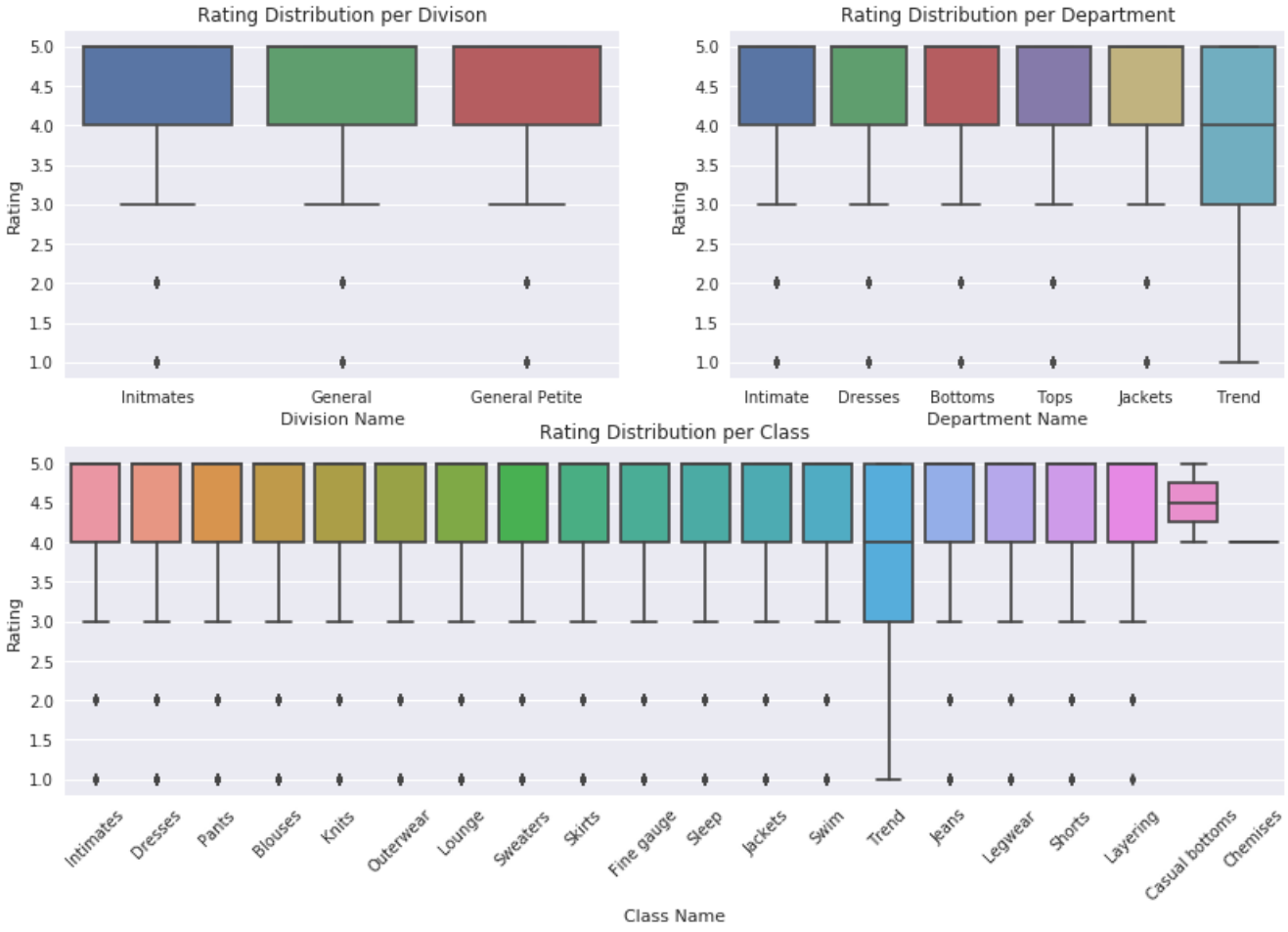
REPORT – CAPSTONE PROJECT II

e. Top 50 popular items:



This is an amazing plot, providing great information for the different branches of the company.

f. Distribution of the ratings per Department, Division and Class



4. NATURAL LANGUAGE PROCESSING

I followed below steps for text pre-processing :

- Removing non-alphabetical words,
- Removing stop words
- Lemmatizing
- Creating bag of words

I also tried to use different tools to compare the performances of each.

REPORT – CAPSTONE PROJECT II

I used word tokenization with NLTK and word tokenization with Regex. They both provided the same sequence of common words but regex tokenization gave higher frequencies.

Top 10 common tokens (NLTK):

[('dress', 11337), ('fit', 10121), ('size', 9360), ('love', 8979), ('top', 8273), ('like', 7032), ('color', 6908), ('look', 6885), ('wear', 6519), ('great', 6094)]

Top 10 common tokens (Regex):
[('dress', 11438), ('fit', 10180), ('size', 9439), ('love', 9004), ('top', 8370), ('like', 7175), ('color', 6987), ('look', 6914), ('wear', 6537), ('great', 6117)]

For this particular project, the difference is not important. I used regex tokenization for the following part of the project.

For the same purpose used Gensim and SpaCy libraries as well.

Used NER (Named Entity Recognition) only to see how it's working.

```
[('absolutely', 'RB'), ('wonderful', 'JJ'), ('silky', 'JJ'),
('sexy','NN'), ('comfortable', 'JJ')]
```

5. MACHINE LEARNING MODELS

I updated the project as to split the ratings as good (4,5) or bad (1,2,3). So, the problem is to make binary predictions either the review text belongs to a good or a bad rating.

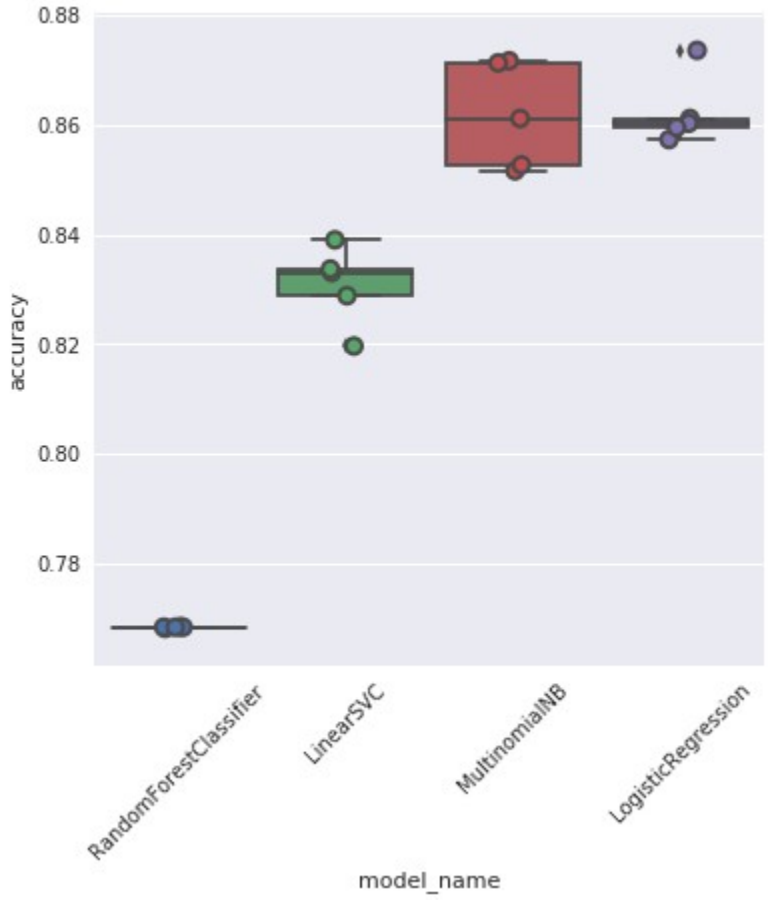
I choosed the easy way and dropped the missing values because I felt comfortable with the sample size.

After splitting the data to train and test sets, I applied 4 different machine learning model with cross validation.

- Random Forest Classifier
- Linear SVC
- Multinomial Naive Bayes Classifier
- Logistic Regression

REPORT – CAPSTONE PROJECT II

Below is the plot of accuracy scores of each ML model.




```
Below is the mean accuracy scores of the model I listed above.
LinearSVC                0.830866
LogisticRegression       0.862371
MultinomialNB            0.861688
RandomForestClassifier   0.768390
```

Logistic Regression and Multinomial Naive Bayes classifiers are providing pretty similar results. Previously, I tried making predictions of 5 classes and the scores of Multinomial Naive Bayes was better than Logistic Regression.

With this project, I also tried TPOT[2] for the first time in my projects. After 1 night (5 hours) of running, TPOT also provided Logistic Regression as the best scoring model.

2 https://github.com/EpistasisLab/tpot

REPORT – CAPSTONE PROJECT II

Below is the list of train and test scores of the ML models:

Model Name	Train Accuracy Score	Test Accuracy Score
Linear SVC	0.830866	0.832948
Naive Bayes	0.861688	0.859608
Logistic Regression	0.862371	0.845893
Random Forest	0.768390	0.774387

I also used Recurrent Neural Networks, and in particular LSTMs, to perform sentiment analysis in Keras. But it didn’t resulted good scores (0.77).