

Information Retrival: Assignment 3

Ankit Kumar 2021015

1. Dataset Details

- Electronic.json Downloaded from:
https://datarepo.eng.ucsd.edu/mcauley_group/data/amazon_v2/categoryFiles/Electronics.json.gz

```
Electronics dataframe:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20994353 entries, 0 to 20994352
Data columns (total 12 columns):
#   Column          Dtype
---  -
0   overall         int64
1   verified        bool
2   reviewTime      object
3   reviewerID      object
4   asin            object
5   style           object
6   reviewerName    object
7   reviewText      object
8   summary         object
9   unixReviewTime  int64
10  vote            object
11  image           object
dtypes: bool(1), int64(2), object(9)
memory usage: 1.7+ GB
```

- Meta Electronics.json Downloaded from:
https://datarepo.eng.ucsd.edu/mcauley_group/data/amazon_v2/metaFiles2/meta_Electronics.json.gz

```
Meta Electronics dataframe:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 786445 entries, 0 to 786444
Data columns (total 19 columns):
#   Column          Non-Null Count  Dtype
---  -
0   category        786445 non-null object
1   tech1           786445 non-null object
2   description     786445 non-null object
3   fit             786445 non-null object
4   title           786445 non-null object
5   also_buy        786445 non-null object
6   tech2           786445 non-null object
7   brand           786445 non-null object
8   feature         786445 non-null object
```

```
9 rank 786445 non-null object
10 also_view 786445 non-null object
11 main_cat 786445 non-null object
12 similar_item 786445 non-null object
13 date 786445 non-null object
14 price 786445 non-null object
15 asin 786445 non-null object
16 imageURL 786445 non-null object
17 imageURLHighRes 786445 non-null object
18 details 785607 non-null object
dtypes: object(19)
memory usage: 114.0+ MB
```

Category: Headphones

- 4
- Report the total number of rows for the product. Perform appropriate pre-processing as handling missing values, duplicates and other.

```
Total Number of Reviews for Headphones: 1553905
Average Rating Score: 3.91
Number of Unique Headphone Products: 26864
Number of Good Ratings: 1098538
Number of Bad Ratings: 455367
Number of Reviews corresponding to each Rating:
overall
5.0 833388
4.0 265150
1.0 199233
3.0 139264
2.0 116870
```

5

Done

6

To extract relevant statistics, perform the following EDA - Top 20 most reviewed brands:

```
brand
Sony 106069
Sennheiser 73590
Bose 33887
Beats 26355
Mpow 24975
Audio-Technica 24869
Koss 22354
```

Bluedio	20893
Symphonized	20579
JVC	20456
Panasonic	20427
Etre Jeune	18310
Philips	18025
EldHus	17610
XBRN	15897
TaoTronics	14658
Plantronics	14486
SoundPEATS	14272
MEE audio	12878
Toysdone	12796




Top 20 least reviewed brands:

Howard O. Pittman	1
Maxpod	1
Stitch	1
DZAT	1
Phshion	1
Matezon	1
EzzMaxx	1
MOMO	1
TKS	1
UNHO	1
E-3LUE	1
Twisters	1
FocalTop	1
Creek	1
U Happy	1
Vicious Vinyl Shack	1
YUIN	1
Blueflame	1
Fosheng	1
April Music	1

- The most positively reviewed headphone is: **Sony MDR7506 Professional Large Diaphragm Headphone**
- Count of ratings for the product over 5 consecutive years:

overall	1.0	2.0	3.0	4.0	5.0
reviewTime					
2014	20525.0	13024.0	18113.0	37500.0	112327.0
2015	42119.0	24582.0	30365.0	57790.0	187244.0
2016	56260.0	30941.0	35178.0	64768.0	210633.0
2017	36387.0	19752.0	21223.0	35761.0	135478.0

2018	18708.0	9601.0	9853.0	15533.0	62241.0
------	---------	--------	--------	---------	---------

- Good Reviews Word Cloud  Good Reviews Word Cloud
- Bad Reviews Word Cloud  Bad Reviews Word Cloud
- Plot a pie chart for Distribution of Ratings vs. the No. of Reviews.  Pie Chart
- Report in which year the product got maximum reviews.
 - Year with maximum reviews: 2016
- Year with the highest number of customers: 2016
- TF-IDF used to extract the most important words from the reviews.

5 Machine Learning Models:

- Logistic Regression
- Naive Bayes
- SVM (Optimized)
- Random Forest (Optimized)
- Gradient Boosting (Optimized)

Model: Logistic Regression				
	precision	recall	f1-score	support
Average	0.44	0.14	0.21	35077
Bad	0.75	0.77	0.76	78806
Good	0.88	0.95	0.92	274594
accuracy			0.84	388477
macro avg	0.69	0.62	0.63	388477
weighted avg	0.82	0.84	0.82	388477
Model: Naive Bayes				
	precision	recall	f1-score	support
Average	0.56	0.00	0.00	35077
Bad	0.84	0.44	0.57	78806
Good	0.78	0.99	0.87	274594
accuracy			0.79	388477
macro avg	0.73	0.48	0.48	388477
weighted avg	0.77	0.79	0.73	388477
Model: SVM (Optimized)				
	precision	recall	f1-score	support
Average	0.50	0.06	0.10	35077
Bad	0.74	0.77	0.76	78806

Good	0.87	0.96	0.91	274594
accuracy			0.84	388477
macro avg	0.70	0.60	0.59	388477
weighted avg	0.81	0.84	0.81	388477
Model: Random Forest (Optimized)				
	precision	recall	f1-score	support
Average	0.00	0.00	0.00	35077
Bad	0.99	0.00	0.01	78806
Good	0.71	1.00	0.83	274594
accuracy			0.71	388477
macro avg	0.57	0.33	0.28	388477
weighted avg	0.70	0.71	0.59	388477
Model: Gradient Boosting (Optimized)				
	precision	recall	f1-score	support
Average	0.50	0.03	0.06	35077
Bad	0.75	0.60	0.66	78806
Good	0.82	0.97	0.89	274594
accuracy			0.81	388477
macro avg	0.69	0.53	0.54	388477
weighted avg	0.78	0.81	0.77	388477

11

User-User Collaborative Filtering MAE Values: [0.7821766302667899, 0.782124134365546, 0.7821157385300801, 0.7820361286578879, 0.7821922991082872]
Item-Item Collaborative Filtering MAE Values: [0.801168583654993, 0.8018119957664103, 0.8016944058826564, 0.8016944058826564, 0.8016944058826564]

12: Also, report the TOP 10 products by User Sum Ratings.

for rating 3, 4 and 5

	ASIN	Positive Review Count
0	B00004T8R2	3757
1	B00001P4ZH	3184
2	B00001WRSJ	2675
3	B000065BPB	1427

4	B00005N9D3	1243
5	B00005N6KG	1194
6	B000065BP9	1022
7	B00005QBU9	956
8	B00001P4XA	863
9	B00005RFD3	840