Product Recommendation in Ecommerce

MSC PROJECT PRESENTATION

AUGUST 2019

Pranjal Jain

180393996

Queen Mary
University of London

Agenda

- What is a recommender system?
- Why ? What is the need ?
- Practical Applications
- Problem Description
- Approach and Methodology
- Results
- Problems associated with Recommendations
- Conclusions

Why did I choose this topic?

- Practical Data Science Application
- Gives a good understanding about the business aspect of a product development

What is a recommender system?

- *Systems that help users **identify products** of particular interest to them.*
- ***A lot of time users don't know what they want until they get a recommendation** – system that does that.*
- *Focuses on the task of **INFORMATION FILTERING***

Source : Web

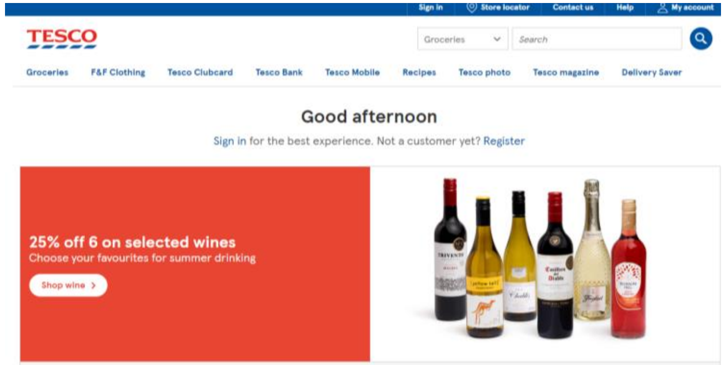
Purpose & Importance

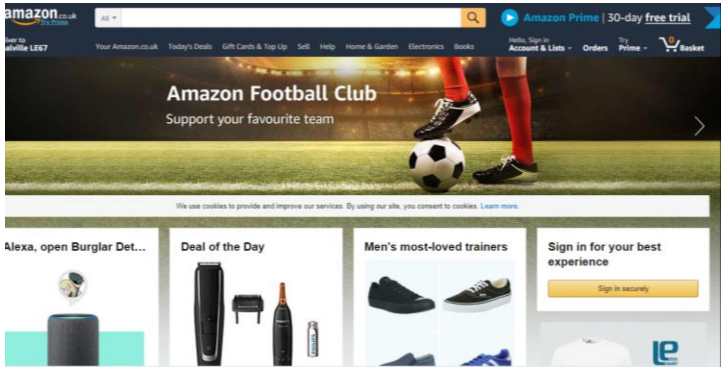
- Gives a tailored experience to the customer.
- Improves customer level satisfaction.
- Offline Recommendation – Mother / Student Example
- Increases the net sales of a merchant.

Applications of a Recommendation System

- Many of the top Ecommerce retailers use recommender systems to improve sales.
- Retail giant Amazon credits about **35%** of its revenues to the recommendation engine in use

Source :Twitter

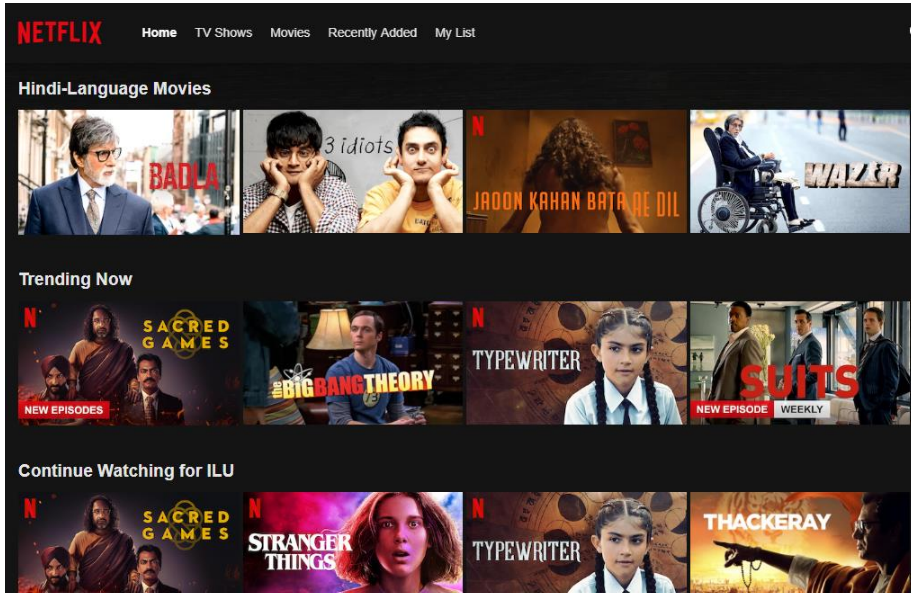




Applications continued..

Users may find new books, music, or movies that was previously unknown to them.

Netflix apparently knows better about what I would like to watch rather that me deciding myself

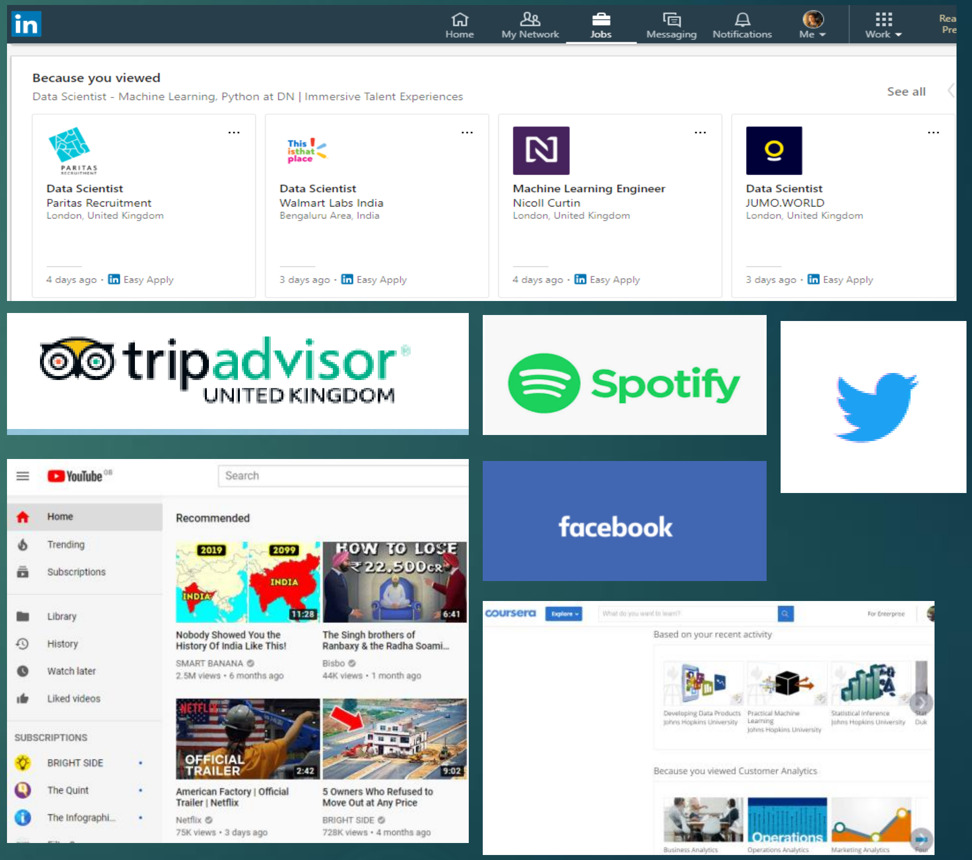


Netflix Prize Competition

- Organized in 2006
- Winning prize : $1,000,000
- Objective : Improved the accuracy of their existing system "*Cinematch*" by 10% using machine learning and data mining
- 100 million ratings of 17,770 movies from 480,189 customers.
- Took 3 years to complete the challenge

Applications continued..

- **LinkedIn** knows which Job should I apply for
- **Facebook** suggests who do I know and who should I add in my network
- **TripAdvisor** tells which place should I visit
- **Coursera** recommends me which course certification will I like based on my recent activity.



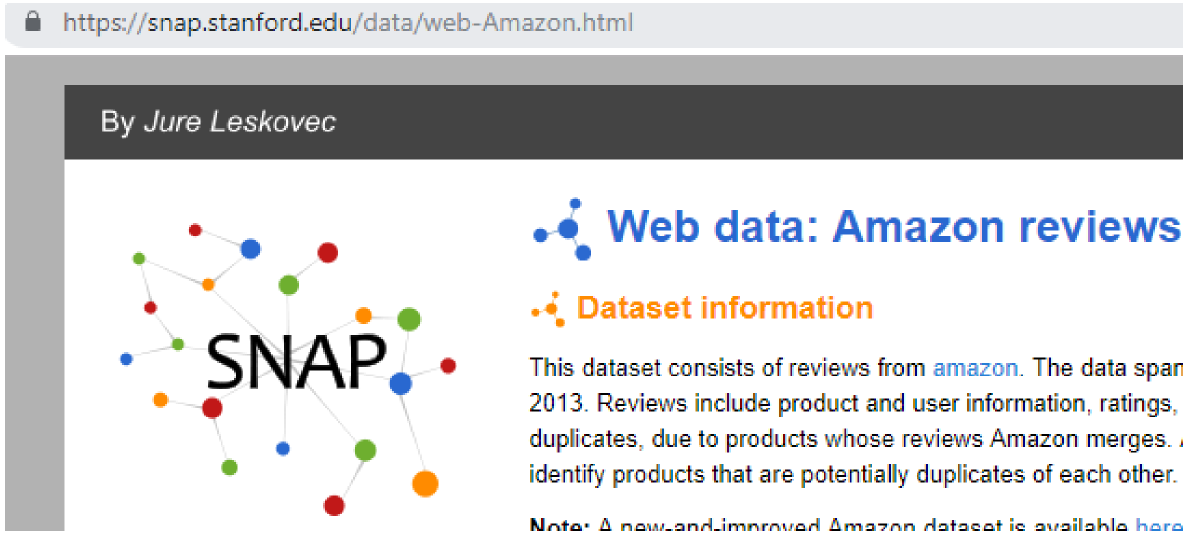
Life Cycle of a Recommendation Engine

Data Collection	Finding patterns in user behaviour and trends;	Extracting valuable insights;	Calculating probabilities or weights	Comparing them with the available item inventory;	displaying the most similar matches.

Data Collection

- Using Amazon's customer level data retrieved from SNAP Julian McAuley, UCSD repository.
- This is customer level from the online retailer www.Amazon.com for a period of 18years.

  (Jun 1995 - Mar 2013)



Data Cleaning

- The obtained dataset is in unstructured format. Hence one of the most important steps is to clean the data in a well structured format.

- The .txt raw files were converted into .csv wide tables

```
data
```


```
['product/productId: B000GKXY4S',
 'product/title: Crazy Shape Scissor Set',
 'product/price: unknown',
 'review/userId: A1QA985ULVCQOB',
 'review/profileName: Carleen M. Amadio "Lady Dragonfly"',
 'review/helpfulness: 2/2',
 'review/score: 5.0',
 'review/time: 1314057600',
 'review/summary: Fun for adults too!',
 'review/text: I really enjoy these scissors for my inspiration books that I am making (like collage, but in books) and using t
hese different textures these give is just wonderful, makes a great statement with the pictures and sayings. Want more, perfect
for any need you have even for gifts as well. Pretty cool!',
 '',
 'product/productId: B000GKXY4S',
 'product/title: Crazy Shape Scissor Set',
 'product/price: unknown',
 'review/userId: ALCX2ELNHLQA7',
 'review/profileName: Barbara',
 'review/helpfulness: 0/0',
 'review/score: 5.0',
 'review/time: 1328659200',
 'review/summary: Making the cut!',
 'review/text: Looked all over in art supply and other stores for "crazy cutting" scissors for my 4-year old grandson. These ar
e exactly what I was looking for - fun, very well made, metal rather than plastic blades (so they actually do a good job of cut
ting paper), safe ("blunt") ends, etc. (These really are for age 4 and up, not younger.) Very high quality. Very pleased with t
he product.',
 '',
```


Unstructured

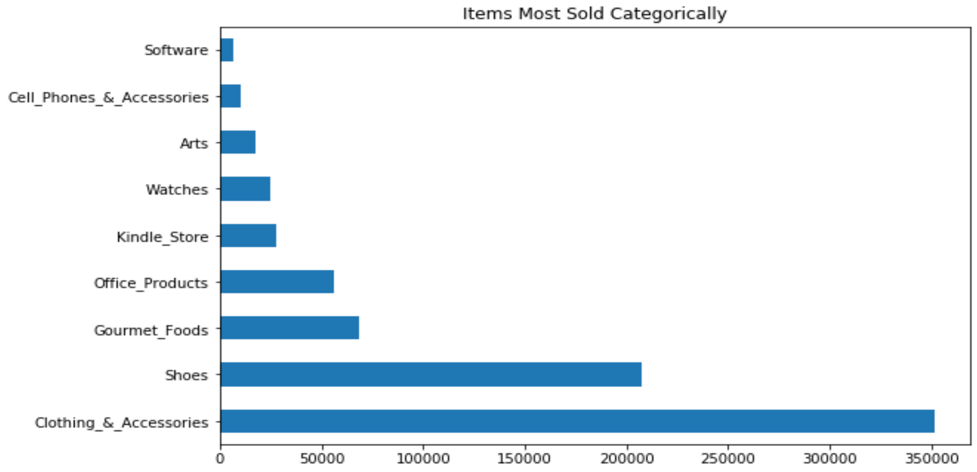
product/price	product/productId	product/title	review/helpfulness	review/profileName	review/score	review/summary	review/text	review/time	review/userId
unknown	B000GKXY4S	Crazy Shape Scissor Set	2/2	Carleen M. Amadio "Lady Dragonfly"	5.0	Fun for adults too!	I really enjoy these scissors for my inspirat...	1314057600	A1QA985ULVCQOB
unknown	B000GKXY4S	Crazy Shape Scissor Set	0/0	Barbara	5.0	Making the cut!	Looked all over in art supply and other store...	1328659200	ALCX2ELNHLQA7
unknown	B000140KIW	Fiskars Softouch Multi-Purpose Scissors 10"	1/1	L. Heminway	5.0	Fiskars Softouch Multi-Purpose Scissors, 10"	These are the BEST scissors I have ever owned...	1156636800	A2M2M4R1KG5WOL
unknown	B000140KIW	Fiskars Softouch Multi-Purpose Scissors 10"	0/0	R. GARCIA	5.0	Best scissors ever	This Fiskars Scissors are the best i've bougt...	1214784000	ARQAQ6ZYMFPCA
unknown	B000140KIW	Fiskars Softouch Multi-Purpose Scissors 10"	0/0	Dea Carey "deacarey"	5.0	A great tool to make your work easier	I finally gave in and bought these after year...	1173484800	A3FPG4LAJ1HOHZ

Clean Data

Understanding the Data

	Feature	Description	Datatype
1	product/productId	Unique ID which is associated with each product	String
2	product/title	Title of the product	String
3	product/price	price of the product	Integer
4	review/userId	Unique ID which is associated with each Customer	String
5	review/profileName	Name of the Customer	String
6	review/helpfulness	fraction of customer who found the review helpful	Integer
7	review/score	Rating of the product	Integer
8	review/time	Time of the review	UNIX time
9	review/summary	Review Summary	String
10	review/text	Text of the review	String

price	productId	title	review	profileName	score	summary	text	userId	purchasedate	Class
17.94	B000CD483K	C-Line Clear 62033 Heavyweigl	0/0	Thomas Perrin "Perrin &	5	Superior produ	Ever since some of m	A1186EZQ23CU4X	23-11-2012	Office_Products
443.04	B0006Q9950	Wasp Barcode Technologies 63	14/14	Handyman	4	Good product,	My boss had us using	A2CW9GKMNFAU	04-11-2011	Office_Products
13.99	B0001YXWV4	Panasonic MARKER ERASER KIT	0/0	C L Huddleston	5	Best markers m	We use our white bo	A14XEQHPPULFD/	01-02-2013	Office_Products
13.99	B0001YXWV4	Panasonic MARKER ERASER KIT	0/0	Eiji Nakamura	5	Good item	Fast shipment and fa	A7YN96KKCI8GO	16-01-2013	Office_Products
17.28	B000GR7OYW	Avery Nonstick Heavy-Duty EZ	0/0	shstric	5	The only brand	I am a college studer	A36BHVA80D0OH	13-11-2012	Office_Products
17.28	B000GR7OYW	Avery Nonstick Heavy-Duty EZ	0/0	v2can5	5	Good quality fo	This 4" folder was pu	AIAOFEPWPX1J8	14-09-2012	Office_Products
17.28	B000GR7OYW	Avery Nonstick Heavy-Duty EZ	0/0	rose312	1	Huge & Unrelia	I don't know if i got a	A11CL4JDRJ8ROZ	19-08-2012	Office_Products
17.28	B000GR7OYW	Avery Nonstick Heavy-Duty EZ	0/0	SeriouslyHappy	5	Good Size, Exa	I am using this to org	A2232SPXNILNBL	15-03-2012	Office_Products
17.28	B000GR7OYW	Avery Nonstick Heavy-Duty EZ	0/0	malhej	5	good product	For the price, this ite	AE10MU3XESM8I	31-12-2011	Office_Products
17.28	B000GR7OYW	Avery Nonstick Heavy-Duty EZ	0/1	LMB "Christmas Nut"	3	big and sturdy	The binder I receivec	AS1OI6YNHH8C0	28-08-2011	Office_Products
17.28	B000GR7OYW	Avery Nonstick Heavy-Duty EZ	9/9	J. Donahoe "Dog, cat, &	5	Huge, high-qua	I bought this binder f	A2TGKNAG87PYX	12-01-2011	Office_Products
17.28	B000GR7OYW	Avery Nonstick Heavy-Duty EZ	3/3	Clmence	4	The biggest Bir	I was looking for a bi	A17HUD2DYQ81U	14-05-2012	Office_Products
17.28	B000GR7OYW	Avery Nonstick Heavy-Duty EZ	1/1	Eva	1	Broke within th	I bought this binder f	A1DVVL2R5YCE4N	07-10-2012	Office_Products
17.28	B000GR7OYW	Avery Nonstick Heavy-Duty EZ	1/1	M. Taylor "Myrna"	5	Same dependa	I keep all of my equi	A18SNZN6Z16NCY	09-05-2012	Office_Products
17.28	B000GR7OYW	Avery Nonstick Heavy-Duty EZ	1/1	Peggy W. Harper "agran	5	Binder Review	Avery is a product w	AGP0OT38AVWKI	22-09-2011	Office_Products
17.28	B000GR7OYW	Avery Nonstick Heavy-Duty EZ	1/1	MsRealMuzik	5	Nice Product	The product was ship	ALF6GZ2700Y6C	07-09-2011	Office_Products
17.28	B000GR7OYW	Avery Nonstick Heavy-Duty EZ	0/0	Cheryl L. Jones "one hot	5	binder	these worked great f	A3HMZHKYT81BE.	23-01-2013	Office_Products
17.28	B000GR7OYW	Avery Nonstick Heavy-Duty EZ	0/0	CRAIG R NUSSBAUM	5	Great - holds a	Well, a binder is a pr	A3Q81RANE2GJ42	13-01-2013	Office_Products



Data Cleaning Tasks

- Loading the raw data
- Converting the data into a list
- Using dict(zip()) function, pairs the list element with other list element at corresponding index in form of key-value pairs. This can be seen from the snippet below.

```
m = df['COL'].str.split(':', expand=True).groupby(0)[1].apply(list).reset_index()
df = pd.DataFrame(dict(zip(m[0], m[1])))
```

Data Cleaning continued..

Once we obtain the structured table –

Feature x Dimension

Converting into **appropriate datatype**

E.g. date into

YY-MM-DD

Finding what **categories** data is distributed

Merging different datasets into one – **class wise**

Exploratory Analysis

- Finding missing values
- Data was generally complete except for product prices.
- ~76% was missing hence I decided to drop it.

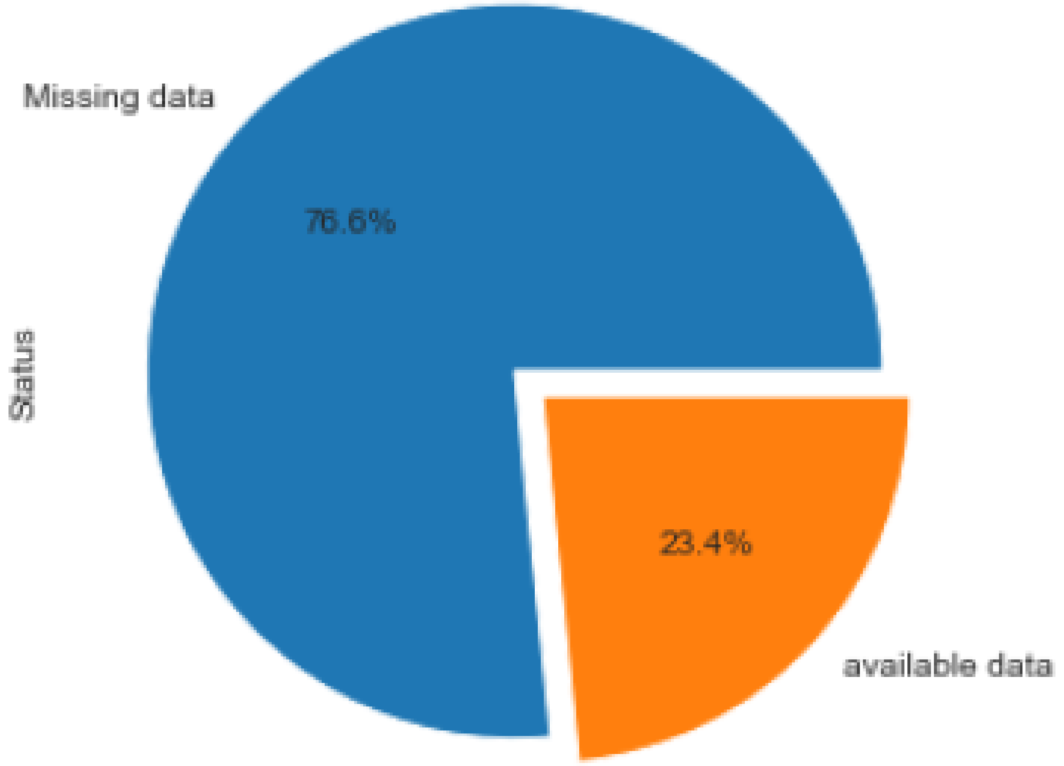


Figure 12 : Pie chart representing missing data in price column

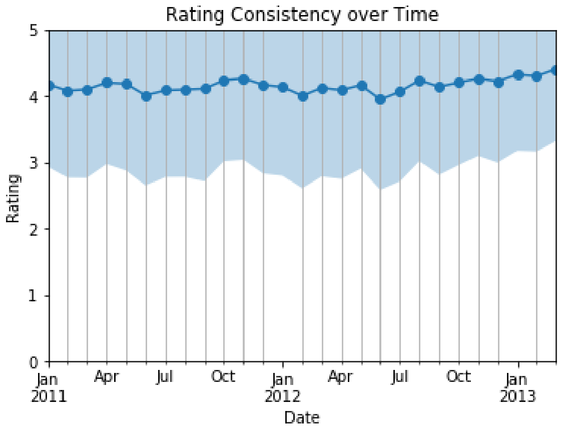
EDA continued

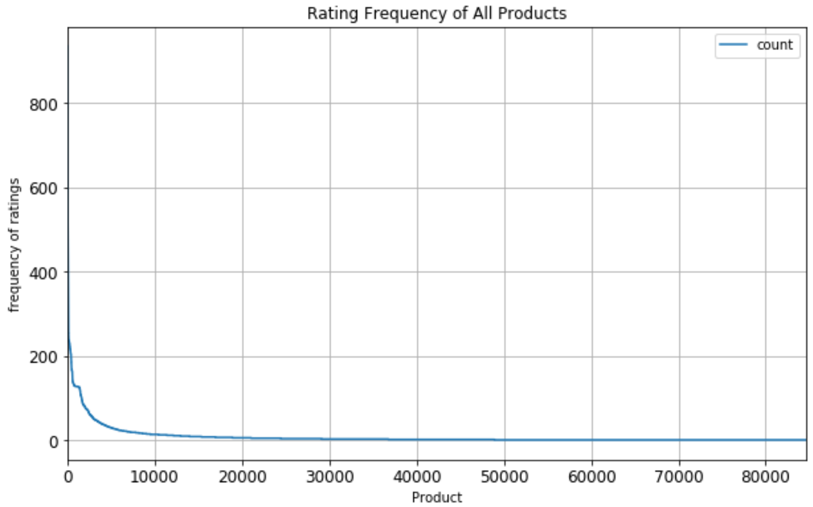
Finding consistency

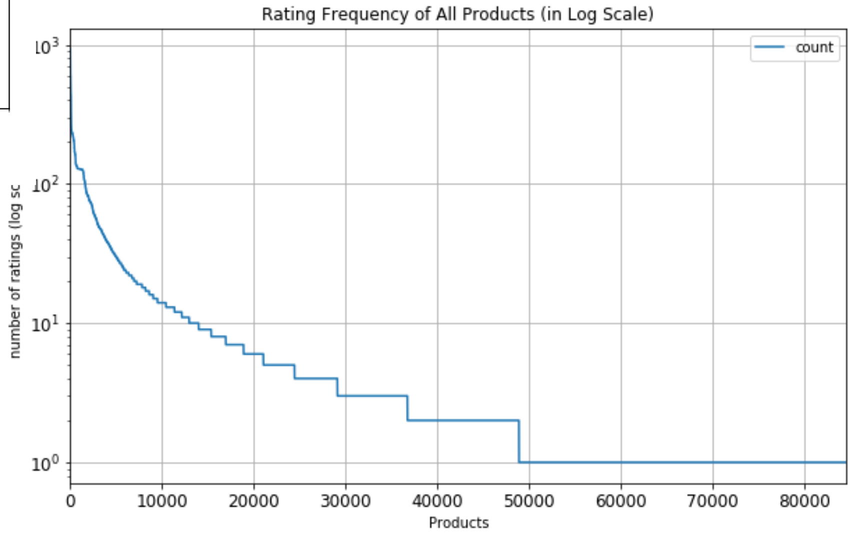
- Fig 1 shows Rating consistency along with the standard deviation over the time

Finding anomalies

- Fig 2 shows Long tail property or Power Law distribution







Feature Engineering

- SELECTING THE RELEVANT FEATURES USED IN THE ANALYSIS.
- DROPPING THE UNUSED FEATURES LIKE PRICE, REVIEW (N /D) AND THE PRODUCT TEXT .

price	dropped
productId	***
title	*
review	unused
profileName	*
score	***
summary	unused
text	***
userId	***
purchasedate	***
Class	**

Problem Formulation

- I tried to recommend product to the customer using three different algorithms.

a. Affinity based analysis

b. Similar Customer based analysis

c. Content based analysis

Affinity based analysis

- Also called *Market basket Analysis* or *association rule learning*
- *Algorithm* : Creating an n dimensional Affinity Matrix
- **Input** : User Id, Product Id
- **Transformation** : Customers who purchased product P1 also purchased product P(x)
- **Output** : List of similar products ranked according to the highest lift value.

Affinity Analysis continued..

- Association rules are defined in terms of two products(A →B)
- This is called **Base Class** and **Associated Class**
- Weight each product on the basis of *Confidence*, *Expected Confidence* and *lift* values.

T : Total number of transactions in a sample data	(Suppose)	1000
X1 : Number of customers that brought A	(Suppose)	100
X2 : Number of customers that brought B	(Suppose)	200
X3 : Frequency of co-occurrence	(Suppose)	50
Support	X3 / T	0.02
C : Confidence	X3 / X1	0.5
Ce : Expected Confidence	X2 / T	0.2
Lift	C / Ce	2.5

Figure 17 : Dummy data table used in the table just to explain the concept of various indices accordingly

	base_class	asso_class	base_class_count	asso_class_count	CO	Confidence	Expected_Confidence	Lift
0	B00004RM25	B00004VYLJ	2	2	2	1.000000	0.000312	3209.000000
1	B003L20ICO	B00004VYLJ	2	2	2	1.000000	0.000312	3209.000000
2	B00004RM25	B003L20ICO	2	2	2	1.000000	0.000312	3209.000000
3	B00004VYLJ	B003L20ICO	2	2	2	1.000000	0.000312	3209.000000
4	B00004VYLJ	B00004RM25	2	2	2	1.000000	0.000312	3209.000000
5	B003L20ICO	B00004RM25	2	2	2	1.000000	0.000312	3209.000000
6	B00006IF4J	B00006IF4L	1	1	1	1.000000	0.000156	6418.000000
7	B00006IF4L	B00006IF4J	1	1	1	1.000000	0.000156	6418.000000
8	B00006M7OC	B00006M7PK	6	6	6	1.000000	0.000935	1069.666667
9	B00006M7Q0	B00006M7PK	6	6	6	1.000000	0.000935	1069.666667
10	B00006M875	B00006M7PK	6	6	6	1.000000	0.000935	1069.666667
11	B00006M883	B00006M7PK	6	6	6	1.000000	0.000935	1069.666667
12	B00006M9TP	B00006M7PK	6	6	6	1.000000	0.000935	1069.666667

Affinity Matrix Output

Affinity Analysis continued..

- For customers with lot less past purchases, Affinity becomes a limitation.
- For these, recommendation is made using a *POPULARITY MATRIX*

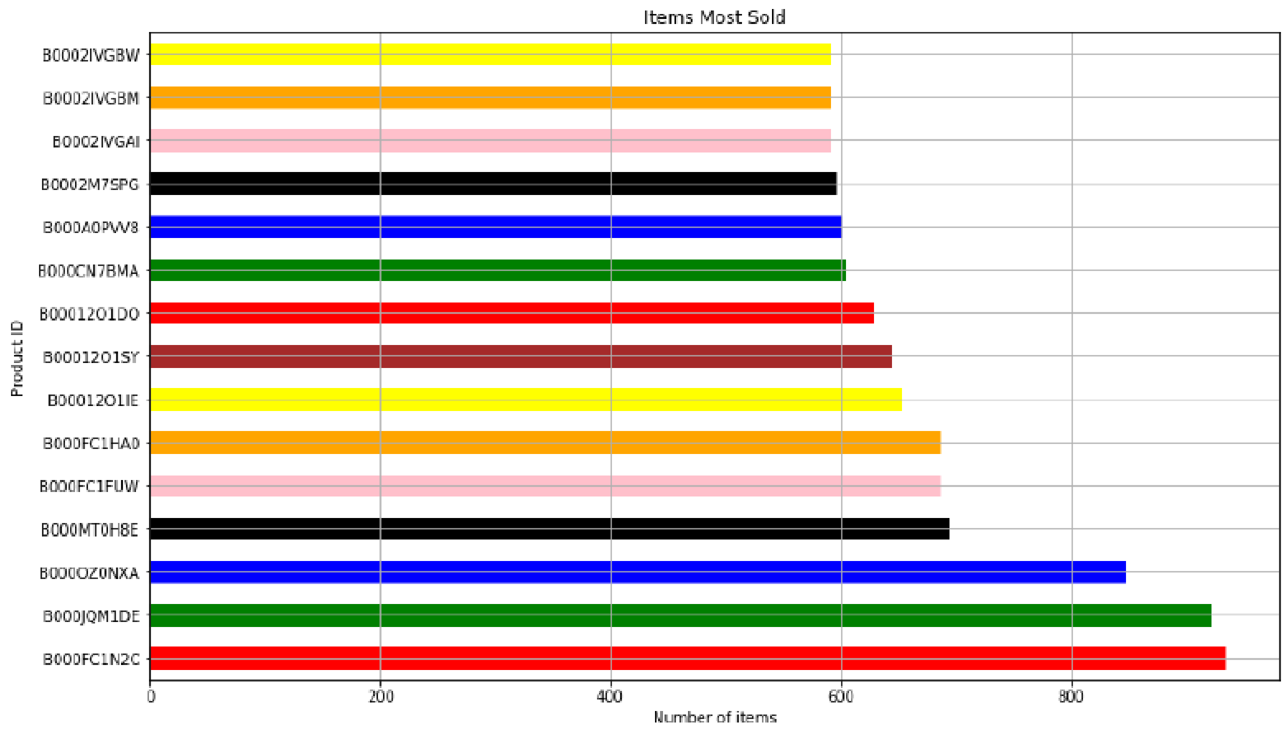


Figure 22:Popular Product frequency in terms of product ID

Affinity Analysis continued..

Other factors which were included were Recency and consumables.

Recency was based on the recent purchases of the customer which were given more weight.

Consumable was given to the products which needed timely replacement like office products, printer cartridge, soap etc.

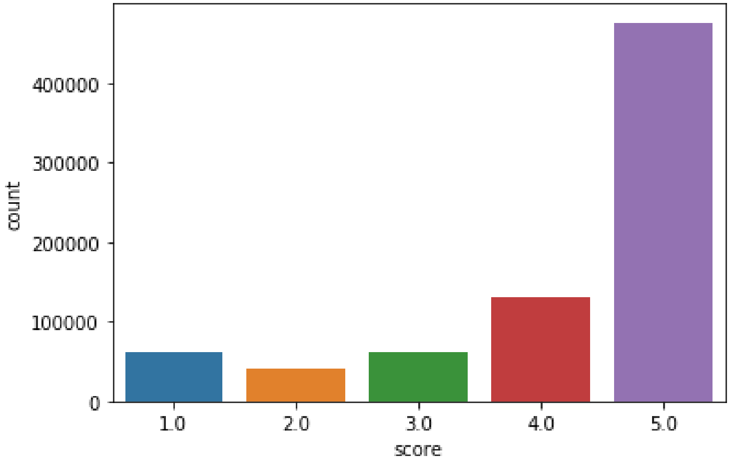
Replenished were products from Food and Gourmet category which could be recommended to the customers again.

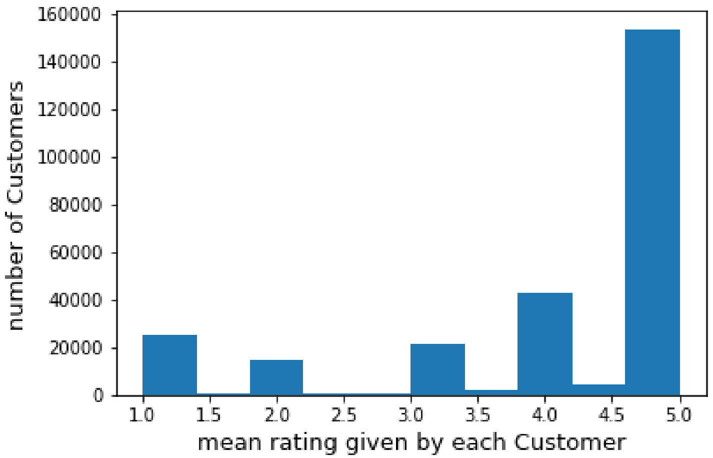
Similar Customer based analysis

- Also known as Collaborative Filtering
- Algorithm : Creating an n dimensional Utility Matrix
- Input : User Id, Product Id, Rating(Score 1-5)
- Transformation : find a group of other customers whose likes and dislikes are similar to customer A and same for products.
- Output : List of similar products/customers ranked according to the highest correlation value.

Collaborative continued..

- SIGNIFICANCE : It does not requires to understand the nature of the items and still can suggest complex products.
- LIMITATIONS : It requires a lot of data to make accurate predictions about a user hence requiring lots of computational power and resources.
- Distribution of score is shown in the histogram plot.





Collaborative continued..

- Jaccard similarity : Sim(C1, C2) = |s(C1) ∩ s(C2) |/|s(C1) U s(C2) |

- Cosine similarity : $$\text{Sim(C1, C2)} = \frac{\sum_1^n s(C1)\, s(C2)}{\sqrt{\sum s(C1)^2}\sqrt{\sum s(C2)^2}}$$

- Pearson similarity : $$\text{Sim(C1, C2)} = \frac{\sum_1^n (s(C1)-\mu)\ (s(C2)-\mu)}{\sqrt{\sum(s(C1)-\mu)^2}\sqrt{\sum(s(C1)-\mu)^2}}$$

  Here, μ is the mean score for the Customer

UTILITY MATRIX	P1	P2	P3	P4	P5	P6
C1		2		4		4
C2	1		3			5
C3		3		1		
C4	4		4		5	

Collaborative continued..

This is a sample view of the Utility Matrix using Pearson metric

----

Some Statistics derived from the analysis :

Unique Number of Customer : 2,65,401

Unique Number of Products : 84,626

	0	1	2	3	4	5	6	7
0	1.000000	-0.000322	-0.000322	-0.000322	-0.000322	-0.000322	-0.000455	-0.000322
1	-0.000322	1.000000	-0.000322	-0.000322	-0.000322	-0.000322	-0.000455	-0.000322
2	-0.000322	-0.000322	1.000000	-0.000322	-0.000322	-0.000322	-0.000455	-0.000322
3	-0.000322	-0.000322	-0.000322	1.000000	-0.000322	-0.000322	-0.000455	-0.000322
4	-0.000322	-0.000322	-0.000322	-0.000322	1.000000	-0.000322	-0.000455	-0.000322
5	-0.000322	-0.000322	-0.000322	-0.000322	-0.000322	1.000000	-0.000455	-0.000322
6	-0.000455	-0.000455	-0.000455	-0.000455	-0.000455	-0.000455	1.000000	-0.000455
7	-0.000322	-0.000322	-0.000322	-0.000322	-0.000322	-0.000322	-0.000455	1.000000
8	-0.000455	-0.000455	-0.000455	-0.000455	-0.000455	-0.000455	-0.000644	-0.000455

```
similarities,indices = findksimilarusers(99,util_df, metric='correlation')
```

```
5 most similar users for User 99:

0: User 3144, with similarity of 1.0
2: User 2104, with similarity of -0.00032175032175030326
3: User 2102, with similarity of -0.00032175032175030326
4: User 2100, with similarity of -0.00032175032175030326
```

Figure 48 : Similar user obtained using Pearson metric

```
similarities,indices=findksimilaritems(99,util_df)
```

```
5 most similar items for item 99:

0: Item Index : 100 , with similarity of 1.0
1: Item Index : 1024 , with similarity of 0.9999999999998896
2: Item Index : 2811 , with similarity of -0.00025335697998477424
3: Item Index : 2809 , with similarity of -0.00025335697998477424
4: Item Index : 2807 , with similarity of -0.00025335697998477424
```

Figure 49 : Similar product obtained using Pearson metric

Output

Content based analysis

- Using the sentiment analysis for textual mining in Product reviews given by the customer in form of written texts
- Algorithm : Creating an n dimensional review Matrix
- Input : User Id, Product Id, Rating(Score 1-5), summary, count
- Transformation : predict a product for a similar customer.
- Output : List of all similar products ranked according to the highest average score obtained by feature analysis.

Content Based continued..

- 3 important steps for this analysis :

  Tokenization

  Sample text : "This product is absolutely amazing, MUST BUY for all the dog lovers!!

  After tokenization : 'this' ,'product', 'is', 'absolutely', 'amazing', 'must' ,'buy', 'for', 'all', 'the', 'dog' 'lovers'

- **Removing *stop words*** : most commonly occurring terms such as "the", "was", "is" etc.

- **Classification** :

Content Based continued..

- I took only those products which were bought 100 times or more.
- Summary of those products were cleaned by tokenizing.
- Splitting data into train and test set.
- Classification Method used KNN Algorithm.
- To find Euclidean distance between the two datapoints.
- Predicted other similar products.
- Same procedure was repeated for predicting similar users.

Output

```
For product :  B000JHCYTE , the average Score is : 4.747747747747748
The 1st Similar product is   B0000224GM  , the average Score is : 4.745454545454545
The 2nd Similar product is   B0001YS61K  , the average Score is :  4.752212389380531
-----------------------------------------------------------
For product :  B000JIKN0A , the average Score is : 3.312820512820513
The 1st Similar product is   B00008ION9  , the average Score is : 4.096153846153846
The 2nd Similar product is   B00008IOOI  , the average Score is :  4.104761904761904
-----------------------------------------------------------
For product :  B000JLHRII , the average Score is : 3.8088642659279777
The 1st Similar product is   B000FS67LS  , the average Score is : 3.9595238095238097
The 2nd Similar product is   B00016QPAW  , the average Score is :  4.285067873303167
-----------------------------------------------------------
```

Two recommended products according to the score

```
Based on  reviews, for Customer   AR3WTWQ4H2GOD
The 1st similar Customer is   A292V24Y5TJIIC .
Customer likes following products
Based on  reviews, for Customer   AY3NS68W1N98P
The 1st similar Customer is   A292V24Y5TJIIC .
Customer likes following products
Based on  reviews, for Customer   AYUCFJMTDAJC3
The 1st similar Customer is   A3MVS23LU1ZC1E .
Customer likes following products
 B0007YX26S
 B000EE9D00
Based on  reviews, for Customer   AZ2X4NOLQ1UNV
The 1st similar Customer is   A3MVS23LU1ZC1E .
Customer likes following products
 B0007YX26S
 B000EE9D00
--------------------------------------------------
```

Figure 53 : Output for similar customer analysis

Results

Results

Affinity

Forming a customer table to recommend 5 products

2 on the basis of lift score

2 on the basis of popularity

1 on the basis of recency

FYI : If product was replenishable, it was given more weight.

Collaborative

Predicting a group of 5 users which are similar to the customer in question.

Predicting 5 products similar to the product in question .

This is based on a score – based system

Text Based

Predicting 2 similar products based on the average score obtained after classification.

Predicting similar customers and their top bought product based on the review analysis.

Customer Table – Priority and less important

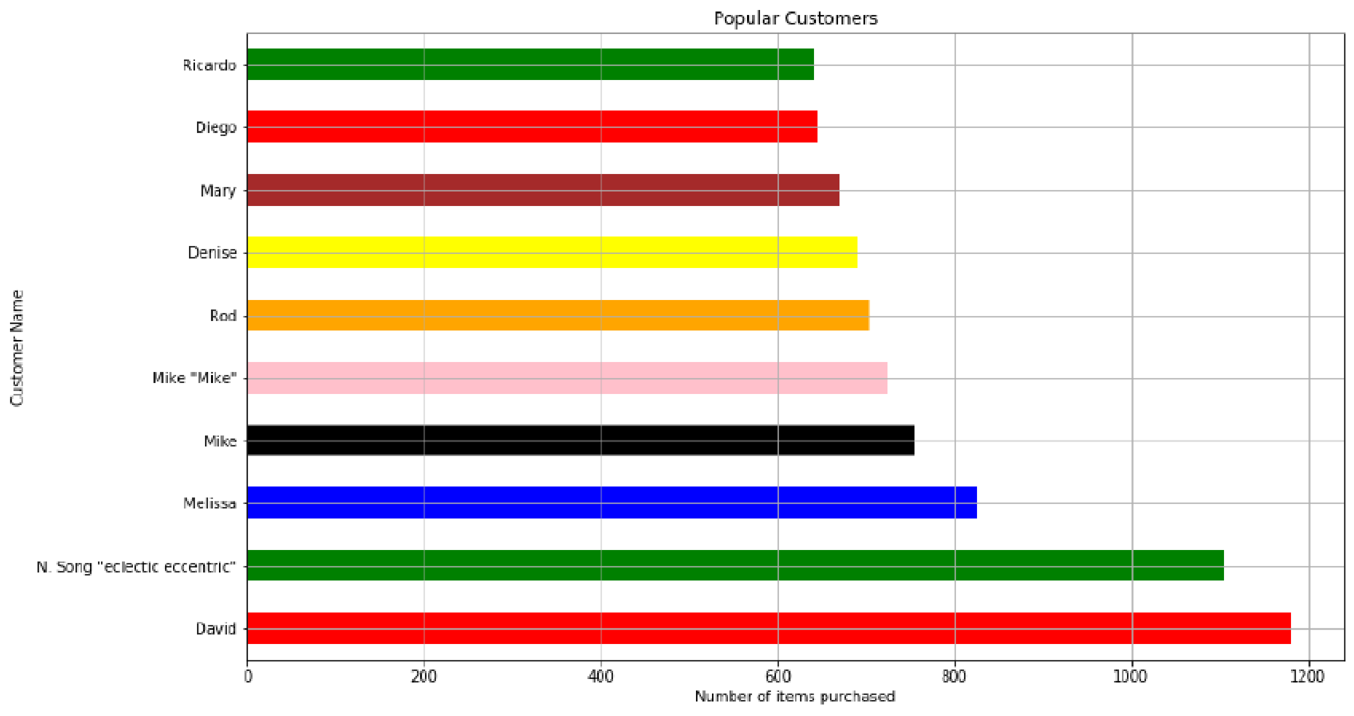


Figure 27 : This bar plot shows the most active customers from our dataset.

```
userId
 AV9NKOINQONGN     1104
 ANDNAFTUKW3D9      708
 A18WDI1W0XJLNL     618
 A1NWLSBO3XE74A     593
 AAXDBRTR04J35      593
 A2KV4LCZMPSIMO     579
 A8QAOMRX0JLXJ      573
 A37LKHEF0ZPVSA     564
 A3RTT2QP2V25QW     563
 A1X553B80L6SD6     560
Name: productId, dtype: int64
```

```
avg_num_reviews3 = main.groupby('userId')['productId'].count()
len(avg_num_reviews3.nsmallest(keep='all'))
```

```
202745
```

There were also about 200K users who had only a single purchase for whom we can predict better using popularity matrix

Accuracy score

Similar Product

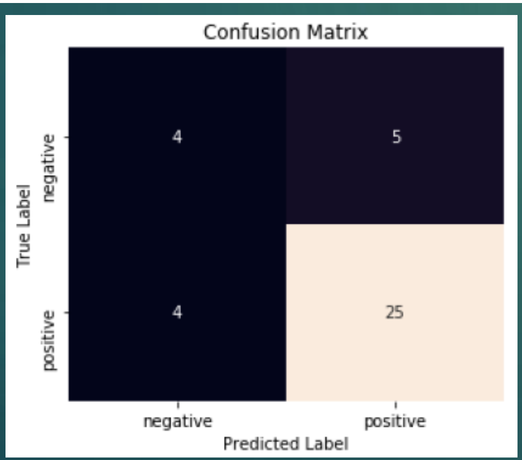
- Accuracy of prediction = ~76%

```
              precision    recall  f1-score   support

           3       0.50      0.44      0.47         9
           4       0.83      0.86      0.85        29

    accuracy                           0.76        38
   macro avg       0.67      0.65      0.66        38
weighted avg       0.75      0.76      0.76        38

Accuracy Score : 0.7631578947368421
```



Similar User

- Accuracy of prediction = ~50%

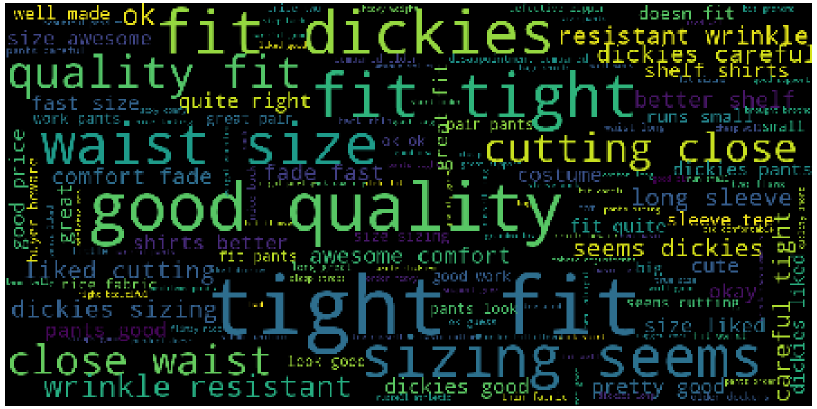
```
print(classification_report(df5_test_target, knnpreds_test))

Predicting review score for test dataset customers are :  [3 3 2 4]
              precision    recall  f1-score   support

           1       0.00      0.00      0.00         1
           2       0.00      0.00      0.00         0
           3       0.50      1.00      0.67         1
           4       1.00      1.00      1.00         1
           5       0.00      0.00      0.00         1

    accuracy                           0.50         4
   macro avg       0.30      0.40      0.33         4
weighted avg       0.38      0.50      0.42         4
```

For Score 1



For Score 3

For Score 5

Word Cloud Representation

Issues associated

Cold Start Problems

New Customers with no purchase history have no recommendations.

New Products with no ratings have no recommendations.

Solution : Recommending using the popularity matrix – Top N products

Sparsity - not every user has rated every other product gives a very sparse matrix.

Fraud Recommendation

Solution : Content based filtering tends to avoid this incorrect way of prediction .

Conclusion

- A concept model successfully implemented after conducting a research through literature review.
- Helped me understand the importance of Data cleaning
- Finding insights and exploring data statistics is as important as implementing a machine learning algorithm.
- Project helps in understanding recommender in both Online markets – Content & Collaborative and offline scenarios – Market basket

Thank you