AMAZON PRODUCT RECOMMENDATION SYSTEM USING CUSTOMER REVIEWS

Category – Grocery

Group Members :

Simran Mhaske – A20552202
Hrishikesh Pawar-A20543095
Harshal Sawant - A20538827
Ishan Prabhune - A20538828
Yashraj Diwate – A20553800
Pratik Patil – A20547062

Introduction

Our project focuses on analyzing Amazon customer reviews using a recommendation system. The motivation behind this project stems from the increasing reliance on online reviews for purchasing decisions. With the vast amount of reviews available on Amazon, it becomes challenging for users to sift through them to find relevant and trustworthy information.

We performed classification on sentiment_label column which was derived after sentiment analysis of text column. We used few columns which can be main factors which impact customer satisfaction at the most as a features to perform predict the sentiment label and perform classification.

We used content based filtering for building recommendation system.

Research Problem

In today's digital marketplace, we're tasked with creating a recommendation system to suggest similar products to users based on their reviews. Our focus is on delivering personalized recommendations while ensuring they meet a certain quality standard. Leveraging the vast Amazon Customer Reviews dataset, we'll preprocess the data and employ collaborative filtering techniques to predict products aligning with user preferences. Crucially, we'll filter out products with lower ratings to maintain recommendation quality and enhance user satisfaction. Our goal is to enrich the online shopping experience and foster long-term customer loyalty through tailored, high-quality recommendations.

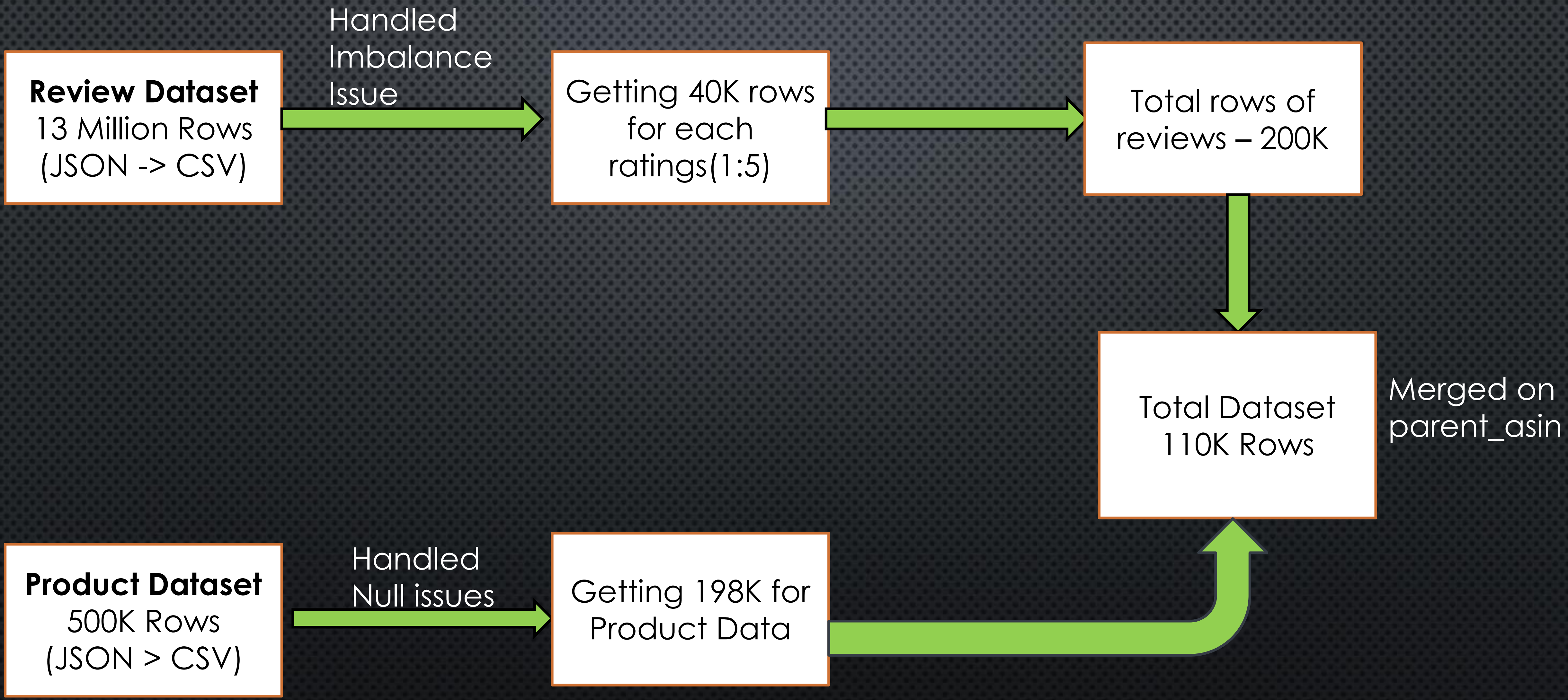
Dataset

Dataset: Obtained from Stanford Network Analysis Project (SNAP), comprising around 13 million records.

Features include review ratings, titles, textual content, images, ASINs, user IDs, timestamps, and helpful votes.

Data Source : https://snap.stanford.edu/data/web-Amazon.html
This is how our data looks:

Data Cleaning Flow



Review Dataset
13 Million Rows
(JSON -> CSV)

Handled Imbalance Issue

Getting 40K rows for each ratings(1:5)

Total rows of reviews – 200K

Total Dataset
110K Rows

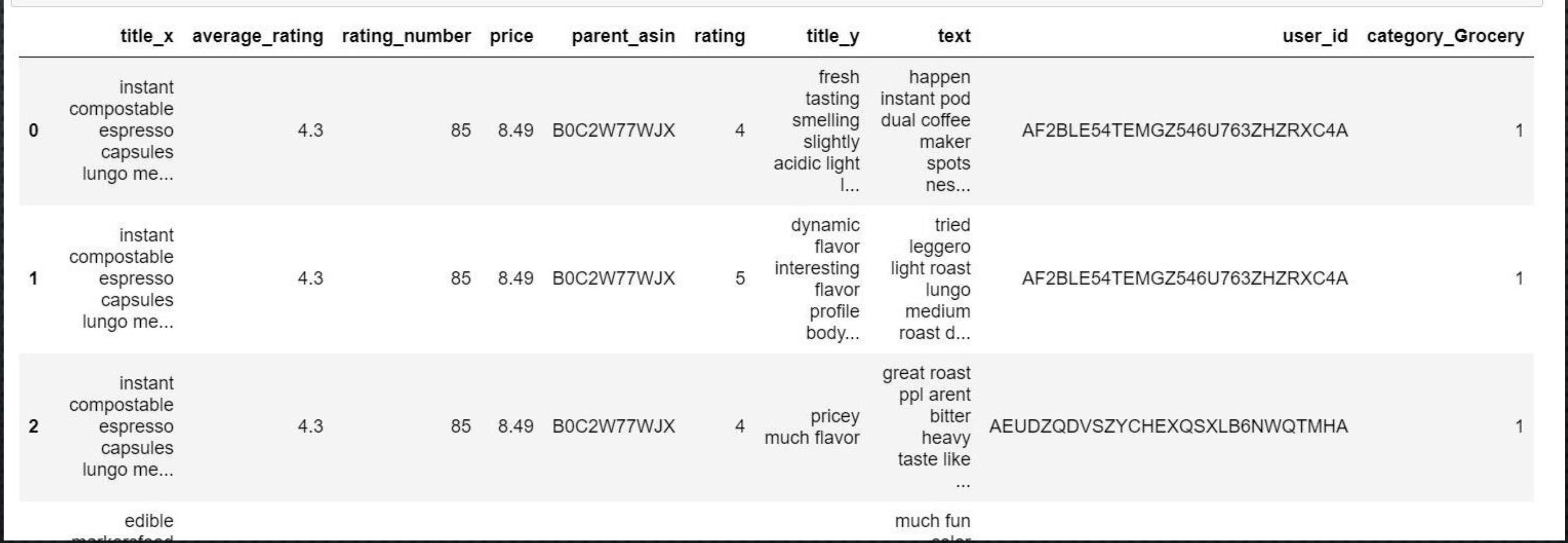
Merged on parent_asin

Product Dataset
500K Rows
(JSON > CSV)

Handled Null issues

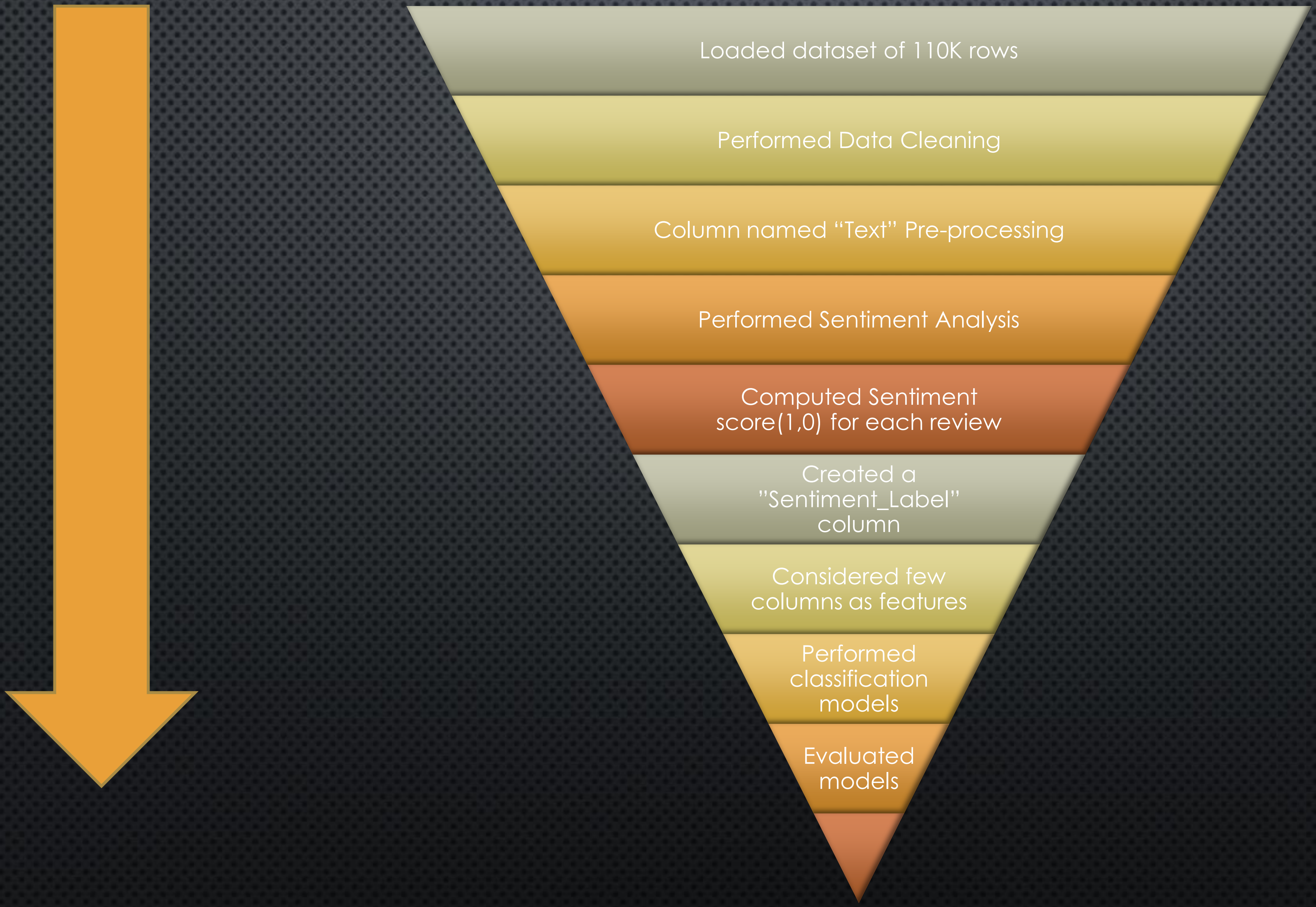
Getting 198K for Product Data

AFTER MERGING THIS IS HOW OUR DATA LOOKS LIKE



	title_x	average_rating	rating_number	price	parent_asin	rating	title_y	text	user_id	category_Grocery
0	instant compostable espresso capsules lungo me...	4.3	85	8.49	B0C2W77WJX	4	fresh tasting smelling slightly acidic light l...	happen instant pod dual coffee maker spots nes...	AF2BLE54TEMGZ546U763ZHZRXC4A	1
1	instant compostable espresso capsules lungo me...	4.3	85	8.49	B0C2W77WJX	5	dynamic flavor interesting flavor profile body...	tried leggero light roast lungo medium roast d...	AF2BLE54TEMGZ546U763ZHZRXC4A	1
2	instant compostable espresso capsules lungo me...	4.3	85	8.49	B0C2W77WJX	4	pricey much flavor	great roast ppl arent bitter heavy taste like ...	AEUDZQDVSZYCHEXQSXLB6NWQTMHA	1
	edible							much fun		

CLASSIFICATION MODEL BASED ON PRODUCT REVIEW SENTIMENT



Loaded dataset of 110K rows

Performed Data Cleaning

Column named “Text” Pre-processing

Performed Sentiment Analysis

Computed Sentiment score(1,0) for each review

Created a ”Sentiment_Label” column

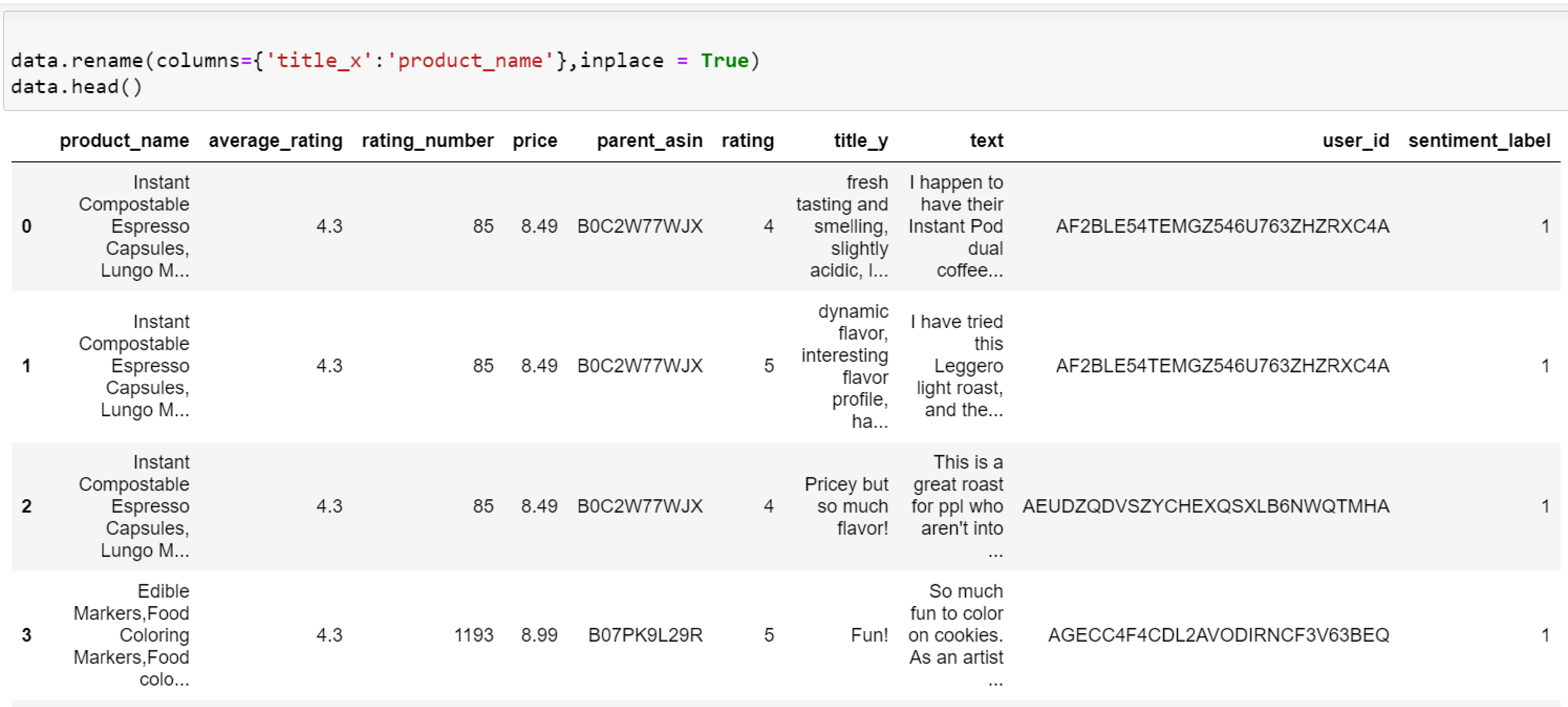
Considered few columns as features

Performed classification models

Evaluated models

OVERVIEW OF DATA AND DATA PREPROCESSING

- Dataset for Classification:
  You can see the '**sentiment_label**' created in the dataset after data pre processing and sentiment analysis of the '**text**' column.



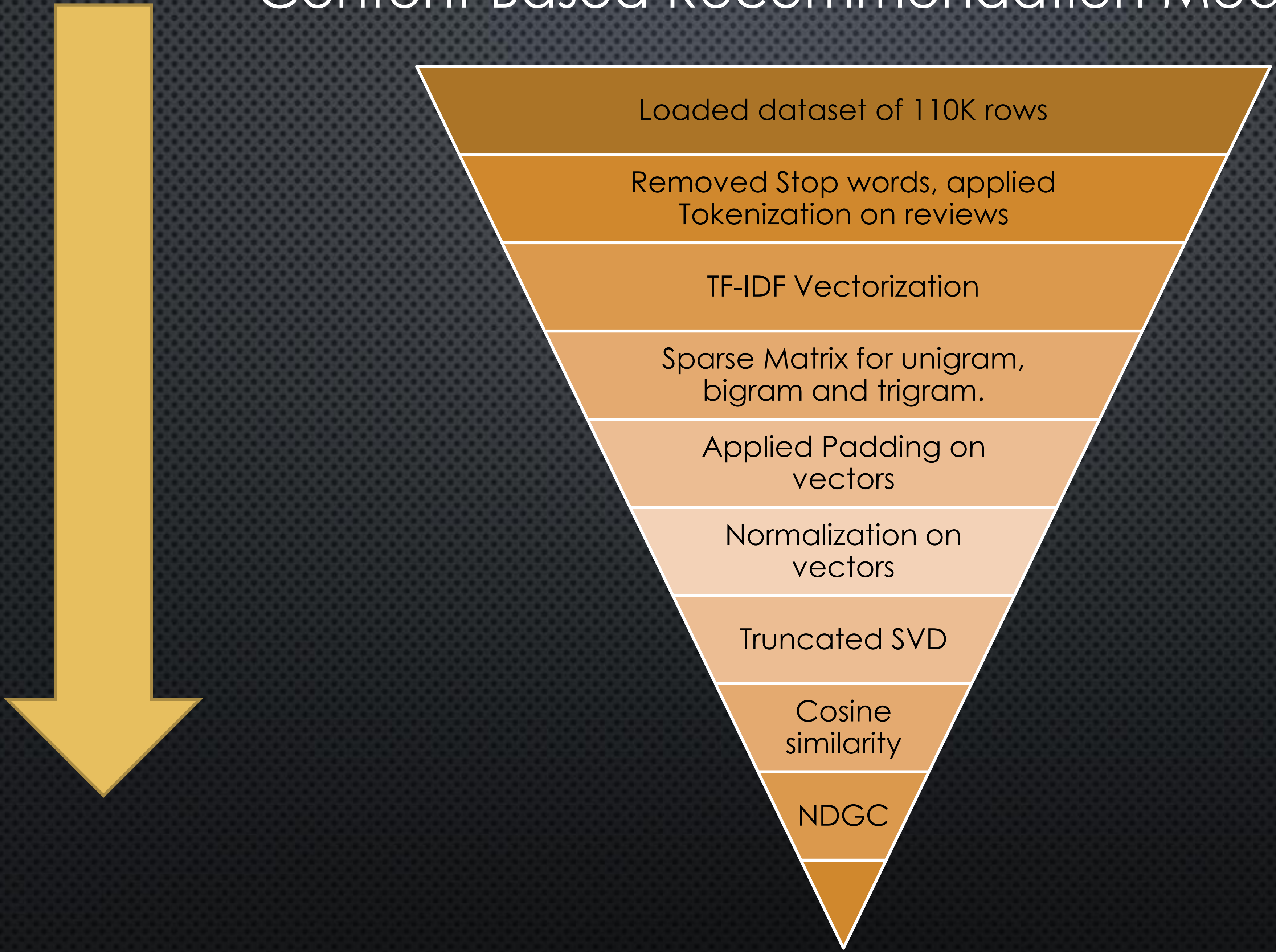
```
data.rename(columns={'title_x':'product_name'},inplace = True)
data.head()
```

	product_name	average_rating	rating_number	price	parent_asin	rating	title_y	text	user_id	sentiment_label
0	Instant Compostable Espresso Capsules, Lungo M...	4.3	85	8.49	B0C2W77WJX	4	fresh tasting and smelling, slightly acidic, l...	I happen to have their Instant Pod dual coffee...	AF2BLE54TEMGZ546U763ZHZRXC4A	1
1	Instant Compostable Espresso Capsules, Lungo M...	4.3	85	8.49	B0C2W77WJX	5	dynamic flavor, interesting flavor profile, ha...	I have tried this Leggero light roast, and the...	AF2BLE54TEMGZ546U763ZHZRXC4A	1
2	Instant Compostable Espresso Capsules, Lungo M...	4.3	85	8.49	B0C2W77WJX	4	Pricey but so much flavor!	This is a great roast for ppl who aren't into ...	AEUDZQDVSZYCHEXQSXLB6NWQTMHA	1
3	Edible Markers,Food Coloring Markers,Food colo...	4.3	1193	8.99	B07PK9L29R	5	Fun!	So much fun to color on cookies. As an artist ...	AGECC4F4CDL2AVODIRNCF3V63BEQ	1

CLASSIFICATION MODEL RESULTS

Model	Accuracy	Recall		F1		Precision	
Sentiment_Label		0	1	0	1	0	1
Decision Tree	0.74	0.32	0.84	0.32	0.84	0.32	0.84
Knearest Neighbour (KNN)	0.78	0.16	0.93	0.22	0.88	0.36	0.82
Random Forest	0.77	0.22	0.91	0.27	0.87	0.36	0.83
Logistic Regression	0.81	0	1	0	0.89	0	0.81
SVM	0.808	0	1	0	0.89	0	0.81
GradientBoosting	0.809	0	1	0	0.89	0.62	0.81
Ada Boost	0.808	0	1	0.01	0.89	0.55	0.81
XG Boost	0.804	0.06	0.98	0.1	0.89	0.41	0.81
Multinomial Naive Bayes	0.71	0.18	0.85	0.19	0.83	0.21	0.81
Categorical Naive Bayes	0.79	0.18	0.94	0.24	0.88	0.4	0.83

Content Based Recommendation Model Flow



Loaded dataset of 110K rows

Removed Stop words, applied Tokenization on reviews

TF-IDF Vectorization

Sparse Matrix for unigram, bigram and trigram.

Applied Padding on vectors

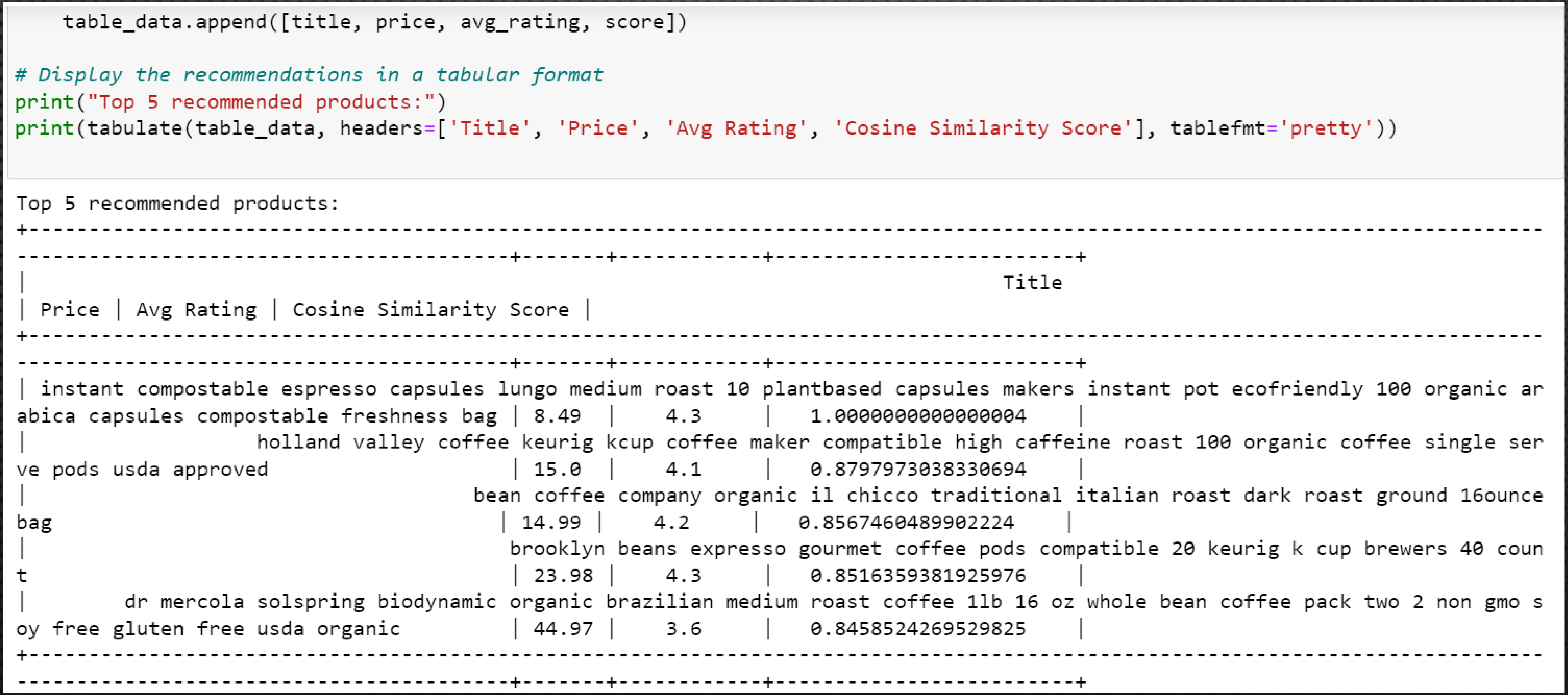
Normalization on vectors

Truncated SVD

Cosine similarity

NDGC

RECOMMENDATION OF PRODUCT



```python
    table_data.append([title, price, avg_rating, score])

# Display the recommendations in a tabular format
print("Top 5 recommended products:")
print(tabulate(table_data, headers=['Title', 'Price', 'Avg Rating', 'Cosine Similarity Score'], tablefmt='pretty'))
```

```
Top 5 recommended products:
+------------------------------------------------------------------------------------------------------------------------------
-----------------------------------------+-------+------------+-------------------------+
|                                                                                 Title
| Price | Avg Rating | Cosine Similarity Score |
+------------------------------------------------------------------------------------------------------------------------------
-----------------------------------------+-------+------------+-------------------------+
| instant compostable espresso capsules lungo medium roast 10 plantbased capsules makers instant pot ecofriendly 100 organic ar
abica capsules compostable freshness bag | 8.49  |    4.3     |   1.0000000000000004    |
|                   holland valley coffee keurig kcup coffee maker compatible high caffeine roast 100 organic coffee single ser
ve pods usda approved                    | 15.0  |    4.1     |   0.8797973038330694    |
|                                     bean coffee company organic il chicco traditional italian roast dark roast ground 16ounce
bag                                     | 14.99 |    4.2     |   0.8567460489902224    |
|                                        brooklyn beans expresso gourmet coffee pods compatible 20 keurig k cup brewers 40 coun
t                                        | 23.98 |    4.3     |   0.8516359381925976    |
|        dr mercola solspring biodynamic organic brazilian medium roast coffee 1lb 16 oz whole bean coffee pack two 2 non gmo s
oy free gluten free usda organic         | 44.97 |    3.6     |   0.8458524269529825    |
+------------------------------------------------------------------------------------------------------------------------------
-----------------------------------------+-------+------------+-------------------------+
```

RECOMMENDATION MODEL EVALUATION

Model	NDGC Score
Cosine Similarity – Unigram	0.98
Cosine Similarity – Bigram	0.99
Cosine Similarity – Trigram	0.97
Cosine Similarity – Genism	0.98

LOGICAL EXPLAINATION - 1

WHAT ITEM ID HAVE YOU TAKEN :

- WE ARE USING TITLE_X AS MY ITEM ID TO PASS INTO THE MODEL TO GET RECOMMENDATION. THIS IS BECAUSE THE TITLE_X WHICH IS PRODUCT TITLE CONTAINS STRING WITH BRIEF INFORMATION AND KEYWORDS SUCH AS 'INSTANT', 'GOOD' ETC. AND HENCE COMPARING THESE WITH THE REVIEWS OF THE PRODUCTS WOULD GIVE BEST RECOMMENDATIONS ON THEM.
- COMPARING THE SIMILARITIES OF THE PRODUCT WITH INDEX PRODUCT_INDEX (REFERRED TO BY ITS TITLE IN TITLE_X) WITH ALL OTHER PRODUCTS BASED ON THEIR REVIEWS.
- WE CREATED A GET_RECOMMENDATION FUNCTION WHICH ITERATES OVER CHUNKS OF THE TF-IDF MATRIX TO COMPUTE COSINE SIMILARITY SCORES BETWEEN THE TF-IDF VECTOR OF THE SPECIFIED PRODUCT (PRODUCT_INDEX) WHICH WOULD RETRIEVE TITLE OF THE PRODUCT AND THE TF-IDF VECTORS OF ALL OTHER PRODUCTS.
- THE FUNCTION RETURNS A LIST OF TUPLES CONTAINING THE INDICES OF RECOMMENDED PRODUCTS ALONG WITH THEIR COSINE SIMILARITY SCORES RELATIVE TO THE SPECIFIED PRODUCT.
- THIS WAS OUR UNDERSTANDING FROM VARIOUS ONLINE SOURCES BASED ON WHICH WE BUILD THE RECOMMENDATION MODEL.

LOGICAL EXPLAINATION - 2

HOW ARE YOU TRYING TO RECOMMEND BASED ON PRODUCT TITLE AND REVIEW TEXT –

- During classification we have different columns to create a label column. But as our problem statement says, we are going to recommend the products based on the reviews posted by customer on provided product name.
- So for this reason, we created sentiment_label column as our label variable based on sentiment analysis from the review's text which is in 'Text' column.
- All other columns such as average_rating, rating_number, rating and price are used as features to predict this sentiment_label column in our different classifications model.
- On which later on, we have evaluated the different models using different evaluation metrics.

LOGICAL EXPLAINATION - 3

- NDCG score explanation -NDCG is a ranking metric that evaluates the quality of the ranked list of recommendations.
- It considers both the relevance of recommended items and their position in the list.Cosine similarity measures the similarity between items based on their feature vectors (e.g., TF-IDF vectors of reviews).
- NDCG can be used to evaluate the quality of recommendations generated using cosine similarity if the relevance of recommended items can be assessed in a graded manner.

THANK YOU !