Kaggle Projects

Quora Insincere Questions Classification

H&M Personalized Fashion Recommendation

Outline

Quora Insincere Questions Classification

- Introduction
- Data
- Exploration and Visualization
- Preprocessing
- Modelling
- Evaluation and Results
- Additional Work

H&M Personalized Fashion Recommendations

- Introduction
- Data
- Exploration and Visualization
- Preprocessing
- Calculating Similarity
- Results
- Future Work

Quora Insincere Questions Classification

Introduction

What is Quora?

Quora is a platform that empowers people to learn from each other

On Quora, people can ask questions and connect with others who contribute unique insights and quality answers

Problem Statement

A key challenge is to weed out insincere questions -- those founded upon false premises, or that intend to make a statement rather than look for helpful answers

The aim of this project is to develop a model that identify and flag insincere questions

What are Insincere Questions?

Insincere question is defined as a question intended to make a statement rather than look for helpful answers

Characteristics of an Insincere Question

- Has a non-neutral tone
- Is disparaging or inflammatory
- Isn't grounded in reality
- Uses sexual content for shock value

Data

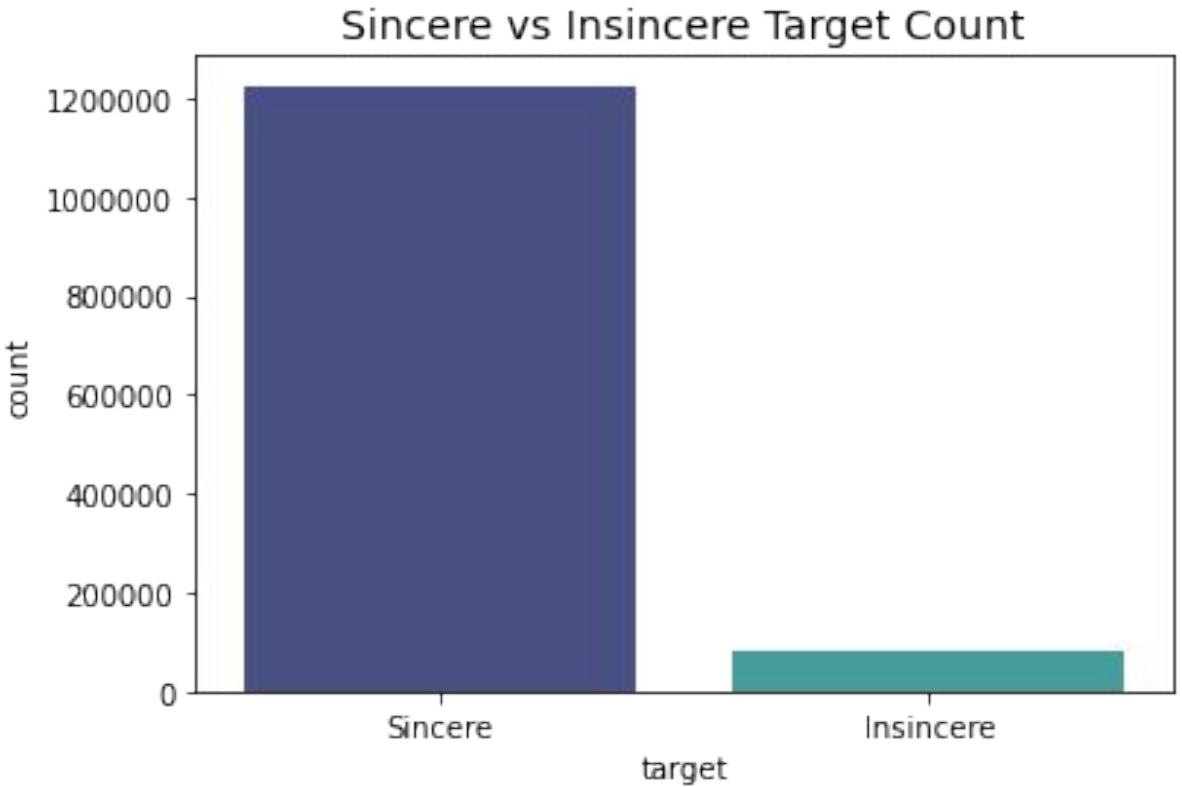
Data

	qid	question_text	target
0	00002165364db923c7e6	How did Quebec nationalists see their province...	0
1	000032939017120e6e44	Do you have an adopted dog, how would you enco...	0
2	0000412ca6e4628ce2cf	Why does velocity affect time? Does velocity a...	0
3	000042bf85aa498cd78e	How did Otto von Guericke used the Magdeburg h...	0
4	0000455dfa3e01eae3af	Can I convert montra helicon D to a mountain b...	0
5	00004f9a462a357c33be	Is Gaza slowly becoming Auschwitz, Dachau or T...	0
6	00005059a06ee19e11ad	Why does Quora automatically ban conservative ...	0
7	0000559f875832745e2e	Is it crazy if I wash or wipe my groceries off...	0
8	00005bd3426b2d0c8305	Is there such a thing as dressing moderately, ...	0
9	00006e6928c5df60eacb	Is it just me or have you ever been in this ph...	0

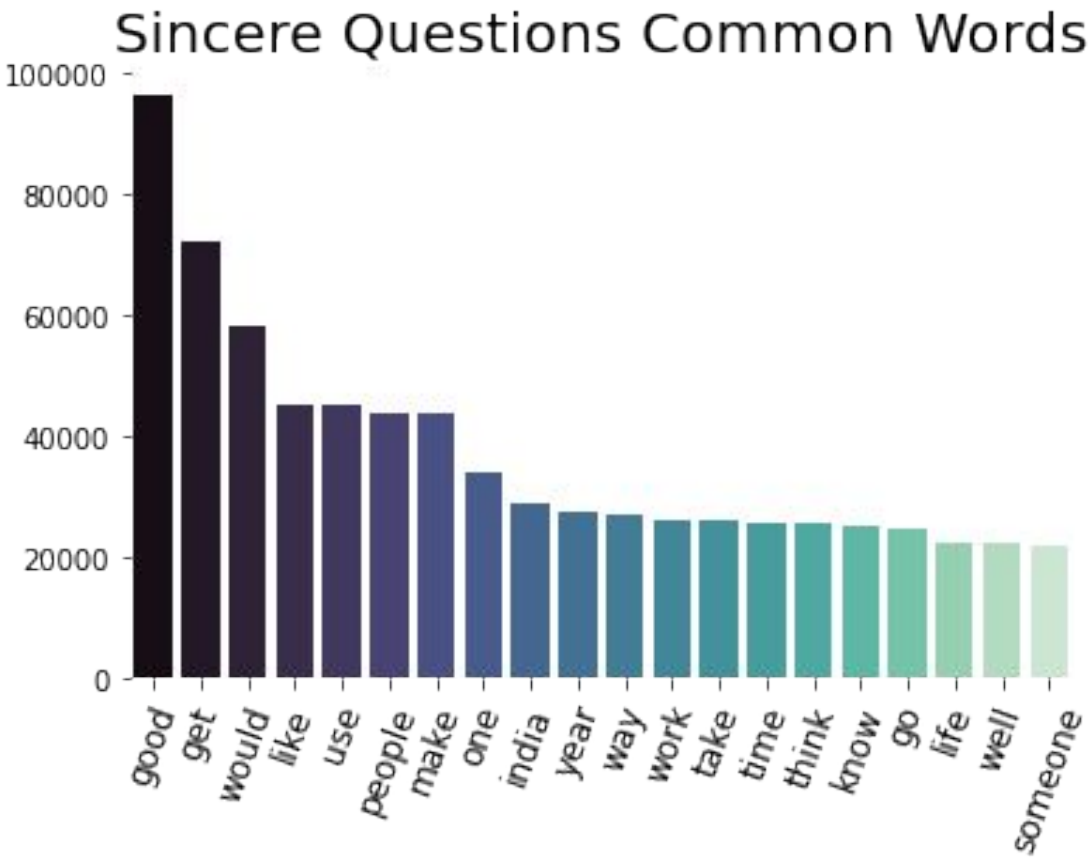
The training file contains the following data fields:

- qid - unique question identifier
- question_text - Quora question text
- target - a question labeled "insincere" has a value of 1, otherwise 0

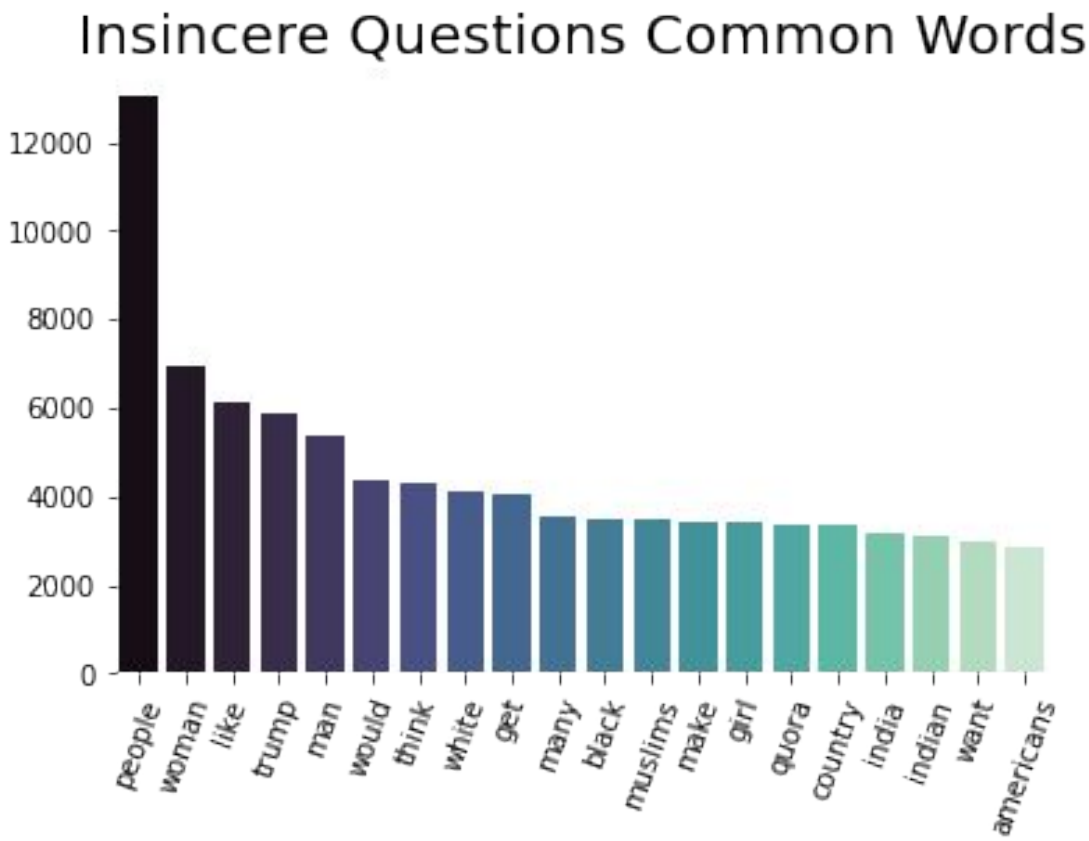
Exploration and Visualization



While exploring the data, I found out the questions distribution was highly imbalanced!



Top occurring keywords in Sincere Questions



Top occurring keywords in Insincere Questions

Preprocessing

Preprocessing Steps

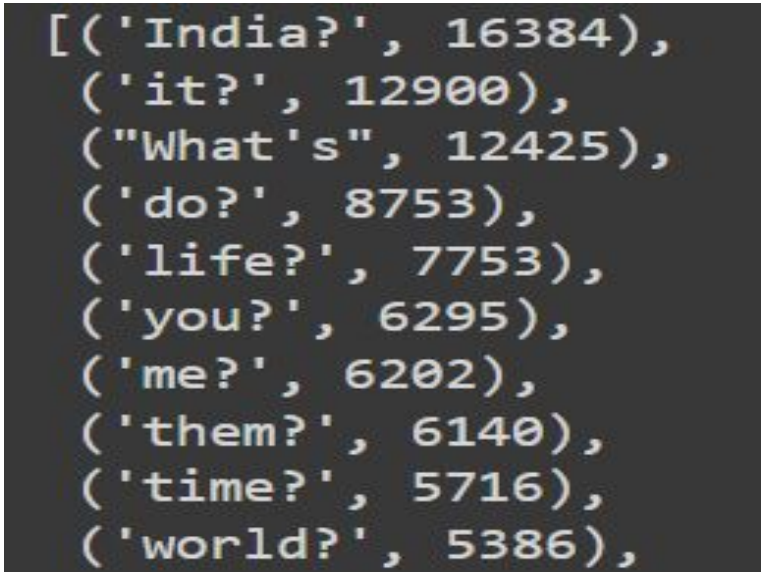
Initially embeddings existed for only 33.16% of the vocabulary, after preprocessing it increased to 74.24%

- Loaded pre-trained **Glove embeddings Common Crawl** (840B tokens, 2.2M vocab, cased, 300d vectors)
- **GloVe** is an unsupervised learning algorithm for obtaining vector representations for words
- Checked the vocab which exists in the embeddings

Preprocessing Steps

Incorporated vocabulary not existing in embedding using different pre-processing techniques like

- Removing punctuations,
- Removing non-english words
- Replacing multiple empty spaces with a single empty space



Words not present in the embeddings

Preprocessing Steps

Created a misspelled dictionary for incorporating words like

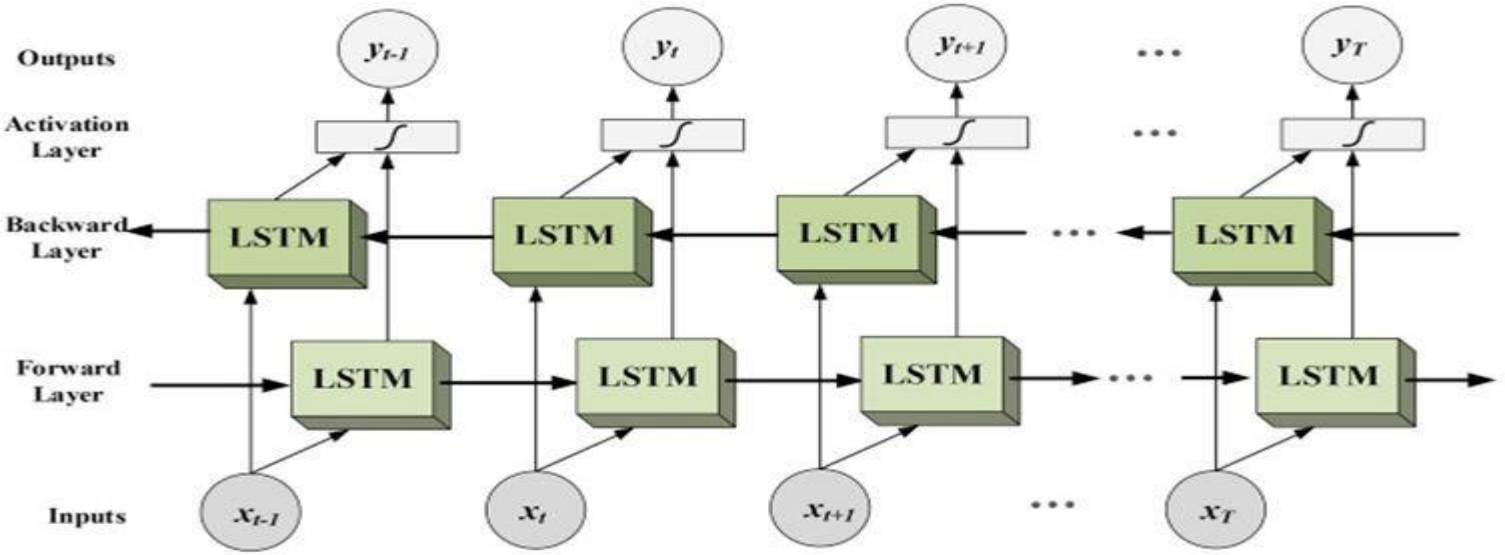
- Joined words
- Acronyms
- Frequently misspelled words

```
mispell_dict = {'Quorans':'Quora users',
                'Brexit':'Britan exit',
                'Quoras':'quora',
                '201718':'2017-2018',
                'Jongun':'Jong-un',
                'nonMuslim':'non-muslim',
                'GDPR':'General Data Protection Regulation',
                'demonitisation':'demonetization',
                'SJWs':'social justice warriors',
                'BNBR':'Be Nice Be Respectful'}
```

Misspelled Dictionary

Bi-directional LSTM Model

Bi-Directional LSTM Model



Hyperparameter Optimization

For hyperparameter optimization, I used **ReduceLROnPlateau**, so it reduces the learning rate when a metric has stopped improving.

Evaluation and Results

Evaluation Metric

Imbalanced classes are a common problem in machine learning classification where there are a disproportionate ratio of observations in each class

I used the **F1 score which balances precision and recall**.

- Precision is the number of true positives divided by all positive predictions. Low precision indicates a high number of false positives.
- Recall is the number of true positives divided by the number of positive values in the test data.. Low recall indicates a high number of false negatives.

Results

I got an **F-1 score of 0.6705** with the Bi-directional LSTM model.

Additional Work

What other approaches did I try?

Other than the results I tried to implement BERT through Tensorflow and Pytorch both

Questions?

H&M Personalized Fashion Recommendations

Introduction

Introduction

H&M Group is a family of brands and businesses with 53 online markets and approximately 4,850 stores.

But with too many choices, customers might not quickly find what interests them or what they are looking for, and ultimately, they might not make a purchase.

Aim

To enhance the shopping experience, product recommendations are key

The aim of this project is to develop product recommendations based on data from previous transactions, as well as from customer and product meta data.

Data

Data

- articles.csv - detailed metadata for each article_id available for purchase
- customers.csv - metadata for each customer_id in dataset
- transactions_train.csv - the training data, consisting of the purchases each customer for each date, as well as additional information.

Articles File

```
[5]:
```

	0	1	2	3	4
article_id	108775015	108775044	108775051	110065001	110065002
product_code	108775	108775	108775	110065	110065
prod_name	Strap top	Strap top	Strap top (1)	OP T-shirt (Idro)	OP T-shirt (Idro)
product_type_no	253	253	253	306	306
product_type_name	Vest top	Vest top	Vest top	Bra	Bra
product_group_name	Garment Upper body	Garment Upper body	Garment Upper body	Underwear	Underwear
graphical_appearance_no	1010016	1010016	1010017	1010016	1010016
graphical_appearance_name	Solid	Solid	Stripe	Solid	Solid
colour_group_code	9	10	11	9	10
colour_group_name	Black	White	Off White	Black	White
perceived_colour_value_id	4	3	1	4	3
perceived_colour_value_name	Dark	Light	Dusty Light	Dark	Light
perceived_colour_master_id	5	9	9	5	9
perceived_colour_master_name	Black	White	White	Black	White
department_no	1676	1676	1676	1339	1339
department_name	Jersey Basic	Jersey Basic	Jersey Basic	Clean Lingerie	Clean Lingerie
index_code	A	A	A	B	B
index_name	Ladieswear	Ladieswear	Ladieswear	Lingeries/Tights	Lingeries/Tights
index_group_no	1	1	1	1	1
index_group_name	Ladieswear	Ladieswear	Ladieswear	Ladieswear	Ladieswear
section_no	16	16	16	61	61
section_name	Womens Everyday Basics	Womens Everyday Basics	Womens Everyday Basics	Womens Lingerie	Womens Lingerie
garment_group_no	1002	1002	1002	1017	1017
garment_group_name	Jersey Basic	Jersey Basic	Jersey Basic	Under-, Nightwear	Under-, Nightwear
detail_desc	Jersey top with narrow shoulder straps.	Jersey top with narrow shoulder straps.	Jersey top with narrow shoulder straps.	Microfibre T-shirt bra with underwired, moulde...	Microfibre T-shirt bra with underwired, moulde...

Customers File

:

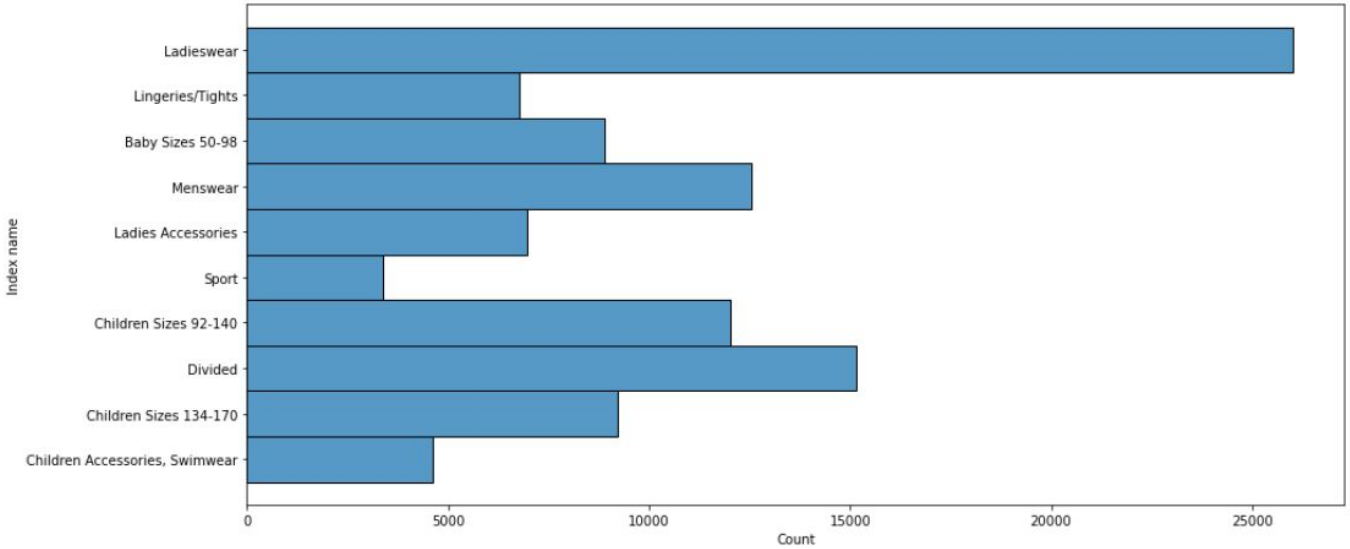
	0	1
customer_id	00000dbacae5abe5e23885899a1fa44253a17956c6d1c3...	0000423b00ade91418cceaf3b26c6af3dd342b51fd051e...
FN	NaN	NaN
Active	NaN	NaN
club_member_status	ACTIVE	ACTIVE
fashion_news_frequency	NONE	NONE
age	49.0	25.0
postal_code	52043ee2162cf5aa7ee79974281641c6f11a68d276429a...	2973abc54daa8a5f8ccfe9362140c63247c5eee03f1d93...

Transactions File

	t_dat	customer_id	article_id	price	sales_channel_id
0	2018-09-20	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	663713001	0.050831	2
1	2018-09-20	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	541518023	0.030492	2
2	2018-09-20	00007d2de826758b65a93dd24ce629ed66842531df6699...	505221004	0.015237	2
3	2018-09-20	00007d2de826758b65a93dd24ce629ed66842531df6699...	685687003	0.016932	2
4	2018-09-20	00007d2de826758b65a93dd24ce629ed66842531df6699...	685687004	0.016932	2
5	2018-09-20	00007d2de826758b65a93dd24ce629ed66842531df6699...	685687001	0.016932	2
6	2018-09-20	00007d2de826758b65a93dd24ce629ed66842531df6699...	505221001	0.020322	2
7	2018-09-20	00083cda041544b2fbb0e0d2905ad17da7cf1007526fb4...	688873012	0.030492	1
8	2018-09-20	00083cda041544b2fbb0e0d2905ad17da7cf1007526fb4...	501323011	0.053373	1
9	2018-09-20	00083cda041544b2fbb0e0d2905ad17da7cf1007526fb4...	598859003	0.045746	2

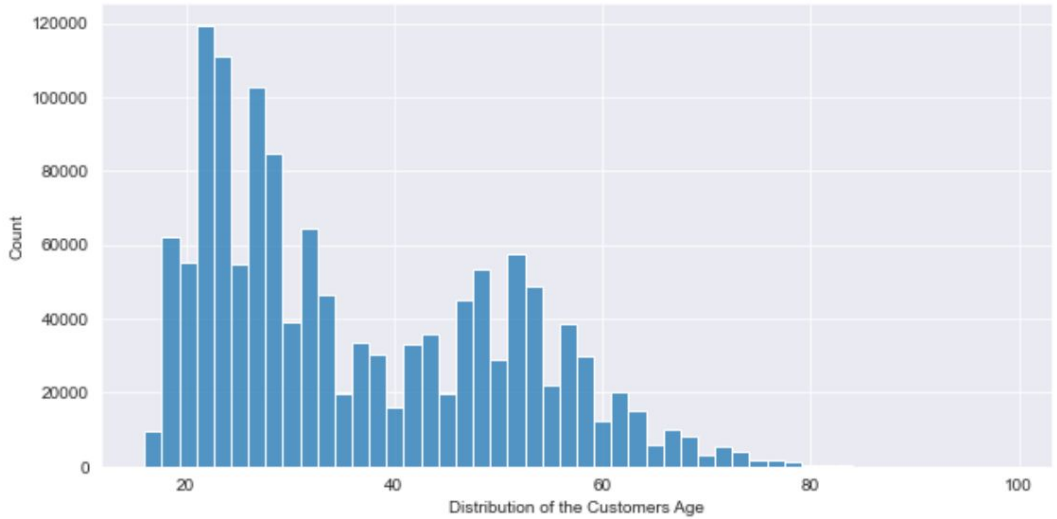
Exploration and Visualization

Articles Index name Count



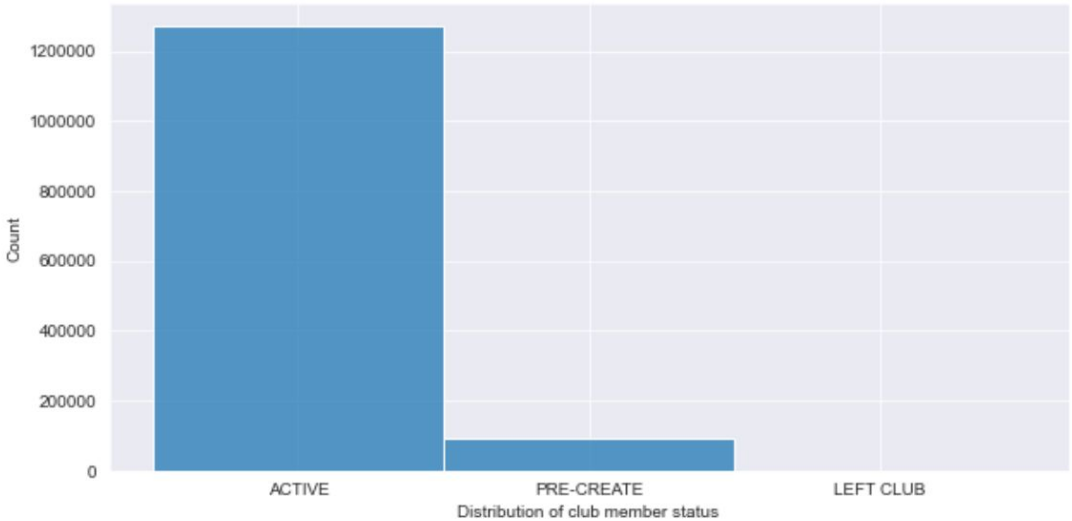
Articles with the Index Name Ladieswear have the highest count

Distribution of Age of Customers



The number of customers in the Age Range 25-35 are very high

Club Member Status Count



There are over one million Active Customers

Preprocessing

Preprocessing Steps

Combined all the text data of articles including description into a new column called text

Preprocessed the text column

- Removing stopwords
- Removing punctuations
- Performed Lemmatization

Converted preprocessed words into vector form using TF-IDF

Preprocessing Steps

- Merged the article table with the transactions table
- Grouped the customer’s id with the text details of the articles they have bought

Calculating Similarity

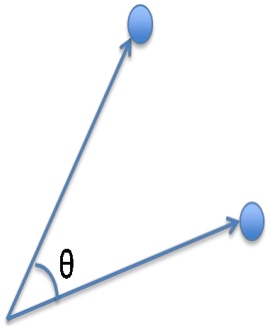
Cosine Similarity

Cosine Similarity measures the similarity between two vectors of an inner product space

Calculated cosine similarity between the User bought articles and the Article TF-IDF

Based on the scores derived from cosine similarity, generated the recommendations.

$$sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$



Results

Transactions for one particular Customer ID

37]:

	t_dat	customer_id	article_id	price	sales_channel_id
24773518	2020-04-18	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	858883002	0.030492	2
24773520	2020-04-18	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	750424014	0.042356	2
24773521	2020-04-18	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	750424014	0.042356	2
24773522	2020-04-18	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	870304002	0.033881	2
24773523	2020-04-18	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	870304002	0.033881	2
24773524	2020-04-18	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	852643001	0.025407	2
24773525	2020-04-18	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	852643003	0.025407	2
21953952	2020-02-03	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	351484002	0.022017	2
21953951	2020-02-03	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	351484002	0.022017	2
24773519	2020-04-18	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	851400006	0.059305	2
21953950	2020-02-03	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	723529001	0.025407	2
0	2018-09-20	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	663713001	0.050831	2
6827145	2019-03-01	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	578020002	0.013542	2
31521960	2020-09-15	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	794321007	0.061000	2
1	2018-09-20	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	541518023	0.030492	2
23934158	2020-04-01	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	727808007	0.067780	2
23934157	2020-04-01	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	727808001	0.067780	2
165807	2018-09-24	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	663713001	0.050831	2

Results

]:

	article_id	detail_desc	score
0	821648004	Quilted top in sturdy sweatshirt fabric with a...	0.557127
1	821648003	Quilted top in sturdy sweatshirt fabric with a...	0.544008
2	662879008	Fancy dress cape in jersey with a concealed ho...	0.536211
3	721991003	Long-sleeved top in cotton jersey with a print...	0.535286
4	721991004	Long-sleeved top in cotton jersey with a print...	0.523437
5	458428031	5-pocket jeans in washed stretch denim with a ...	0.516344
6	458428037	5-pocket jeans in washed stretch denim with a ...	0.516344
7	865034001	Short, wide dress in airy, patterned chiffon w...	0.506835
8	589832001	Short, fitted off-the-shoulder dress in stretc...	0.50571
9	756859003	Romper suit in soft cotton jersey with a print...	0.505697
10	721991006	Long-sleeved top in cotton jersey with a print...	0.505181
11	865030001	Fitted top in ribbed cotton jersey with a roun...	0.503169
12	865033003	5-pocket slim-fit trousers in washed, stretch ...	0.503169

Top 12 Recommendations for one customer

Future Work

Future Work

Use BERT Embeddings and calculate Cosine Similarity based on it

Explore Collaborative Filtering and Hybrid Techniques to generate better recommendations

Questions?

Thank You!