Recommendation Engine - Web App

Ebru Zümrüttaş

Project Definition:

The main goal of this project is creating **Recommendation engine** for e-commerce website which aims to *reach out needs of users while creating a new way of demand in the perspective of user.*

The objectives are with the scope of the data on hand, understanding the insights of user and products relations ,predicting what user demands and ***assuring user met with these products***.

Scope Of The Project:

Processes managed by using a specific data frame which is received from **Trenyol** E-Commerce Company.

Prototype for Web Application is created through the created recommendation engine with this data only.

Data covers the orders by users and the features of the products in this order with a specific timeline.

Time Constraint for that project is two weeks.

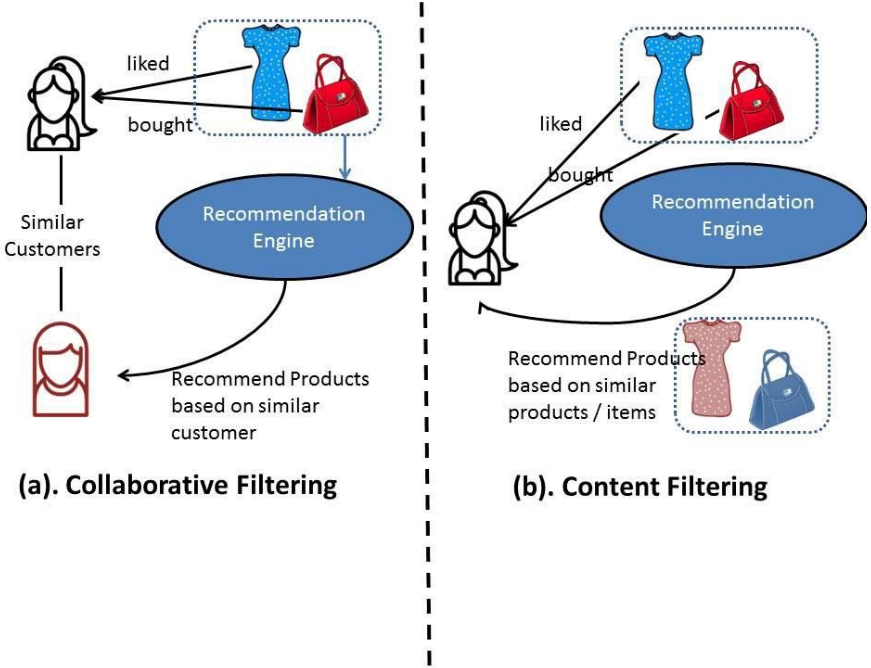
What Is a Recommendation Engine

A recommendation engine is a system that suggests products, services, information to users based on analysis of data.

Collaborative Filtering essentially means that similar users like similar things.

Content-based Filtering essentially considers items/users features.

Multiple methods are used in this project as the need of the concept. Each step will be examine and the results will be shared.

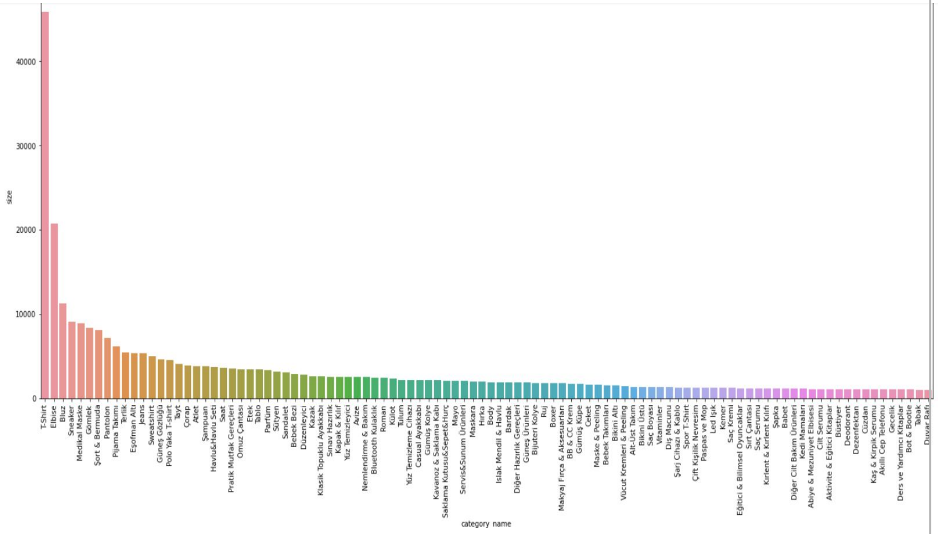


Insight of Data

- **partition_date**: The date user bought the product.
- **orderparentid**: The order id that all the product user bought in one order.
- **user_id**: Id of the user.
- **productcontentid**: Id of the product.
- **brand_id**: Id of the brand that products have.
- **category_id** : Id of the category that product included.
- **category_name**: Name of the category that product included.
- **gender**: Gender assigned to the product.
- **price**: Price of product.
- **color_id**: Id of the color that products have.
- **business_unit**: Business unit that products have
- **ImageLink**: URL of the photograph of the product.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 508228 entries, 0 to 508227
Data columns (total 12 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   partition_date    508228 non-null  object
 1   orderparentid     508228 non-null  int64
 2   user_id           508228 non-null  int64
 3   productcontentid  508228 non-null  int64
 4   brand_id          508228 non-null  int64
 5   category_id       508228 non-null  int64
 6   category_name     508228 non-null  object
 7   gender            475493 non-null  object
 8   price             508228 non-null  float64
 9   color_id          375670 non-null  float64
 10  business_unit     508228 non-null  object
 11  ImageLink         508228 non-null  object
dtypes: float64(2), int64(5), object(5)
memory usage: 46.5+ MB
```

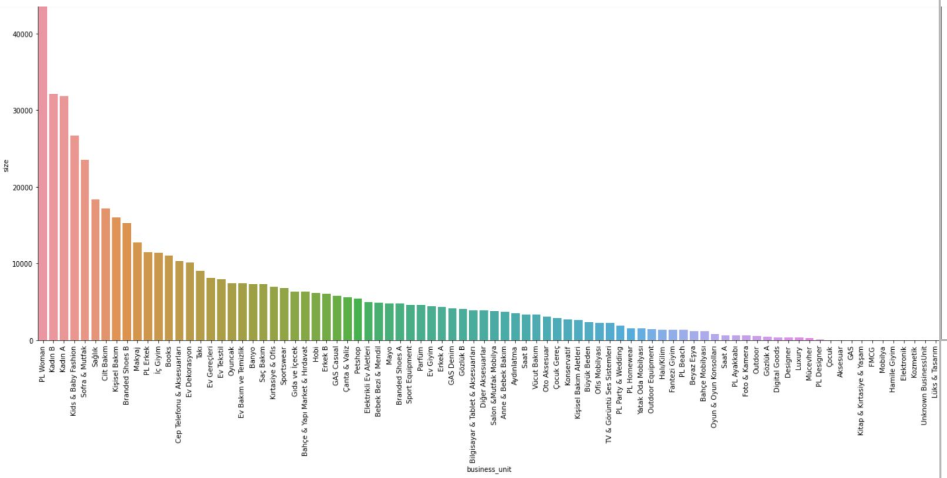
Insight of Data - Category



In statistics and business, a **long tail of distributions** of numbers is the portion of the distribution having many occurrences far from the "head" or central part of the distribution.

There are only 100 category name displayed in this graph.Therefore the most of the sales comes from non-popular goods.This is an important part while creating a recommendation engine. **In user based recommendation engine category variety must be increased.**

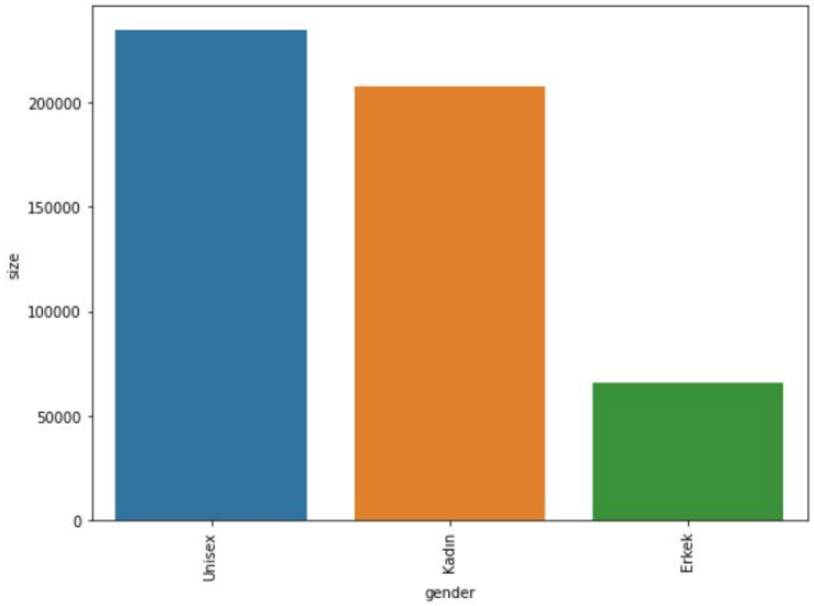
Insight of Data - Business Unit



Long Tail statistical pattern can be shown Previous slide.The hits are in the head. There are only 100 business_unit is shown in the graph. ***Therefore the variety of business unit must be increase in the recommendations that showed to the user.***

Insight of Data - Gender

In examination of bussiness_unit of two gender type Unknown and Unisex are mostly similar. **Due to that fact Unknown gender turn to Unisex.**

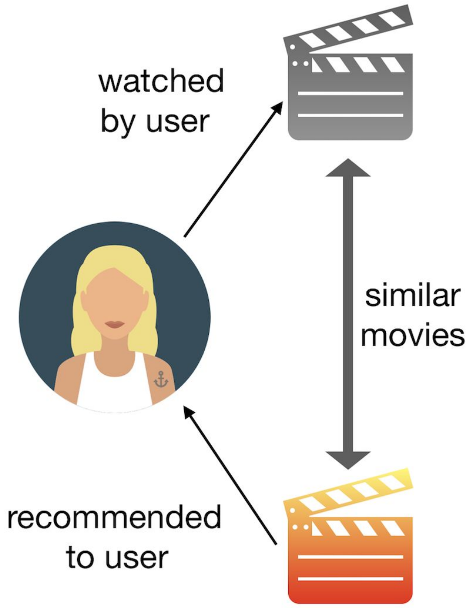


Content Based Recommendation Engine

Content-based filtering is based on a single user's interactions and preference. **Recommendations are based on the metadata collected from a user's history and interactions**.

For example, recommendations will be based on looking at established patterns in a user's choice or behaviours.

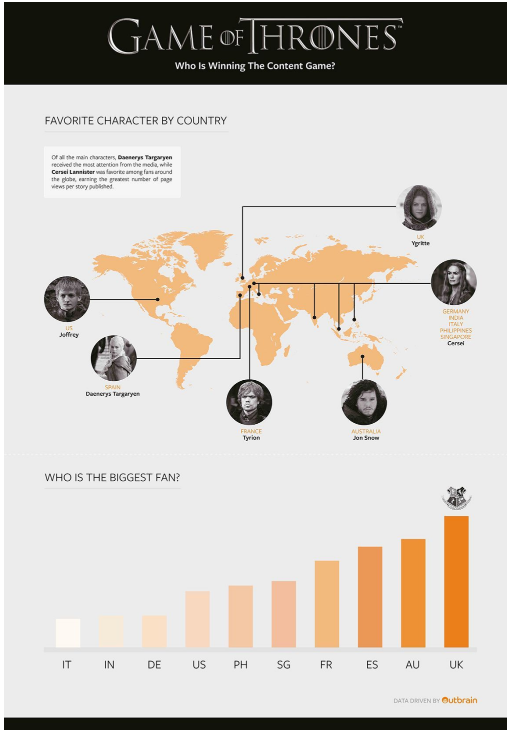
Returning information such as products or services will relate to your likes or views. With an approach like this, the more information that the user provides, the higher the accuracy.



Cold Start

Cold start occurs when a recommender system cannot draw inferences for a query due to lack of sufficient information. A particular form of the content-based recommendation system is a case-based recommender. **These evaluate items' similarities and have been extensively deployed in e-commerce.**

Determining Product Popularities

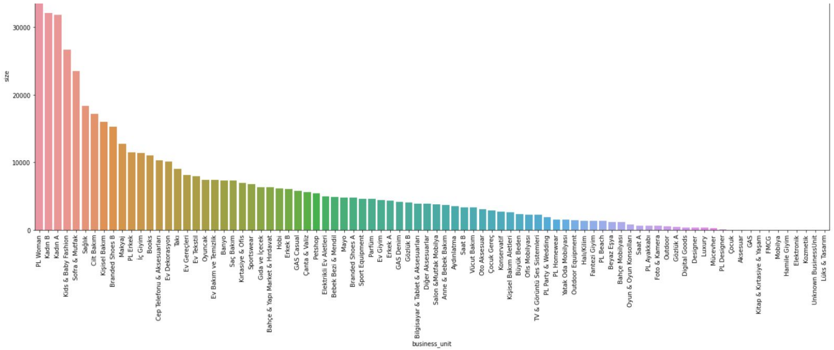


1. **Determining Weight For Gender** :Basically all gender of products size divided by all products size individually.The weights are included to main data frame.

2. **Determining Weight For Category:** In this section all the business unit weights are calculated by their weight on all the orders.

3. **Determining Weight For Business Unit:** In this section all the category weights are calculated by their weight on all the orders.

Construct Main Weight For Products



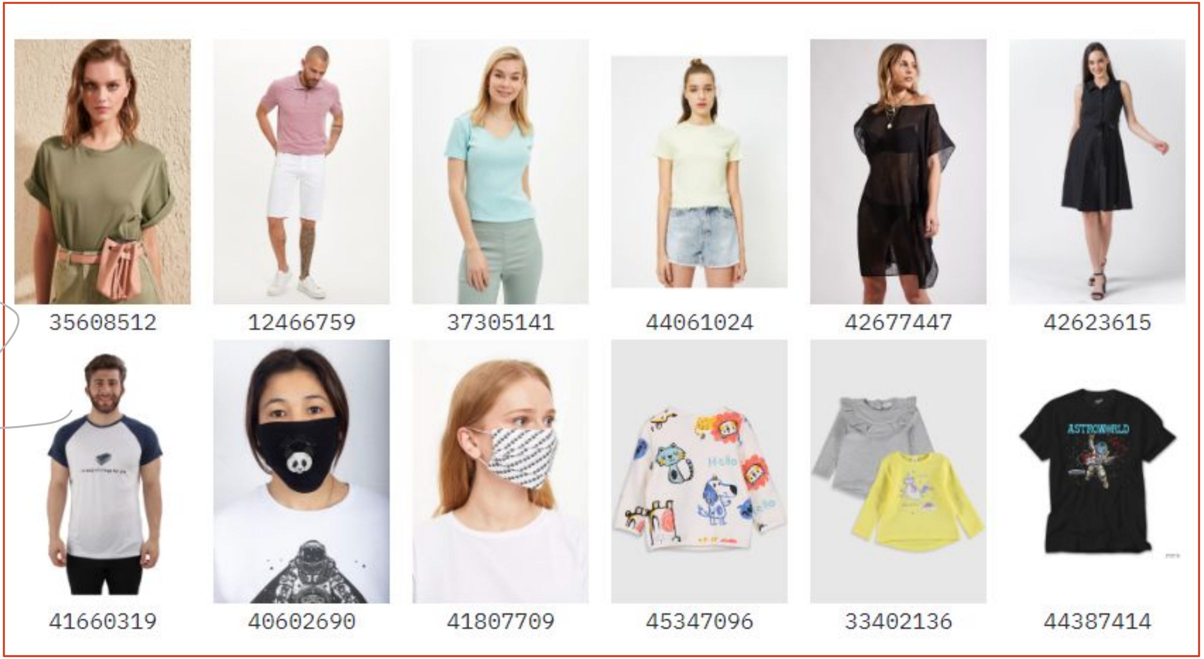
The weights are manipulated by the importance of weight as to reach desired recommendation engine.

Due to make variety in the range of all the parts that we worked on data grouped by within a strategy. ***The main purpose is to show a user who has no behavioural data in the database, to show variety of products but in the range of popular items.***

All the weights are explicitly gained from users not in the sense of user ranking,but in the sense of popularity ranking which is gained from the sales.

BaseLine Model *Outcome*

Seriously though,
that T-Shirt..!!



Product Similarity: Convert Textual Data Into a Vector Matrix

Tf in **tf-idf** weight measures frequency of terms in a document. And **idf** measure importance of that given term in given corpus.

TF(t)= (**Number of times term 't' appears in a document**) / (**Total number of terms in the document**)

IDF(t) = log_e(**Total number of documents / Number of documents with term 't' in it**)

Convert Textual Data Into a Vector Matrix Product

At first description column has been constructed for united the textual data from main database.

Stop words has been used for ignoring the widely used words such as the or a. Because of the data has Turkish words, Turkish words are included to stop words.

tfidf_matrix constructed by fit_transform method.

tokens

documents	face 0	person 1	guide 2	lock 3	cat 4	dog 5	sleep 6	micro 7	pool 8	gym 9	...
D1		0.05					0.25				
D2	0.02			0.32					0.45		
...											

TF-IDF sparse matrix example

Then **similarity between all products** is computed using **SciKit Learn's linear_kernel** .

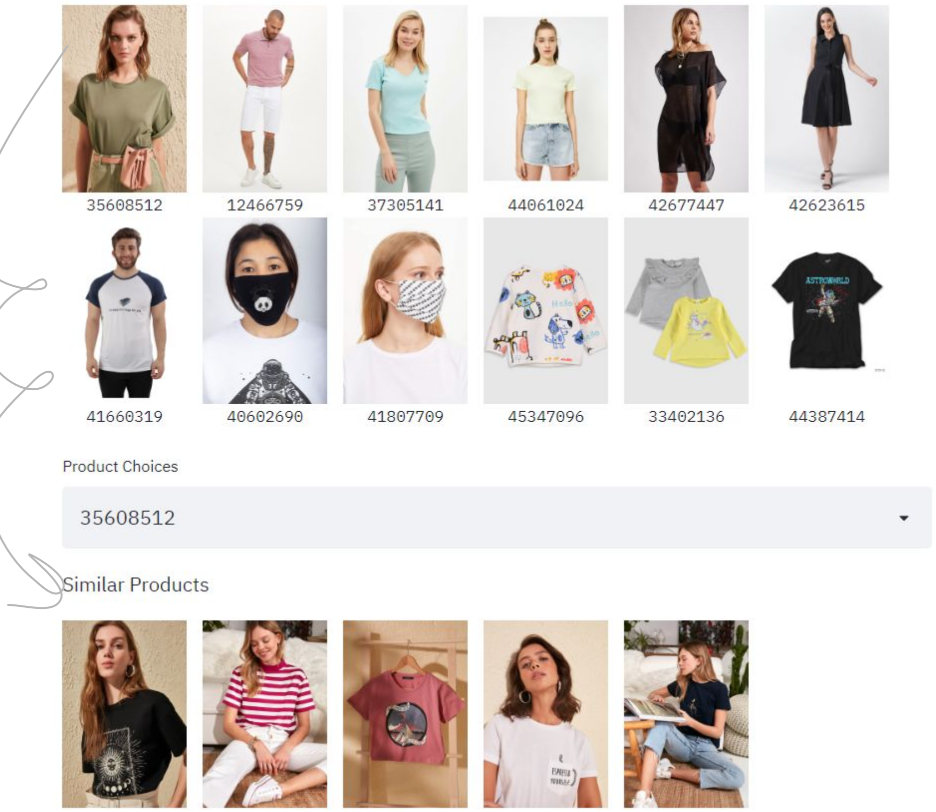
Prediction

To find the most related documents, we can use **cosine_similarites.argsort()** to get the most related document similarities values.

Similar items sorted ,first one removed due to its own similarity rate with itself

Function code to get similar items product id from the description field is written.

Function code to listing the products with the higher similarity with URL given in the ImageLink column is written.



Collaborative Recommendation Engine

This module shows how to retrieve the top-10 items with highest rating prediction.

Firstly **train an SVD algorithm on the dataset,** and then **predict all the weight for the pairs (user, item)** that are not in the training set.

Then retrieve the **top-10 prediction for each user.**

Map the predictions to each user.

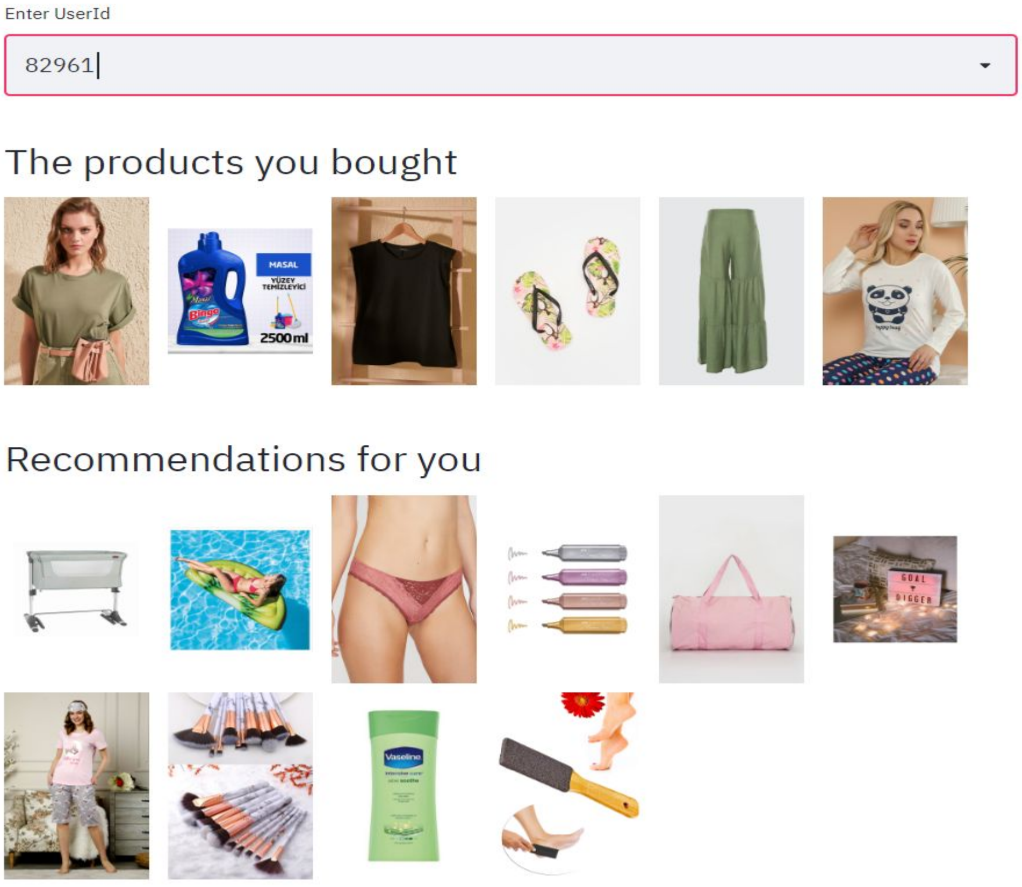
Then sort the predictions for each user and retrieve the highest ones.

Accuracy *is calculated by RMSE metric.*

RMSE: 0.0833

0.08334527017878225

Outcome


Enter UserId
82961
The products you bought
MASAL
YÜZEY TEMİZLEYİCİ
2500 ml
Bingo
Recommendations for you
GOAL
DIGGER
Vaseline

Conclusion

Content Based Recommendation

- **PRO**: Because of the weights given based on the popularity of the product and the data only based on the orders tf-idf vectorizer and cosine similarities generated satisfying outcomes when it comes to the content based recommendation.
- **CON**: Due to the recommendation is not specific to the user the recommendation scope is larger than expected, that might mislead the marketing strategies.

Collaborative Recommendation

- **PRO:** This engine specified for user and covers the popularity of product but in this case ratings given based on behaviour of user.
- **CON:** Not having sufficient amount of data that covers that user behaviours mislead the recommendation.

Excitements

- ❑ Freedom to design a path
- ❑ Seeing user behaviours behind data
- ❑ Marketing point of view

& Challenges

- ❑ Time Management
- ❑ Problem to see bigger picture

UP School Journey

FeedForward = Move Forward

COMMUNITY

WITH

A

SHARED

VISION