Recommender System User Analysis and System Building using KASANDRA

Donna I. Baret

December 2019

Abstract

This paper explores click behaviours of users of the www.kelkoo.com (and partners) in Germany. We explored if the advertisements are given attention on, and during what day of the week and time of day should the businesses employ and advertisement. Moreover, we're going to explore how often do customers provide implicit feedback. To add to that, we're going to devise collaborative algorithms based on the offers and the users. Results of our analysis show that a user is offered 25 advertisements and clicks on one recommended advertisement on average. We see an influx of website activity from 9 pm to 4 am especially on weekends. Matrix Factorization condensed the variables in to 30 factors for model building. Moreover, we developed an algorithm that made use of User-based Collaborative Filtering, Popular Method, and Alternating Least Square approach based on our data.

I. Introduction

Recommendation Systems and Customer Behavior

As more people have easier access to the internet, the emergence of online shopping is becoming more popular process (Zuroni Md Jusoh, 2012). With this popularity, businesses are becoming more competitive in capturing the interest of its customers, one example is the rise of personalized advertisements. A service wherein businesses use the profile of the user on posting targeted content on their websites. (USA Patent No. US 2008.0015878A1 , 2006). Online shopping not only affects the business' reach but also benefits the customers as there is a reduced time in traffic, they have better buying decisions because they can compare items, the processes are documented, and it is available for 24 hours. As a result, it is more environmentally friendly because of less transportation footprint and convenient (Zuroni Md Jusoh, 2012).Recommender systems (RSs) are defined as a system that filters information that predicts a user's preferences and recommend an item that might interest the users. (Adomavicius & Tuzhilin, 2005) (Adomavicius G., 2005)

In this paper, we are going to analyze the recorded customer behavior of Kelkoo – a European leader in eCommerce advertising. The process wherein: The User visits Kelkoo's website and enters a search keyword; then user browsing through Kelkoo's or partner's website is shown an ad; the user enters search keywords in Kelkoo's partner's website which does not cache offers; and the user enters search keywords in Kelkoo's partner's website on which offers are cached (Sumit Sidana, 2017).

Moreover, we're going to explore if the advertisements are given attention on, and during what day of the week and time of day should the businesses employ advertisement. Moreover, we're going to explore how often do customers provide implicit feedback.

If the advertisement is not given attention to, we should improve on recommendation systems and improve the user interface. Moreover, if there is spike of response on certain days of week, we should give focus to certain advertisements.To add to that, we would explore different techniques in creating a user based collaborative filtering procedure and explore different results.

II. Review of Related Literature

Information Systems

Recommender Systems (RS) aim to capture a personalized preference by suggesting a themed list of items that might be of their interest. From this suggested list, the users provide various types of feedback on specific items that have been presented to them, allowing the system to learn and improve the quality of future recommendations. The feedback given by a user can be of different nature, and it has evolved over time from explicit feedback, given in the form of ratings on a numerical scale, to mostly implicit feedback inferred from user's behavior, such as clicking on items, bookmarking a page or listening to a song. Implicit feedback presents several challenging characteristics such as the scarcity of negative feedback. Kasandr (Kelkoo lArge ScAle juNe Datafor Recommendation) is a collection that gathers one month of Kelkoo's data collected from 20 European countries. This dataset contains 16 million clicks given by 123 million customers over 56 million offers that have been displayed to them during their surf sessions. These clicks come along with contextual information, such as the geographical location of users or the hierarchical taxonomy of offers, which make the collection challenging for the design of efficient recommender systems.

A study by Zhang, 2017 on the customer behavior analysis reveals that customers shop more during the start of the week and online shopping prime starts at 8:00 P.M. This can reveal that websites can charge more for traffic hours for advertisement as well as make offers more aggressive during this time.

III. Methods

A. Data Understanding

a. Data Description

In this study we're going to focus on the dataset records interactions of Kelkoo's customers between June, 1st 2016 and June, 30th 2016.

Table 1. Basic Statistics

Label	Summary
# of Offers Shown	4,433,789
# of Click Throughs	197,337
# of Unique Offers	1,220,622
# of Unique Categories	271
# of Unique Merchants	703
# of Unique User IDs	172,021

Table 2. Aggregate Statistics per User

Average Numbers of Offers Shown to 1 user	25.77
Max Number of Offers Shown to 1 user	18646
Min Number of Offers Shown to 1 user	1
Average Number of Clicks Shown to 1 User	1.1472

b. Explore Data

We can see that the incident number is the unique identifier and we will assign counts per one-way, two-way, and three-way table

i. Univariate Analysis

Table 3. Top 10 Categories that has the highest click through rate

	Rating	
Category	0	1
100020813		100.00%
128101		100.00%
121201		100.00%
100345723	5.56%	94.44%
100333423	12.70%	87.30%
100485423	15.56%	84.44%
100472123	19.35%	80.65%
100556113	27.54%	72.46%
100295823	28.77%	71.23%
100345823	31.25%	68.75%

We can see that categories 100020813,128101, and 121201 had the highest clickthrough rate

Table 4. Top 10 Merchants that has the highest click through rate

	Rating	
Merchant	0	1
ea0c486f7e2afe..		100.00%
cde1bb72b28d4..		100.00%
c0d3746533193..		100.00%
a7192675d3c81..		100.00%
245d2c7b8e6fc4..		100.00%
36e2130a3c070..		100.00%
3e0c8ff0db6c0b..		100.00%
afd6221dc4627..	13.04%	86.96%
251703522c9c0..	14.29%	85.71%
05631b2b32600..	17.86%	82.14%

Table 4 shows that there are 7 merchants that have 100 % percent click through rate

Table 5. Top 10 Users that has the highest clicks

	Rating	
Userid	0	1
7625efac4a89c4..	1,258	552
314dc010def12..	1,336	541
747d97bdf66ece..	1,145	510
817d96c958dd9..	1,112	472
ac47f1435465a..	1,170	471
f0fbac7eb4c2c0..	1,110	469
a333d3bed4caf8..	1,161	451
b4380e45e42d8..	1,138	433
d119005c254ce..	1,188	426
4ba6cb76318d7..	933	417

Table 5 shows on the other hand that the top users with highest click throughs click on offers almost half of the time.

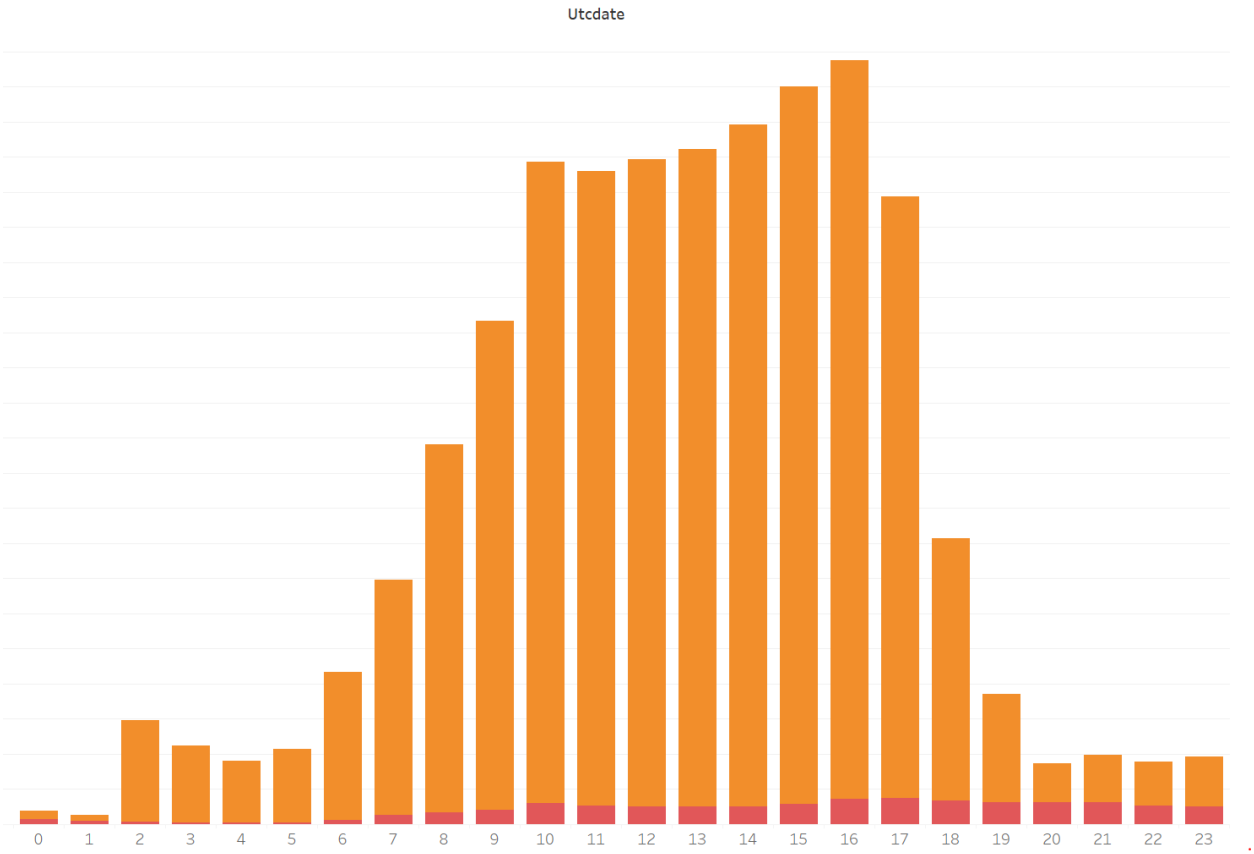

Figure 1. Website Activity per Hour by Number of Visits

Figure 1 shows that website activity spikes 10 PM to 4 am,

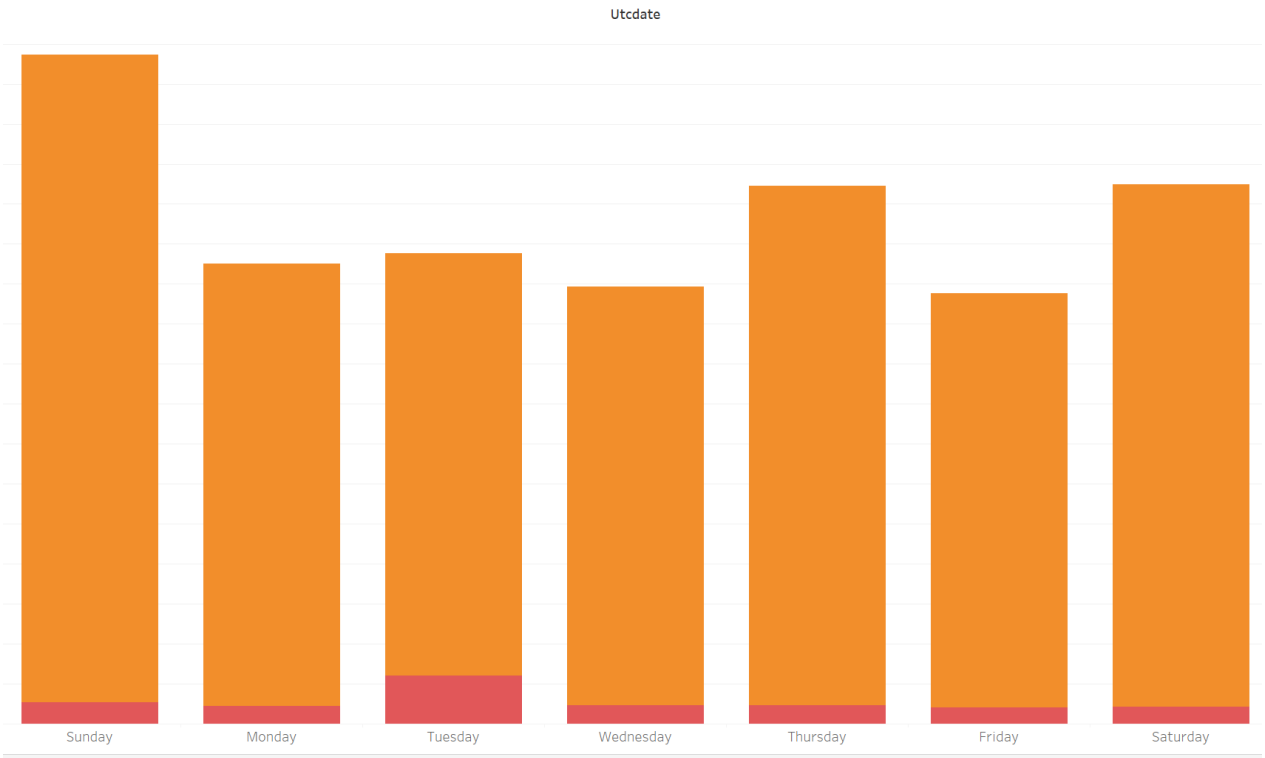

Figure 2. Website Activity per Day of Week

Figure 2 shows that Sunday had the highest website activity followed by Thursday and Saturday. However, Tuesday had the highest number of click throughsa

B. Data Mining Analytics/Business Process Discovery

a. Data Preparation

Looking at the data, we can see that there are offers that has been clicked a few times. Therefore, there might be bias in this data extreme values. As prelimenary analysis and due to computing power, we will limit the data to the number of users who has clicked at an offer more than 30 times and offers that has been clicked at least 100 times.

b. Matrix Factorization

Matrix capture patterns in rating data in order to learn certain characteristics, aka latent factors that describe users and offers.

Using matrix factorization condensed the variables in to 30 factors for model building

Table 6. Fitted Model

Fitted Model	Count
Number of users	14040
Number of items	1505
Number of factors	30

c. Recommender System

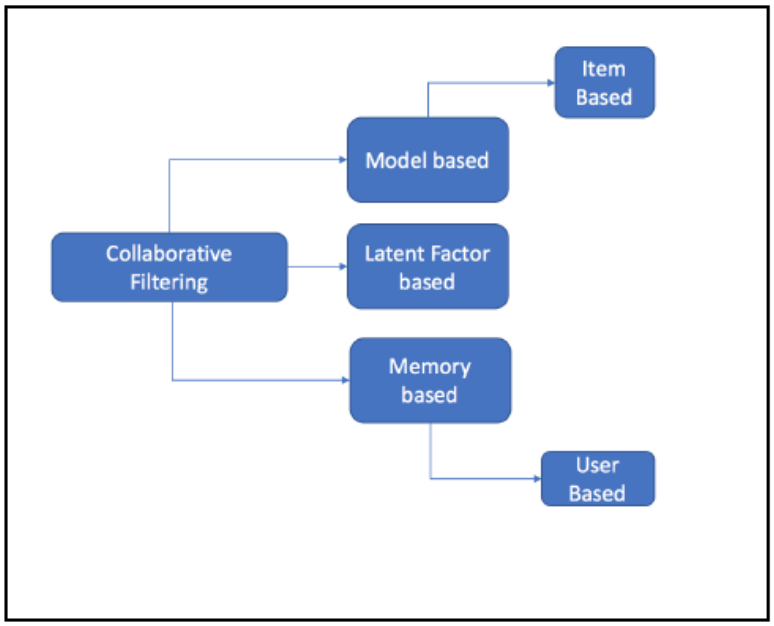

Figure 3. Different Approaches to Collaborative Filtering (Subramanian, 2017)

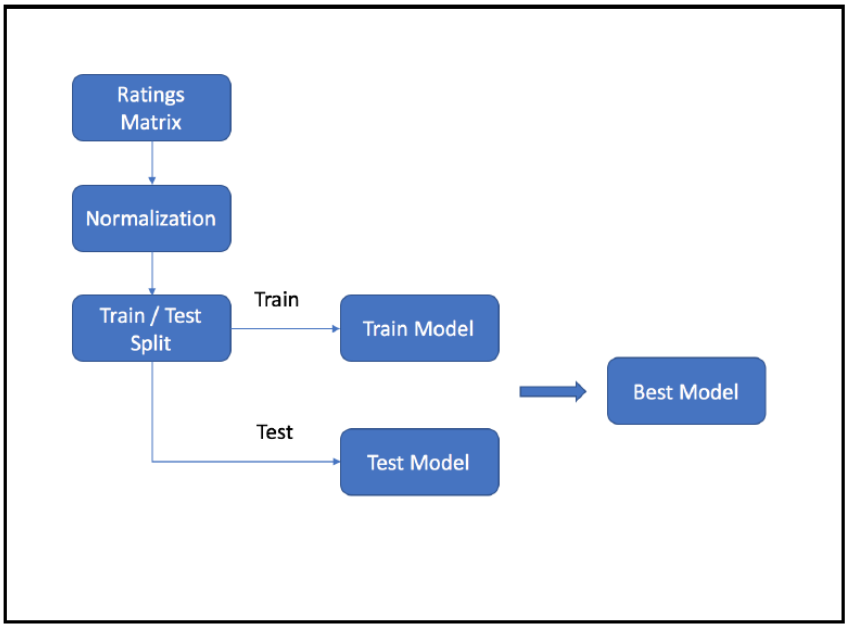

Figure 4. Steps in Designing our recommendation Project (Subramanian, 2017)

i. User-Based Collaborative Filtering
The recommender is based on user-based collaborative filtering

Table 7. Top 6 Recommended offers for 3 of the users Using Collaborative Filtering

		Users	
	`0dd0e66789b8fa299b47e2bbc94a168627c91580864843	$`0dd1f22ca6088bbf9845597ec6856e4ddc9cf811b1	$`0dd239ece907b1f272524efb39c9017813b442b25dd8fa6cbf320645c3b4b820`
	19754ec121b3a99fff3967646942de67	19754ec121b3a99fff3967646942de67	19754ec121b3a99fff3967646942de67
Offrers	f678c08c9eed563c6380ec1e9791106b	f678c08c9eed563c6380ec1e9791106b	f678c08c9eed563c6380ec1e9791106b
	837464e90d3d28e060969cee5b143c97	837464e90d3d28e060969cee5b143c97	837464e90d3d28e060969cee5b143c97
	5ac4398e4d8ad4167a57b43e9c724b18	5ac4398e4d8ad4167a57b43e9c724b18	5ac4398e4d8ad4167a57b43e9c724b18
	5670cce5b15cde9f10991a59086d3f2a	5670cce5b15cde9f10991a59086d3f2a	5670cce5b15cde9f10991a59086d3f2a
	75c247f3113a727067abdcd9eafe3b64	75c247f3113a727067abdcd9eafe3b64	75c247f3113a727067abdcd9eafe3b64

ii. Popular Approach
This answers the question, among the offers not clicked by the users, the system will recommend one that that is most popular among users

Table 8. Top 3 Recommended offers for 3 of the users Using Popular Approach

		Users	
	$`0c549a64984d63718eb6d77993d5bfc68a4d4833664d392d3bf40f6953	$`0c55868c02fb8dfb5120cba1ef51cf87fbb	$`0c6735433b2da68045df3aad00168ce39db5df
	ccbdecfb71d4a0a7e836a4a4b1e69c97	a5fc37404646ac3d34118489cdbfb341	a5fc37404646ac3d34118489cdbfb341
Offrers	fe8efbbd8879b615478cf7314b3b87ba	3c9af92d575a330167fb61dda93b5783	3c9af92d575a330167fb61dda93b5783
	b8e18f806c165f9c316e1eadd12200aa	241145334525b9b067b15de4fd7a0df1	241145334525b9b067b15de4fd7a0df1

iii. Alternating List Square
Recommender for implicit data based on latent factors calculated by alterating least square algorithm.

Table 9. Top 3 Recommended offers for 3 of the users Using ALS

		Users	
	$`0dd239ece907b1f272524efb39c9017813b442b25dd8fa6cbf320645c3b	$`0c55868c02fb8dfb5120cba1ef51cf87fbb	$`0dd926881e9f8994dbaf56d22766dba64a84bb
	f678c08c9eed563c6380ec1e9791106b	a5fc37404646ac3d34118489cdbfb341	837464e90d3d28e060969cee5b143c97
Offers	19754ec121b3a99fff3967646942de67	3c9af92d575a330167fb61dda93b5783	75c247f3113a727067abdcd9eafe3b64
	3c9af92d575a330167fb61dda93b5783	5670cce5b15cde9f10991a59086d3f2a	5ac4398e4d8ad4167a57b43e9c724b18

IV. Results

A. Discussion/Conclusion

In this paper, we have analyzed the logs of the www.kelkoo.com (and partners) in Germany. We have seen that website activity is busiest during 9 PM to 4 am. This shows that off-hours where malls are usually closed give online businesses opportunities to sell more and advertise because customers are more likely to depend more on deliveries and online shopping. Moreover, a distinctively high activity during Saturdays and Sundays provide opportunities for users to spend time at home more through online shopping while businesses can grab this opportunity to sell. We can see however, that Tuesdays have a higher click-through rate which means that users are more likely to click through an ad on Tuesdays. This implies that there's a chance that users are more likely to consider recommendations during this time. Moreover, website owners can benefit from the prime-time slots by charging more to the advertisements and less on less busy days and hours.

It is also recommended for management to analyze the high click through rates for merchants and categories and discover the reason behind. The best practices discovered on this can be applied on ads that has low click-through rate.

We also conducted a User-Based Collaborative Filtering wherein the recommender is based on user-based collaborative filtering. On the other hand, the popular approach which shows that most popular offers by the use. Moreover, ALS which is for implicit data based on latent factors calculated by alternating least square algorithm.

Collaborative filtering is useful for both user and business at it makes product comparison for the user easier and at the same time gives a business an opportunity to advertise and be known by the users. Due to machine memory and speed, cross validation techniques were not explored. It is highly recommended to perform an evaluation procedure for the filtering techniques. Moreover, compared to explicit feedback, less research has been explored on implicit feedback. Therefore, more procedure should be done regarding this. One limitation of this paper is that the implicit feedback was only measured through the condition of if the user clicked on the offer or not. It is recommended to explore ways on how the analysis will be affected on the amount of clicks the user performed on the same offers