BUSINESS CASES WITH DATA SCIENCE

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS – MAJOR IN BUSINESS ANALYTICS

ManyGiftsUK Recommender System

Group F

Lorenzo Pigozzi	m20200745
Nguyen Huy Phuc	m20200566
Ema Mandura	m20200647
Xavier Gonçalves	m20201090

May 3rd, 2021

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

CONTENTS

1. INTRODUCTION

As a fundamental strategy for any online retail company, *ManyGiftsUk* has the need of the creation of a recommendation system that makes possible to showcase products to meet the customer's needs. The recommendation systems are nowadays a very important component in all the retail industry, because when consumers are faced with products they like, they statistically tend to watch more closely and buy more. Just to make an example, more than the 30 % of the sales of Amazon are due to the recommendations displayed at the end of the page, as all of us already know. Another example could be Netflix, that it's able to "manipulate" the views of the users thanks to the strong recommender system that was developed. The importance of the recommendation systems is even increasing in the last period, due to the increasing number of transactions that are computed through the web. These patterns of tastes on the part of customers are defined both by past options, the history of purchases and interactions, and by previous interactions among the products themselves.

Considering this situation, the necessity for *ManyGiftsUk* for the implementation of a recommender system is clear and justified, and on this project weteam will analyze the data provided in order to build the necessary model and try to achieve the ambitious but relevant business objectives that were listed out.

2. BUSINESS UNDERSTANDING

2.1. Background

ManyGiftsUk is a retail company based in the United Kingdom that does not have any physical store, but only a website and around 80 employees. With nearly thirty years of existence, it is mainly dedicated to selling unique all-occasion gifts. For years in the past, the merchant relied heavily on direct mailing catalogs, and orders were taken over phone calls. It was only two years ago that the company launched its own website and shifted completely to the web. Since then, the company has maintained a steady and healthy number of customers from all parts of the United Kingdom and the world and has accumulated a huge amount of data about many customers. The company also uses Amazon.co.uk to market and sell its products.

ManyGiftsUk needs to use the data collected on transactions in order to build a recommender system that can skillfully make valuable and relevant suggestions to users to improve their shopping experience on the website.

2.2. Business Objectives

The overall goal of the company, considering the business situation, is to improve the experience of the consumers and at the same time increase ManyGifts' revenues, as the two topics are related. But how to do it? Which are the specific objectives of the company?

Entering a bit further in the specific situation of the company, the most important challenge that our team will try to achieve along this project is the implementation of a recommender system that will be able to make good suggestions to new and already registered customers. For the new customers, the business objective is obviously to catch their attention and showcase directly the top products of the company, instead for the already known customers the challenge is even harder but more interesting, indeed the objective is to improve the customer experience during the shopping phase, providing them useful suggestions that could potentially increase the customer satisfaction and at the same time the economic profit of the company. The suggestions will be obviously made based on the data collected.

Furthermore, *ManyGifts* has specifically requested to make particular attention for the cold-start problem, that consists of the challenge of recommending items for the new customers. The challenge is usually difficult because for the new customers the company doesn't have historical data, and thus is more difficult to try to estimate which products could be the new customer be interested in, but this could be very relevant for the overall experience of the customer.

2.3. Business Success criteria

The success of the project can be measured by the available evaluation measures. The recommendation system should therefore be as effective and accurate as possible for users, considering the past of the transactions available in the data.

In order to have an effective and successful recommender system, several measures can serve as a reference to meet the business criteria. One is the Conversion Rate (CVR), which measures the proportion of people who are persuaded by the recommendation (announcement) to add a product to the cart and make its purchase. Closely related is also the concept of Click-Through Rate (CTR), which represents a medium that tracks the number of website users who engage with an ad or item recommendation, such as viewing the object or adding it to the cart.

In addition, the Return on Investment (ROI) is also a possible measure to understand the dimension of the success of the built model, a ratio computed using gains and the investment previously made. In the business context, the Customer Lifetime Value (CLV) can also be used to perceive the measure of the success of the recommender system, also describing value a given customer can have throughout his life for *ManyGifts*, considering its purchasing pattern data.

2.4. Situation assessment

The dataset contains data about retail transactions, such as the type of product through code and description, date, identification and country of the customer, as well as the type of event created by a customer's interaction, the type of item that was interacted with and the identification of the transaction, if it happened, among other data.

The data provided was mostly clean, with few insignificant missing values. It is mostly accurate and relevant.

2.5. Determine Data Mining goals

As already explained, in this context, the strategy used in terms of analysis and data mining will have to do an exploration of the data and resolve some issues such as providing a recommender system capable of effectively suggesting products to users, to increase the chances of the interaction resulting in a purchase. For this point, one of the challenges will be to propose relevant products to new customers.

In addition, it will be essential to be able to measure the effectiveness of the models created through evaluation and selection strategies of an appropriate quality measure. Finally, in the deployment phase, provide an explanation of the challenges and recommendations made in the context of the implementation of the recommender system.

3. DATA MINING PROCESS

3.1. DATA UNDERSTANDING AND EDA

The customer transaction dataset held by the merchant has 8 variables as shown below, and it contains all the transactions occurring between 01/12/2010 and 09/12/2011. Over that period, there were 25900 valid transactions in total, associated with 4070 unique items and 4372 customers from 38 different countries. The dataset has 541909 instances, each for a particular item contained in a transaction. Also, it is important to note that many of *ManyGiftsUK* customers are wholesalers.

For each row of the dataset, the information stored are 3 variables that are id of the invoice, item and the customer, respectively. Furthermore, other 5 variables give us information about the country, the description of the product, the invoice date, the price and the quantity purchased.

Starting to explore further the dataset, we discovered that the dataset contains only 1454 missing values for the Description variable, nevertheless we still have the '*StockCode'* for these products and so we don't need to drop those records. The dataset also contains 135080 missing values for the Customer id, but in the metadata is specified the reason for that, indeed these records are transactions done by new customer, and for this reason they still don't have an ID.

Another problem that is mentioned in the metadata regards the cancellation of an order. In fact, the dataset provided contains some records with negative values for price and quantity, associated with an ID of the invoice starting with the letter 'C', that means that the order is cancelled and explain the reason of the negative values, used to compensate the positive ones of the transactions already stored in the transactional system.

Regarding the Country variable, we decided to do not consider it at all for the creation of the model and for the following analysis as it is really unbalanced, because almost all the transactions are registered in the UK, as it's possible to see in the plot below.

Furthermore, in order to better understand the behavior of the customers we carried out a RFM analysis, that will be particularly useful for the implementation of the model and for improving the results, as we will explain later.

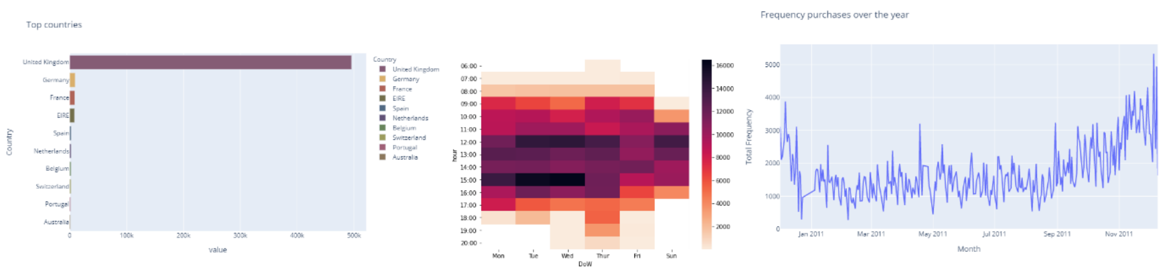

Figure 1. EDA plots

3.2. DATA PREPARATION

The Exploratory Data Analysis has shown that in order to proceed to the modeling phase, a data pre-processing step is necessary.

First, we had to deal with the problem of the cold start, related with the missing values of the '*CustomerID'*, and this problem has been solved simply splitting the dataset in 2: transactions of new customers and transactions of already known customers, and we will have different solutions for them.

Sparse matrix creation

In addition, in order to proceed to the modeling-phase we focused only on 3 of the variables, CustomerID, StockCode and Quantity. The reason is that it's necessary to focus only on them for the creation of the matrix that will be used for the development of the model; for solving the problem of the cancellations, we grouped by CustomerID and StockCode summing up the Quantity, then we set the Quantity = 0 to 1 and finally we filtered out all the records with a negative quantity amount. The reason is that the Quantity = 0 for a product appears also in the case of cancellations that we're previously stored and subsequently "deleted" adding a row with same product and negative values. We decided to keep also those records because even though the order was cancelled and not actually purchased, it was completed and so it's still possible to have some useful information about it.

Then the next step was the creation of the interaction matrix, and we ended up with a sparsity level of 98.23%. A rule of thumb tells us that a collaborative filtering works well with a maximum sparsity about 99.5% or so. We are well below this, so we should be able to get decent results.

Train-Test split

Subsequently we split the dataset into 2 smaller dataset, training and test sets in order to build a model with the training set and testing the result with unseen data. We used the train test split method from the Implicit package to randomly split the transactional data. The training dataset is 80% of the transactional data of user-item, while the test dataset is 20% of the transactional data. We do not account for the datetime when splitting because we believe the time is too short to make any difference in the buying sequence of the customer. Thus, now we are having the shape of both sets is (4338, 3664) incorporating 4338 customers and 3664 products.

	CustomerID	StockCode	Quantity
0	12346	23166	1
1	12347	16008	24
2	12347	17021	36
3	12347	20665	6
4	12347	20719	40

RFM Analysis and Clustering

To have better analysis of the customer of ManyGiftsUK, the RFM (Recency – Frequency – Monetary) analysis is carried out from the data engineered from the retail dataset.

	Recency	Frequency	Monetary
CustomerID			
12346	325	1	77183.60
12347	1	7	4310.00
12348	74	4	1797.24
12349	18	1	1757.55
12350	309	1	334.40

The suitable number of clusters founded from K-means clustering algorithm is 2. Also, the intention was to use small number of clusters to keep the recommender system applicable with sufficient data. The profile of each cluster is also clearly profiled into 2 customer types: normal users (cluster 0) and best buyers (cluster 1) with the quantity of 2593 and 1746 respectively for each cluster.

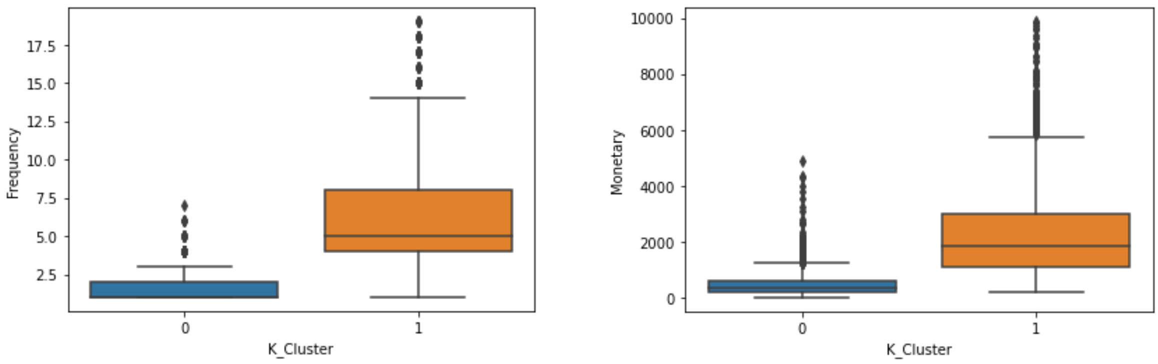

Figure 2. Frequency and Monetary value box-plots distribution for the 2 clusters

The clustering result will be further applied for separating the customer transactional database into respective dataset. Then, each of the cluster dataset will be used to build a different recommender system for that specific customer cluster.

3.3. Recommender System

3.3.1. Cold-start problem and Popularity model as baseline model:

The collaborative filtering method always faces a major challenge which is called the "Cold-start" problem. This problem encounters when the system has no information to make recommendations for the new users. As a result, the matrix factorization techniques cannot apply for these users. Thus, several solutions should be applied to overcome this problem from new users and new items.

Popularity-based recommender

The top 10 best-selling products will be used to create a recommender system for new users. This baseline model suggests the most popular products with the highest number of purchases in the dataset. It is necessary that the top selling products list should be updated regularly in order to keep the recommendation list updated. The creation of this model will also be used as a baseline model that acts as a baseline for evaluation the effectiveness of the recommender systems that are going to be built.

For newly launched products of the company, it is suggested that the administrator of the recommender system occasionally add the newly promoted products to the recommendation list. By this way the new products will have the opportunity to appear for customers' consideration along with normal recommendation.

Moreover, as the dataset currently only contains implicit data (historical transactional data), the only method suitable to build the recommender system is limited to collaborative filtering. It is suggested that the company should collect more data about products such as categories, ratings and about customer such as demographic and interests. Having those data will enable the ability to build a content-based recommender system and hybrid recommender system for the company.

3.3.2. Recommender system with matrix factorization algorithm

With the Popularity model as the base line, we continue to evaluate 4 matrix factorization models: the Bayesian Personalized Ranking, the Logistic Matrix Factorization and the Alternative Least Square model applying on the whole dataset and for the clustered ones.

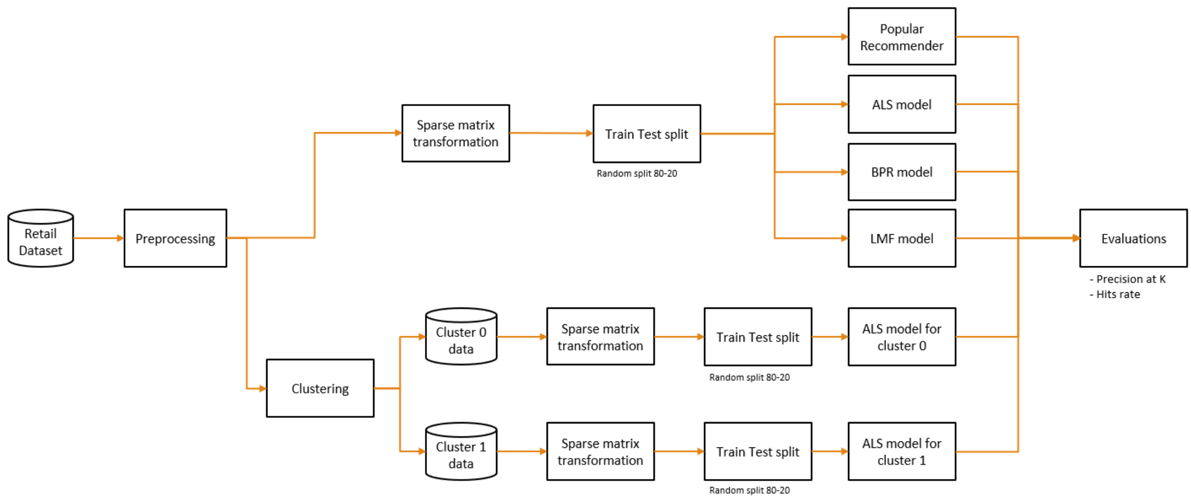

Figure 3. Matrix Factorization Pipeline

The evaluation metrics used are precision at K and Hits rate. Firstly, Precision at K is chosen for this case because it corresponds well to the setting of an e-retailer. According to Shani, and Gunawardana (2011) this metric is best fitted to business that only allows a limited number of items can be recommended to a user. Since a website only has a limited space to list a limited number of options (for example, only five to ten items can be recommended), the precision at K comes as the most suitable evaluation metrics for the recommender system. From our own experience and analysis from top e-commerce websites, we will use a value of 10 for K, which means that the recommender system will be evaluated by average number of products that are going to be purchased in the 10 products that are recommended to the user. Secondly, Hits-rate is the number of times that the recommender successfully recommended at least 1 product that is bought by the customer later. Since any purchase from customer as a result of the recommender is considered a success, this metric will be useful in evaluating the system without penalizing on irrelevant recommendation, which is not a major problem in e-commerce business.

$$\textbf{Precision at K} = \frac{\text{Number of products recommended that are relevant}}{\text{Number of products recommended (10)}} \qquad \textbf{Hits-rate} = \frac{\text{Number of customers found at least 1 recommendation useful}}{\text{Total number of customers that got the recommendation}}$$

4. RESULTS EVALUATION

Beside the preprocessing of data, in order to improve recommender system performance, several hyperparameter tunning processes are carried out, the best hyperparameter for ALS model is alpha = 1, factors = 20, regularization = 0.1, iterations = 50.

Using the ranking_metrics_at_k method provided by the Implicit package, the precision-at-K results are calculated for the recommender system for final evaluation. The hits-rates are calculated manually from the prediction of the models on the test set

	Popularity model	*BPR model*	*LMF model*	*ALS model*	***ALS Cluster_0 model***	***ALS Cluster_1 model***

P@K	0.03387	0.0469	0.04171	0.18412	**0.12408**	**0.22449**
Hits-rate	N/A	0.2428	0.2181	0.5986	**0.4297**	**0.7720**

The result showed that the ALS model have the best overall performance. However, we can see that the performance increase significantly for the Cluster 1 recommender system. And as know that they are our top buyers, to make the best to the profit of the business, we decide to make the trade off to choose this last solution even though the result of the respective cluster 0 model is not good as good as the general model.

Thus, to favor the more important customers to the business, we will implement the Recommender system in combination with Clustering method. Having this model applied, the top buyers will have the hit rate of nearly 80% and 22% of precision.

5. DEPLOYMENT AND MAINTENANCE PLANS

The deployment phase of every company is crucial for the business context. It provides a better idea about the effective usage of the machine learning algorithm created, designing a possible bridge between the business goals and the machine learning goals, applying the algorithm in a real-world context.

Example of recommendation lists

User – Item recommendation:

```
Find related items for user 12346

[('22961', 'JAM MAKING SET PRINTED', 0.013814082),
 ('23307', 'SET OF 60 PANTRY DESIGN CAKE CASES ', 0.010468042),
 ('23167', 'SMALL CERAMIC TOP STORAGE JAR ', 0.010374689),
 ('22722', 'SET OF 6 SPICE TINS PANTRY DESIGN', 0.009967294),
 ('23296', 'SET OF 6 TEA TIME BAKING CASES', 0.009858806),
 ('23236', 'STORAGE TIN VINTAGE DOILY ', 0.00941932),
 ('22919', 'HERB MARKER MINT', 0.00940817),
 ('23294', 'SET OF 6 SNACK LOAF BAKING CASES', 0.009407425),
 ('23165', 'LARGE CERAMIC TOP STORAGE JAR', 0.009369746),
 ('23295', 'SET OF 12 MINI LOAF BAKING CASES', 0.009238686)]
```

Item – Item recommendation:

```
Find related items for MEDIUM CERAMIC TOP STORAGE JAR

[('23166', 'MEDIUM CERAMIC TOP STORAGE JAR', 1.0),
 ('23167', 'SMALL CERAMIC TOP STORAGE JAR ', 0.89451414),
 ('23324', 'RUSTIC STRAWBERRY JAM POT LARGE ', 0.82094103),
 ('23165', 'LARGE CERAMIC TOP STORAGE JAR', 0.82045496),
 ('23512', 'EMBROIDERED RIBBON REEL ROSIE', 0.7972661),
 ('23285', 'PINK VINTAGE SPOT BEAKER', 0.70843875),
 ('23299', 'FOOD COVER WITH BEADS SET 2 ', 0.6790141),
 ('22653', 'BUTTON BOX ', 0.6638413),
 ('23326', 'HANGING MINI COLOURED BOTTLES', 0.66304946),
 ('23236', 'STORAGE TIN VINTAGE DOILY ', 0.65319353)]
```

To better understand the recommender system, a Business Process was designed in order to be able to effectively contemplate the possible and differentiated types of users.

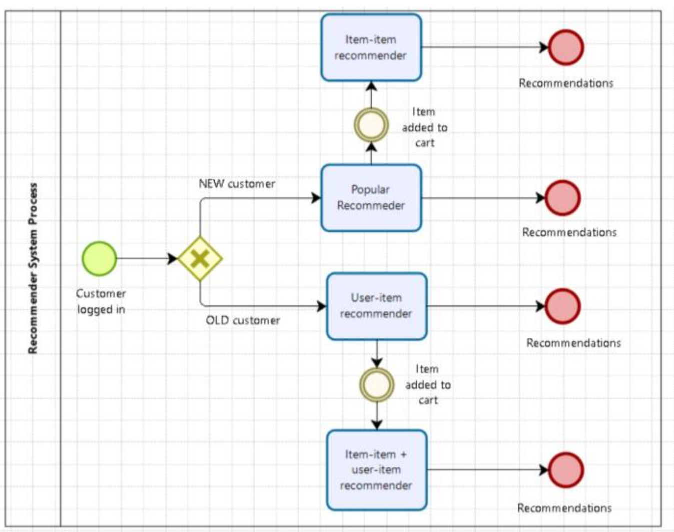

Figure 4. Recommendation Process

When entering the website to search for items of interest, users can be considered as a new customer or an old customer, depending on whether they already have past interactions or not. For a new customer, the cold start problem will be solved using the "Popular Recommender", a content-based filtering system. Then, if an item is added to the cart, the recommendation will be made based on that item, in more concrete terms. On the other hand, if no item is added to the cart, the suggestion will continue to be according to the popular recommender.

On the opposite side, for an older customer, the gift suggestion will be made based on the customers' history, according to a collaborative filtering system, at least until an item is added to the cart. In this situation, the system will become hybrid, providing item-item and user-item suggestions in order to increase the chances of a purchase.

Overcoming the cold start problem and having the model able to effectively recommend attractive gifts to new users of the online platform, there are other challenges and recommendations that must be faced in relation to the recommender system. One of the main challenges is the lack of data related to the business component, as it would be very useful to employ information on economic aspects of products in suggestions made to users of the platform, creating a tendency for potential customers to be tempted to purchase more expensive goods from among the most recommended, in order to maximize revenue in general.

One of the challenges in such a system is overspecialization, which occurs mainly when there is still not much data available on part of the customers. The somewhat limited availability of data will cause only a range of items with relative similarity to appear in the recommendations for some customers, not suggesting novel items. To this end, diversity and some randomness must also be introduced.

A suggestion for the model could be to obtain more information about some items, such as books, in which the category and content is essential to make a good suggestion, or clothes, where fashion style is crucial.

Another of the main challenges in implementing a recommendation system is also managing its ratings and opinions about products. If there were data on this issue in the database, in order to be considered for potential future suggestions, this variable could influence the model. However, and considering the sparsity problem, the users of the platform most of the time, even if they make a purchase there, will not assign a review, or will only do it in case the experience is bad, obtaining biased data.

The recommendation system should also filter out obsolete and older items and is therefore a challenge to maintain. For this, a more effective time threshold can be used in order to increase the appearance of more recent items in the recommendations to users.

6. CONCLUSIONS

In order to support the *ManyGiftsUk*, our team analyzed a dataset containing all retail transactions of the company. After a brief exploration and insights of the data, in this project it was carried out a preprocessing and manipulation of the data in order to achieve the expected results that the company requested: the creation of a recommender system, in order to increase the overall sales amount of the company.

To be able to achieve this result, we firstly compute a data preprocessing, in order to prepare the data and build the necessary interaction matrices that were used for the development of different models. As mentioned, the model trained were the Popular recommender, the Alterning Least Squares (ALS), the Bayesian Personalized Ranking, the Logistic Matrix Factorization and the LightFM. The evaluation of the model was carried out using different metrics, such as the AUC score, the Precision at K and the hits score. Based on those metrics, the model selected was the ALS. The hyperparameter and training phase led us to perform a further engineering of the records, and we increased the results significantly. In particular, the best results were achieved computing techniques connected with the clustering and RFM analysis, and those techniques proposed the creation of the two clusters, one of which is the normal user and the other the best buyer. Each dataset resulting from clustering was used as a constituent of a different recommendation system, suitable for this type of consumer. This analysis was made with the intention of maintaining a low number of clusters so that the recommendation system would be more easily adapted.

6.1. CONSIDERATIONS FOR MODEL IMPROVEMENT

The data we were provided was of good quality and allowed us to preform valuable analysis. With more time at hands, more insights could be gathered from the data.

The dataset lacks data on economic information that could help to create a model that would add more expensive products within the recommended products, as an effective way to maximize potential revenue.

Also, there are some inconsistencies in the data that should be fixed on the existing records and added correctly for new records. For example, the items with same stock codes have different descriptions, and the same items sometimes have different prices without a visible pattern of price change.

In addition, some variables that would add quality to the model would be ratings and opinions, whether they were not excessively biased and that are representative of the population. The model should include some diversity and randomness in the recommended products in order to avoid limited content and overspecialization. Moreover, the model can be improved if it is able to filter outdated and old items in the future and propose more recent items through an effective time threshold and always keep up to date with the latest data.

7. REFERENCES

[1] Khusro, Shah; Ali, Zafar; Ullah, Irfan (2016). Recommender Systems: Issues, Challenges, and Research Opportunities, Information Science and Applications (ICISA 2016), pages 1179-89.
https://fardapaper.ir/mohavaha/uploads/2019/03/Fardapaper-Recommender-Systems-Issues-Challenges-and-Research-Opportunities.pdf

[2] Oliveira, Jorge (2016). Event Recommendation Engine, IST, pages 1-4.

[3] Shani, G., & Gunawardana, A. (2011). Evaluating Recommendation Systems. In F. Ricci, L. Rokach, & B. Shapira, Recommender systems handbook (pp. 257-297). New York: Springer