REPUBLIC OF TURKEY

YILDIZ TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

DEVELOPING A PERSONALIZED RECOMMENDER SYSTEM IN AN E-COMMERCE FIRM

Aslı LEVENT

MASTER THESIS

Department of Statistics

Statistics Program

Supervisor

Assoc. Prof. Dr. Gülder Kemalbay

July, 2022

REPUBLIC OF TURKEY

YILDIZ TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

DEVELOPING A PERSONALIZED RECOMMENDER SYSTEM IN AN E-COMMERCE FIRM

A thesis submitted by Aslı LEVENT in partial fulfillment of the requirements for the degree of MASTER of SCIENCE is approved by the committee on 19.07.2022 in Department of Statistics, Statistics Program.

Assoc. Prof. Gülder KEMALBAY
Yıldız Technical University
Supervisor

Approved By the Examining Committee

Assoc. Prof. Gülder KEMALBAY, Supervisor
Yildiz Technical University _____

Asst. Prof. Erhan ÇENE, Member
Yildiz Technical University _____

Asst. Prof. Arzu BAYGÜL EDEN, Member
Koc University _____

I hereby declare that I have obtained the required legal permissions during data collection and exploitation procedures, that I have made the in-text citations and cited the references properly, that I haven't falsified and/or fabricated research data and results of the study and that I have abided by the principles of the scientific research and ethics during my Thesis Study under the title of Developing a Personalized Recommender System In an E-commerce Firm supervised by my supervisor, Assoc. Prof. Dr. Gülder KEMALBAY. In the case of a discovery of false statement, I am to acknowledge any legal consequence.

Aslı LEVENT

Signature

Dedicated to my family

ACKNOWLEDGEMENTS

I would like to thank my supervisor Assoc. Prof. Gülder KEMALBAY and my family for their support and their help throughout this project.

Aslı LEVENT

TABLE OF CONTENTS

LIST OF SYMBOLS vi

LIST OF ABBREVIATIONS vii

LIST OF FIGURES viii

LIST OF TABLES ix

ABSTRACT x

ÖZET xii

1 INTRODUCTION 1

1.1 Literature Review 1

1.2 Objective of the Thesis 4

1.3 Hypothesis 4

2 THEORICAL BACKGROUND 5

2.1 Recommender Systems 5

2.2 Content-Based Filtering 6

2.3 Collaborative Filtering 7

2.4 Hybrid Filtering 13

3 METHODS 15

3.1 Alternating Least Squares 15

3.2 Singular Value Decomposition 16

3.3 Data Transformation 18

3.4 Evaluation Metrics 18

4 RESULTS AND DISCUSSION 20

4.1 Results 20

5 CONCLUSION 26

REFERENCES 27

PUBLICATIONS FROM THE THESIS 31

LIST OF SYMBOLS

R	Rating Matrix
U	User Latent Factor
Q	Item Latent Factor

LIST OF ABBREVIATIONS

ALS	Alternating Least Squares
CF	Collaborative Filtering
CBF	Content Based Filtering
MAE	Mean Squared Error
RMSE	Root Mean Square Error
SVD	Singular Value Decomposition

LIST OF FIGURES

Figure 2. 1 Recommender Systems 5
Figure 2. 2 Content-based filtering (CBF) dataflow 6
Figure 2. 3 Collaborative filtering (CF) techniques 7
Figure 2. 4 User-Item Matrix 9
Figure 2. 5 User-based Collaborative Filtering 9
Figure 2. 6 Item-based Collaborative Filtering 10
Figure 2. 7 Model based Collaborative Filtering 11
Figure 2. 8 Matrix Factorization 12
Figure 2. 9 Hybrid Filtering 14
Figure 4. 1 Top 20 Most Selling Products 21
Figure 4. 2 Top 20 Categories 21
Figure 4. 3 Rating Distribution 23
Figure 4. 4 Collaborative Filtering w/Cosine Similarity RMSE 23
Figure 4. 5 Collaborative Filtering w/Pearson Similarity RMSE 23
Figure 4. 6 Evaluation Results of ALS 24
Figure 4. 7 Top 20 Most Recommended Products 25

LIST OF TABLES

Table 4. 1 Descriptive Statistics 22
Table 4. 2 Evaluation Results of SVD..... 24

ABSTRACT

Developing A Personalized Recommender System in An E-commerce Firm

Aslı LEVENT

Department of Statistics

Master Thesis

Supervisor: Assoc. Prof. Dr. Gülder Kemalbay

Nowadays it is so important for almost every medium/big sized e-commerce companies to create personalized experience on their websites to increase the interaction with users. Personalized experience can be increased by different ways. However, recommender systems are commonly utilized because of their very handful and effective implementation. It uses users' transactional or behavioral data to suggest the products that are likely related to users' historical data. These recommendations not only help users but also aid companies to increase product variety in the basket and make them reach the growth that they aim. For this reason, the application part of this thesis aims to develop a product recommender system for an e-commerce company. Recommender systems can be developed either memory based, or model based. While memory-based approaches mostly work better and give good results with small data, model based approaches work better with big and sparse data and perform well in production. Since handling big sparse data in production is difficult to establish, we must use an approach which overcomes this problem. Therefore, model based collaborative filtering approach, Singular Value Decomposition and Alternating Least Squares matrix

factorization methods have been applied and the results have been compared to pick the best model. In addition, to be able to apply models to our data, we firstly focused on the transition of transactional data to explicit data. In final, Alternating Least Squares matrix factorization method with the RMSE value of 0.79 and MAE value of 0.51 is picked as best model to integrate into the main recommendation engine. Based on additional analysis the product categories have been recommended for users are different than most selling products and have %43 of coverage of users' main categories. In conclusion, we decided to make use of Alternating Least Squares matrix factorization method in production.

Keywords: Recommendation Systems, Model-based Collaborative Filtering, Singular Value Decomposition, Alternating Least Squares, Matrix Factorization Method, Transactional Data

ÖZET

Bir E-Ticaret Firmasında Kişiselleştirilmiş Öneri Sistemi Geliştirilmesi

Aslı LEVENT

İstatistik Anabilim Dalı

Yüksek Lisans Tezi

Danışman: Doç. Dr. Gülder KEMALBAY

Günümüzde hemen hemen her orta/büyük ölçekli e-ticaret şirketinin, kullanıcılarla etkileşimi artırmak için web sitelerinde kişiselleştirilmiş deneyim oluşturması çok önemlidir. Kişiselleştirilmiş deneyim farklı yollarla artırılabilir. Tavsiye sistemleri, çok az ve etkili uygulamaları nedeniyle yaygın olarak kullanılmaktadır. Kullanıcıların geçmiş verileriyle ilişkili olabilecek ürünleri önermek için kullanıcıların geçmiş veya davranışsal verilerini kullanır. Bu tavsiyeler, kullanıcıların olduğu kadar şirketlerin de sepetteki ürün çeşitliliğini artırmalarına ve hedefledikleri büyümeye ulaşmalarına yardımcı olur. Bu nedenle bu tezin uygulama kısmı, bir e-ticaret firması için ürün tavsiye sistemi geliştirmeyi amaçlamaktadır. Öneri sistemleri, Bellek tabanlı veya Model tabanlı olarak geliştirilebilir. Bellek tabanlı yaklaşımlar küçük veri setleriyle daha iyi çalışır ve iyi sonuçlar verirken, Model tabanlı yaklaşımlar büyük ve seyrek verilerle daha iyi çalışır ve web siteleri gibi canlı ortamlarda daha iyi performans gösterir. Bu tezde kullanılan veriler büyük veri olduğundan ve geliştirilen model web sitesinde canlı ortamda çalışacak şekilde kullanılacağından dolayı gelecekte bu iki durumu da

çözen bir yaklaşım kullanmalıyız. Bu nedenle, model tabanlı işbirlikçi filtreleme yaklaşımı, Tekil Değer Ayrışımı ve Değişen En Küçük Kareler matris ayrıştırma yöntemleri denenmiş ve en iyi modeli seçmek için sonuçlar karşılaştırılmıştır. Ayrıca elimizdeki veri bir satın alma verisi olduğu ve herhangi bir model uygulanmaya hazır olmadığı için, öncesinde işlemsel verileri açık verilere çevirecek yaklaşım uygulanmıştır. Son olarak, RMSE değeri 0.79 ve MAE değeri 0.51 olan Değişen En Küçük Kareler matris çarpanlarına ayırma yöntemi, ana öneri motoruna entegre edilecek en iyi model olarak seçilmiştir. Tüm sonuçlara ve kullanıcılara önerilen ürün kategorilerindeki çıktılara dayanarak, Değişen En Küçük Kareler yöntemi ile çalışan matris çarpanlara ayırma modelinin kullanılmasına karar verilmiştir.

Anahtar Kelimeler: Öneri Sistemleri, Modele Dayalı İşbirlikçi Filtreleme, Tekil Değer Ayrışımı, Değişen En Küçük Kareler, Matris Çarpanlara Ayırma Yöntemi, İşlemsel Veri

1 INTRODUCTION

1.1 Literature Review

The term recommendation system may seem new, but it appeared in 90s by the project named Tapestry [1]. In the mid 90s, what most researchers described the recommender system was a way to provide recommendations by aggregating the results from people who provide recommendations as an input and deliver the results to people who are looking for recommendations. In a short term, this was most likely a collaboration between users rather than creating personalization by using user historical data. Later on, this definition has been expanded as selecting the most suitable products by optimizing the results on users with personalization [29] [4].

Nowadays, most of the industries but especially e-commerce tries to find a way to keep their users online. The idea behind this is to spend less time and money on users while increasing their time spent and attractions on the web page and at the same time get more information about them. Therefore, the information can be used in future to create user-oriented and so personalized experience [19].

Personalization has become very important over the years due to the increasement of web page usages. Having huge variety of products in the same category sometimes makes some vendors or products invisible by users. Some of the products may be suitable for some users but also those products may not be suitable for other users who are looking for similar products but not interested in any specific feature of it. Therefore, to be able to increase the usability of the web pages and make users comfortable with their searches, personalization has become one of the fundamental aims for e-commerce systems. To be able to create this personalization, recommender system techniques have been used by many of the largest e-commerce web sites such as Amazon.com, eBay [30] [2].

Recommended products can either be custom or not for each user because it depends on the approaches and the platform that it has been implemented. There are different kind of techniques that have been used to develop Recommender Systems depending on the given information. However, they are commonly classified into three categories: Collaborative Filtering (CF), Content Based Filtering (CBF) and Hybrid Filtering [4].

Content based Filtering is preferred well when the topic is recommending a movie or a music because this method uses more information about the item and it produces sensible recommendations [10] [11].

However, it cannot generate suitable suggestions if the description (content) does not contain sufficient information for categorization. To solve this problem, feature weighting has been used for content-based filtering, which assigns different levels of importance to different features. But also, this can cause biased results to users when there are only a few attributes [14].

On the other hand, Hybrid filtering combines more than two techniques to achieve better results. In some articles the main idea is described as combining two or more methods helps to achieve better results. Most commonly, Collaborative Filtering is combined with Content Filtering to achieve high accuracy. Because, in some cases, each of the techniques have their own strengths and weaknesses so that researchers combine these both techniques to achieve more reliable results [12].

In addition, sometimes combining both filtering techniques covers the lack of information from both methods and produce better results. For instance, in Collaborative Filtering the information of user is lost whereas in Collaborative Filtering the item information is lost.[28].

However, Collaborative Filtering is one of the most popular techniques that is adopted to recommend products to users in most of the industries but especially in e-commerce industry. For example, there is a successful implementation of recommending articles by predicting rating scores per user with Collaborative filtering. It discusses how optimizing results on the web page had positive impact on the users [5].

Besides having so much impact on users, Collaborative Filtering is also effective on sparse and large scale of data. In the article from G. Linden, B. Smith and J. York explain how Amazon.com created high quality recommendations to their users with using item-to-item Collaborative Filtering on their web page. The system basically calculates the results every time when a user has made a purchase or rated a product. By using this method, they also tried to reduce recommending most popular product which is also very important point in recommender systems [6].

Collaborative filtering can be categorized into Model Based and Memory Based approaches and each approach has its own strengths and weaknesses. While Memory Based approaches have some issues on sparse and large data, Model based approaches overcomes this problem [31].

In addition, the main limitation of Collaborative Filtering is that they perform badly when there is fraud or profile injection attacks, which might have significant negative effects on the robustness of such systems. Because of the changes in user behavior, Collaborative Filtering algorithms must be updated continually over time. Therefore, new approaches are required that can rapidly produce high-quality recommendations, even for very large-scale problems.[13]

But also, research shows that Model Based Collaborative Filtering approaches more likely to overcome this situation [32].

Since Memory based methods have the computational problems on large data sets and sparse data, researchers direct themselves to Model based approaches to reach efficient results. Comparison of two approaches show that Model based approach on large e-commerce data was more impactful and efficient than Memory based approach [23].

Besides how all the techniques are used, the input data type is also important to be able to implement these algorithms. Recommender System algorithms accepts different kind of data inputs such as explicit, implicit or in some cases hybrid feedback [3].

The raw data we have is transactional data and as mentioned above, it is not recommended to apply algorithms to transactional data because the algorithms

like Collaborative Filtering are not working well. Therefore, it is recommended to derive transactional data to any rating related data input to be able to create real user-item interaction based on scores or ratings. There are some ways to do that. For example, from raw transactional data the information of if the user has bought an item or not can be obtained and implicit data can be derived. But this also may cause lack of information of how much the user is the most interested in which product. To prevent this, transactional data can also be derived into explicit data which gives more information about how much the user is interested in which item [33].

1.2 Objective of the Thesis

Based on the above information, the aim of thesis is to create personalized recommender systems for one of the biggest e-commerce companies in Turkey.

The successful model will be applied in production. So that, besides checking the evaluation metrics such as RMSE and MAE, the computation power, cold start problems and any other cost of the development issues will be considered while finalizing the application.

Another point to mention is that the data is an original data set of the company and has not been used before.

In the technical perspective, this thesis also aims to make comparison of Model and Memory based methods in Collaborative Filtering and tries to find best and cost-effective solution for the company.

1.3 Hypothesis

This thesis defends that RMSE score of Alternating Least Squares (ALS) method will be 30 percent less than Singular Value Decomposition (SVD) method or Memory based Collaborative filtering model results as well as the computational power.

2
THEORICAL BACKGROUND

2.1 Recommender Systems

Recommender Systems are one of the machine learning methods which generally uses ranking/rating data to predict user ratings for a specific item. In other words, it recommends new products to users with using different approaches.

Recommender Systems are commonly classified into three categories: Collaborative Filtering (CF), Content Based Filtering (CBF) and Hybrid Filtering (which is also shown in Figure 1 Recommender Systems Schema).

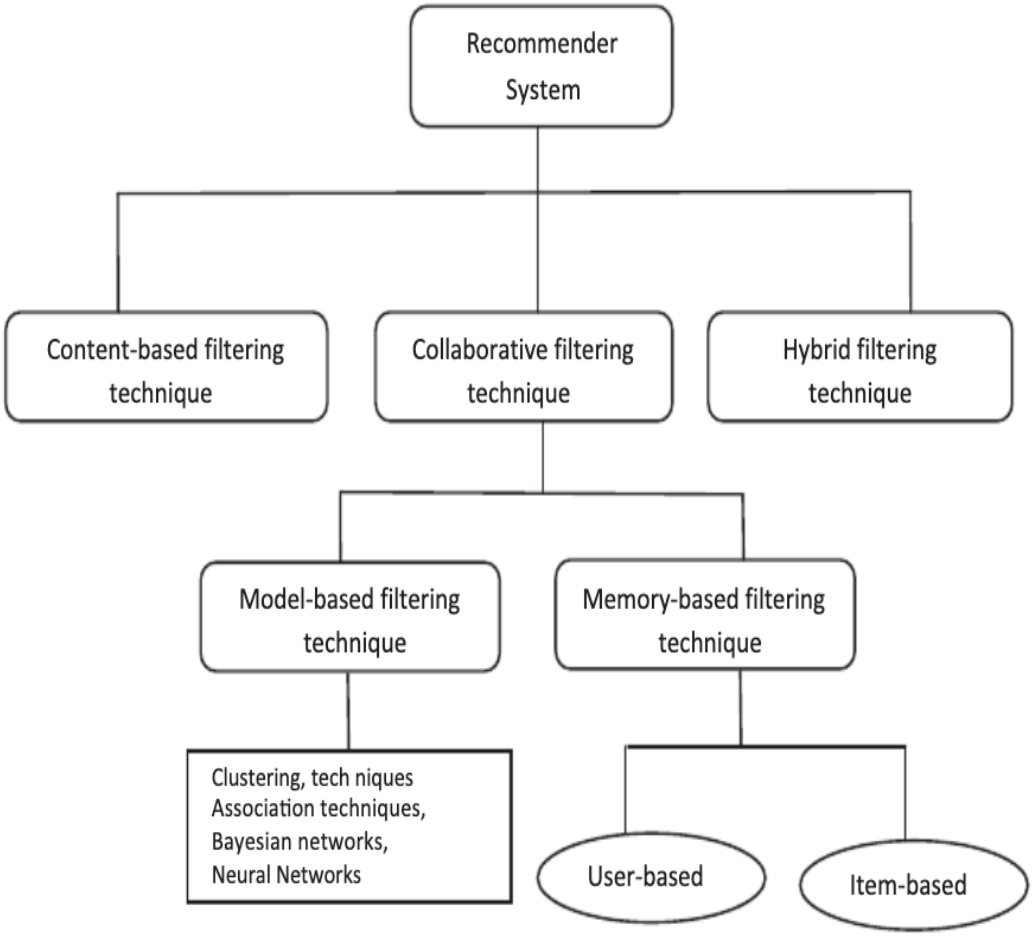

Figure 2. 1 Recommender Systems [3]

2.2 Content-Based Filtering

Content Based Collaborative Filtering techniques make use of product descriptions (contents) to generate recommendations. It creates utility matrix in the other words feature vectors by using the information of user preferences and item attributes. This matrix represents the preference of a user with a specified item by the usage of cosine or any other distance (i.e., Euclidean). Moreover, it uses user historical data to produce results from the items that user has interacted before [8] [9].

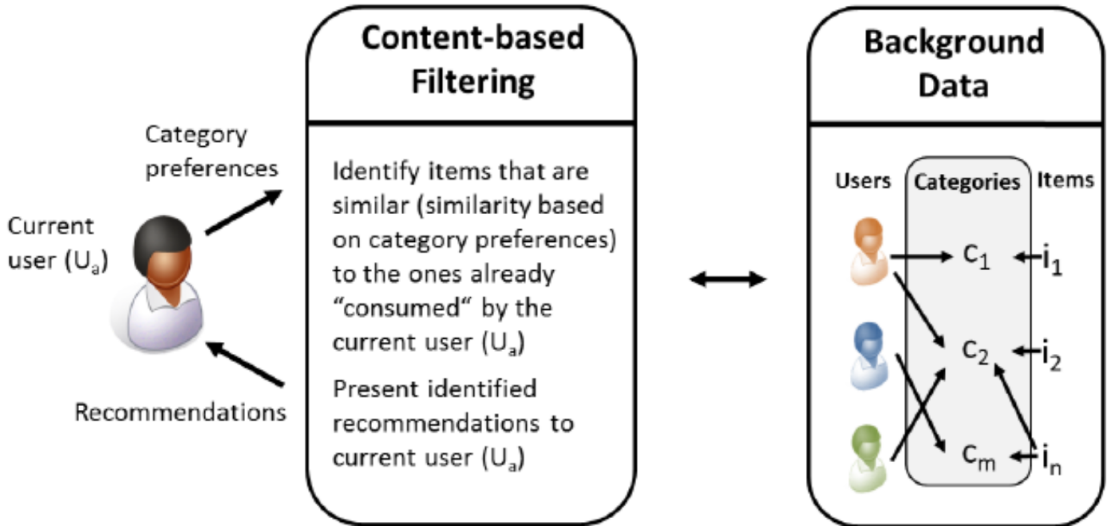

Figure 2. 2 Content-based filtering (CBF) dataflow [15]

As it illustrated in Figure 2, based on user category preferences, similar items from the same category are obtained by using historical data and later these similar products are recommended to user. To be able to apply this algorithm, products details should have been predetermined.

2.3 Collaborative Filtering

Collaborative filtering (CF) is a popular recommendation algorithm that uses the information about the past or existing user's behavior to predict which items the user will be interested in. These systems aggregate data about users' purchasing - rating- behavioral habits and make recommendations to other users based on similarity in overall patterns. The main assumption behind this method is predictions can be provided for active users rely on the information of other users' opinion. So, if users have purchased or rated some items, they will likely purchase the items which are more similar what they have purchased before or the items from other users who have similar taste with them.

Even though there are lots of use cases for all techniques, according to researches, Collaborative Filtering is one of the most popular technique that is adopted to recommend products to users in e-commerce industries [5] [6] [7]

CF can be divided into two techniques. The first one is Memory Based filtering and the second one is Model Based collaborative filtering. [16][17][18].

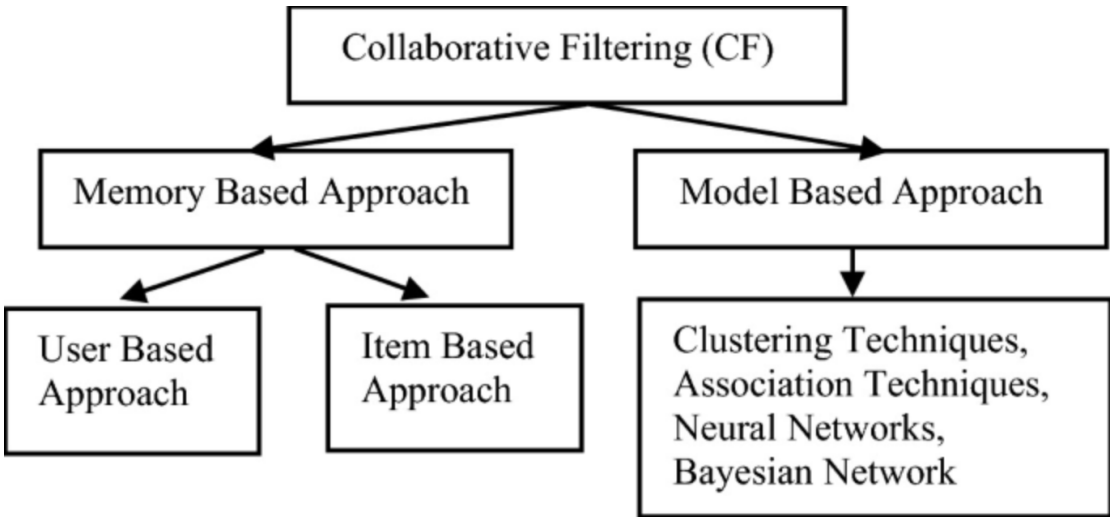

Figure 2. 3 Collaborative filtering (CF) techniques [15]

2.3.1 Memory Based Collaborative Filtering

This technique makes use of similarity algorithms to generate recommendations for users. These similarities can be either based on items or users. It gathers similar users or items with similar rankings and produce recommendations of unseen products by user. This technique uses different similarity metrics to define nearest neighbors, but the most preferred methods are Cosine Similarity and Pearson Correlation Coefficient. These methods can be applied either user-item matrix or item-item matrix depends on content of the application. Hence this technique splits into two approaches. The first approach is *User based* and the second approach is *Item based*.

Similarity Measures

1. **Cosine Similarity**

Cosine Similarity uses cosine of the angle between users or items. And the equation is:

$$similarity\ (x, y) = \cos(\theta) = \frac{x.y}{|x|.|y|} \tag{2.1}$$

2. **Pearson Correlation Coefficient**

$$r = \frac{n(\sum x.y) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2]}\sqrt{[n\sum y^2 - (\sum y)^2}} \tag{2.2}$$

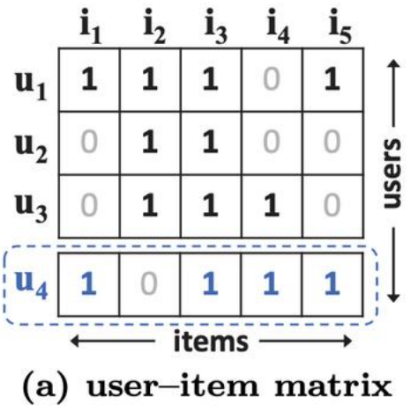

Figure 2. 4 User-Item Matrix [35]

Figure 2.4 shows how user-item matrix look like. While each row represents users, each column represents items. As it is seen from the matrix, some of the products are not purchased by users whereas some of them are. The last row shows the predicted results for User 4. To be able to make predictions for User 4, nearest neighbors should be found by using similarity measures which are mentioned above.

2.3.2 User Based Collaborative Filtering

In this technique, similarity methods are applied to the users to predict the recommended items rely on the ratings given to the specific item by the other users who have similar taste as the specific user.

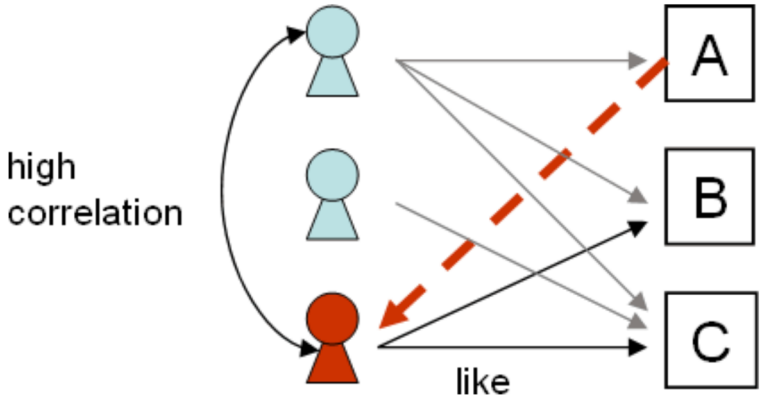

Figure 2. 5 User-based Collaborative Filtering [36]

It basically follows these main steps:

- It looks for group of users who are similar by using similarity measures

- Their ratings are used to calculate recommendations
- Based on results, recommendations are shown to specific user

2.3.3 Item Based Collaborative Filtering

In this technique, similarity methods are applied to the items instead of users to find most similar products as a first step. Then, items are recommended to the user that are most similar what user has purchased before.

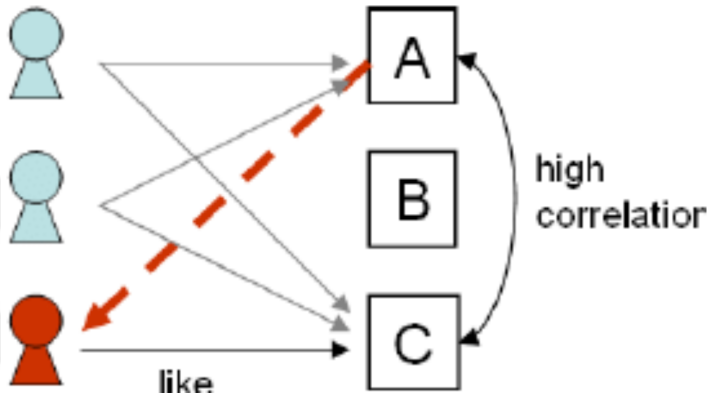

Figure 2. 6 Item-based Collaborative Filtering [36]

It basically follows these main steps:

- Opposite to User based filtering it generates item-item matrix from user-item matrix
- It looks for group of items which are similar by using similarity measures
- Based on historical data of user, new similar products are recommended.

However, the larger data becomes the harder to maintain memory based collaborative filtering approaches due to the large scale of the user-rating matrix. Because this method needs whole data to make recommendations. It faces us the real cold start problems such as speed and scalability [24]. Therefore, companies are more likely to use Model Based Collaborative Filtering technique in order to overcome this situation.

Memory-based recommendation systems are not always as fast and scalable as we would like them to be, especially in the context of actual systems that generate real-time recommendations on the basis of very large datasets. To achieve these goals, model-based recommendation systems are used.

2.3.4 Model Based Collaborative Filtering

This technique makes use of a probabilistic or predictive approach to compute the expected value of a user prediction based on the given ratings on other items. This process is performed by different Machine Learning Algorithms such as Clustering, Association Rules, Bayesian Networks or Matrix Factorization [23]. In other words, it extracts information from the given dataset and uses models to make recommendations without having to use whole dataset unlikely Memory Based approaches [24].

This technique uses three different approaches such as Clustering based, Matrix factorization based and Deep Learning.

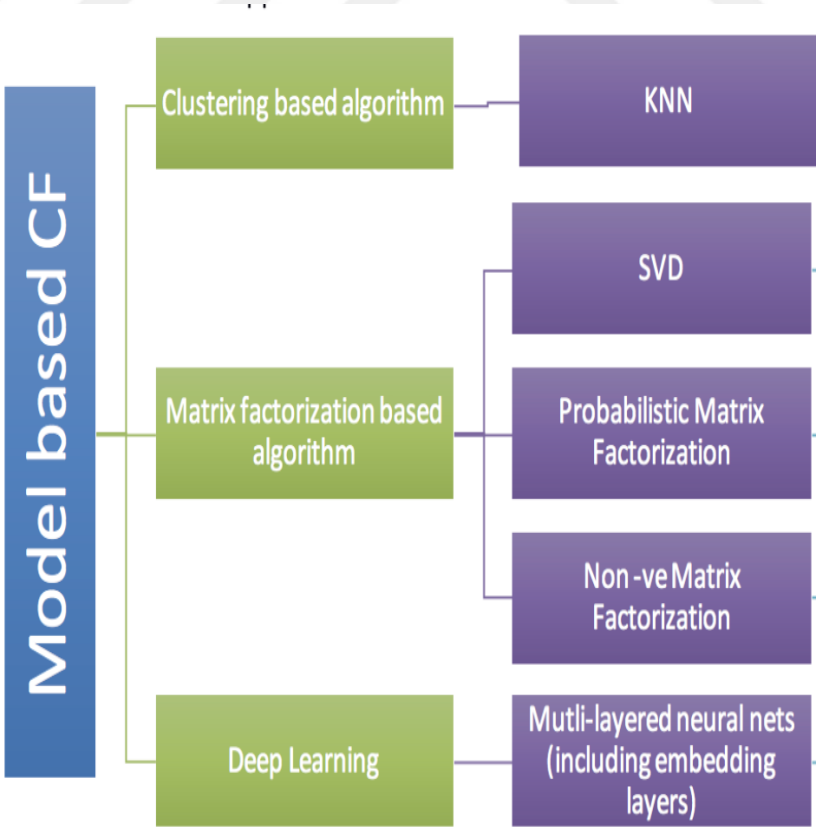

Figure 2. 7 Model based Collaborative Filtering [37]

In this thesis we will focus on Matrix Factorization Based Algorithms, and we will deep dive into ALS and SVD methods. And later on, we will see the applications of these algorithms on the data.

2.3.5 Matrix Factorization

Matrix Factorization, or in the other words matrix decomposition, is a way to reduce the main matrix into smaller matrixes and makes easier to perform on it. In recommender systems, the fundamental assumption behind matrix factorization is that each item is associated with an item vector and each user is associated with a user vector. And when these two vectors are multiplied, high interest on the specific item can be measured for users.

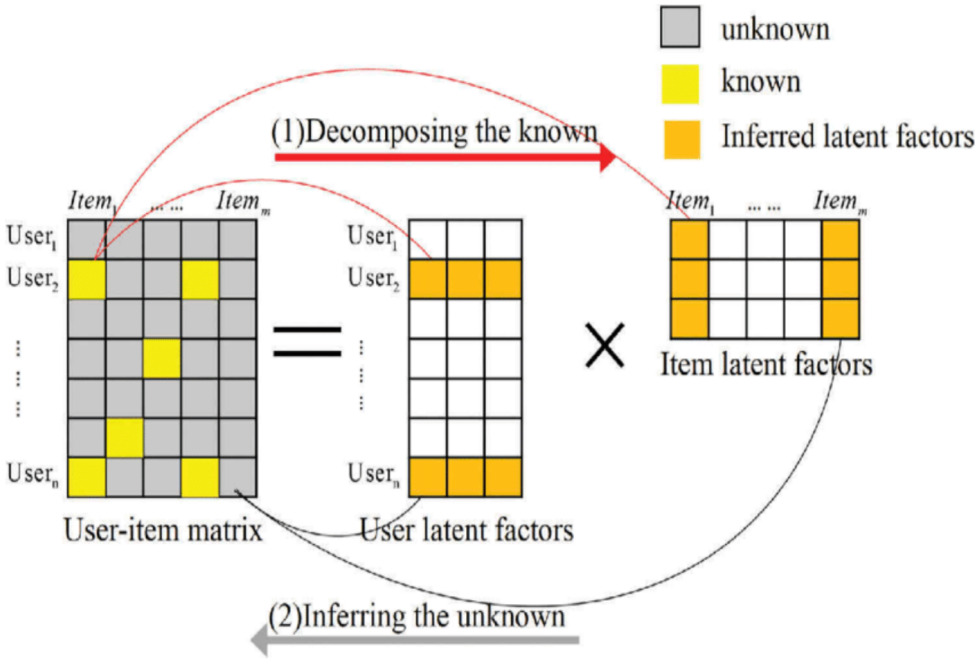

Figure 2. 8 Matrix Factorization [38]

According to the above information and illustration in Figure 2.8:

- User-item matrix consist of both observed and unobserved data, and it is decomposed into two matrices by using matrix factorization.
- Each user in user matrix is expressed with k number of factors. This matrix says that each user how related to the specific factor.

- Each item in item matrix is also expressed with k number of factors. Similarly, item matrix says how related is the item with the specified factor.
- As a third step, if we multiply each factor of the user by the related factor of the item from the item matrix, we can reach the most related items for the user.

The problem is that how to form these factors efficiently. Because the number of factors is unknown and selected number of factors might not cover the effect on overall ratings. This number can be set at the beginning and generally it is preferred to fix to a small number of k (e.g., k=10) to be able to minimize the cost and loss function. In addition to defining number of factors, it is also important to minimize the Root Mean Square Error (RMSE) along with "regularization term". Regularization term should be picked carefully to be able to prevent from overfitting.

In final, matrix starts learning ratings for all the factors for both user and item matrices after defining the number of factors and regularization parameter. In attempt to reach best model, the process should be optimized. And this optimization can be done by minimizing the loss function. Derivatives are used to minimize this loss function. This function can be minimized by using either Alternating Least Square (ALS), Gradient Descent or Singular Value Decomposition (SVD) approaches. However, both approaches have their own strengths and weaknesses. For instance, while ALS is hard to optimize, Gradient Descent can be slow [25]. Since the cost of time also important in the learning process we will only cover Alternating Least Squares in this thesis.

2.4 Hybrid Filtering

Hybrid recommender systems combine two or more recommendation methods to gain better performance.

The most common techniques that are combined are Content Based Filtering and Collaborative Filtering.

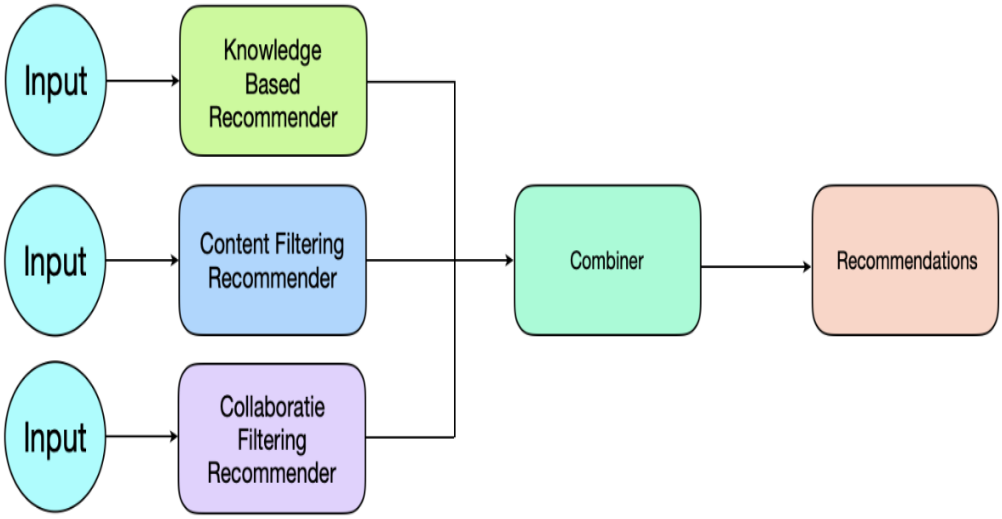

Figure 2. 9 Hybrid Filtering [39]

3
METHODS

3.1 Alternating Least Squares

ALS Matrix Factorization attempts to estimate the ratings matrix R as the product of two lower-rank matrices.

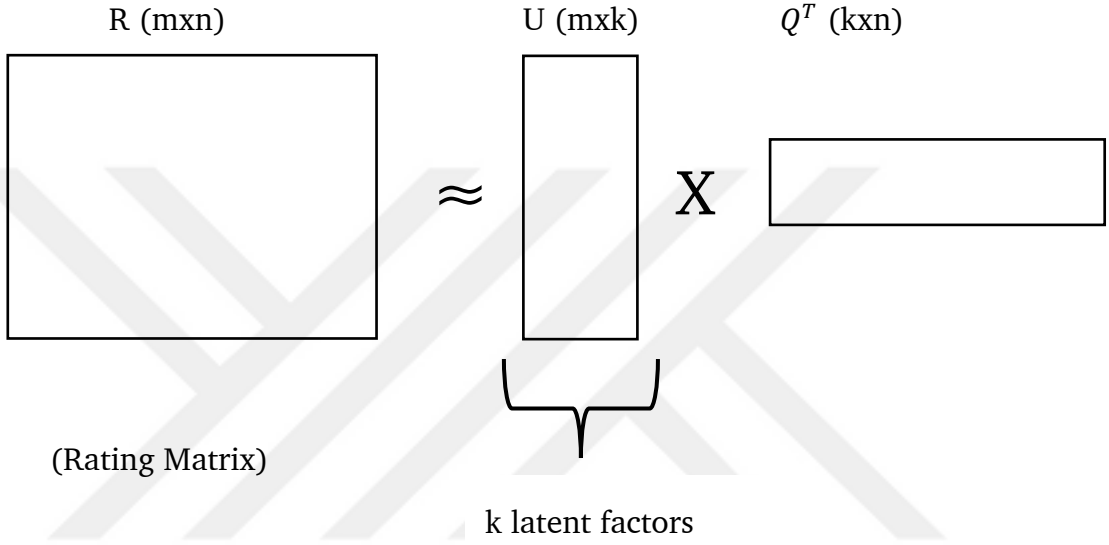

Figure 3. 1 Matrix Factorization

Assume that overall user-item rating matrix is represented as R, user latent factor is represented as U and item latent factor is represented as Q.

This brings us the equation of:

$$R \approx U \, x \, Q^T \tag{3.1}$$

And the Error function of this equation is:

$$Error_{ij} = \underbrace{\sum w_{ij}}_{(1)} \underbrace{(R_{ij} - u_i . q_j^T)}_{(2)} + \underbrace{\lambda(||U||_2 + ||Q||_2)}_{(3)} \tag{3.2}$$

(1) Completion Term

where w_{ij} $\begin{cases} 1 \; R_{ij} \text{ is known} \\ 0 \; R_{ij} \text{ is unknown} \end{cases}$

(2) Cost Function

(3) Regularization Term

ALS is applied to minimize the error function. It iterates two step optimizations. For instance, starting with iteration steps for U, it first fixes Q and solves for U, and in the second iteration, this time it fixes U while solving for Q. In every iteration each step the cost function can either decrease or stay unchanged, but never increase [26].

As soon as the error term is minimized U and Q matrixes are multiplied back together and the predictions are made.

3.2 Singular Value Decomposition

Singular value decomposition (SVD) is known as one of the matrix factorization and unsupervised machine learning methods that reduces dimensionality and decomposes the matrices [27].

It decomposes the matrix into three matrices. Therefore, SVD can be formulated as:

$$A = U \sum V^T \tag{3.3}$$

In the equation above where A represents overall user-item matrix, U represents user matrix and V represents item matrix. In addition to general Matrix Factorization formula in Equation 1, SVD weights the factors as in the Equation 2 which is represented as $\sum$.

Any pair of orthogonal vectors in two-dimensional space forms a basis for that space. In this case, if we can call the orthogonal vectors in the input space v_1 and v_2. After applying a matrix transformation, A to these two vectors, we get Av_1 and Av_2.Furthermore, let's decompose these two transformed vectors

into unit vectors, u_1 and u_2, times their respective magnitudes, $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$. In other words:

$$Av_1 = u_1\boldsymbol{\sigma}_1 \tag{3.4}$$

$$Av_2 = u_2\boldsymbol{\sigma}_2 \tag{3.5}$$

Secondly, any vector x can be described using the basis vectors v1 and v2 in the following way,

$$x = (x.v_1)v_1 + (x.v_2)v_2 \tag{3.6}$$

and this brings us to the similar formulation with the basis vectors:

$$x = x_1\begin{bmatrix}1\\0\end{bmatrix} + x_2\begin{bmatrix}0\\1\end{bmatrix} \tag{3.7}$$

Next, left-multiply both sides of Equation 3.4 by the matrix transformation A:

$$Ax = (x.v_1)Av_1 + (x.v_2)Av_2 \tag{3.8}$$

Next, we can rewrite Av_i as $u_i\boldsymbol{\sigma}_i$:

$$Ax = (x.v_1)u_1\boldsymbol{\sigma}_1 + (x.v_2)u_2\boldsymbol{\sigma}_2 \tag{3.9}$$

Now we switch ordering of the above equation by the following equation:

$$Ax = u_1\boldsymbol{\sigma}_1 v_1^T x + u_2\boldsymbol{\sigma}_2 v_2^T x \tag{3.10}$$

We can remove **x** from both sides of the equation

$$A = u_1\boldsymbol{\sigma}_1 v_1^T + u_2\boldsymbol{\sigma}_2 v_2^T \tag{3.11}$$

In final, it gives us the SVD equation in the equation 3.2:

$$A = \underbrace{[u_1 \quad u_2]}_{U}\underbrace{\begin{bmatrix}\boldsymbol{\sigma}_1 & 0\\ 0 & \boldsymbol{\sigma}_2\end{bmatrix}}_{\sum}\underbrace{\begin{bmatrix}v_1^T\\ v_2^T\end{bmatrix}}_{V^T} \tag{3.12}$$

3.3 Data Transformation

To be able to apply algorithms, the data will be derived into explicit data with using the equation below [33]:

$$AP(u,i) = \ln\left(\frac{Purchase\ count\ of\ the\ item\ by\ spesific\ user}{The\ overall\ number\ of\ transactions\ of\ user} + 1\right) \tag{3.13}$$

$$RP(u,i) = \frac{AP(u,i)}{Max(AP(c,i))} \tag{3.14}$$

$$Implicit\ rating(u,i) = Round\ up\ (5\ x\ RP(u,i)) \tag{3.15}$$

As a result of this transformation, our transactional data is derived into rating data in a scale between 1-5. So, the higher purchase count, the higher the rating based on user preferences.

3.4 Evaluation Metrics

Since the data has the information of explicit ratings, the evaluation metrics for the model comparison will be Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

3.4.1 Root Mean Square Error

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors.

RMSE can be formulized as:

$$RMSE = \sqrt{\sum_{i}^{n} \frac{(X' - X)^2}{n}} \tag{3.16}$$

It is important have smaller RMSE to be able to define the results are reasonable and model is successful.

3.4.2 Mean Squared Error

The Mean Absolute Error (MAE) is the average of all absolute errors.

MAE can be formulized as:

$$MAE = \frac{1}{n}\sum_{i}^{n} |x_i - x| \tag{3.17}$$

It is also should be a small number as RMSE to be able to call the results are successful.

4
RESULTS AND DISCUSSION

4.1 Results

In this thesis, we aim to develop personalized recommender systems for users of an e-commerce company. As mentioned above sections we will apply different recommender system approaches and examine the results. According to results, the decision of which model will be made, and that model will be applied in the production.

4.1.1 Data

For this examination, two months of transactional data has been used. The data has been collected by the dates between 2021-07-01 and 2021-08-31 and it contains approximately 1,4 M unique users and 4M transactions with the product details. There are 1606 unique products from 137 categories.

Figure 4.1 shows the top 20 most selling products in the baskets. Since these products tend to be suggested more, we will take it consideration when the results are being evaluated. In addition, the products which only ordered once will be removed from the dataset in order to reduce sparsity. According to this information sparsity is calculated as 0.68% which says that our data is sparse as it is expected but not as much as movie recommender systems. So, the results expected to be more accurate.

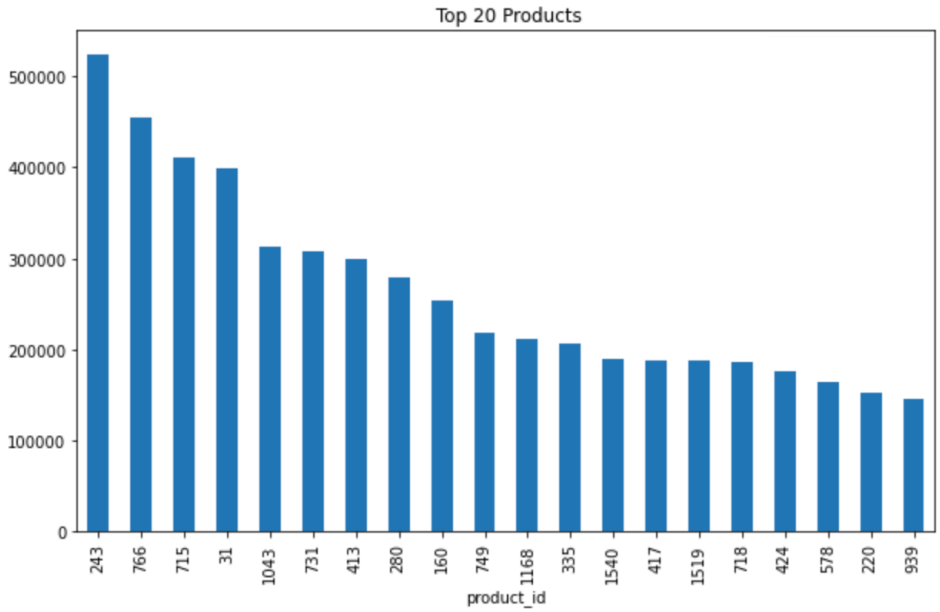

Figure 4. 1 Top 20 Most Selling Products

Figure 4.2 shows the most preferred categories under transactions. There is a huge difference between the top 10 categories and the rest categories can be obtained. Therefore, we can say that the distribution is biased.

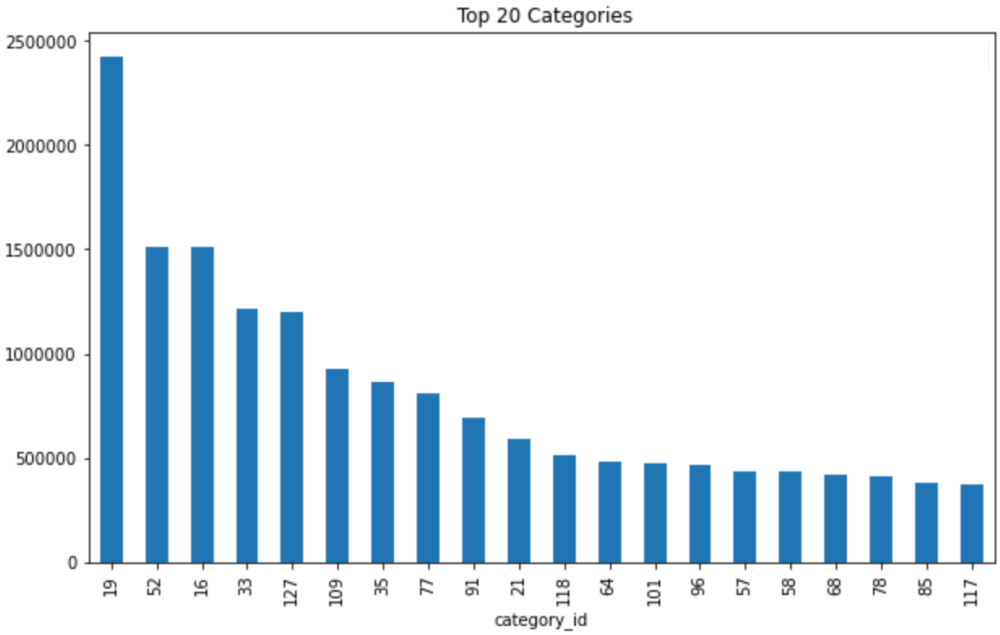

Figure 4. 2 Top 20 Categories

Table 4.1 contains descriptive statistics to get a better understanding of the data. These numbers show that users have 3 orders in average while users who have

more than 1 orders (repetitive users) have 3.04 orders in this specified time period. This also demonstrates that users who only gave 1 order represents only small amount (%2.03) of all users. Therefore, they will not be eliminated from our data set. In addition, there are 4.53 distinct products in each basket and 10.85 products have been consumed in average per user.

Table 4. 1 Descriptive Statistics

Average Order Per User	3.00
Average Order Per Repetitive User	3.04
Average product In Basket	4.53
Average product per User	10.85

Since the raw data does not have any information about ratings and only consists of user purchase data, it is needed to be transmitted into rating data to be able to apply algorithms properly. After transforming purchase data to explicit data with the equations Eq1, Eq2 and Eq3 in chapter 3, the rating data distribution is illustrated in Figure 4.3. Even though specified equation has been applied to purchase data, the ratings distributed more around 1, 2, 3, 5, but 4. We do not take it consideration for now.

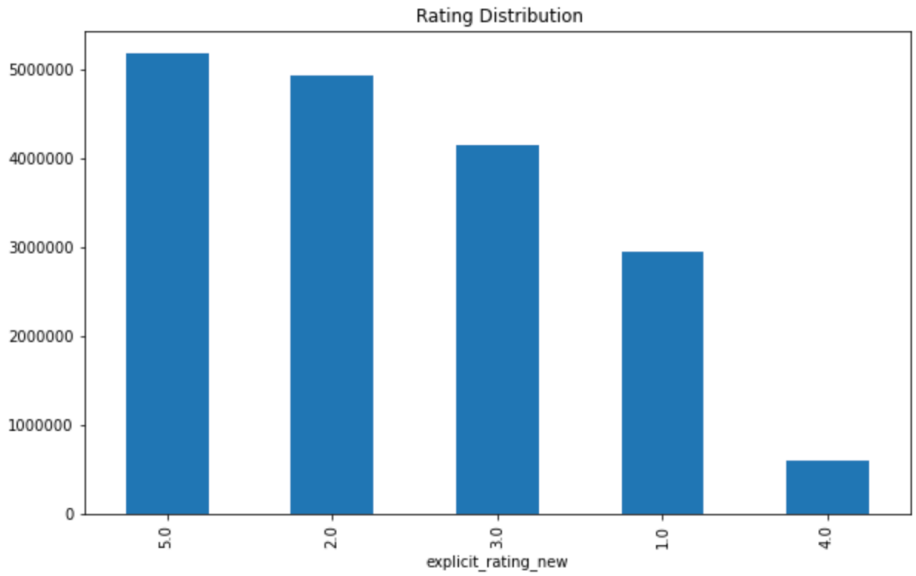

Figure 4. 3 Rating Distribution

After having information about the data and transforming the purchase data to explicit rating data, methods become available to apply. Firstly, the data is devided into train and test data set with the percentages of 70% of train and 30% of test.

4.1.2 Application

As mentioned in section 1 we will compare Model and Memory based approches and select the best one to implement in production. The first approach is Memory based Collaborative Filtering. We will compare two results from different similarity measures and interpret the results. According the results the calculated RMSE for the Collaborative Filtering with using Cosine Similarity is calculated as 3.24 (Figure 4.4) and for Pearson smilarity it is calculated as 1.39 (Figure 4.5).

```
Overall RMSE: 3.245170385202053
```

Figure 4. 4 Collaborative Filtering w/Cosine Similarity RMSE

```
Overall RMSE: 1.3986988674546674
```

Figure 4. 5 Collaborative Filtering w/Pearson Similarity RMSE

According to both Collaborative Filtering results, RMSE values are higher than 1 and this makes the results unreliable and inaccurate. And since the picked model will become live in production, the computation power and the duration of the calculation of these models are another important factors. Therefore, even though the RMSE results were meaningful, we wouldn't move on with Memory based algorithms because it takes lots of time and requires lots of memory to calculate and get results.

Secondly, after having bad results from Memory based approaches, Model based approaches are applied and the results are interpretted. To be able to understand which Memory Based approach is giving better results, SVD and ALS Matrix Factorization algorithms are applied and results are interpretted. But, since the regression parameters needs to be defined in order to pick the best model, K-fold Cross-Validation method has been applied to be able to interpret results.

SVD Matrix Factorization Results;

Table 4. 2 Evaluation Results of SVD

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std
RMSE	1.0427	1.0449	1.0446	1.0431	1.0448	1.0440	0.0009
MAE	0.8324	0.8348	0.8339	0.8328	0.8347	0.8337	0.0010

ALS Matrix Factorization Results;

```
RMSE = 0.7983350032996621

MAE = 0.5151503773359896
```

Figure 4. 6 Evaluation Results of ALS

In Table 4. 2 and Figure 4.6, SVD and ALS matrix factorization methods' results are illustrated, respectively. According to results, while ALS method has RMSE

value with 0.79 and MAE value with 0.51, SVD has 1.04 RMSE valuse and 0.83 MAE value.

As a result from the information above, ALS Matrix Factorization method has performed better than SVD based on RMSE and MAE scores. But still, there is another step to check before we say that these results are menaingful and ready to production. Because of that highly recommended products are also checked if they are overlapping with the most selling products in the raw data and following results have ben examined;

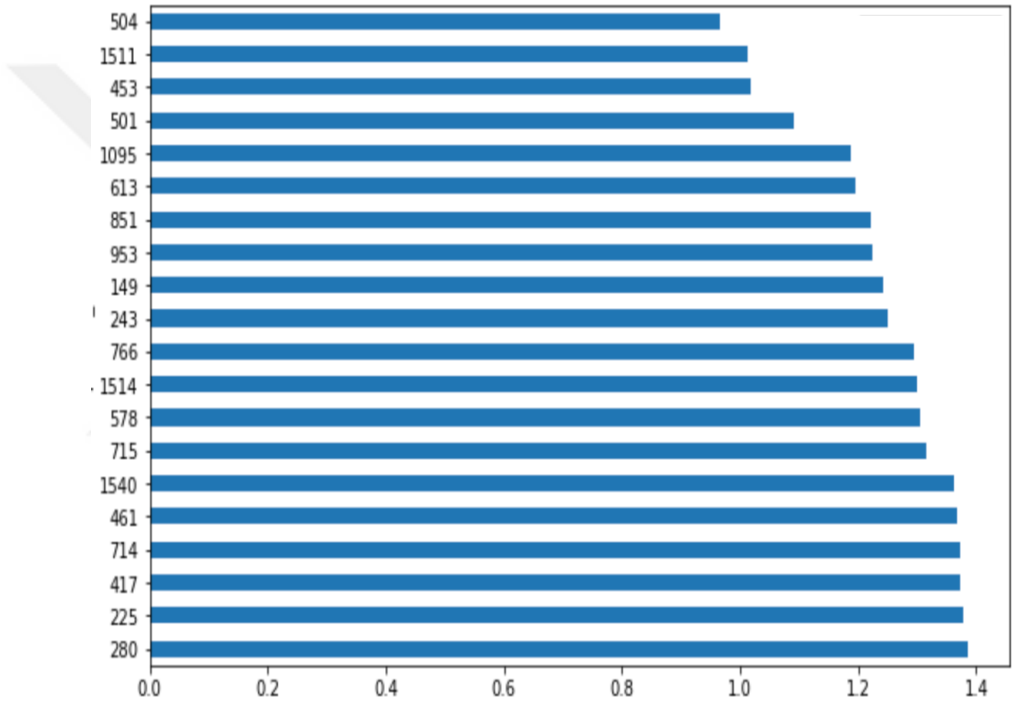

Figure 4. 7 Top 20 Most Recommended Products

In the figure 4.7, Recommended top 20 products are different from top 20 popular products and 43% of recommended products are from the same categories but different products that users have ordered before.

5
CONCLUSION

According to results, Model based approaches worked better on big and sparse data than Memory based approaches. As discussed in the introduction part, even though Memory Based Collaborative Filtering performs well in some cases they fail on big and sparse data. In addition, they need high memory and good computation power to perform calculations. On the contrary, Model Based approaches do not need the whole data to calculate results that is why it works faster even the data big and sparse. When we compare the RMSE and MAE results of each Model and Memory Based Collaborative results, ALS Matrix Factorization method has the lowest RMSE and MAE results with the numbers 0.79 and 0.51 while all the other methods have higher RMSE which are higher than 1. But as far as it is also discussed in chapter 1, RMSE and MAE is not enough to say the results are successful. We also need to be sure not to recommend most selling products and recommend the products from the same category of particular user. As it can be observed from Figure 4.8, most recommended 20 products are different from most selling products in Figure 4.1. In addition, 43% of recommended products are from the preferred categories by users. In conclusion, ALS Matrix Factorization method gives better results based on the recommendation results and RMSE, MAE values as well as the less computational power.

REFERENCES

[1] Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. Communications of the ACM, 35(12), 61-70.

[2] Schafer, J. B., Konstan, J., & Riedl, J. (1999, November). Recommender systems in e-commerce. In Proceedings of the 1st ACM conference on Electronic commerce (pp. 158-166).

[3] Isinkaye, F. O., Folajimi, Y. O., & Ojokoh, B. A. (2015). Recommendation systems: Principles, methods and evaluation. Egyptian informatics journal, 16(3), 261-273.

[4] Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE transactions on knowledge and data engineering, 17(6), 734-749.

[5] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994, October). Grouplens: An open architecture for collaborative filtering of netnews. In Proceedings of the 1994 ACM conference on Computer supported cooperative work (pp. 175-186).

[6] Linden, G., Smith, B., & York, J. (2003). Amazon. com recommendations: Item-to-item collaborative filtering. IEEE Internet computing, 7(1), 76-80.

[7] Greenstein-Messica, A., & Rokach, L. (2018). Personal price aware multi-seller recommender system: Evidence from eBay. Knowledge-Based Systems, 150, 14-26.

[8] Barragáns-Martínez, A. B., Costa-Montenegro, E., Burguillo, J. C., Rey-López, M., Mikic-Fonte, F. A., & Peleteiro, A. (2010). A hybrid content-based and item-based collaborative filtering approach to recommend TV programs enhanced with singular value decomposition. Information Sciences, 180(22), 4290-4311.

[9] Pal, A., Parhi, P., & Aggarwal, M. (2017, August). An improved content based collaborative filtering algorithm for movie recommendations. In 2017 tenth international conference on contemporary computing (IC3) (pp. 1-3). IEEE.

[10] Chen, H. W., Wu, Y. L., Hor, M. K., & Tang, C. Y. (2017, July). Fully content-based movie recommender system with feature extraction using neural network. In 2017 International conference on machine learning and cybernetics (ICMLC) (Vol. 2, pp. 504-509). IEEE.

[11] Van den Oord, A., Dieleman, S., & Schrauwen, B. (2013). Deep content-based music recommendation. Advances in neural information processing systems, 26.

[12] Burke, R. (2007). Hybrid web recommender systems. The adaptive web, 377-408.

[13] Anam-Ro, Seoungbuk-Gu, Anam-dong, Seoul, Content-based filtering for recommendation systems using multiattribute networks

[14] Debnath, S., Ganguly, N., & Mitra, P. (2008, April). Feature weighting in content based recommendation system using social network analysis. In Proceedings of the 17th international conference on World Wide Web (pp. 1041-1042).

[15] Felfernig, A., Jeran, M., Ninaus, G., Reinfrank, F., Reiterer, S., & Stettinger, M. (2014). Basic approaches in recommendation systems. In Recommendation Systems in Software Engineering (pp. 15-37). Springer, Berlin, Heidelberg.

[16] Recommender Systems: An Introduction By Dietmar Jannach,Markus Zanker, Alexander Felfernig,Gerhard Friedrich

[17] Integrating Knowledge-based and Collaborative-filtering Recommender Systems, Robin Burke

[18] Collaborative Filtering Recommender Systems By Michael D. Ekstrand, John T. Riedl, Joseph A.Konstan

[19] Aly, M., Hatch, A., Josifovski, V., & Narayanan, V. K. (2012, April). Web-scale user modeling for targeting. In Proceedings of the 21st international conference on world wide web (pp. 3-12).

[20] S. Breese, D.Heckerman, and C.Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In Proceedings of the Fourteenth Conference on Uncertainty in Artifical Intelligence, 1998.

[21] M.Deshpande and G. Karypis. Item-based top-n recommendation algorithms. ACM Trans. Inf. Syst., 22(1):143-177, 2004.

[22] B.M. Sarwar, G. Karypis, J.A. Konstan, and J. Reidl. Item-based collaborative filtering recommendation algorithms. In Proceedings of the 10th International World Wide Web Conference, pages 285-295, 2001.

[23] Aditya, P. H., Budi, I., & Munajat, Q. (2016, October). A comparative analysis of memory-based and model-based collaborative filtering on the implementation of recommender system for E-commerce in Indonesia: A case study PT X. In 2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS) (pp. 303-308). IEEE.

[24] Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001, April). Item-based collaborative filtering recommendation algorithms. In Proceedings of the 10th international conference on World Wide Web (pp. 285-295).

[25] Zadeh, R., Li, H., He, B., Lublin, M., & Perez, Y. (2015). Cme 323: Distributed algorithms and optimization, spring 2015. University Lecture.

[26] Pilászy, I., Zibriczky, D., & Tikk, D. (2010, September). Fast als-based matrix factorization for explicit and implicit feedback datasets. In Proceedings of the fourth ACM conference on Recommender systems (pp. 71-78).

[27] Chai, Z. Y., Li, Y. L., Han, Y. M., & Zhu, S. F. (2018). Recommendation system based on singular value decomposition and multi-objective immune optimization. IEEE Access, 7, 6060-6071.

[28] Geetha, G., Safa, M., Fancy, C., & Saranya, D. (2018, April). A hybrid approach using collaborative filtering and content based filtering for recommender system. In Journal of Physics: Conference Series (Vol. 1000, No. 1, p. 012101). IOP Publishing.

[29] Resnick, P. and Varian, H.R., Recommender Systems, in Resnick and Varian (eds.), 56-58.

[30] Goy, A., Ardissono, L., & Petrone, G. (2007). Personalization in e-commerce applications. In The adaptive web (pp. 485-520). Springer, Berlin, Heidelberg.

[31] Gong, S., Ye, H., & Tan, H. (2009, May). Combining memory-based and model-based collaborative filtering in recommender system. In 2009 Pacific-Asia Conference on Circuits, Communications and Systems (pp. 690-693). IEEE.

[32] Cheng, Z., & Hurley, N. (2009, October). Effective diverse and obfuscated attacks on model-based recommender systems. In Proceedings of the third ACM conference on Recommender systems (pp. 141-148).

[33] Choi, K., Yoo, D., Kim, G., & Suh, Y. (2012). A hybrid online-product recommendation system: Combining implicit rating-based collaborative filtering and sequential pattern analysis. electronic commerce research and applications, 11(4), 309-317.

[34] Najmani, K., Benlahmar, E. H., Sael, N., & Zellou, A. (2020, December). Collaborative Filtering Approach: A Review of Recent Research. In International Conference on Advanced Intelligent Systems for Sustainable Development (pp. 151-163). Springer, Cham.

[35] He, X., Liao, L., Zhang, H., Nie, L., Hu, X., & Chua, T. S. (2017, April). Neural collaborative filtering. In Proceedings of the 26th international conference on world wide web (pp. 173-182).

[36] Kalz, M., Drachsler, H., Van Bruggen, J., Hummel, H., & Koper, R. (2008). Wayfinding services for open educational practices. International Journal of Emerging Technologies in Learning (iJET), 3(2), 24-28.

[37] Mohamed, M. H., Khafagy, M. H., Elbeh, H., & Abdalla, A. M. (2019). Sparsity and cold start recommendation system challenges solved by hybrid feedback. International Journal of Engineering Research and Technology, 12(12), 2735-2742.

[38] Cai, L., Xu, J., Liu, J., & Pei, T. (2018). Integrating spatial and temporal contexts into a factorization model for POI recommendation. International Journal of Geographical Information Science, 32(3), 524-546.

[39] Lesmeister, C., & Chinnamgari, S. K. (2019). Advanced machine learning with R: tackle data analytics and machine learning challenges and build complex applications with R 3.5. Packt Publishing Ltd.

PUBLICATIONS FROM THE THESIS

Conference Papers

1. A. Levent and G. Kemalbay, “Developing A Personalized Recomender Sytem in an E-Commerce Firm”, 7. International Congress on Engineering and Technology, Proceedings Book, page 823, 16-18 April 2022, Istanbul, Turkey.