MDSAA

Master Degree Program in Data Science and Advanced Analytics

Major in Business Analytics

Online Retailer Recommender System

Group C

Catarina Moreira, number: 20201034

Luísa Barral, number: 20201045

Madalena Valério, number: 20200657

Yu Song, number: 20200572

May, 2021

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

INDEX

INTRODUCTION **1**

BUSINESS UNDERSTANDING **1**

- Background 1
- Business Objectives 1
- Business Success criteria 1
- Situation assessment 2
- Determine Data Mining goals 2

PREDICTIVE ANALYTICS PROCESS **2**

- Data understanding 2
- Data preparation 7
- Modeling 8
- Evaluation 8

RESULTS EVALUATION **9**

DEPLOYMENT AND MAINTENANCE PLANS **13**

CONCLUSIONS **13**

- Considerations for model improvement 13

REFERENCES **14**

1. INTRODUCTION

During the last few decades, with the rise of Youtube, Amazon, Netflix and many other such web services, recommender systems have taken more and more place in our lives. In a very general way, recommender systems are algorithms aimed at suggesting relevant items to users (for example, items being movies to watch, text to read, products to buy or anything else depending on industries). So, from providing advice on songs for people to try, suggesting books for them to read, or finding clothes to buy, recommender systems have greatly improved the ability of customers to make choices more easily.

Given the number of possible choices available, especially for online shopping, it can really make a difference.

As a proof of the importance of these systems, we can mention that firstly, in Netflix, 2/3 of the movies watched are recommended, secondly, in Google, news recommendations improved click-through rate by 38% and lastly, for Amazon, 35% of sales come from recommendations.

This project is a business case of practicing recommender systems for an online shop in the UK to build a recommendation system to improve the shopping experience of customers and solving a cold start problem which suggests possible products to new customers. As a result we are invited to develop a Recommender system based on the transaction dataset from Nov 2010 to Nov 2011, we will design a recommender system with implicit feedback and provide recommendations.

2. BUSINESS UNDERSTANDING

2.1. Background

ManyGiftsUK started in 1981 by selling gifts on direct mailing catalogues which were ordered by phone. The company changed its selling and distributing to online sales 2 years ago that ships all over the world. With more and more customers buying a huge variety of goods, the company hopes to recommend relevant products to customers in a better way according to their purchasing behavior to have an obvious increase in sales.

2.2. Business Objectives

The company hopes to have a recommendation system of its own to provide its customers with more convenient shopping experience and increase sales. And they hope to increase profits by recommending products which new customers might purchase when visiting the website.

2.3. Business Success criteria

A useful outcome of this project would be to design a recommender system with implicit feedback and to provide recommendations. So that when customers view the page of items, our system would successfully recommend most related items for them to choose and buy together, which saves a lot of

time and improves the shopping experience. Also, we need to provide a solution to the cold start problem: offer relevant products to new customers.

2.4. Situation assessment

Our team was given 1 csv file that contains all the transactions occurring between 01/12/2010 and 09/12/2011. The dataset has simply 8 variables of 25900 valid transactions of 541909 instances, which include 4070 distinct products provided to 4372 customers from 38 countries. Specifically, the variable 'CustomerID' , 'InvoiceNo' and 'StockCode' show a 5-digit integral number uniquely assigned to each customer, each transaction, and distinct product respectively.

2.5. Determine Data Mining goals

AS for explicit and implicit data for each user/item interaction, we found no explicit ones to use. So we will focus on implicit feedback. In Data Mining terms, we intend to apply the alternating least squares (ALS) algorithm for collaborative filtering. Our goal is to find out the hidden features of a large matrix of user/item interactions and connect them to each other in a much smaller matrix of user features and item features.

3. PREDICTIVE ANALYTICS PROCESS

3.1. Data understanding

We started by reading the description of each variable from the dataset provided to us. After that, we loaded the data file (retail.csv) to a pandas Dataframe. Our dataset has 541909 rows and 8 columns, and it contains all the transactions occurring between 01/12/2010 and 09/12/2011. We have 4372 Customers from 38 different countries, 4070 item codes and 25900 distinct Invoice numbers that are assigned to each transaction. With some exploration, we identified that a couple of data types were incorrect ('InvoiceDate' and 'CustomerID'), we found that we have missing values in 'Description' and 'CustomerID', we also observed some duplicates and there were negative values for the 'Quantity' and 'UnitPrice' variables, which happen for items that were returned.

After doing this exploration, we decided to separate the non_metric_features from metric_features. In this way, we were able to do some visualizations in order to verify the correlation between the metric features.

As we can see in the pairplot of ***Figure 1*** and in the correlation matrix of ***Figure 2***, we can verify that there are not any high correlation between the metric features.

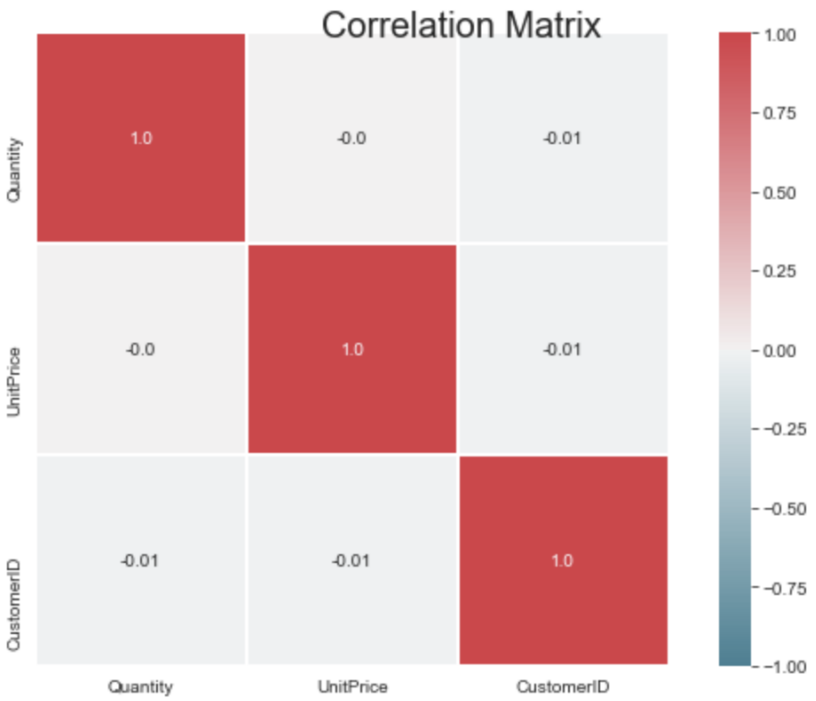

Figure 1 - Correlation Matrix with the Metric Features

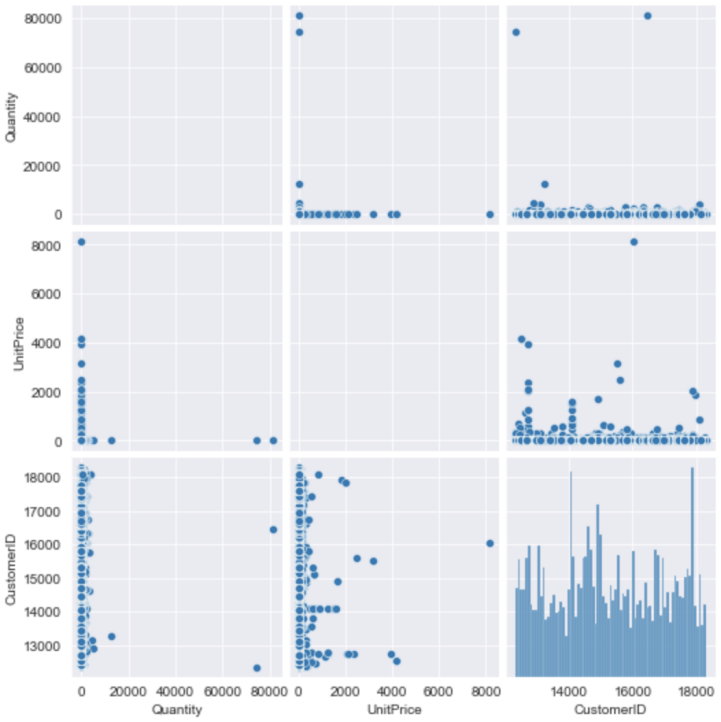

Figure 2 - Matrix with the Metric Features

Following this, we began the exploratory data analysis to get our initial insights of the data. Firstly, in ***Figure 3,*** we found the top 10 that were refunded(these items have the quantity below 0).

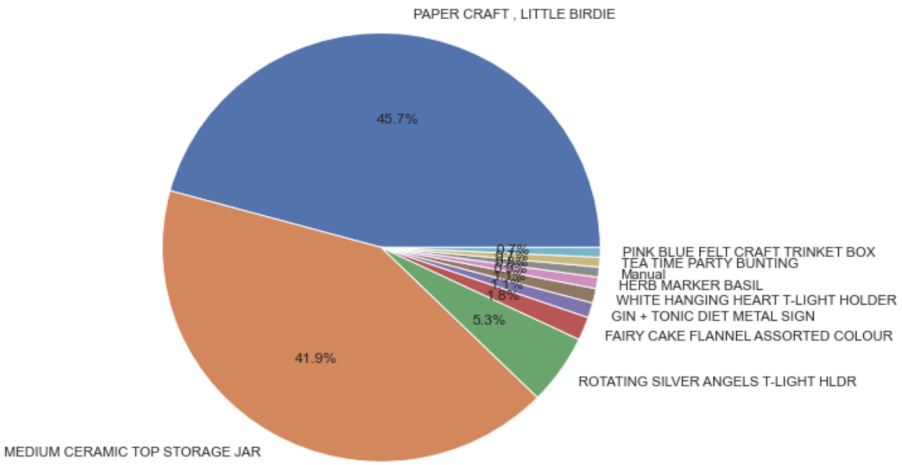

Figure 3 - Top Items that were Refunded

Through this pie chart, we can observe that the Paper Craft, Little Birdie along with medium ceramic top storage jar are the most refunded.

After exploring this pie chart, we decided to create some auxiliary columns. We create the 'Country_map' column to see the total number of people per country, the 'Price' column that was the multiplication between 'Quantity' and 'UnitPrice', the 'Date' column with each date (without the time), the 'Time' column with each time, the 'Month' column that was a column with the months (integer) and the 'MonthName' column with the months names.

Firstly, we started to create a barplot where it was possible to see the quantity of unique descriptions in each stock code.

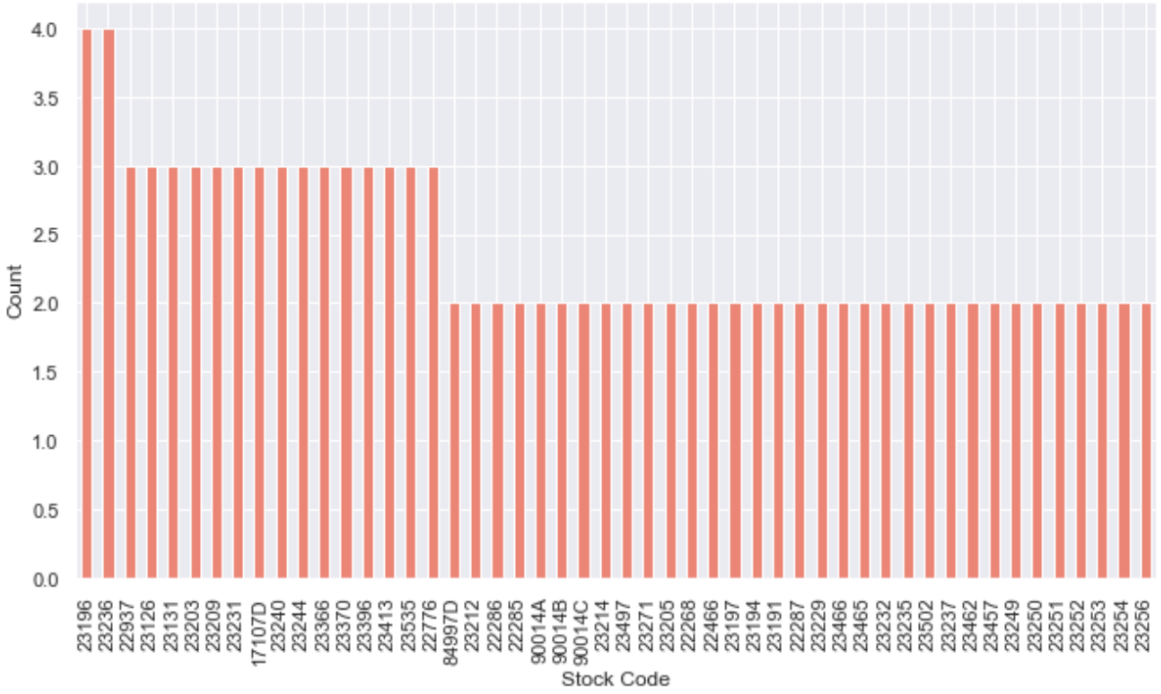

Figure 4 - Number of Unique Descriptions in each Stock Code

Secondly, in ***Figure 5***, using wordcloud library, we can see the most common gifts description in the dastaset.

The words that appear with the largest size, are the ones that are the most common.

Figure 5 - The Most Common Gifts Description

After doing this last visualization, we also wanted to explore the data grouping it into groups.

So, to do Figure 6, Figure 7, Figure 8, we grouped the data by country and we did these 3 visualizations. Through the first one, we could see the revenue per country, through the second one, we could see the number of customers per country and through the last one we could see the quality of gifts sold per country.

For easy reading, exploration and to be able to compare countries better, we used the log10 scale.

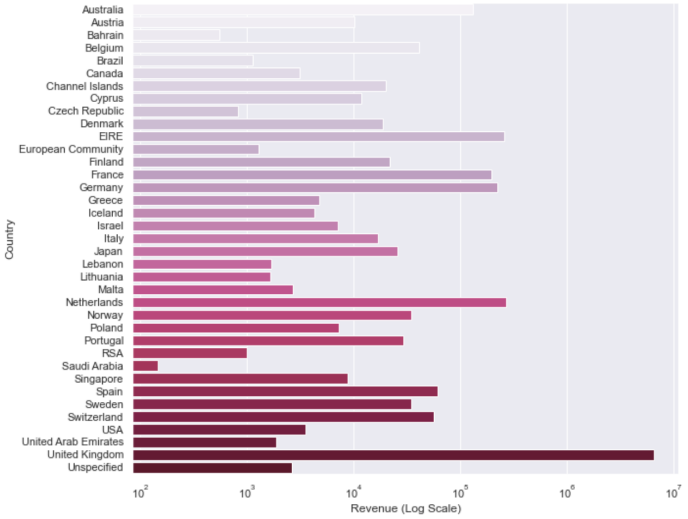

Figure 6 - Revenue by Country

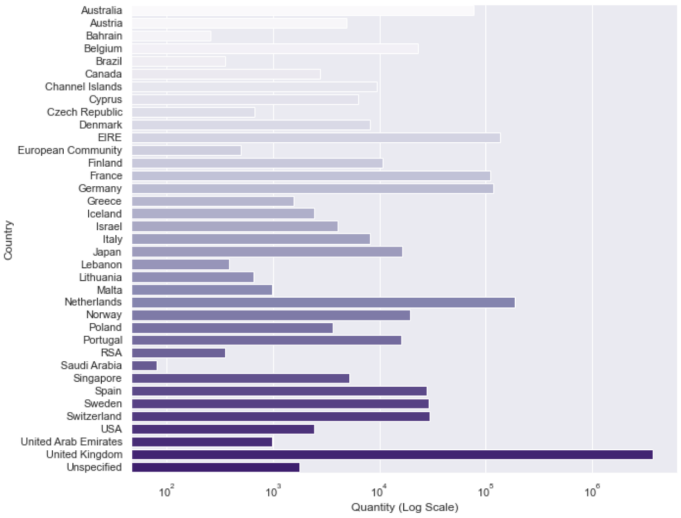

Figure 7 - Quantity Sold by Country

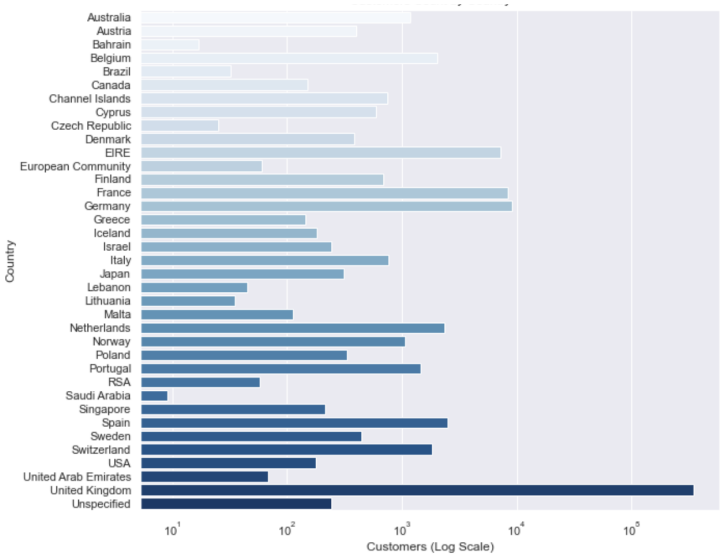

Figure 8 - Customers Count by Country

As we can see, United Kingdom is the country that has more customers. As such, and as expected this implies a greater number of gifts sold and, in turn, a greater revenue.

Lastly, we did two more visualizations where we grouped the data by month.

In ***Figure 9***, we wanted to understand what the months were with the most customers.

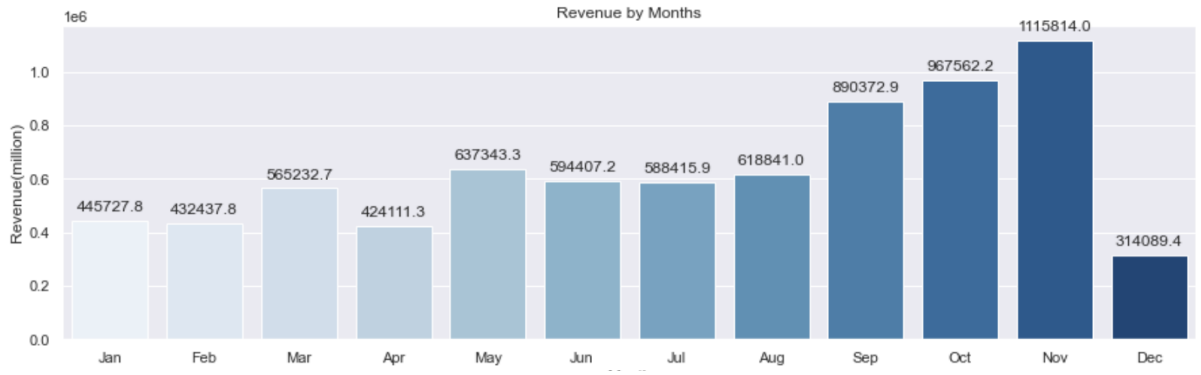

Figure 9 - *R*evenue by Months

In ***Figure 10***, we wanted to know what was the month with the highest revenue.

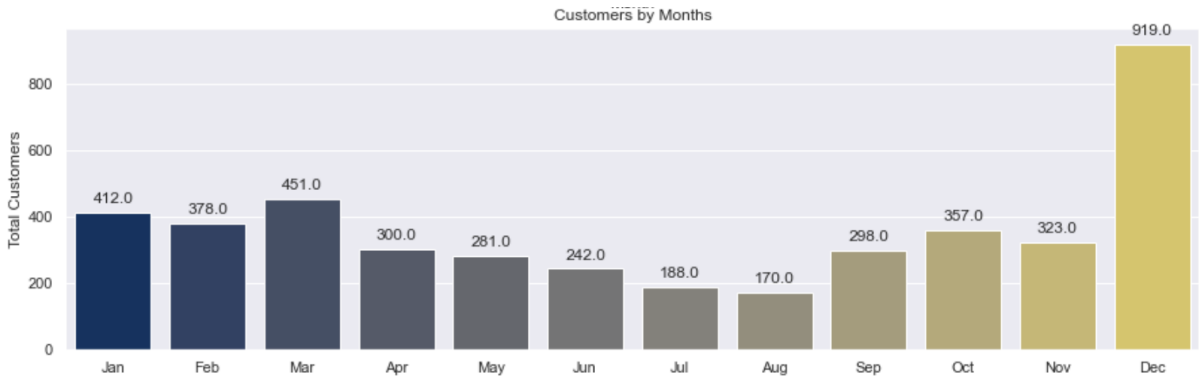

Figure 10 - *R*evenue by Months

As we can observe, although December is the month with the most customers, the month in which there is the highest revenue is November.

3.2. DATA PREPARATION

When checking for missing values, we found that 'CustomerID' is missing in approximately 25% of the rows. Since we need to know who bought each item in our recommender system, we decided to keep only the rows that have a customer ID. To answer the cold start problem, that we'll

address later, we will use those customers that don't have ID and treat them as new customers, to which we recommend only the m*ost popular products.

After removing the rows where Customer ID is null, we no longer have missing values. At this point, we changed the data types of 'InvoiceDate' to datetime and 'CustomerID' to integer. Next we removed the rows where 'Quantity' is negative, because we are not interested in knowing if a product is returned, if they bought it in the first place then that describes their purchase intent. Then, we verified that the dataset had 5225 duplicated values and the 'Quantity' and 'UnitPriceColumns' contained some outliers. In this way, we decided to drop the duplicates and remove the outliers.

Lastly, in order to apply the model we need to put all the transactions of the users into a matrix, where we will put each unique customer ID into a row and each unique product ID into a column. The values of the matrix should be the total of purchases of every item by every client. When checking the sparsity of this matrix we got a value of 97.519%, which we decided to reduce by only keeping users and items with frequency higher than 7.

3.3. Modeling

Collaborative Filtering doesn't require information about the users or items, so in order to find out how they are related to each other, we will have to use a factorization matrix. The group decided to apply the Alternative Least Square model to our user_item matrix in order to discover the hidden features that relate users and items to each other in a smaller matrix.

Our matrix will be quite sparse, since most users only interact with some items, so we can factorize it into separate matrices, one with the user feature vectors for each user and other with the item feature vectors for each item. We will use the ALS to discover the product of the user feature vectors and the item feature vectors which represents the predicted rating for a specific interaction between a user and an item. This will work even if the user and the item haven't had any prior interaction.

Besides the ALS model we also implemented the popularity based recommender model which is a common baseline approach for recommending systems, due to the fact that it simply recommends to the user the most popular items that the user has not yet consumed/purchased.

The Group decided to develop two different ways of fitting our matrix through the Alternative Least Square model. The first one, when training our dataset we take in consideration the time of the purchase, we then train the model based on earlier dates and test based on more recent purchases. Since we had poorer results in this first try, we implemented the ALS model described above in our solution.

3.4. Evaluation

When evaluating the models on the metrics, Precision at K, Mean Average Precision at K, Normalized Discounted Cumulative Gain at K and AUC at K, with K being 10, we found that in the first approach the ALS model performed worse than the baseline popular model, which only recommends the most popular items to every user and tends to be hard to beat in most recommender systems.

	pop_model	als_model
precision	0.086421	0.053672
map	0.039436	0.021821
ndcg	0.090338	0.054276
auc	0.512220	0.507087

Figure 11 - Evaluation of models on 1st approach

Therefore we decided to go with the second approach ALS model, that beats the benchmark of popularity in every metric.

	pop_model	als_model
precision	0.013257	0.049786
map	0.005832	0.018044
ndcg	0.015772	0.050664
auc	0.499150	0.504914

Figure 12 - Evaluation of models on 2nd approach

For this model we also calculated the mean Area Under the Curve that gave us a result of 0.851, while the Popular Item benchmark had lower AUC of 0.771. This means the system is recommending items the user in fact had purchased in the test set far more frequently than items the user never ended up purchasing.

4. RESULTS EVALUATION

To better understand how the recommendation system works, we chose to examine two particular users. We assessed the items they have bought in the past and checked the products the system recommends for them.

	StockCode	Description
45	POST	POSTAGE
98	21977	PACK OF 60 PINK PAISLEY CAKE CASES
99	84991	60 TEATIME FAIRY CAKE CASES
298	21980	PACK OF 12 RED RETROSPOT TISSUES
406	21213	PACK OF 72 SKULL CAKE CASES
409	84992	72 SWEETHEART FAIRY CAKE CASES
649	22616	PACK OF 12 LONDON TISSUES
1263	21982	PACK OF 12 SUKI TISSUES
1264	21981	PACK OF 12 WOODLAND TISSUES
1265	21967	PACK OF 12 SKULL TISSUES
3889	22437	SET OF 9 BLACK SKULL BALLOONS
3998	21211	SET OF 72 SKULL PAPER DOILIES
7347	84988	SET OF 72 PINK HEART PAPER DOILIES
11981	21725	SWEETIES STICKERS
131403	23077	DOUGHNUT LIP GLOSS
131472	23076	ICE CREAM SUNDAE LIP GLOSS
131649	23078	ICE CREAM PEN LIP GLOSS

Figure 13 - Items Customer ID 12348 bought

	StockCode	Description
0	84692	BOX OF 24 COCKTAIL PARASOLS
1	23119	PACK OF 6 LARGE FRUIT STRAWS
2	23309	SET OF 60 I LOVE LONDON CAKE CASES
3	21974	SET OF 36 PAISLEY FLOWER DOILIES
4	23155	KNICKERBOCKERGLORY MAGNET ASSORTED
5	84596B	SMALL DOLLY MIX DESIGN ORANGE BOWL
6	22197	SMALL POPCORN HOLDER
7	23154	SET OF 4 JAM JAR MAGNETS
8	21212	PACK OF 72 RETROSPOT CAKE CASES
9	84987	SET OF 36 TEATIME PAPER DOILIES

Figure 14 - Items recommended to Customer ID 12348

For the customer with ID 12348 we can see that the products they bought are different types of Cake Cases, Tissues, Doilies and Lip Gloss. So the top 10 recommended products for this customer ended up being items similar to those they have purchased. For example, the system recommends other Cake Cases and Doilies, but also other Kitchen Accessories like Cocktail Parasols, Fruit Straws and Popcorn Holder.

	StockCode	Description
117	21169	YOU'RE CONFUSING ME METAL SIGN
119	21175	GIN + TONIC DIET METAL SIGN
261	85152	HAND OVER THE CHOCOLATE SIGN
334	21463	MIRRORED DISCO BALL
335	21464	DISCO BALL ROTATOR BATTERY OPERATED
341	82580	BATHROOM METAL SIGN
343	82581	TOILET METAL SIGN
344	22413	METAL SIGN TAKE IT OR LEAVE IT
639	21790	VINTAGE SNAP CARDS
689	21892	TRADITIONAL WOODEN CATCH CUP GAME
1203	21179	NO JUNK MAIL METAL SIGN
1386	48138	DOORMAT UNION FLAG
2155	21181	PLEASE ONE PERSON METAL SIGN
2301	21864	UNION JACK FLAG PASSPORT COVER
3504	82582	AREA PATROLLED METAL SIGN
3636	21163	DO NOT TOUCH MY STUFF DOOR HANGER
3638	21161	KEEP OUT BOYS DOOR HANGER
4214	21911	GARDEN METAL SIGN
4704	82551	LAUNDRY 15C METAL SIGN
4721	21174	POTTERING IN THE SHED METAL SIGN
6571	21164	HOME SWEET HOME METAL SIGN
6635	21906	PHARMACIE FIRST AID TIN
6736	21165	BEWARE OF THE CAT METAL SIGN
8694	22412	METAL SIGN NEIGHBOURHOOD WITCH
33900	20619	TROPICAL PASSPORT COVER
503722	21175	GIN AND TONIC DIET METAL SIGN

Figure 15 - Items Customer ID 18230 bought

	StockCode	Description
0	85150	LADIES & GENTLEMEN METAL SIGN
1	82600	NO SINGING METAL SIGN
2	22467	GUMBALL COAT RACK
3	82578	KITCHEN METAL SIGN
4	21908	CHOCOLATE THIS WAY METAL SIGN
5	72741	GRAND CHOCOLATECANDLE
6	22115	METAL SIGN EMPIRE TEA
7	82583	HOT BATHS METAL SIGN
8	21166	COOK WITH WINE METAL SIGN
9	20963	APPLE BATH SPONGE

Figure 16 - Items recommended to Customer ID 18230

The Customer with ID 18230 has bought a lot of Metal Signs and other decorative items like Door Hangers and Disco Ball, as well as small random items like Passport Cover and First Aid Tin. So the top 10 items outputted by the recommender system are many different Metal Signs the client hasn't bought, in addition to an Apple Bath Sponge and a Gumball Coat Rack.

Cold Start Problem

For first time users the system we presented isn’t sufficient. Since we don’t have any information about new customers preferences and purchases, we propose the application of a popularity based strategy. In this way Trending items for a certain month or time of day will be recommended to new clients.

	Description	Quantity
586	CHARLOTTE BAG SUKI DESIGN	9177
2274	POPCORN HOLDER	5830
2391	RED RETROSPOT CHARLOTTE BAG	4962
3364	WOODLAND CHARLOTTE BAG	4073
2032	PAPER CHAIN KIT 50'S CHRISTMAS	3738
2327	RABBIT NIGHT LIGHT	3586
3445	came coded as 20713	3100
2057	PARTY BUNTING	2995
2994	STRAWBERRY CHARLOTTE BAG	2951
1995	PACK OF 72 RETROSPOT CAKE CASES	2630

Figure 17 - Top 10 items for Customers with no ID

Month	Description	
1	CHARLOTTE BAG SUKI DESIGN	629
	STRAWBERRY SHOPPER BAG	523
	JUMBO BAG CHARLIE AND LOLA TOYS	323
2	CHARLOTTE BAG SUKI DESIGN	549
	STRAWBERRY CERAMIC TRINKET BOX	241
	WOODLAND CHARLOTTE BAG	196
3	CHARLOTTE BAG SUKI DESIGN	847
	SMALL POPCORN HOLDER	769
	WOODLAND CHARLOTTE BAG	549
4	CHARLOTTE BAG SUKI DESIGN	387
	PARTY BUNTING	326
	WOODLAND CHARLOTTE BAG	250
5	ASSORTED COLOURS SILK FAN	796
	PARTY BUNTING	540
	CHARLOTTE BAG SUKI DESIGN	462
6	CHARLOTTE BAG SUKI DESIGN	439
	PARTY BUNTING	394
	SET/6 RED SPOTTY PAPER PLATES	304
7	PARTY BUNTING	453
	CHARLOTTE BAG SUKI DESIGN	438
	JUMBO BAG CHARLIE AND LOLA TOYS	288
8	CHARLOTTE BAG SUKI DESIGN	605
	RED RETROSPOT CHARLOTTE BAG	374
	PARTY BUNTING	316
9	CHARLOTTE BAG SUKI DESIGN	570
	POPCORN HOLDER	424
	RED RETROSPOT CHARLOTTE BAG	360
10	CHARLOTTE BAG SUKI DESIGN	760
	RED RETROSPOT CHARLOTTE BAG	480
	PAINTED METAL STAR WITH HOLLY BELLS	349
11	POPCORN HOLDER	4002
	RABBIT NIGHT LIGHT	2561
	PAPER CHAIN KIT 50'S CHRISTMAS	1989
12	CHARLOTTE BAG SUKI DESIGN	2108
	PAPER CHAIN KIT 50'S CHRISTMAS	1459
	HOT WATER BOTTLE TEA AND SYMPATHY	993

Figure 18 - Top 3 items for Customers with no ID by Month

As we can see above, it's possible for us to calculate the most popular items for new customers, which we defined as those without ID in our dataset. We can also find the top items sold by month. For new users that visit the store, these items can be recommended.

5. DEPLOYMENT AND MAINTENANCE PLANS

For the deployment, the group suggests the development of a Flask App where it can be seen the recommendations for a given customer.

Firstly, there are a few things that need to be done in order to obtain our recommender system deployed with the flask app. First we recommend the creation of a pickle file containing the top 10 products, that is, the most sold products by Many Gifts UK so far. This will be used later in the recommendations for new customers. We also suggest the implementation of the functions "get_item_purchased" and "rec_item", which will return a list of the purchased items and the recommended items, respectively. After implementing these 3 important steps we can now build our flask app.

For the flask app we suggest the development of a home page where the user can input his customer id number and see the recommended items given its previous purchases. The items recommended will be items that the customer has never bought. As for the new customers, our cold start problem, we constructed a database that recommends the top 10 most sold products for that month.

To maintain a good performance from our application, we suggest a continuous update of the data and transactions so that we can be one step ahead of our customers and foresee their needs.

6. CONCLUSIONS

In this project, we have shown how to design a recommender system with implicit feedback and extract the information from the user-item interactions to provide the best recommendations. After training and testing our model we fit it using our second attempt of the ALS, which uses the user and item feature vectors . With this ALS model we were able to build the functions that will recommend the best items to a given customer, one that has previous purchases. As for the new customers, our cold start problem, we solve it by recommending the top 10 sold items for that month. Lastly, for the deployment, we suggest the implementation of a flask app where we make use of the previously created functions, like the "recommend_item" and the "top10_products", to show our recommendations.

6.1. Considerations for model improvement

For model improvement we suggest a more intensive search of other collaborative filtering models, which will improve our recommendations. The group also suggests a bigger dataset extracted from the Many Gifts UK, since more data, whether is from more customers, longer time interval or more features, could tell us more information about our customers, items, interactions and most of all improve our recommendations

7. REFERENCES

[1] Hu, Y., Koren, Y., & Volinsky, C. (n.d.). *Collaborative Filtering for Implicit Feedback Datasets.*

[2] *Steinweg-Woods, J. (2016, May 30). A Gentle Introduction to Recommender Systems with Implicit Feedback. Retrieved from https://jessesw.com/Rec-System/*

[3] *Introduction to Recommender Systems*. (n.d.). Retrieved from tryolabs: https://tryolabs.com/blog/introduction-to-recommender-systems/

[4] Rocca, B. (2019, June 3). *Introduction to recommender systems*. Retrieved from towards data science: https://towardsdatascience.com/introduction-to-recommender-systems-6c66cf15ada

[5] Shaw, A. (n.d.). *Product Recommendation System for e-commerce*. Retrieved from Kaggle: https://www.kaggle.com/shawamar/product-recommendation-system-for-e-commerce