BUSINESS CASES WITH DATA SCIENCE

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

ManyGiftsUK- Recommender System

Group X

Beatriz Chumbinho, number: R20170867

Inês Costa number: R20170775

Mª Leonor Morgado, number: R20170871

Rodrigo Matias, number: R20170880

May, 2021

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

INDEX:

1. INTRODUCTION.....2

2. BUSINESS UNDERSTANDING.....2
   - 2.1. Background.....2
   - 2.2. Business Objectives.....2
   - 2.3. Business Success Criteria.....2
   - 2.4. Situation Assessment.....3
   - 2.5. Costs and Benefits.....3
   - 2.6. Risk & Contigency.....3
   - 2.7. Determine Machine Learning Goals.....3

3. RECOMMENDER SYSTEM.....3
   - 3.1. Data understanding.....4
     - 3.1.1.Data Description.....4
       - Dataset.....4
       - Variables.....4
       - Exploratory Data Analysis.....4
   - 3.2. Data preparation.....5
     - 3.2.1. Missing Values.....5
     - 3.2.2. Noisy Data.....5
     - 3.2.3. Reduce Sparsity.....6
     - 3.2.4. Split the data.....6
   - 3.3. Modeling.....6
     - 3.3.1. For existing Customers.....6
     - 3.3.2. Approaching the cold-start problem.....7
   - 3.4. Evaluation.....7

4. RESULTS EVALUATION.....8

5. DEPLOYMENT AND MAINTENANCE PLANS.....9

6. FURTHER ACTIONS.....9

7. CONSLUSION.....10

8. REFERENCES.....10

1. INTRODUCTION

Recommender Systems (RS) have become very popular because they aim to provide personalized recommendations to users for specific items which leads to an increase in sales. Some studies suggest that when faced with easier choices customers tend to buy more and this is the main reason for RS success, because they facilitate decision making for customers. One of the biggest advantages of RS is that users discover patterns that they might not be able to find otherwise.

Our study is based on the collaborative filtering methodology, due to the characteristics of the data we have in hands. This method uses data about the user - item interactions (not only the user's past item purchases but also similar purchases of other users). The standard method of collaborative filtering is known as the Nearest Neighborhood algorithm. This way, the similarities between target user X and all other users (e.g. user Y) are calculated and the most similar users are selected. Then, the algorithm takes the weighted average of data from those users with similarities as weights. Fig.1 presents the main idea of the collaborative filtering methodology.

The company, ManyGiftsUK, asked our team for help to build a RS to improve their user experience, and to improve their revenue as well based on the data that they have been collected.

In this analysis, our team of data scientists will deal with a particular challenge for ManyGiftsUKy and a very important problem in RS: The cold start problem. This problem is related to recommendations for novel users, it states in the way that the company is able to suggest relevant items for a new customer that does not have any prior purchase history.

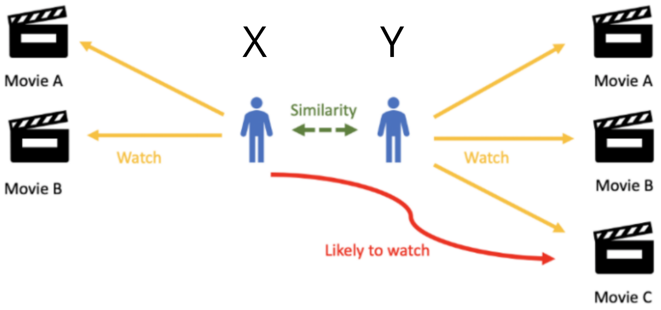


Fig. 1

2. BUSINESS UNDERSTANDING

2.1. Background

ManyGiftsUK is a UK-based non-store online retailer, operating since 1981, and counting with a staff of 80 members. This company operates worldwide and it is important to refer that many of ManyGiftsUK customers are wholesalers. Previously, the main sales channels were the mail, catalogs, and the phone. Recently, ManyGiftsUK launched a website where people can buy online, now that website and Amazon.co.uk are the only channels. Nevertheless, the company has accumulated data about customers, namely from 2010 and 2011.

2.2. Business Objectives

The goal of this company is to build a recommender system. Once it facilitates people's choices the user experience is also benefited and so the sales tend to increase. In a nutshell:

- Recommend items that a specific user appreciates
- Improve user experience when making purchases on the website
- Solve the cold start problem - how can we suggest items to new customers?
- Increase sales

2.3. Business Success criteria

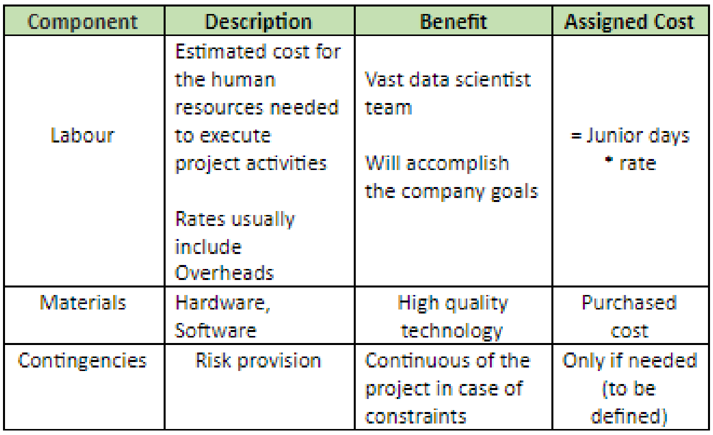
Our team's goal in the present project is to provide to ManyGiftsUK the following, shaped as a report and a presentation:

- Obtain a reliable and efficient recommender system able to increase the sales
- Captivate new customers in their first interaction with the web page due to the suggestions made by the system - cold start problem
- Reach a significant value of the system in the total sales - 10% in the beginning

2.4. Situation assessment

To develop this project ManyGiftsUK hired four data scientists for the time of 3 weeks. To the data scientists, a dataset from 2010/2011 was made available in order to perform exploration and develop a recommendation system capable of presenting to the customers the items that they most identify with.
The platform available to build the system in order to accomplish the objective is Python Jupyter Notebook, using the implicit library.

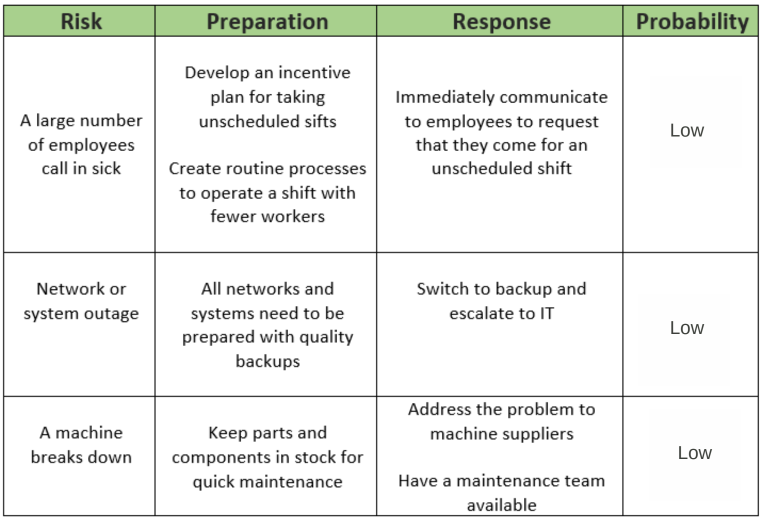
2.5. Costs and Benefits:



Component	Description	Benefit	Assigned Cost
Labour	Estimated cost for the human resources needed to execute project activities Rates usually include Overheads	Vast data scientist team Will accomplish the company goals	= Junior days * rate
Materials	Hardware, Software	High quality technology	Purchased cost
Contingencies	Risk provision	Continuous of the project in case of constraints	Only if needed (to be defined)

Table 1

2.6. Risk & Contigency:



Risk	Preparation	Response	Probability
A large number of employees call in sick	Develop an incentive plan for taking unscheduled sifts Create routine processes to operate a shift with fewer workers	Immediately communicate to employees to request that they come for an unscheduled shift	Low
Network or system outage	All networks and systems need to be prepared with quality backups	Switch to backup and escalate to IT	Low
A machine breaks down	Keep parts and components in stock for quick maintenance	Address the problem to machine suppliers Have a maintenance team available	Low

Table 2

2.7. Determine Machine Learning goals

The main goal of this project is to build a recommender system for ManyGiftsUK customers, including both usual customers and new ones - cold start. The used methodology was the Collaborative filtering. This methodology helped our team to reach the goal of building the recommendation system. In a nutshell:

- Predict the items that a specific user will like the most
- Suggest appropriate items to new customers
- A precision equal or greater than 0.06
- An AUC (Area Under The Curve) equal to or greater than 0.5

NOTE: AUC tells how much the model is capable of distinguishing between classes and is the area under the ROC curve, which is plotted with sensitivity against the specificity. The greater the AUC, the better the model.

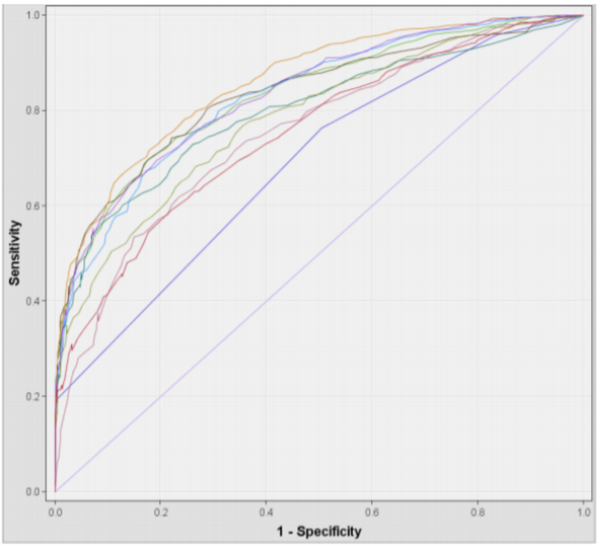


Fig. 2

3. RECOMMENDER SYSTEM

In order to develop the present recommendation system, several steps were performed. Our team started by understanding the data we had in hands, taking into consideration the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology - a standard process model that describes common approaches used to conduct data mining studies. CRISP-DM methodology can be understood by looking at the image on the right. This way, we started by understanding the business, the project objectives, the requirements and the data itself, provided by ManyGiftsUK in a circular, iterative and interactive perspective. The data was prepared, in the way presented lately in the present section. Then, the data was the input for our model using the Implicit library with the ALTERNATING LEAST SQUARES - ALS - model.

This model enabled us to generate suggestions specifically for each customer, including a metric of confidence on the suggestion. Also, it is possible to insert an item and find similar ones. Having done all these steps, our team evaluated the achieved results comparing them with your business needs. line 24 of Jupyter notebook

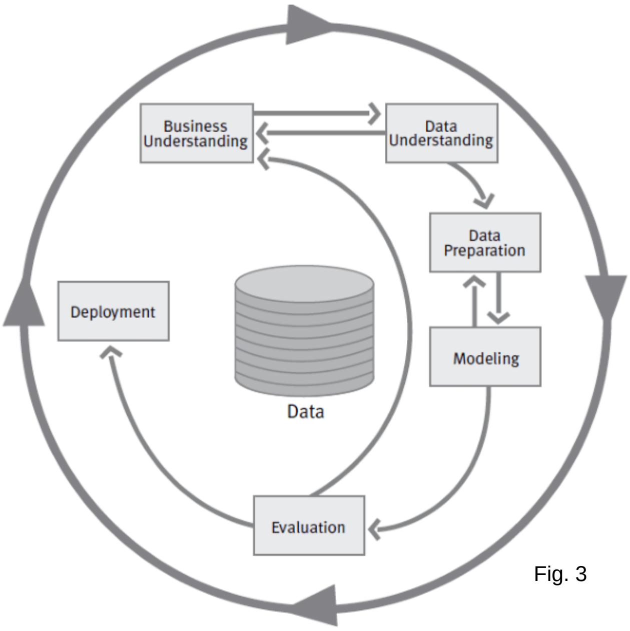


Fig. 3

3.1. Data understanding

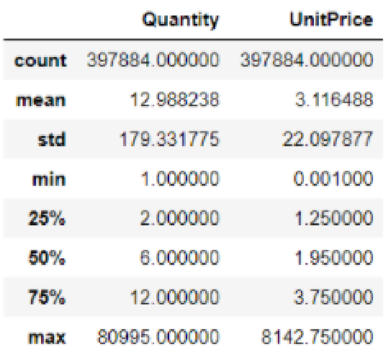
3.1.1.Data Description: line 10 of Jupyter notebook

Dataset:

The dataset provided relates to the purchases made in ManyGiftsUK - before moving to 100% online - for the period between 01/12/2010 and 09/12/2011, which corresponds to 373 days or 53 weeks. It contains 8 variables and 541909 rows each for a particular item in a specific transaction, from those instances only 25900 were valid. The dataset refers to 4070 unique products, 4372 different customers from 38 countries around the world.

Variables:

Our team worked on understanding each variable taking into account their meaning in the process of reaching both business and data mining goals the best we can. From all the available variables ('InvoiceNo', 'StockCode', 'Description', 'Quantity', 'InvoiceDate', 'UnitPrice', 'CustomerID', 'Country') we gave special attention to the StockCode and CustomerID, in order to connect these two features. Although there is a particularity that is important to refer about the InvoiceNo variable, some of these codes contain a 'C', meaning that the order has been canceled. Our team handled this situation by splitting the data. In a furthe section of this document, there is a deepen explanation of this topic.
In table 3 it is possible to observe some descriptive statistics on the Quantity and UnitPrice.



	Quantity	UnitPrice
count	397884.000000	397884.000000
mean	12.988238	3.116488
std	179.331775	22.097877
min	1.000000	0.001000
25%	2.000000	1.250000
50%	6.000000	1.950000
75%	12.000000	3.750000
max	80995.000000	8142.750000

Table 3

Exploratory Data Analysis - EDA:

1. How much do the products cost?
   Observing the bar plot it is possible to confirm that the majority of the products' cost fall in a range of 0.5€ to 5€.

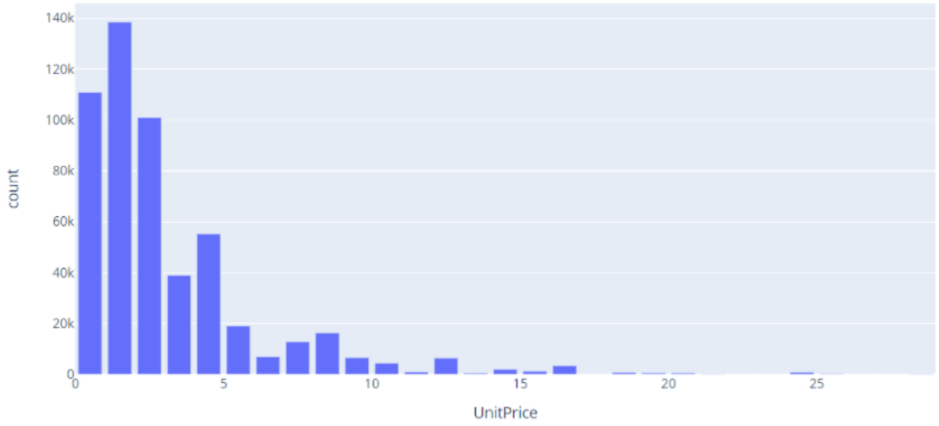


Fig. 4

2. How many products do the customers buy per order?

The majority of the clients buy in a range of 1 to 20 products, but a considerable number of customers buy until 200 items in only one transaction.

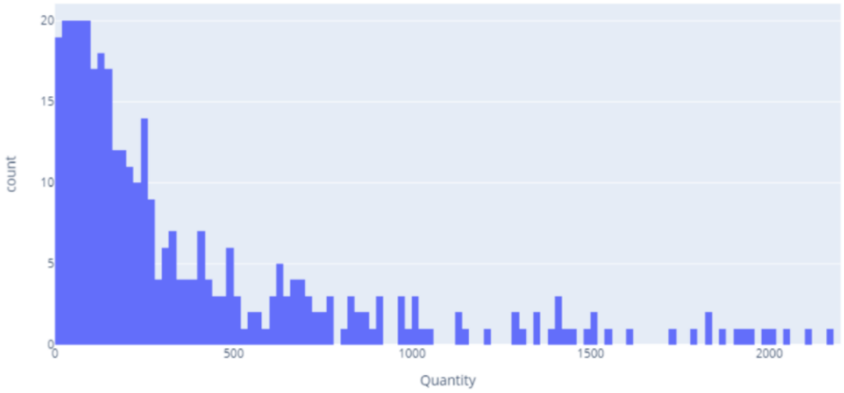


Fig. 5

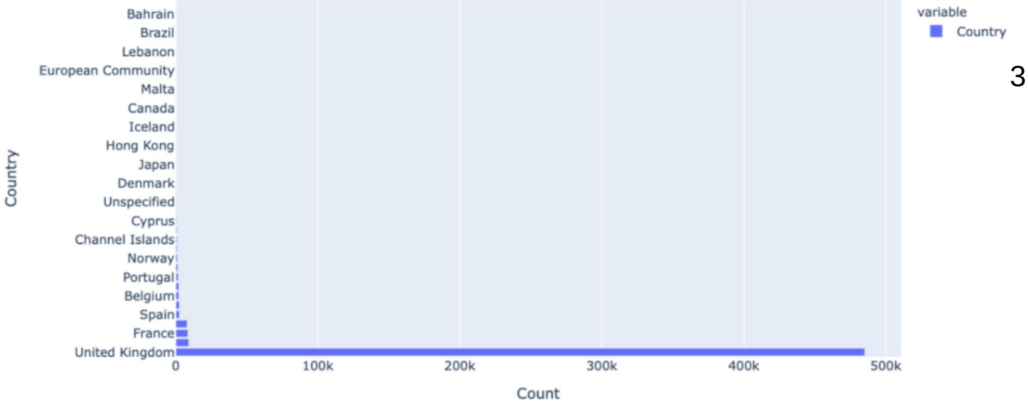


Fig. 6

3. Where are the customers from?

As it is possible to see, the greatest part of the customers are from the UK, but there is also a considerable number of consumers mainly in Europe.

3.2. Data preparation

In order to ensure data quality to make correct suggestions and consequently, obtain accurate results we had to deal with data that was corrupted or missing.

3.2.1.Missing Values: After analyzing the dataset we noticed that the variables Description (of the item) and CustomerID contained null entries. About the CustomerID, those nulls have a reason to be since they refer to new customers, it is, people without an account on the website. To handle this situation, our team created the dataset df_new containing only rows with a null CustomerID and here we will only apply the methodology directed to overpass the cold start problem. Regarding the Description variable, this problem was directly related to the cancelled orders. Since the canceled orders represented less than 2% of all the dataset, we opted to remove them. To accomplish this task our team created a new boolean column indicating if the order contained a ‘C’ in the InvoiceNo, and if True the row was removed. All the null entries on Description were included on the deleted rows. line 6 and 15 of Jupyter notebook

3.2.2. Noisy data:

In order to find the noisy data, our team searched for inconsistencies, manually both using code to perform verifications and serving of visualizations. Having done this analysis, we ended up facing UnitPrice values equal to or below 0 and we removed those transactions from the dataset. Regarding Quantity, our team found values that represent outliers, this way the transactions with a Quantity value greater than 2200 were deleted, which corresponds to 13 rows. The deleted irregularities from UnitPrice and Quantity correspond to less than 1% of the data.

These deleted rows together with the deleted cancellation rows correspond to less than 3% of the original data, being in compliance with the best practices.

More inconsistencies were found - as the Quantity below 0 - but were only contained in the rows corresponding to cancellations, which were deleted as previously mentioned. line 6 of Jupyter notebook

3.2.3. Reduce Sparsity: When dealing with this type of Recommender systems (collaborative filtering) and implicit datasets, often what is intended is to reduce the sparsity, because the more sparsity exists the less information we have. Having a high sparsity value, there will be more missing values and less information and therefore it is harder for the model to provide good recommendations. So what our team started to do in the preprocessing step is try to reduce sparsity. In this step, it was decided that items that are not purchased very frequently are to be eliminated and also the users that have not had a lot of purchasing history.
By doing this it is possible to reduce the sparsity, therefore, improve the recommendations obtained. In order to obtain a limited dataset, our team decided that the minimum number of occurrences that each item must have is 5 and the same for the users. By applying these restrictions the number of missing values in the matrix has decreased.

Our 'Raw dataset' has 4336 users, 3664 items and a sparsity of 2.504%.
The 'Limited dataset' has 4110 users, 3191 items and a sparsity of 3.022%.
line 17 of Jupyter notebook

3.2.4. Splitting the data: A recommender system is itself a prediction since it aims to estimate which products will a certain customer appreciate based on past events. This way, it is necessary to split the dataset into a train set and a test set. Since we are dealing with dates from purchases, the proper way to handle the split is to guarantee that all test events occur after all train events in order to avoid anomalies, e.g. bias. To split the dataset, our team provided 29% of the data for the test and the remaining 71% for the train. In the image it is possible to observe how the split looks like:
line 19 of Jupyter notebook

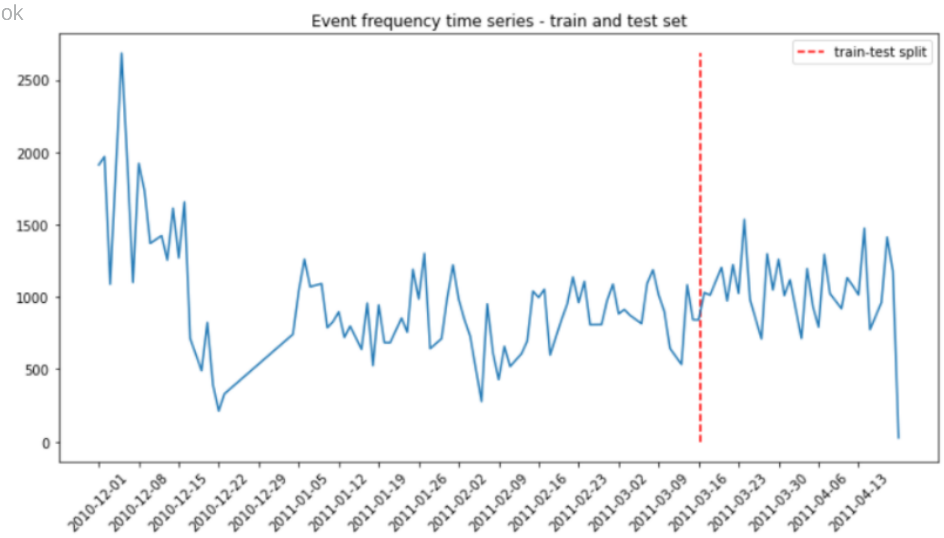
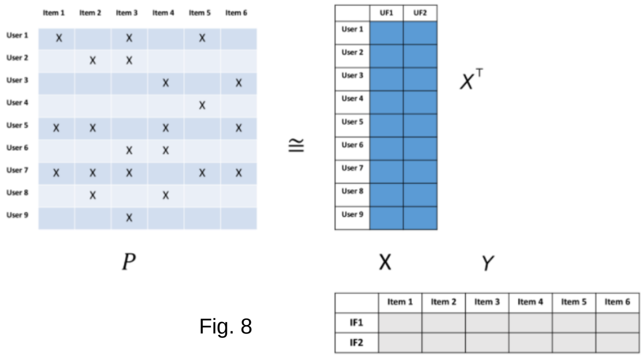


Fig. 7

3.3. Modeling

3.3.1. For existing customers: Our team chose a path based on trial and error in order to compare results and pick up the best model solution. For this purpose, the Implicit library from python was used. Our team applied the Alternating Least Squares -ALS- algorithm to the data, which was built for large-scale collaborative filtering problems. This algorithm performs a matrix factorization algorithm, similarly to a PCA or a Correspondence Analysis. ALS decomposes a Preference Matrix into the combination of some other matrices. In this case, it will decompose P into X and Y. Where X is a matrix with a vector for each user and Y is a matrix in which we have the data for each item. This can be observed in the following figure:



	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
User 1	X		X		X	
User 2		X	X			
User 3				X		X
User 4					X	
User 5	X	X		X		X
User 6			X	X		
User 7	X	X	X		X	X
User 8		X		X		
User 9			X			

	UF1	UF2
User 1		
User 2		
User 3		
User 4		
User 5		
User 6		
User 7		
User 8		
User 9		

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
IF1						
IF2						

Fig. 8

This matrix gives information on whether a user would like or not an item and the level of confidence of that assumption.
The classical latent factor model associates each user with a user-factors vector and each item with an item-factors vector. The more factors we have, the better it can explain the original data. However, if we maintain too many factors the dimensionality is strongly reduced and consequently, there is a loss of interpretability. The score in the ALS model between a certain pair of user-item is obtained by taking the inner product between the respective vectors of user-factors and the item-factors. ALS learns by minimizing the loss function, the model minimizes two loss functions alternatively: It first holds user matrix fixed and runs gradient descent with item matrix and then it holds item matrix fixed and runs gradient descent with user matrix. The loss function can be observed in the figure below.

$$\min_{x_\star, y_\star} \sum_{u,i} c_{ui}(p_{ui} - x_u^T y_i)^2 + \lambda \left(\sum_u \|x_u\|^2 + \sum_i \|y_i\|^2 \right)$$

Fig. 9

Besides the ALS, our team also applied the Bayesian Personalized Ranking - BPR- which optimizes to predict if an item is selected by a user or not in order to come up with more personalized rankings for each user. Also, our team tested the Logistic Matrix Factorization - LMF - algorithm, which is a collaborative filtering recommender model that learns probabilistic distribution whether users like it or not.

Model Fine Tuning: The parameters elected by our team were factors=20,regularization=0.1,iterations=40, random_state = 0

line 24 of Jupyter notebook

3.3.2. Approaching the cold-start problem: There is not a scientific unanimous opinion in a matter of which are the best practices to overpass this problem. A famous solution is to use hybrid systems - as the LightFM library - that is able to combine both Collaborative and Content-Based filtering. Unfortunately, to use this type of systems, more data features both about users and items are required (e.g. item category, customer´s age).
In this analysis, our team considered as new customers the ones without the *'CustomerID'* attribute - no account created in the ManyGiftsUK website.
To overpass this problem an algorithm named Popular Recommender was built. This algorithm verifies which are the most common items over all the transactions in the dataset. Those items are the ones we suggest ManyGiftsUK to recommend to their new customers.
In order to provide a cold-start recommender system of higher confidence more data will be needed. This requirement will be better explained in a further stage of the present document- section 5. line 35 of Jupyter notebook

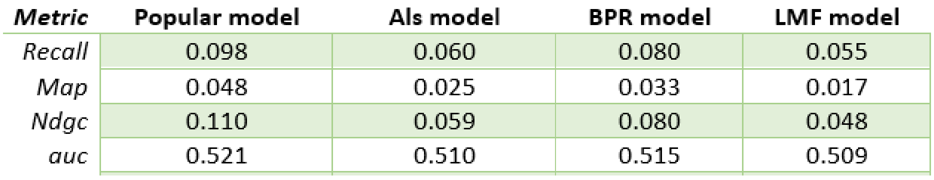
3.4. Evaluation:

line 27 of Jupyter notebook

Taking into consideration the metric of Precision and AUC, in this section the performance of the different algorithms will be evaluated.
Regarding the Precision, this metric measures the proportion of correctly positive events from all events identified as positive. This is: TP/(TP+FP)
As present previously, AUC tells how much the model is capable of distinguishing between classes.
Another used metric was the Mean Average Precision - MAP - which the general definition is finding the area under the precision-recall curve above.

NOTE: $Recall = \frac{TP}{TP + FN}$

Regarding the Normalized Discounted Cumulative Gain - NDCG - indicator, it is important to refer that the Cumulative Gain corresponds to the sum of all the relevance scores in a recommendation set. Thus, the Normalized Discounted Cumulative Gain involves discounting the relevance score by dividing it with the log of the corresponding position. Also, NDCG provides a score which has upper and lower bounds, enabling to take the mean across all the recommendations' score to report a final score. This way, the obtained scores for the ALS, BPR and LMF were:



Metric	Popular model	Als model	BPR model	LMF model
Recall	0.098	0.060	0.080	0.055
Map	0.048	0.025	0.033	0.017
Ndgc	0.110	0.059	0.080	0.048
auc	0.521	0.510	0.515	0.509

Table 4

The elected model:

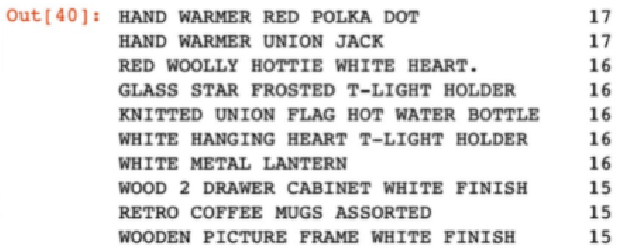
Taking only into consideration the metrics explained above, the rational choice should fall into the BPR model. However, our team works under the principle of professional skepticism, this way we verified the suggestions made by this model. There should be similarities between the recommended items for a specific customer in this type of study and during this task, our team noticed that the suggested items were very different among them. For that reason, our final choice is the ALS model which was also minutely tested and explored, presenting more reliable outputs.

4. RESULTS EVALUATION line 39 of Jupyter notebook

To meet the business objectives the Recommender System was developed with success for both existing and new customers, as specified in the business objectives. The data scientists team had complied with the programmed schedule - 3rd May.

The recommender system for the existing customers relies on the Collaborative Filtering methodology and presents for each user 10 products he/she will probably appreciate, as well as a measure of confidence for each of those products. Here, it is possible to observe a sample recommendation made by our system as well as the last 10 products and the quantity purchased by this client. This recommendation is specific for the Customer with the ID number 17850.0.

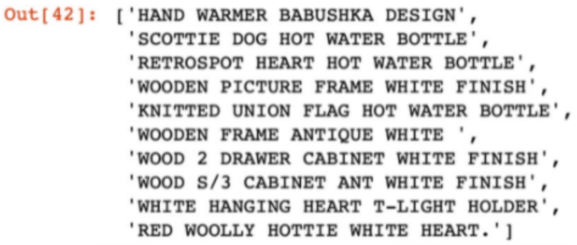
Last 10 purchases:



```
Out[40]: HAND WARMER RED POLKA DOT                17
         HAND WARMER UNION JACK                   17
         RED WOOLLY HOTTIE WHITE HEART.           16
         GLASS STAR FROSTED T-LIGHT HOLDER        16
         KNITTED UNION FLAG HOT WATER BOTTLE      16
         WHITE HANGING HEART T-LIGHT HOLDER       16
         WHITE METAL LANTERN                      16
         WOOD 2 DRAWER CABINET WHITE FINISH       15
         RETRO COFFEE MUGS ASSORTED               15
         WOODEN PICTURE FRAME WHITE FINISH        15
```

Fig. 10

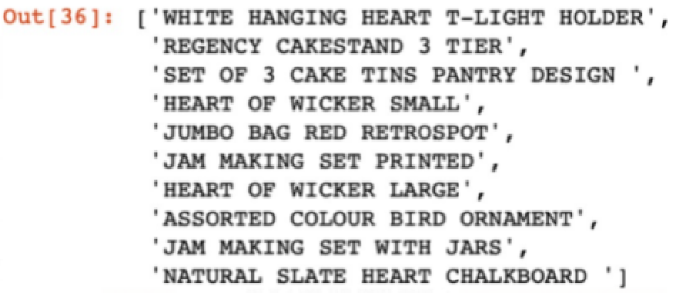
10 item recommendations:



```
Out[42]: ['HAND WARMER BABUSHKA DESIGN',
          'SCOTTIE DOG HOT WATER BOTTLE',
          'RETROSPOT HEART HOT WATER BOTTLE',
          'WOODEN PICTURE FRAME WHITE FINISH',
          'KNITTED UNION FLAG HOT WATER BOTTLE',
          'WOODEN FRAME ANTIQUE WHITE ',
          'WOOD 2 DRAWER CABINET WHITE FINISH',
          'WOOD S/3 CABINET ANT WHITE FINISH',
          'WHITE HANGING HEART T-LIGHT HOLDER',
          'RED WOOLLY HOTTIE WHITE HEART.']
```

Fig. 11

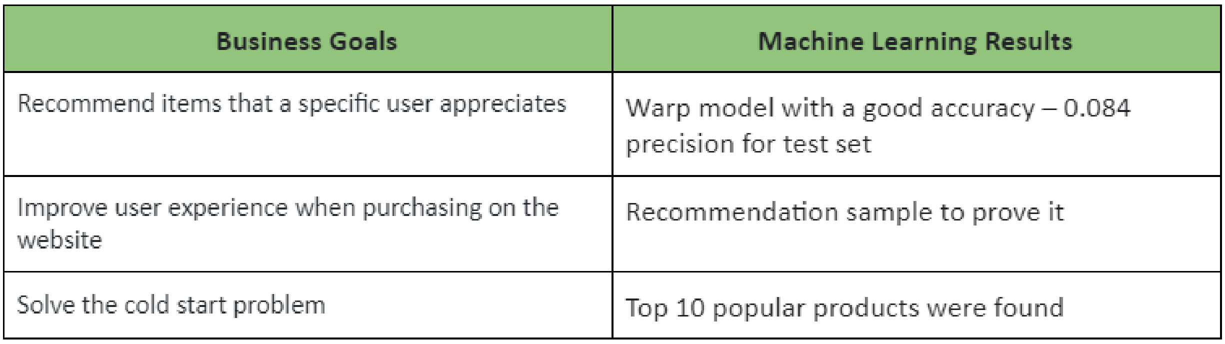
On the other hand, the recommender system for the new customers relis in an algorithm built during the development of the present analysis. Similarly, it also offers 10 item suggestions for the new users. At right, it is illustrated the output of our Popular Recommender:



```
Out[36]: ['WHITE HANGING HEART T-LIGHT HOLDER',
          'REGENCY CAKESTAND 3 TIER',
          'SET OF 3 CAKE TINS PANTRY DESIGN ',
          'HEART OF WICKER SMALL',
          'JUMBO BAG RED RETROSPOT',
          'JAM MAKING SET PRINTED',
          'HEART OF WICKER LARGE',
          'ASSORTED COLOUR BIRD ORNAMENT',
          'JAM MAKING SET WITH JARS',
          'NATURAL SLATE HEART CHALKBOARD ']
```

Fig. 12

Business Goals vs Machine Learning Results:



Business Goals	Machine Learning Results
Recommend items that a specific user appreciates	Warp model with a good accuracy – 0.084 precision for test set
Improve user experience when purchasing on the website	Recommendation sample to prove it
Solve the cold start problem	Top 10 popular products were found

Table 5

5. DEPLOYMENT AND MAINTENANCE PLANS

A way to motivate and serve customers better is to suggest relevant products for them to buy. Obviously, in order to do that it is important, in some way, to have an idea about what are the preferences and the behavior of customers.

With the evolution of time also customers' habits and preferences tend to change and evolve. This way, it is imperative to keep monitoring the transactions' data in order to keep the recommender system updated. A great upgrade in the available data that we strongly recommend to ManyGiftsUK is to start tracking also other user-website interactions and not only the transactions. Could be beneficial for the company to have knowledge about which products do people add to the cart and end up not buying, as well as the products that each client spends time viewing, the number of times that a user interacted with a specific item would also be welcome.

Our deployment suggestion in a matter of new customers is to test the solution first with existing customers in order to perceive their acceptance of our recommendations. The greater the adherence the greater the confidence level in presenting those products to new customers. We believe the best approach is to define a mark. This way you can present the suggested products to all your clients with an account on the website and if 15% of them decide to buy the confidence level is feasible, in the other hand, if the confidence falls in a value smaller than 5%, maybe those products are not the ones to solve the cold-start problem.

Besides this, our team also advises you to create a 'welcome' simple form in your website where new users can choose the fields of products where they have a major interest.
The objective is to find users with similar features and use this information as input to produce a predictive model in order to then build a Recommender system specifically to solve the cold-start problem.

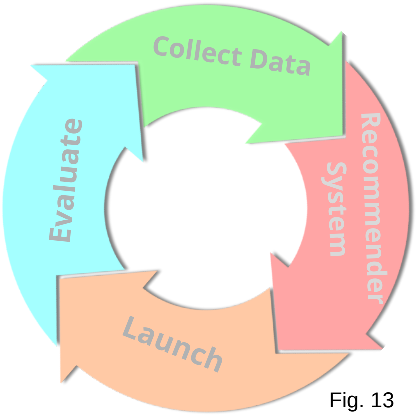


Fig. 13

6. FURTHER ACTIONS

Further Actions	Pros	Cons
Collect more data about user´s features	Better insights and opportunity to develop Content-Based filtering	Might be invasive
Include a 'welcome' form for new customers	Insights about their purchase areas of interest	Must be very small
Test the suggestions of cold-start with existing customers	Further study on the accuracy/opportunity for improvement	------------------------
Track customer´s actions within the website	Insights about customer´s 'wish list' and interests	Might be redundant

Table 6

7. CONCLUSION

Recommender Systems -RS- nowadays are a new important area of research in machine learning. The main idea of RS is to build relationships between the items, users and make the decision to select the most appropriate item for a specific user. In this analysis, and according to the data provided, a collaborative filtering methodology was performed and an ALS model was applied in order to obtain good quality predictions and to solve the cold start problem.

In the near future, our team will be glad to help ManyGiftsUK with a new RS analysis to be able to address the content-based filtering part of the analysis and with a more accurate analysis for the new customers, having the right data.

Finally, all the expected outcomes of this project have been achieved in consideration of the data limitations.

8. REFERENCES

- https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5
- https://towardsdatascience.com/essentials-of-recommendation-engines-content-based-and-collaborative-filtering-31521c964922
- https://towardsdatascience.com/prototyping-a-recommender-system-step-by-step-part-2-alternating-least-square-als-matrix-4a76c58714a1
- https://github.com/joaopfonseca/business-cases/tree/master/BC4_recommendation_system
- Lucas Bação, Fernando - Business Cases with Data Science- April 11, 2021.
- Henriques, Roberto- Machine Learning - September 28, 2020.