NOVA IMS Information Management School

Business Case #4

MASTER DEGREE PROGRAM, DATA SCINCEAND ADVANCED ANALYTICS – MAJOR IN BUSINESS ANALYTICS

Online Retailer Recommender System

Ana Sofia Silva, number: 20200220
José Francisco Alves, number: 20200653
Miguel Nunes, number: 20200615
Mohammadali Gharghi, number: 20200997

Index

INTRODUCTION

Several studies have proved that when the customer is faced with easy choices, they tend to buy more. Recommendation systems have the potential to change the way websites communicate with users and to allow companies to maximize their profit based on the information they can gather on each customer's preferences and purchases.

ManyGiftsUK is a UK-based non-store online retailer that is focused on selling unique all-occasion gifts. ManyGiftsUK wants to implement a recommendation system in their website and hired a data science team to do it. The company possessed records of client's transactions of one year.

The implemented models were based on collaborative filtering with Alternative Least Squares and Apriori methodologies.

BUSINESS UNDERSTANDING

BACKGROUND

ManyGiftsUK is a UK-based non-store online retailer that is focused on selling unique all-occasion gifts. The company has recently become completely online.

The company expects to build a recommender system with a collected data about transactions occurring between 01/12/2010 and 09/12/2011 that can facilitate the customer's choice with and without the specific customer historic – The recommender system and the cold start problem, which can suggest relevant items to new customers.

BUSINESS OBJECTIVES

The objective of this business passes by implementing a recommender system in the ManyGiftsUK store and an initial product suggestion to the customer. The recommender system is expected to appear on the website homepage and offer a wide range of relevant products to each customer and the Cold start is supposed to offer relevant products to new customers.

With the provided dataset, the team could come up with the business solution for the problems mentioned below:

- Help users discover items they might not have found otherwise
- Provide an easy choice that leads to a greater purchase
- Increase the company's sales.

BUSINESS VALUE

Customers who are prompted with personalized product recommendations drive 24% of the orders and 26% of the revenues. This signifies how much significance product recommendation has on order volume and overall sales revenue. A product recommender system is a system with the goal of predicting and compiling a list of items that the customer is likely to purchase.

The expected outcome of this study case includes implement adequate evaluation strategies and select an appropriate quality measure for the recommendation method and consequently for the business itself.

DETERMINE DATA MINING GOALS

One of the key techniques used by large retailers to facilitate user choices is called Recommendation System (RS).
In our case, the case study had two main problems:

- Create a RS that appears in the website homepage and offers a wide range of relevant products to each user
- Develop a Cold start model that offers relevant products to new customers

To do this, the data science team have done the following steps:

- **Collaborative Filtering for Implicit Datasets using Alternative Least Squares (ALS)** – These systems passively track different sorts of user behavior, such as purchase history, watching habits and browsing activity, to model user preferences. In our case, this method includes an event, such a click, view, or purchase that indicates confidence.
- **Apriori RS and Association Rules –** With the Apriori method determine which item are often bought together. The last step is to build the association rules, establishing the most important parameter, which in our case is confidence. Confidence reflects the degree of similarities between products and help reflects the value of the similarities. Therefore, it is clear this should be very relevant for making recommendations.
- **Implement adequate evaluation for the RS model –** With the Logistic Regression model to predict future visitor purchase behavior.

ANALYTICS PROCESS

DATA UNDERSTANDING

All the information presented consists of the analysis of two datasets that were provided.
First, the dataset "*events_example*" is made up of the variables: "*timestamp*" describes, in seconds, the time at which a certain item was triggered; *"visitor_id"* corresponds to the id of each visitor to the platform being used; *"event"* can take 3 different values, which correspond to the possible actions by each user. This corresponds to the items viewed, the items added to the cart, and the items that resulted in a transaction*; "itemid"* which corresponds to the id of each product*. "transactionid"* which corresponds to the id of each transaction that was performed. If this transaction has not been performed, a value is returned to NaN. This dataset consists of 2756101 rows.
The second dataset "*retail*" is made up of the variables: *"InvoiceNo"* which corresponds to the invoice number. This consists of a 6-digit integral number uniquely assigned to

each transaction. If this code starts with letter 'c', it indicates a cancellation; “*StockCode”* which corresponds to the corresponding number of each item (5-digit integral number); “*Description”* which corresponds to the description of each product; “*Quantity”* which corresponds to the number of items for each product; *“InvoiceDate”* which corresponds to the day and time when each transaction was generated; “*UnitPrice”* which corresponds to the product price per unit in pounds; “*CustomerID”* which corresponds to the 5-digit integral number uniquely assigned to each customer; “*Country”* which corresponds to the name of the country where each customer resides. This dataset consists of 541909 rows.

Data Preparation

For better analysis and consecutive interpretation of the possible results, some data preparation was done. One of the first situations was to change the column *"timestamp"* from seconds (difficult to interpret) to year-month-day-hour-minute-second format. Another piece of data that was considered relevant to analyze was to realize the maximum amount of items purchased in a single purchase, with 559 items having been reached.

Data Analysis

For data analysis, we used mostly 3 columns that show part of ManyGiftsUK's business, which are: *'Country'*, *'Quantity'* and *'InvoiceDate'*.

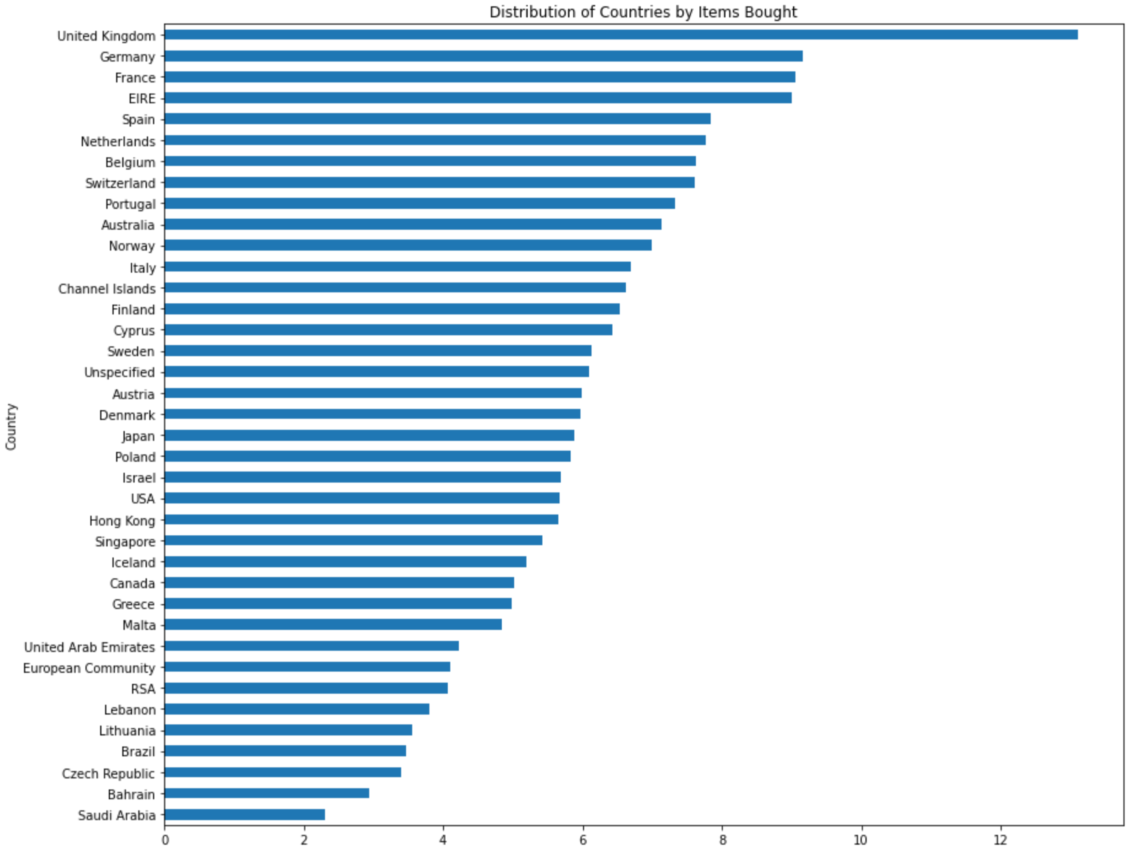

Figure 1 - Distribution of countries by items bought.

According to Figure 1, we can see that the United Kingdom corresponds to the country, with a considerable difference, with more products purchased by customers, this being the place of origin and foundation of the company.

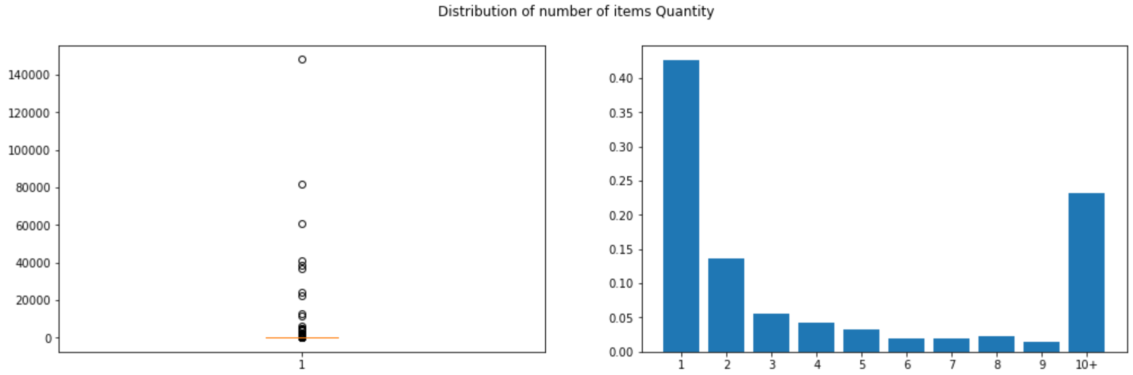

Figure 2 - Distribution of number of items quantity.

According to Figure 2, represented in the bar chart, it is possible to conclude that, sensibly, more than 40% of customers purchased only 1 product. Interestingly, it is also possible to verify that approximately 25% of customers purchased more than 10 items. For customers who buy only 1 product, it is essential to have a recommendation system to promote them to buy more than one item.

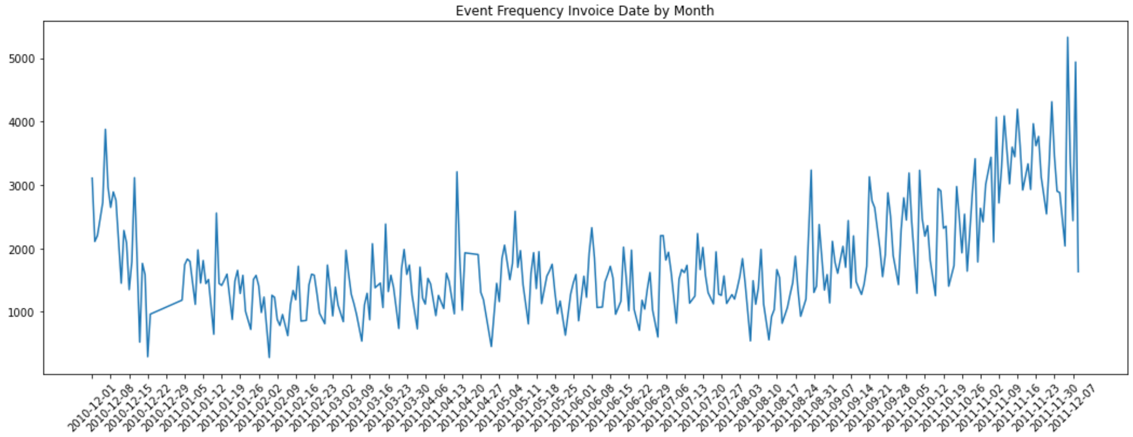

Figure 3 - Time series of Invoice date by month.

In order to obtain the result shown in Figure 3, it was necessary to apply a normalization so that it was possible to represent the clear form of information about the frequency of the 'InvoiceDate' event. From what we can see, from September to December there was an exponential increase in the issuance of invoices, which consequently led to an increase in product purchases during this period. Between January and August, there is a slight stability in demand.

CASE STUDY TOPICS:

1) Recommender system that appears in the website homepage and offers a wide range of products

Two models are created and ready to predict based on user's experience. Once a user reaches the website, based on their login or cookie, the customer ID will be detected and predicts items based on users previous bought and visit items. The data science team used two techniques to develop a recommender system:

- Apriori recommender system with Association Rules

  The Apriori Algorithm is a basic algorithm for the determination of the frequent itemset for Boolean association rules. The principles of Apriori state that "if an itemset is frequent, then all its subset items will be frequent". If the support for the itemset is higher than the support level, the itemset is "frequent". The algorithm is based on the prediction of items, which move from the previous stage on a regular basis. Apriori algorithm includes the type of association rules in data mining.

  The model must derive information from historical data and represent it in such a way as to be able to adapt the resulting model to new situations. Recommendation system can be split into four stages:

  1. The collection and pre-processing of raw data.
  2. Convert pre-processed data into an easily achievable form using the selected method such as Apriori algorithm.
  3. Use the previously developed set of association rules to report recommendations to the user with the function *"recommender_bought_bought"*.

  The model suggests complementary and similar items with recent bought items with a score that can be used for sorting suggested items.

- Collaborative Filtering for Implicit datasets using Alternating Least Square (ALS)

  Alternating Least Squares (ALS) is the model we'll use to fit our data and find similarities. ALS is an iterative optimization process where we for every iteration try to arrive closer and closer to a factorized representation of our original data.

  This method was implemented in the class and used in this project.

2) Cold Start problem: offer relevant products to new customers

There are always new users on the website, so we need to recommend based on cold start system. A model will be created and ready to predict items based on other items. The website waits for the user to choose an item, then suggests similar predicts based on selected item. The methods used are the same for the first case study topic.

3) Evaluation strategies and quality measure selection

In this section, it was created a new dataset based on the visitors dataframe with the information based on the total views of each visitor id, the number of items viewed and the total transactions each one did. After, this we added this information with the purchased items dataframe.

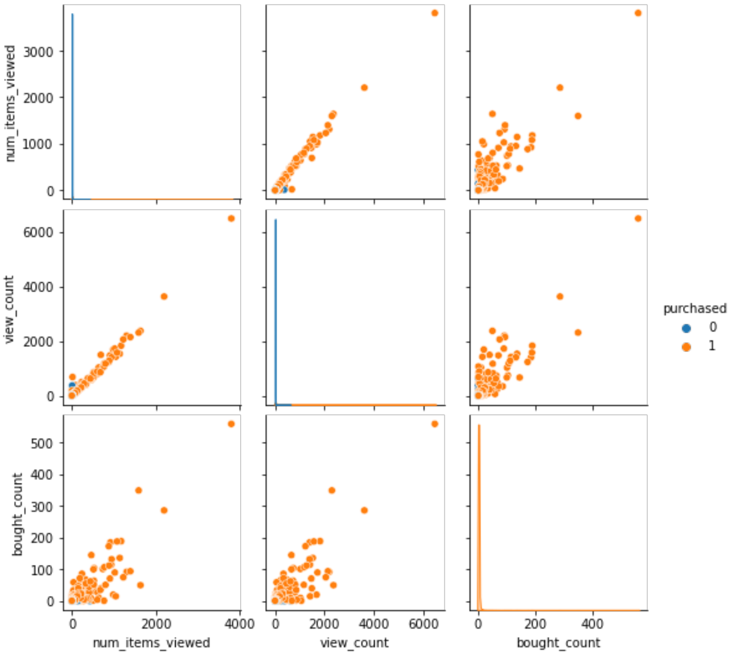

Figure 4 - Scatter plot between the dataframe variables.

The plot above clearly indicates that the higher the view count, the higher the chances of that visitor buying something. With this, the logistic regression was implemented to predict future visitor purchase behavior. The final accuracy with this model was 80.36%.

DEPLOYMENT AND MAINTENANCE PLANS

A set of challenges and recommendations have been made to implement this recommendation system more efficiently and functionally. The challenges are:

- Detailed exploration of the consumption information of each client;
- Assign relevant products to new customers (cold start);
- There was quite a lack of data for customer detection, as the dataset provided did not contain any metadata regarding each customer's visit;
- Changing user’s preference. We suggest based on user's experience, but the preference may change over the time.

Regarding the recommendations:

- Create a recommendation system based on the ratings of each product purchased by customers;
- Have a recommendation system that adapts to seasonal periods;

CONCLUSIONS

From the regression model, it was possible to calculate accuracy of 80.36%, allowing the prediction of customer behaviour when buying products. Thus, from the methods used it is possible, in an efficient way, to recommend to new customers, all the articles that relate to the search that is made by the customer (cold start). It is also possible to recommend to users with a history in the store, the best products that relate to all the purchases that were made as well as all the items that were searched for by the same in the store.

REFERENCES

- B. (n.d.). benfred/implicit. GitHub. https://github.com/benfred/implicit
- C. (2019a, May 7). Recommender System Project. Kaggle. https://www.kaggle.com/ccromer/recommender-system-project
- K. (2019b, June 12). Recommend LigthFM. Kaggle. https://www.kaggle.com/khacnghia97/recommend-ligthfm
- V. (2020a, September 29). Apriori Recommender System. Kaggle. https://www.kaggle.com/victorbonilla/apriori-recommender-system
- V. (2020b, October 3). Factorization Machine recommender in Sagemaker. Kaggle. https://www.kaggle.com/victorbonilla/factorization-machine-recommender-in-sagemaker