CSCM 35 - Big Data and Data Mining Coursework 1

Thomas Tasioulis - 998273

April 2, 2020

1 Introduction

The current big data era that currently we are living through has developed a large amount of data which are stored in chaotic databases. A big research field of Data Science is dealing with Data Mining and Knowledge Discovery in Databases (KDD) with primary purpose to extract knowledge from a database by using different techniques [5]. The research area of data mining is quite wide and have focused on a lot of different areas such as e-commerce, online retail, business intelligence, potential voters in elections, medicine and a lot of other areas [3, 5]. All those approaches tend to cluster the data into segments in order to give a better understanding of the database [2].

Retail data and business data are the predominant fields that are using data mining in order to extract knowledge and to transform this knowledge into profit and benefit based on the type of each one needs. The main reasons that those areas are highly interested in data mining is because the generated data is immense due to the high number of transactions that take place daily [4], leading to chaotic datasets with a lot of noise [3] that need interpretation in order to be useful.

In this paper we tried to analyze an online retail dataset in order to create some insights regarding to it and tried to extract knowledge about the clients of the online company and between the association between the different items. Moreover, we applied some supervised and unsupervised learning techniques to distinct different groups among the customers of the online retail store in order to acquire better knowledge about the dataset. Lastly, we provide a demo code which calculates - illustrates the former procedure in order to make the problem more comprehensive.

2 Methodology

2.1 Dataset and data preprocessing

Our dataset is consisted of 8 features “InvoiceNo, StockCode, Description, Quantity, Invoice Date, Unit Price, CustomerID” and “Country”. Among those features 3 of them “InvoiceNo, StockCode, CustomerID and Country” are nominal - categorical, 2 are quantitative (“Quantity and Unit Price”), Description is text and Invoice Date is date object. Regarding to the preprocessing, firstly we removed all the rows that contained *NA* at any column cell, secondly we removed also all the data that had the word “wrongly” in the description as they would create more noise and we could not rely on them. Moreover, we followed the approach of [3] and we removed the data that contained “Postage” in the description, while all rows that have had negative values in “UnitPrice” or “Quantity” removed as well. Nevertheless, those cells could be treated as “returning packages” or “disfunctional” ones, however we did not want to make this assumption in order to avoid potential fallacies. Lastly, from the date objects we extracted the day, month and year instead of using the whole object which included time to simplify the date objects.

2.2 Data Analysis

For the data analysis, we applied the apriori algorithm [1] in order to calculate the association rules between products that have been sold and to identify relationships between the products and how strong those relationships are. Firstly, we separated the dataset into 4 baskets based on “Country”, “United Kingdom, France, Germany and EIRE” and then we calculated the *Support*, $Lift$ and $Confidence$ for the frequent itemsets from apriori algorithm. By using the association rule mining we calculated its metrics as:

$$\mathit{Support}(X) = \frac{|\{t \,\epsilon\, T; X \subseteq t\}|}{|T|} \tag{1}$$

$$Confidence(X \implies Y) = \frac{\mathit{Support}(X \cup Y)}{\mathit{Support}(X)} \tag{2}$$

$$Lift(X \implies Y) = \frac{\mathit{Support}(X \cup Y)}{\mathit{Support}(X) \times \mathit{Support}(Y)} \tag{3}$$

Additionally to the association rule mining, we performed a clustering by using k-means and Gaussian Mixture Model (GMM) to a subgroup of customers which had bought less than 70 total gifts and less than 250£ total money. Moreover, we created some extra features for the unsupervised learning such as $TotalPurchases$ which is the number of transactions for each customer, $PaidMoney$ which is the total amount of money each customer paid to the store and $Money/Purchase$ which corresponds to the average money per transaction. By taking this subgroup we assumed that those customers consist typical customers instead of businesses and wholesalers. Hence, we can observe how those people can be separated. Additionally to that, we trained a single-tree classifier to fit the data in order to explain the relationships that led to each segment.

3 Results and Discussion

Findings showed that overall, EIRE had the most association rules compared to the other 3 countries with 1132 total rules 2 orders of magnitude higher than the other countries. For each country we have chosen minimum support to 5% instead of UK where minimum support was set to 2% due to UK had higher transaction number by 2 orders of magnitude which would led to insufficient number of rules if the minimum support was set to 5%. All the association rules were calculated for $Lift \geq 1$. Regarding to total rules Germany had 17 rules, France 86 and the UK 76 rules, respectively.

Figure 1 illustrates the histogram of confidence and lift for each different country. It is shown clearly that EIRE has the highest amount of confident rules besides the size of the total rules while Germany seems to lack in confidence for its rules as it descends vastly after 50%. In Germany basket, both confidence and lift tend to have small values and their total rules are not sufficient

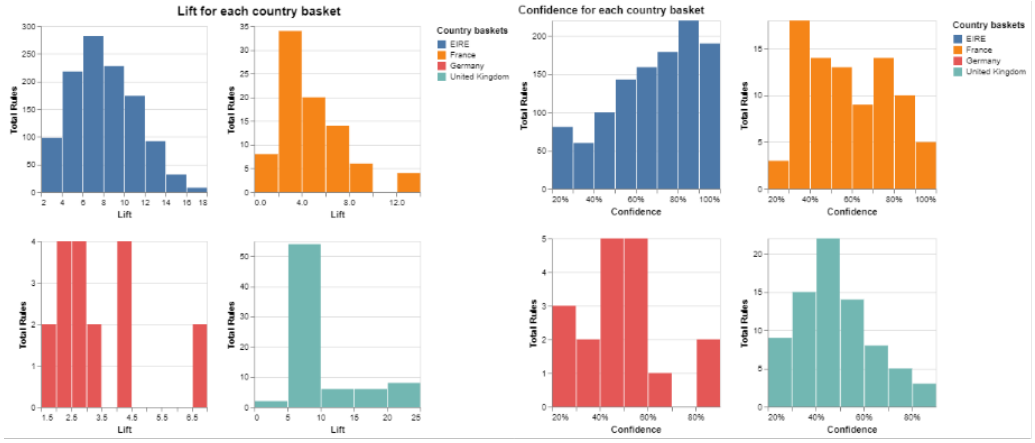

Figure 1: Histogram of Confidence and Lift for the 4 different basket sets

enough, leading to a result that this market place does not have a lot of contribution to our online store so we might not need to focus on this. It is largely instrumental to compare rules by taking into consideration both lift and confidence due to we could be led to inaccurate results if only confidence is concerned as lift evaluates how real and accurate is the confidence level of each rule.

Thus, for all 4 countries we have gathered all rules that have had $Conf \geq 70\%$ and $Lift \geq 10$ to see those relations. For this condition, the top rules that exist is “Tea sets”; more specifically, “Tea saucer”, “Tea plate” and “Teacups” tend to be always together in different colours with probability bigger than 70% at all cases and sometimes people tend to buy different colours of teacups or tea plates, followed up by gardeners-kit with a confidence level of 70% as well. The former countries seemed to have strong rules and weak rules in terms of confidence and lift however, it would be reasonable to see if there is overlaps of rules between countries even if those rules are not so strong in terms of metrics. Thus, in Figure 2, two Venn diagrams are illustrated which show how those rules are shared between the different countries. In this associations we have excluded Germany of our comparison as the rule number is low and their confidence meters are low as well, leading to more noise for our comparison.

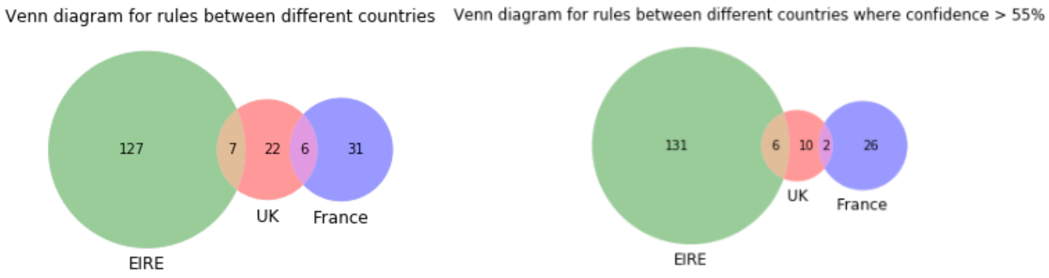

Figure 2: Venn Diagrams for rules between UK, EIRE and France

As Figure 2 shows, UK is the common denominator for France and EIRE, having 7 and 6 sets in common with the former countries. When confidence is concerned and more specifically for 55% or above, UK still have common sets with EIRE and France as well 2 and 6, respectively. For an in-depth investigation of the rules between countries and for the identification of the common sets we have produced a framework which is interactive and visualizes confidence, support, lift and Country. The framework can be found in the demo code that we have provided and a brief illustration is shown in Figure 3.

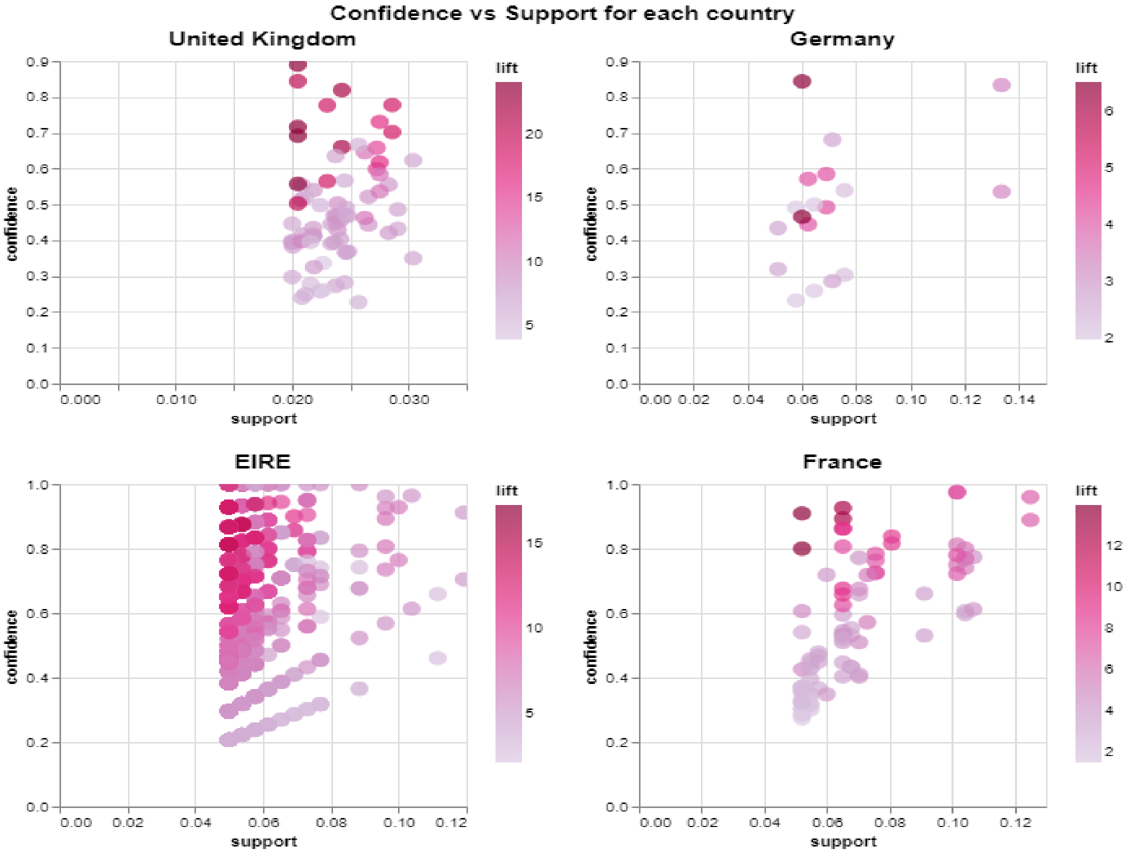

Figure 3: Interactive Visualization Framework for the rules among each country

From Figure 3, we can observe a correlation between support and confidence which is reasonable and expected based on the way it has been calculated. On the other hand, lift does not follow any correlation pattern but instead we can observe some clusters of items with higher lift which tend to have a decent amount of confidence. In the UK we observe that approximately 10 rules have enough lift and confidence, while in France this number is significantly lower (almost 6 or 7 rules); in Germany no more than 1 rule has the strong association we are looking for. Lastly, EIRE seems to have all different rule and when confidence is above 60% lift reaches a decent number as well, creating strong association rules. This could lead to results such as "*EIRE might have more stable* *customer patterns instead of UK besides tea and the latter 2 countries*". France seems to have associations around dolls while EIRE and UK have had similar association rules and every each one was around "Tea-sets" and "Garden-kits". On the other hand, in Germany, the strong rule was about different colour in Charlotte bags.

4 Conclusion

To sum up, we have seen how the associations among the 4 countries are presented and overall, UK and EIRE tend to have same customer habits in comparison with EIRE and France. However, the UK seems to have a lot of habits in common with France and EIRE. The strongest rules in terms of confidence and lift were "Tea sets" for EIRE and the UK with confidence rates of 85% or more especially when 2 items of the "tea set" were already bought. We also observed that the size of each sample did not really affect the size output as France, Germany and EIRE had almost the same sample but instead they led to completely different results when association mining rule was applied on them. Moreover, we have seen that despite the correlation between support and confidence, lift does not tend to follow this pattern and because of that we can have rules that are actually strong instead of statistically produce rules that have no causation. Of course at any case do we not know the causation of those rules, however we can assume some of them as the "tea set rules" or different colours in the same items. Lastly, due to the size of this report the clustering and the tree-classifier results were not included in this report however, they can be found in the demo-code with explicit details as well.

5 Future work recommendations

From this dataset we observed a lot of association rules; nonetheless, those rules could be optimized if we had the exact item instead of the whole description. For example, colour is a noise factor that interrupts us from creating even better rules or discover rules that we have not found them yet. Thus, we propose a possible solution for that is to use the algorithm "Bag of Words" or "Word2Vec" in order to extract the key-tags of each description and reduce the dimension of "Description" immensely. Lastly, the creation of clustering by applying different cluster algorithms and then the association rule mining rule on each different clusters might have interesting results as well.

References

[1] Agrawal, R., Srikant, R., et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (1994), vol. 1215, pp. 487–499.

[2] Berkhin, P. A survey of clustering data mining techniques. In *Grouping multidimensional data*. Springer, 2006, pp. 25–71.

[3] Chen, D., Sain, S. L., and Guo, K. Data mining for the online retail industry: A case study of rfm model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management 19*, 3 (2012), 197–208.

[4] Kohavi, R., Mason, L., Parekh, R., and Zheng, Z. Lessons and challenges from mining retail e-commerce data. *Machine Learning 57*, 1-2 (2004), 83–113.

[5] Olson, D. L., and Delen, D. *Advanced data mining techniques*. Springer Science & Business Media, 2008.