Demand Forecast for The Product Market Based on Online Resources

Zijie Chen[1, *, †], Jingqi Liu[2, †], Yingxu Wang[3, †], Siyao Wu[4, †]

[1]Business school, University of Leicester, LE17RH, Leicester, England

[2]Management school, University of Sheffield, S10 2TN, Sheffield, United Kingdom

[3]NIST International School, Sukhumvit Soi 15, Bangkok, Thailand

[4]International Business School, Beijing Foreign Studies University, 100089, Beijing, China

*Corresponding author. Email: jliu241@sheffield.ac.uk

†These authors contributed equally.

Keywords: Medium Size Company, Behavior Predicting, Customers Classification.

Abstract: As a medium-sized company, the article built a model based on the purchasing behavior of 4,000 customers over the course of a year and used it as the basis for an analysis to predict the purchasing behavior of new customers over the next year. The article first category all the products sold into 6 categories by product classification, then the article category the customers into 11 categories based on the type of product purchased, the number of visits, and the amount spent. Finally, a trainer was used to categorize consumers into eight categories. It was found that 75% of the customers were classified into the correct category but the remaining error was caused by a flaw in the model. The error can still be minimized by extending the test time and using more data.

1. Introduction

E-commerce has become a common way for businesses to expand their customer base beyond their local vicinity. Oftentimes, this results in a customer database being created, which can store all the purchases and orders made over a certain time period. This can be used to track sales and manage inventory. The notebook explores a more sophisticated use: Making a customer value measure with the data in order to numerically rate each customer based on their past orders. We collect lots of data to present the situation.

The data is a random sample of 5 entries in the customer database. The entire database explored is a set of purchases from 2010/12/01 to 2011/12/09.

Essentially, this is a database of invoices, which is common in any type of business nowadays. Each order has a unique invoice number, a description of the product, the quantity of the product sold, the price per unit, country, and a customer ID. The customer ID is generated within this python program and is not part of the original database.

An invoice number is a 6-digit number uniquely assigned to each transaction. If this code starts with 'c', it means that the order was canceled. Stock code is a unique 5-digit number assigned to each distinct product. The description is a string variable that briefly outlines the name and function of a product. Quantity refers to how many products were ordered in that specific transaction. Invoice Date is the time and date when a transaction was generated. Unit price refers to the cost of one unit of a product. Multiplying unit price by quantity will yield the final revenue of a transaction. Customer ID is a unique 5-digit code that is generated for each customer. Lastly, a country is simply the country in which an order was made. This is assumed to be where the customer also resides.

There is also data related to the country where purchases are made from. This figure is a tally that shows that the majority of the customer base is in the UK, followed by EIRE, Hong Kong, etc. From the figure, it can be seen that most of the customer base is in Europe. Later this will be displayed graphically and considered when analyzing customer value.

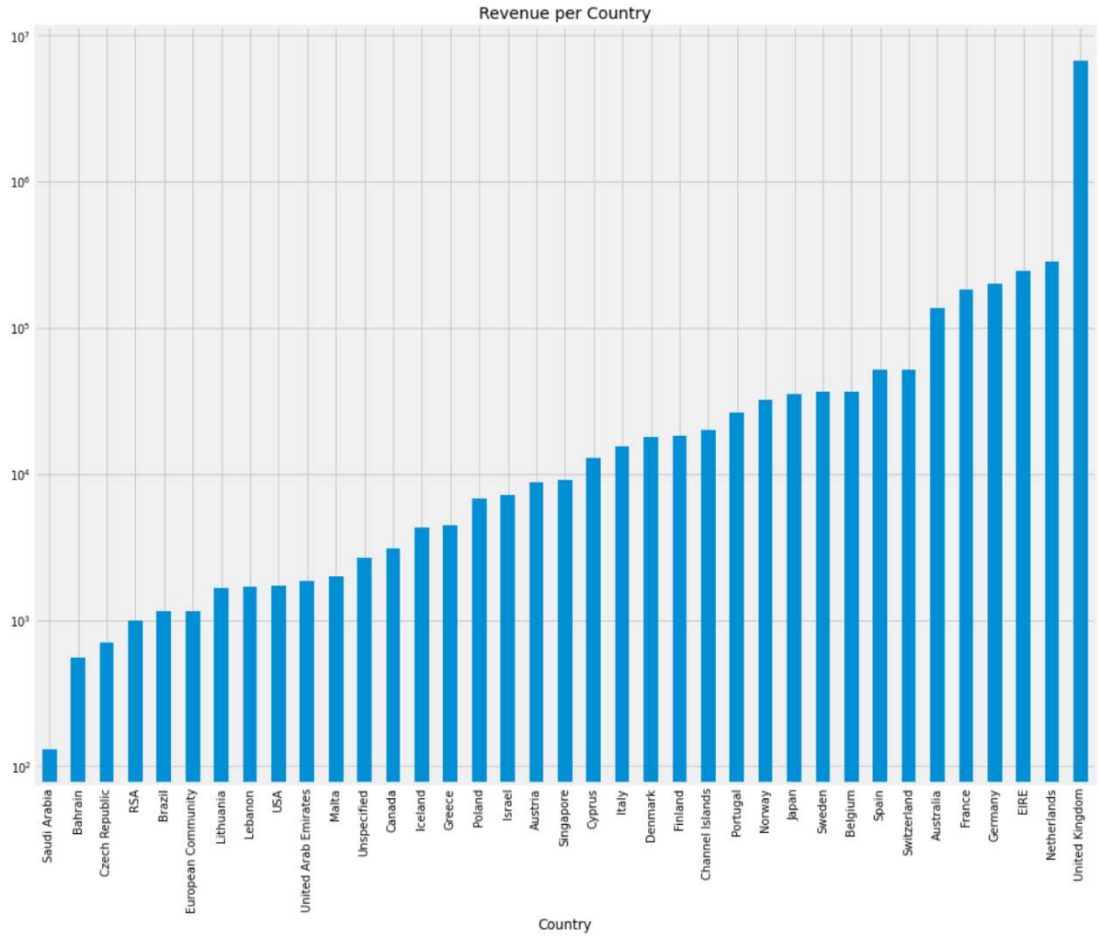

Figure 1. The different product revenue based on country

Figure 1 shows how total revenue is distributed across countries. The UK is responsible for the most revenue, being significantly higher than the other countries. A faraway second in the Netherlands, followed by EIRE and Germany. By grouping transactions by country, we can calculate how much revenue each country has generated. This can be used to target specific demographics and prioritize certain markets.

This database cannot be used immediately. Although the end goal is not to perform extremely sophisticated analysis, the data still needs to be cleaned. The main priority is to count and remove canceled orders. Canceled orders will only be needed later when creating customer value and will be retrieved then. The other step is to create a "basket price" which groups individual transactions of the same order together, just like a shopping basket. This is more practical to analyze since customers rarely order one item, rather it is a common occurrence to purchase multiple products, forming a "basket".

The thesis mainly analyzes the content of an E-commerce database that contains about 4000 customers' purchase lists. To build the recommendation model, we did literature research on machine learning and e-commerce topics, practiced quantitative analysis on product categories and customer segmentation, and trained the model with categorized data. Based on our research, we test the model's validity and it came out that 75% of customers were awarded the right classes.

This research examine and read data in the first phase, then go over the remaining entries and delete any canceled orders. Create Product Groups afterward when. The fourth section is called "Create Customer Groups." In the sixth section, Determine the worth of the customers. Finally, Analyze Item Co-Occurrence and Recommendations at the conclusion.

2. Product CATEGORIES

Filter all canceled orders starting with the letter c, and remove them to get a better analysis. Statistics of the number of orders with different time granularity: according to Month, Weekday, Day, Hour, to show the number of distinct values, and to make the corresponding histograms:

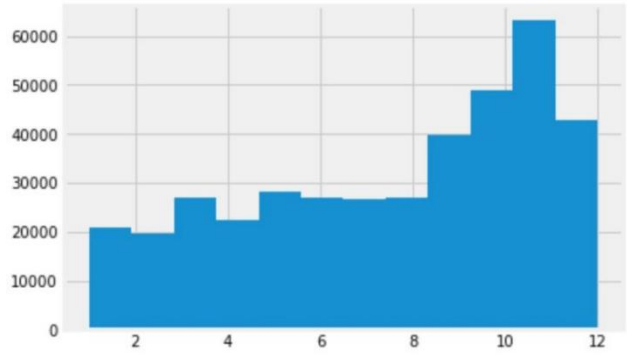

Figure 2. Month situation

The number of orders in November is the largest, and the number of orders in October is second only to November. The lowest order quantity is in February, and the average order quantity months are June, July, and August. The overall order quantity is rising.

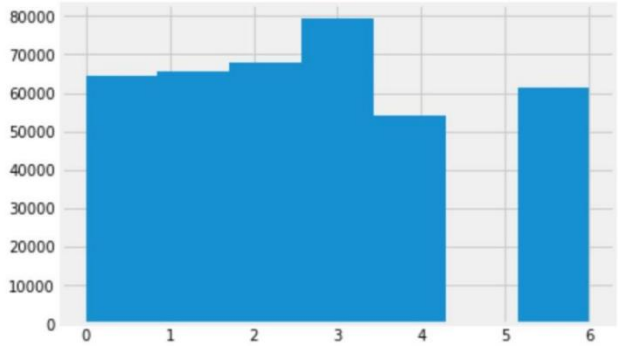

Figure 3 weekday situation

The number of orders placed on the third day is the highest, while the number of orders placed on the second day is only second to the third. Day 4 has the lowest order quantity, whereas days 0 and 2 have the highest order quantity. However, there are no orders were generated on the fifth day.

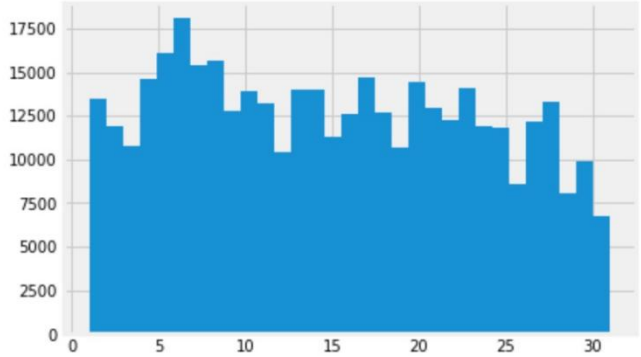

Figure 4. Day situation

The number of orders in 5th is the highest, while 4th is second only to 5th in terms of the number of orders. The lowest order quantity is in the 30th position, and order amounts are continually changing.

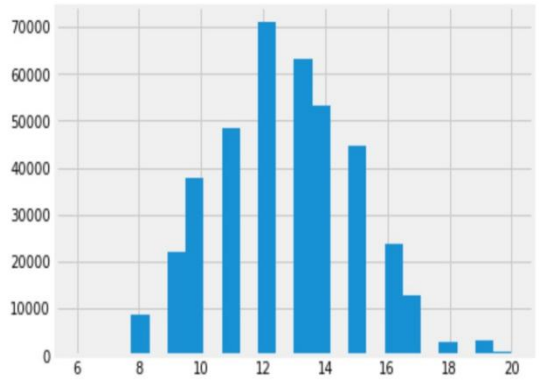

Figure 5. Different hour situation

The distribution of order quantities resembles a normal distribution. Early in the morning, the number of orders is the smallest, then gradually increases, and then gradually decreases in the afternoon. At midday, most orders are placed, while the least is placed early in the morning and late in the evening.

To do a product segmentation, filter out the stock code and calculate the actual price of each product based on the cleaning results. Then collect all prices according to country; each data frame item represents the pricing of a specific product type. As a result, orders are spread across many lines. And gather prices for each order to observe the value generated from each order.

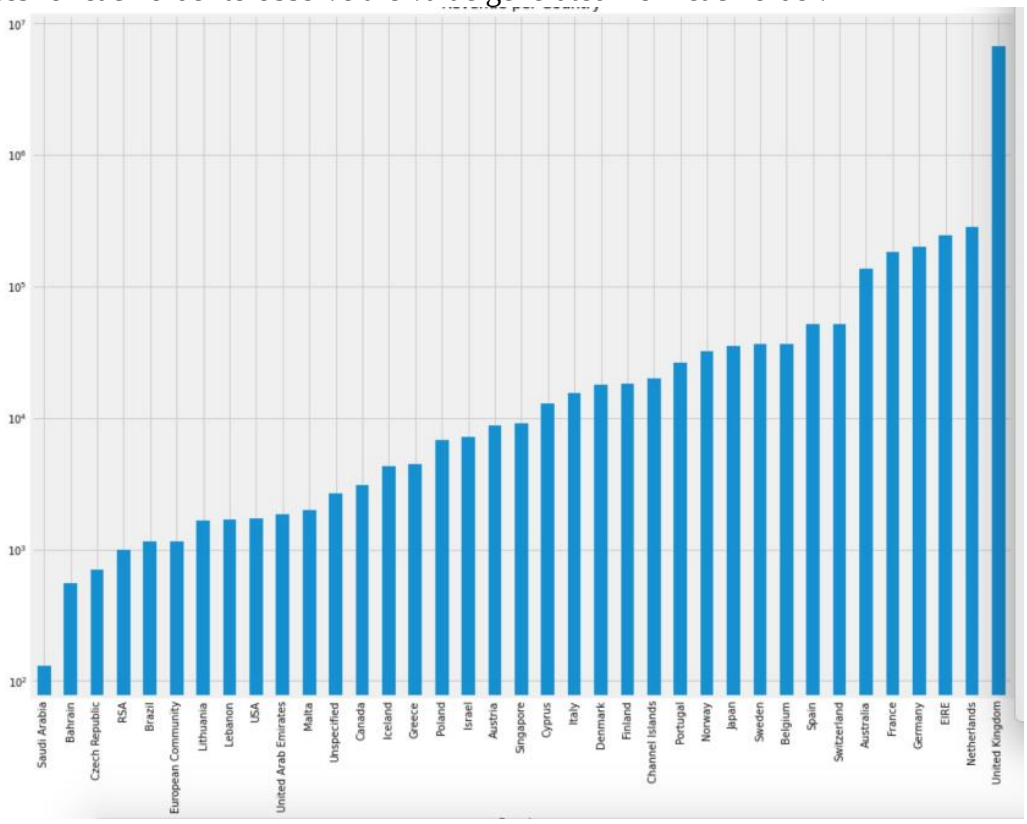

Figure 6. Country revenue condition

This graph displays each country's income, which is ranked from low to high. The United Kingdom has the greatest revenue, while Saudi Arabia has the lowest.

The price of each order is divided to show the proportion of orders in different intervals to the total order volume, shown below:

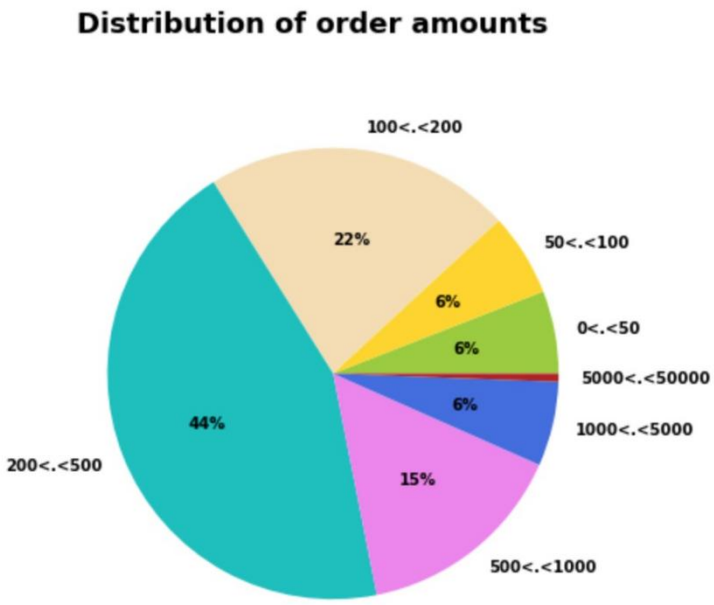

Figure 7. Pie chart of product

The first step in the analysis was to retrieve the product list; the number of words used was converted to a list and then sorted by the number of keyword appearances.

Given that 65 percent of orders had values over £ 200, it is clear that the great majority of orders are for quite substantial items. This distribution has some intriguing features, such as a spike of about 300 and a climb around 100.

There are more than 1400 keywords, but some of them are useless, only appearing a few times. The solution is to clear these words and only consider those that manifest more than 13 times as shown below:

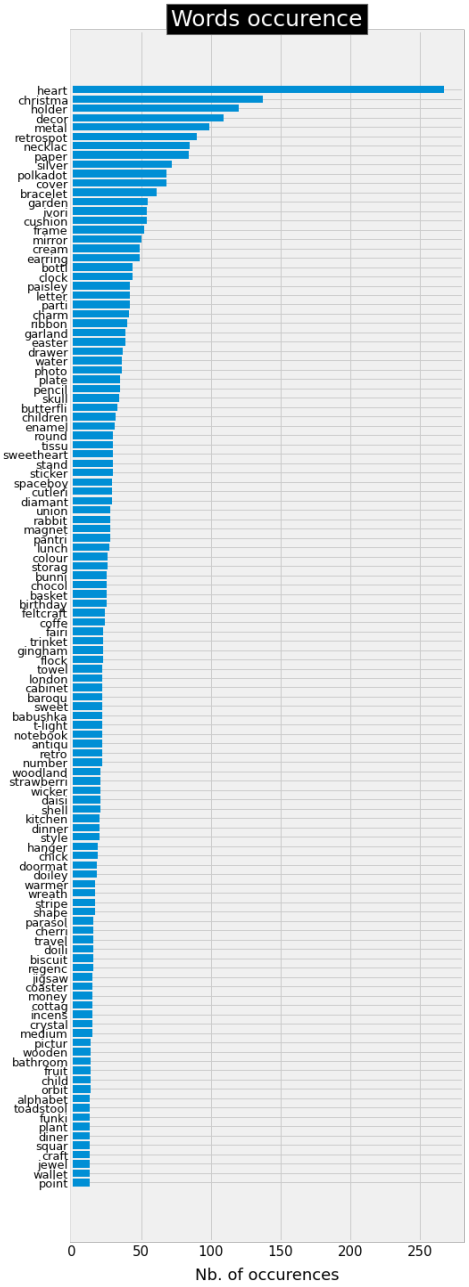

Figure 8. Word occurrence

The most commonly appearing word is heart, while the second most frequently occurring word is Christmas. The frequency of other terms does not change significantly, but the word with the lowest frequency is the point.

Then, using these keywords to form a product group, it is discovered that introducing a price range results in a more balanced number of pieces. In one extreme iteration, the price range precisely divides the six groups. We can express the profile percentage of each element of the different clusters to obtain insight into the categorization quality. As an example, consider the following:

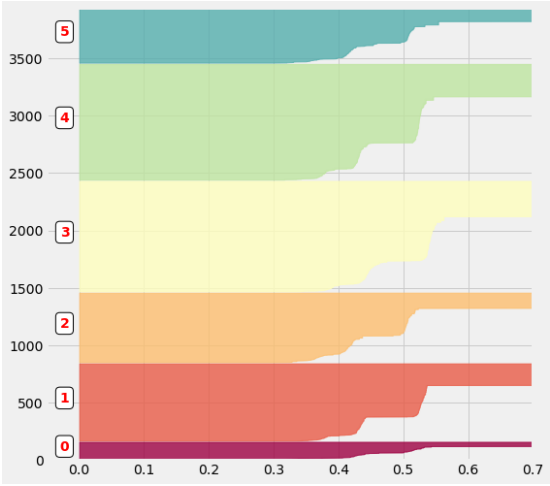

Figure 9. PCA word graph

In reality, the PCA did not perform sufficiently. Therefore the number of components was regulated. The decomposition was maintained for the purpose of visualizing the data.

3. Make Customer Value Measure

In order to measure the value of the customer, the following types of data are used in this paper. "number of visits" "max order" "unique product" and the "canceled%" max order is the biggest number the customer buy in a certain period . after that we define “f1” “f2” “f3” “f4” as the variance of these four numbers we used before

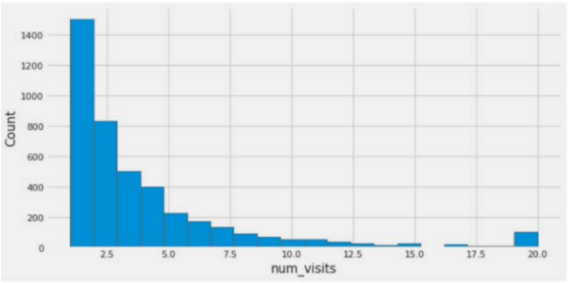

Figure 10. The count of the customer visit

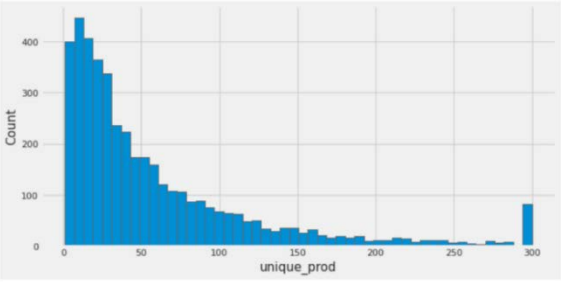

Figure 11. The count of the customer period

CVM (customer value measure) is an indicator for how valuable a customer is to a business. According to the customer value measurement, we divide customers in which we recorded internet into 4 groups, abcd, which is the value of the customers but not very accurate, there are any steps left before this part ends.

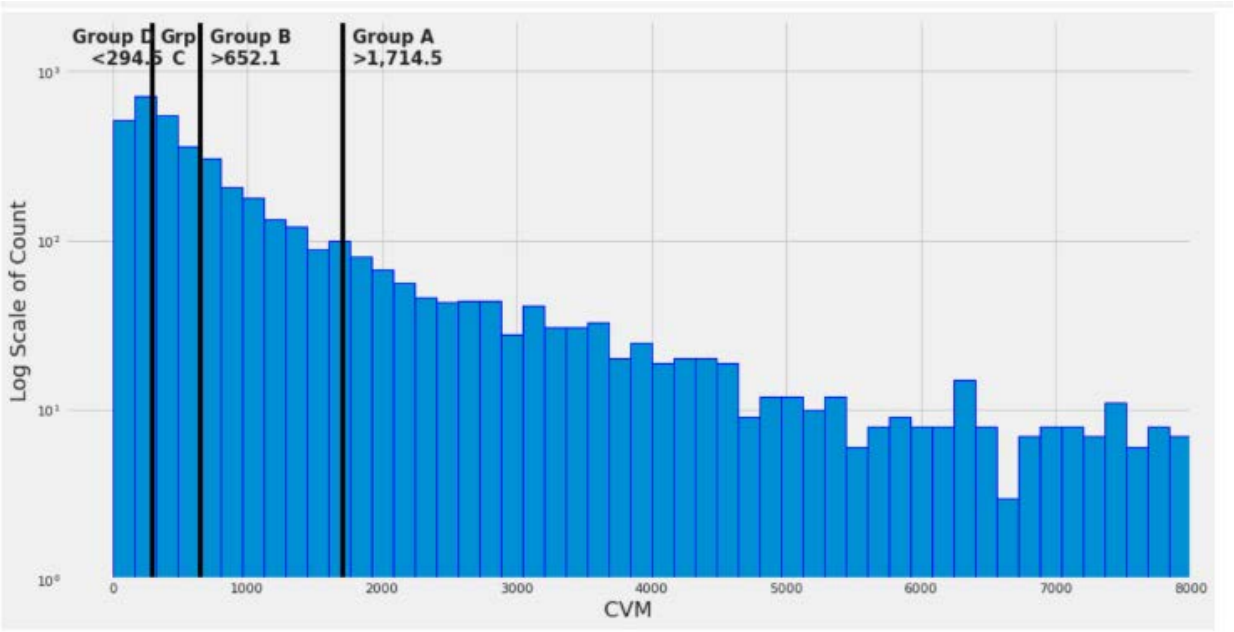

Figure 12. The count of the customer

The new variable being created is adjusted total, the gap between the amount of non-canceled total and the number of canceled total

After a series of calculating like analyzing the customer situation and buying frequency.

Then the customer is divided into 2 groups, repeat customers and one-time customers, in general, repeat customers are considered more valuable than a one-time customer because they can buy a lot of times and create more value than the majority of one-time customers. But the consequence shows that the value of lots of one-time customers is bigger than a repeat customer

The country is also important, it helps the enterprise to locate the target customer and make better decisions as an international company because Internet technology is also effectively used to help the company better understand its customers.

After these, we got the final cvm, it reminds us that the business of the firm spreads within many kinds of people, because the gap between the cvm seems very large, but it's every week in keeping the repeat customer, because the value of the one-time customer is large also, but it because the goods they sell are durable. The majority of customers are divided into the last group, the lowest customer value, which provides us with the last probability: this company lacks high-value trade.

4. Recommendations

The recommendation system aims to discover items that people will most likely buy given the goods in the previous purchase. To achieve the goal, the notebook first mined the hidden relationship between the goods under the direction of association rules. Later, it built a recommender system with the help of Python to generate predict ratings for each good given a certain customer ID after data training.

4.1. Association rules

Association rules mean that there is a strong relationship between customers that purchased goods A and also purchased goods B in the same transaction. Support, confidence, lift, and conviction are the four major perspectives to describe the association between two objects. Support is the relative frequency that the rules show up. Confidence is a measure of the reliability of the rule. Lift equals (observed Support/expected Support) if X and Y are independent. Conviction is (Expected X occurs without Y/ observed X occurs without Y) if X and Y are independent.

4.2. Building Recommender System

Scikit-Surprise and Spotlight are used to help find the association rule. After splitting the data into train set and test set, the notebook performed the SVD algorithm to train the system. Based on the predicted ratings, the system can make recommendations for a specific customer.
The recommendation model can help pinpoint up to 10 specific items to recommend to a specific customer.

5. Conclusion

E-commerce digitalizes the purchasing process of customers, thus allowing people to predict future purchases and provide recommendations based on shopping records. Our paper gathered more than 4,000 shopping records with 8 fields from an e-commerce platform, known as InvoiceNo, StockCode, Description, Quantity, Invoice Date, UnitPrice, CustomerID, Country. To start with, the notebook observed the shape and distribution of our data, clean the null value and remove canceled orders to laid a good foundation for further exploration. Then, the notebook used product description to group products into 6 different categories by performing K means clustering using cosine distance and PCA. Later, the notebook made customer categories based on their visiting habits and number of the purchase and it turned out 8 clusters.

After clustering all the products and customers, the notebook built a recommendation system by finding the relationship between products categories and that of customers. The ultimate goal of the system is to identify consumers' purchasing preferences and product demands and provide behavior patterns for subsequent product sales and sales forecasts. It shows that 75% of customers were awarded the right classes by our system.

The model didn't take seasonal effects into consideration, such as Black Friday and Christmas. Since the data is from 2010/12/01 to 2011/12/09, it doesn't support to make a year-on-year comparison to improve the model and reduce bias.

References

[1] Combe C. Introduction to E-business [M]. Routledge, 2012.

[2] Hoyer V, Janner T, Mayer P, et al. Small and medium enterprise's benefits of next-generation e-business platforms[J]. The Business Review, Cambridge, 2006, 10(2): 8.

[3] Liebowitz, J. ed., 2013. Big data and business analytics. CRC press.

[4] Duan, L. and Xiong, Y., 2015. Big data analytics and business analytics. Journal of Management Analytics, 2(1), pp.1-21.

[5] Chiang, R.H., Grover, V., Liang, T.P. and Zhang, D., 2018. Strategic value of big data and business analytics.

[6] Davidson J, Liebald B, Liu J, et al. The YouTube video recommendation system[C]//Proceedings of the fourth ACM conference on Recommender systems. 2010: 293-296.

[7] Smith B, Linden G. Two decades of recommender systems at Amazon. Com [J]. Ieee internet computing, 2017, 21(3): 12-18.

[8] Kraus, M., Feuerriegel, S. and Oztekin, A., 2020. Deep learning in business analytics and operations research: Models, applications and managerial implications. European Journal of Operational Research, 281(3), pp.628-641.

[9] Kalakota R, Robinson M. e-Business [J]. Roadmap for Success, 2000.

[10] Cao, G., Duan, Y. and Li, G., 2015. Linking business analytics to decision making effectiveness: A path model analysis. IEEE Transactions on Engineering Management, 62(3), pp.384-395. pp. 135–148. DOI: https://doi.org/10.1007/978-3-540-70545-1_14