

Using **Data Science** to make Your Online Retail a **Success**

BAX 453 Application Domains

Final Project Report

Team Members: Leo, Maggie, Nicholas, Siyu, Zimei

May 30th, 2018

Table of Contents

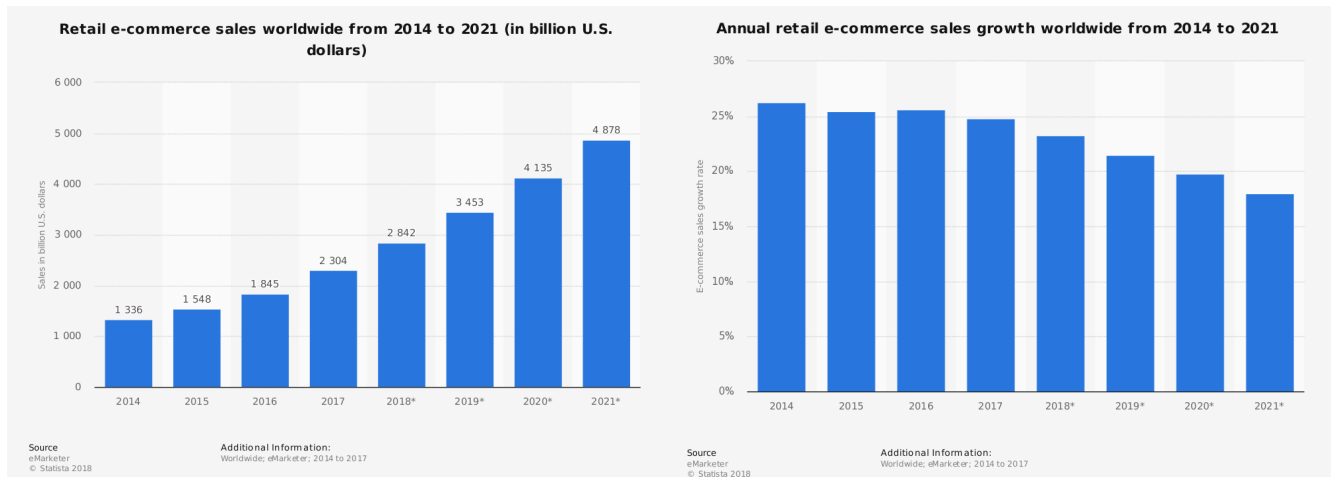
Table of Contents	2
The Current E-Commerce Landscape	2
A Recommendation Engine can be Helpful	4
A Use Case	5
Types of Data Required	5
Metrics Measured	6
Algorithm	7
Measure 1: Support	8
Measure 2: Confidence	9
Measure 3: Lift	9
Results	9
Application	11
Conduct Cross-Sell and Up-Sell Campaigns to Increase Sales	11
Design Effective User Interface and Product Combo Offers	12
Build Shoppers Profile to Optimize Customer Experience	12
Challenges of Recommendation Engine in E-Commerce	13
Data and Product Dependency	13
Lack of Innovation	14
Changing User Preferences	14
Filtering Useful Rules	14
Reference	15

The Current E-Commerce Landscape

As an online retail business owner, if you still think that your only competitor is Amazon, you are totally wrong! According to GoDataFeed, there are over 20 competitive marketplaces and shopping engines in the U.S. that can steal customers from you. Looking at the international arena, there are many more lower cost options which local consumers eventually will get exposed to. **The question is: how do you keep yourself in the loop?**



First of all, do not be intimidated! Despite the competitive landscape, retail e-commerce is still a fast expanding industry. Based on Statista's analysis, worldwide retail e-commerce sales will double from 2017 to 2021 and the annual growth rate will stay over 15% in the next 3 years. If you have already set up the infrastructure for building a online retail business, you should definitely work on improving it, but how?



Think of your shopping experience in a traditional retail store. You walk in, get greeted by a salesperson, browse the items on the shelf and pick up the one that you like. This is very similar to what you would do online: open the browser, search for what you want and click on your favorite item on the page. However, after you eventually decide if you will buy the item, things work differently online and offline. In a physical store, the salesperson will likely bring another similar or matching item to see if you like it; but over the internet, nobody does that job for you. I believe that most of us have bought the recommended items that were completely out of our original plan at least once in a retail store. To a business owner, this is also a significant amount of revenue. So how can you do the same in an online setting?

A Recommendation Engine can be Helpful

A salesperson can only make recommendations based on his/her personal experience. This means, when a new customer walks in, it is likely that the recommendations he/she makes are wrong and it is almost impossible to recommend things before the customer makes a choice. However, if your website can build a profile of each returning customer based on his/her past

shopping behaviors, the chances of making the correct recommendations will increase by a large amount. To do this, you need to conduct **market basket analysis**, and build a **recommendation engine**.

A recommendation engine is enabled by machine learning technologies. It collects data from previous shopping behaviours and search out the most relevant product in the entire catalog to present to consumers¹. Some of the major benefits of a recommendation engine in online retail are²:

- Boost Revenue
- Increase Customer Satisfaction
- Able to Generate Shopping Behavior Reports to Both Sellers and Buyers

A Use Case

Implement Recommendation Engine for UK Online Retailer

In the following report, we will demonstrate how a market basket analysis is conducted and implemented for a UK online retail. We will use past transaction level data to feed in to the machine learning algorithms and provide recommended items for each unique customer. In this way, we hope to improve customer satisfaction and boost sales.

Types of Data Required

Before digging into our recommendation engine analysis, let's first look at the types of data in UK online retail dataset. In general, we have 8 dimensions of data, respectively, the invoice number, the product code in stock, the product name description, the quantities of each product per transaction, the data and time each invoice is generated, unit price, customer ID who makes the purchase, and the name of the country.

¹ <https://www.omniconvert.com/what-is/product-recommendation-engine-ecommerce/>

² <https://www.business2community.com/ecommerce/benefits-recommendation-engines-ecommerce-sector-01985848>

- InvoiceNo:
 - Invoice number.
 - Nominal, a 6-digit integral number uniquely assigned to each transaction.
 - If this code starts with letter 'c', it indicates a cancellation.
- StockCode:
 - Product (item) code.
 - Nominal, a 5-digit integral number uniquely assigned to each distinct product.
- Description:
 - Product (item) name.
 - Nominal.
- Quantity:
 - The quantities of each product (item) per transaction.
 - Numeric.
- InvoiceDate:
 - Invoice Date and time.
 - Numeric, the day and time when each transaction was generated.
- UnitPrice:
 - Unit price.
 - Numeric, Product price per unit in sterling.
- CustomerID:
 - Customer number.
 - Nominal, a 5-digit integral number uniquely assigned to each customer.
- Country:
 - Country name.
 - Nominal, the name of the country where each customer resides.

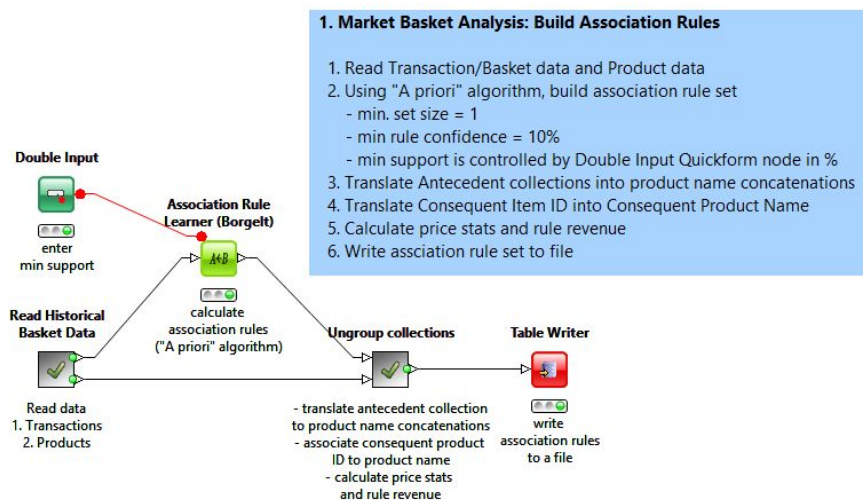
Metrics Measured

The proposed metrics below should be used after implementing the recommendation engine. They will help the online retail owner to assess if the recommendation engine is accurate and effective.

- Support, Confidence, and Lift of the Association Rule
- Dummy variable indicating whether the customer bought the recommended item
- Gross Merchandise Volume
- Revenue Gain
- Number of Active User
- User Purchase Frequency
- Average Basket Size (in quantity and in \$)
- Average spending in \$ per user (to find the most valuable customers)

Algorithm

In order to build a recommendation system, the core part is the Association Rule Learner node, which uses the Apriori algorithm³.



The apriori algorithm is able to decrease the number of items that are needed to examine. It simply means that if an itemset is not frequent, then all its subsets cannot be frequent. That is to

³ <https://www.knime.com/knime-applications/market-basket-analysis-and-recommendation-engines>

say, if we observe {apple} to be infrequent, {apple, beer} can be reasonably expected to be not frequent as much. So we may not take {apple, beer} into consideration when it comes to the list of popular itemsets.






















We can reduce the number of itemsets which need examination and get the list of popular itemsets using the apriori algorithm in the following steps:

- Step 0. Select itemsets which contain just a single item, such as {beer} and {apple}.
- Step 1. Calculate the support for itemsets. Keep the itemsets that meet your minimum support threshold, and remove itemsets that do not.
- Step 2. Generate all the possible itemset combinations using the itemsets you have kept from Step 1.
- Step 3. Repeat Steps 1 & 2 until there are no more new itemsets.

Association rules analysis is used to uncover how items are associated to each other. Three common ways are often used to measure association.

Measure 1: Support

This measure focuses on how popular an item is, measured the proportion of transactions where it exists. In the table below, the support of beer is 4 out of 8, or 50%. And the support of combination of apple and beer is 2 out of 8, which is 25%.

Transaction1	   
Transaction2	 
Transaction3	 
Transaction4	 
Transaction5	  
Transaction6	 
Transaction7	  
Transaction8	  

In this case the sales of beer could take a large proportion in the profit, and 50% could be considered as a support threshold, which means any items with support values larger than this threshold should be identified as significant items.

Measure 2: Confidence

This measure focuses on how likely item Y is purchased when item X is purchased, denoted as {X → Y}. This is calculated by the proportion of transaction with item Y, in which item X also appears. In the table above, the confidence of {beer → apple} is 2 out of 4, or 50%.

$$\text{Confidence}\{\text{beer} \rightarrow \text{apple}\} = \frac{\text{Support}\{\text{beer}, \text{apple}\}}{\text{Support}\{\text{beer}\}}$$

One drawback of the confidence measure is that it only takes how popular beer is into consideration, but not apples. Other transactions, like transaction 7 may contain apple. If apples are also very popular in general, chances are that transactions containing beer will also contain apples, thus inflating the confidence measure.

Measure 3: Lift

This measures how likely item Y is purchased when item X is purchased, while controlling for how popular item Y is. In the previous table, the lift of {beer → apple} is 8/6, according to the formula. A lift value larger than 1 means then item Y is likely to be bought if item X is bought, while a value less than 1 means that item Y is unlikely to be bought if item X is bought.

$$\text{Lift}\{\text{beer} \rightarrow \text{apple}\} = \frac{\text{Support}\{\text{beer}, \text{apple}\}}{\text{Support}\{\text{apple}\} * \text{Support}\{\text{beer}\}}$$

Results

After implementing the Apriori algorithm, we got a list of association rules like {beer -> apple} and its association metrics including support, confidence, and lift.

Here is the table we sampled from the results:

X (antecedents)	Y (consequents)	support	confidence	lift
HERB MARKER MINT, HERB MARKER THYME	HERB MARKER ROSEMARY	0.01	0.95	75
KNITTED UNION FLAG HOT WATER BOTTLE	RED WOOLLY HOTTIE WHITE HEART	0.06	0.65	5.58
SET OF 3 CAKE TINS PANTRY DESIGN	JAM MAKING SET WITH JARS	0.02	0.36	5.02



Pictures above show the items for the second column. And here are some explanation for the support, confidence and lift number.

- Support of 0.06 shows that this combination of the two items takes 6% part of all the itemsets, which is remarkable. This combination is potentially a popular itemset and we need to further dig into some insights.

- Confidence of 0.65 shows that the white heart bottle is 65% likely to be purchased when a union flag one is purchased. One thing should be noticed is that we did not take it into consideration how popular white heart one is in the market, thus let's dig into the lift.
- Lift shows 5.58, which is greater than 1. That means if we control how popular white heart one is in the market, we can still say it's very likely to be bought if union flag one is bought. So in this case, we would highly recommend that these two hot water bottle can be put together for sale, in order to increase the benefits.

Application

Almost everything we read, see, or buy on the internet these days has been selected by a recommendation engine, for instance, news articles on Google, products shown on Amazon, or movies on Netflix. Recommendation engines are big business. The Netflix recommendation engine reportedly makes the company \$1 billion a year by recommending videos and keeping viewers watching so they don't cancel their accounts⁴. Amazon, meanwhile, which has woven product recommendations throughout its site, saw net sales rise about \$20 billion from 2015 to \$136 billion⁵ in 2017.

All these amazing results brought by recommendation engines are products of personalization of service empowered by recommendation algorithms. Personalization is the act to provide products and services customized to individuals on the basis of information of their needs and behavior. Recommendation engines are cultured and learned from explicit or implicit feedback of users. Personalization can be achieved by the aid of market basket analysis. Every shopper's basket has a story to tell and market basket analysis is a common retail, analytic and business intelligence tool that helps retailers to know their customers better. Using the recommendation engine we built based on advanced market basket analysis, online retailers can deploy the model in plan and implement more effective marketing efforts.

⁴<https://www.fool.com/investing/2016/06/19/how-netflixs-ai-saves-it-1-billion-every-year.aspx>

⁵<https://www.statista.com/statistics/266282/annual-net-revenue-of-amazoncom/>

Conduct Cross-Sell and Up-Sell Campaigns to Increase Sales

Cross-sell and up-sell campaigns show the products purchased together, so customers who purchase the Samsung monitor can be persuaded to pick up high quality HDMI cable.






Frequently bought together



- ✓ **This item:** Samsung UE510 LED DISPLAY Monitor, Black, 28" 4K (Certified Refurbished) **\$229.99**
- ✓ **HDMI Cable 6ft - HDMI 2.0 (4K @ 60Hz) Ready - 28AWG Braided Cord - High Speed 18Gbps - Gold Plated...** **\$9.99**

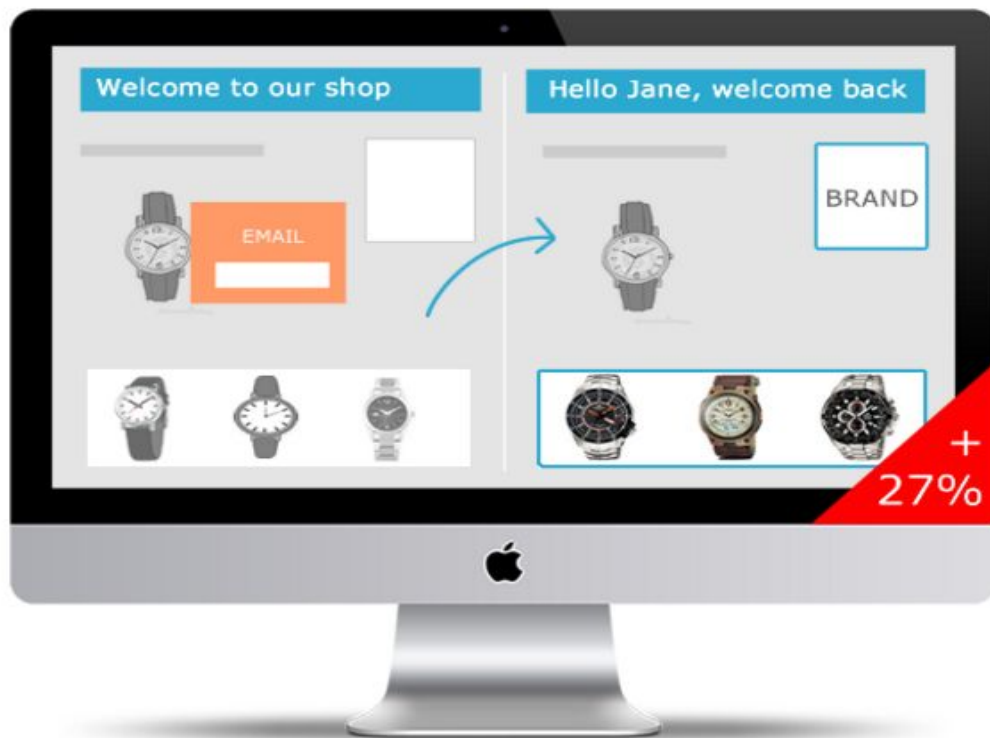
Design Effective User Interface and Product Combo Offers

Product combos are based on product affinities, developing combo offers and design effective user friendly interface by focusing on products that sells together.

Quantity	Buy 1 Get 1	Combo	Free Gift
BIG 6 PCS  50 x 50 x 5cm (19.6 x 19.6 x 1.9 in)	 Buy one, get one free Apple iPhone 6 Tempered Glass Screen Protector - intl 452 B -29% 633-8	 5 x CAPPUCCINO + 3 x CAFÉ AU LAIT (Bundle) NESCAFÉ Dolce Gusto Cappuccino x 3 Boxes + Café Au Lait... ★★★★★ (2 reviews) SGD 59.40 -17% SGD 71.40	 +FREE 1 X Sunglasses 

Build Shoppers Profile to Optimize Customer Experience

Shoppers profile in analyzing market basket with the aid of data mining over time can help retailers to get a glimpse of who their shoppers really are, gaining insight to shoppers' spending range, buying habits, likes and dislikes, and purchase preferences, and optimizing customer experience based on the information retrieved.



Challenges of Recommendation Engine in E-Commerce

Data and Product Dependency

Because of data and product dependency, recommendation engines won't be able to understand every single customers. Their knowledge of a person is based on user activity within

an interface and purchase records, the efficacy of the algorithm itself, and the wisdom of the crowds. Facebook recommendation algorithms, for example, have been shown to poorly reflect a user's actual preferences. Although some users' behavior can be modeled, other users do not exhibit typical behavior. These users can skew the results of a recommender engine and decrease its efficiency.

Lack of Innovation

In trying to select a winner product, recommendation engines tend to reproduce stereotypes and reinforce existing practices. Recommendation algorithms usually don't support the Long Tail enough and just recommend obvious items. Moreover, while algorithms can learn from weighted variables in a product database, the listing itself often has a human touch, including the product description, image, and overall web design. A recommendation engine may thus favor one product over another simply due to better imagery and not product quality.

Changing User Preferences

User preferences might change rapidly. while today some users have a particular intention when browsing on an online retailing website, tomorrow they might have another intention. Especially, in some item categories, for example, fashion clothing, user preference can be totally different after a short period of time, due of the change of fashion trends. Recommendation engines are usually behind the changing trends and take quite some time to update and pick up the changing user preferences.

Filtering Useful Rules

A problem with recommendation engines based on market basket analysis is that sometimes too many rules are generated and it becomes important to filter these rules to select the strongest or

the most relevant. The bottom line is that for performing an efficient market basket analysis, simply applying the algorithm on the available data may result in a profusion of association rules. If we are not careful about how we apply these rules we may lose some valuable information.

Reference

IndraStra Global. (2016, March 08). IT | Benefits and Challenges of Data Mining in E-Commerce. Retrieved from <https://www.indrastra.com/>

Five Problems of Recommender Systems. (2009, January 29). Retrieved from https://readwrite.com/2009/01/28/5_problems_of_recommender_systems/

Deshpande, B. (n.d.). Challenges in filtering useful rules from a market basket analysis. Retrieved from <http://www.simafore.com/blog/bid/113493/Challenges-in-filtering-useful-rules-from-a-market-basket-analysis>

Market Basket Analysis and Recommendation Engines. Retrieved from <https://www.knime.com/knime-applications/market-basket-analysis-and-recommendation-engines>