Association Rule Mining for a UK Online Retail Company

true true true true

Feb 26, 2020

Abstract

The retail industry is one that has changed far beyond recognition with the advent of internet. From once having to make a list, physically travel to a store, buy & haul your purchase yourself, to now simply ordering your needs online and getting it delivered in less than 24 hours, retail buying has become far more convenient. This buying convenience has also introduced challeges on the part of the sellers. The once obvious buying patterns are no longer obvious and requires complex analyses to understand customer preferences. In this project, we attempt to help a UK based Online Retail store understand their customers' buying patterns.

Background

One of the most powerful tools in online retail is a recommender system. Such a system helps sellers mine through their sales and unearth important associations between their products. In turn, such associations can be presented to customers as recommendations. Our client, the online retail store wishes to build a long term strategy based on the understanding this project gives them.

Objective

The objective of our analysis is to develop an unsupervised model using Machine Learning techniques and the CRISP-DM framework on the available online retail data to identify & listdown key product purchase association patterns. This result in turn will help develop better product recommendation.

Data Analysis

The original data set, "Online Retail.csv" is sourced from the UCI Machine Learning Repository. It is a transnational data set which contains all the transactions occurring between 01\12\2010 and 09\12\2011. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

Initial Data Exploration & Cleaning

The original data set in this case is a rather simple data, with features that are obvious & straight forward. A quick look at the header of the data set, helps one understand this.

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 8:26	2.55	17850	United Kingdom
536365	71053	WHITE METAL LANTERN	6	2010-12-01 8:26	3.39	17850	United Kingdom
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 8:26	2.75	17850	United Kingdom
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 8:26	3.39	17850	United Kingdom
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 8:26	3.39	17850	United Kingdom
536365	22752	SET 7 BABUSHKA NESTING BOXES	2	2010-12-01 8:26	7.65	17850	United Kingdom

We learn from the original data that all invoices that start with a "C" are cancelled orders and therefore are removed from the data. Also, the patterns will be interpreted based on the product descriptions and hence all rows with empty descriptions are deleted too. That leaves us with data that has all valid invoices and valid descriptions

Models

From the onset, it was clear that an association rule mining model will answer the question. In this light, it was decided to use the following models for association rule mining

- Apriori algorithm
- FP Growth algorithm
- ECLAT algorithm

All association rule mining algorithms have 3 very important parameters * Support * Confidence * Lift

Support: Support is the proportion that an item represents in the total transaction dataset. Example, *support(PINK REGENCY TEACUP AND SAUCER) = (Trasanctions featuring (PINK REGENCY TEACUP AND SAUCER))/Total Transactions

Confidence: confidence is defined as the probability that an item combination was bought. Example, *confidence(PINK REGENCY TEACUP AND SAUCER & GREEN REGENCY TEACUP AND SAUCER) = (Total transactions with both PINK & GREEN TEACUP & SAUCER)/Total Transactions

Lift: Lift is defined as the increase in the chance that an item combination is bought when a single item in that combination is bought. Example, *Lift(PINK & GREEN REGENCY TEACUP & SAUCER) = Confidence(PINK & GREEN REGENCY TEACUP & SAUCER)/support(GREEN REGENCY TEACUP & SAUCER)

Apriori algorithm

The apriori algorithm is one that is custom built for mining for association rules in a dataset. This algorithm works on datasets that are transactions. In our case, we have already converted the dataset into "transactions" prior to running the model on. The apriori algorithm, as the name suggests works on the basis of previously mined information. It is a breadth first algorithm, where it mines for frequent subsets by traversing across the breadth of each level of transaction.It starts by creating a "candidate set" which is a table of count of each unique item in the dataset. In the next step, it expands by counting each frequently appearing pair and then three items in the subsequent step and so on and so forth. At each step and finally, the stop criterion for the algorithm is decided by the "support" metric explained above

```
#apriori algorithm
apriori.rules <- apriori(trans, parameter = list(supp=0.01, conf=0.8,minlen = 3, maxlen=5))
inspect(apriori.rules)
```