Online Retail Industry

Meghna Venkatesha (venkatesha.m@husky.neu.edu)
Ruchitha M Shanmugha Sundar (midigarahallishanm.r@husky.neu.edu)

Data Mining – Final Project
Northeastern University
Fall 2018

Abstract

Online retail is expanding with the growth of e-commerce. With the growth of e-commerce comes a huge database of customers. It is import for the online retailers to understand what customers' expectations and requirements are. To get a better understanding of what customers' wants online retailers are trying to implement data mining techniques and work around the customers' needs and satisfaction. But the lack of technical knowledge and resource to implement the available technology makes it difficult for the online retailers to provide best customer service. So it is important for the retailers to mine data that fetches information related to products that are bought at a higher frequency, products with good reviews, product that has failed to provide customer satisfaction, etc. With this data they can concentrate on how a product can be marketed and how to improve a product based on user feedback.

In this way retailers can concentrate on each customer and extend the shopping experience to a personal level. We can generate various categories based on features such as place of residence, time of purchase, busiest week, busiest hour, etc. and use this information in marketing strategies. These categories can be achieved by applying k-means clustering algorithm and decision tree. Along with providing best customer services, retailers can also stock their warehouse accordingly to maintain demand-supply balance.

Introduction

Online retail industry is a multibillion industry. With the recent boom in this industry the way customers utilize the facilities provided has also changed. Online retailers generate the atmosphere where each customer feels they are in the lime light.

With World Wide Web it is an easy task to gather information about a customer, get to know their interests, likes and dislikes, etc. With all these details a list of relevant products can be suggested. But along with the growth of online customers, competition among the online retailers has also gone up. Online retailers face different challenges everyday with increasing competition. For an online retailer to be the best in industry, they have to come up with a best marketing strategy to advertise their product as unique and best in market and to attract customers. In order to achieve this online retailers consider many factors as follows,

- Does the sale of a particular product depend on any occasion, season, time, etc.?
- What are the profits gained from a particular customer? How valuable are the products purchased by them and what other products are purchased at the same time?
- For what duration a product was viewed by the customer, and does all the products viewed by the customer during this time span fall under the same category?
- Has the customer responded to any promotional offers in the past? If they have responded, what type of offers do they usually respond to?
- Is it possible to categorize customers as a valuable and loyal?

To identify these factors online retailers have considered data mining techniques. All the data that are customer specific are collected and various data mining techniques are applied to find answers to the above mentioned questions. Online retailers are successful at deriving this solution theoretically, but when it comes to implement this idea into reality they faced the problem of lack of technical knowledge and technical expertise.

This project provides a solution to the above problem. One of the approach to solve this problem is using the RFM model i.e. Recency, Frequency and Monetary Model. Customers can be categorized into various categories using k-means clustering algorithm. Customers can be categorized as valuable, loyal, seasonal shopper, occasional shopper, product bloggers, promotional shoppers, etc. Based on these categories online retailers can strategize a customer specific marketing technique and recommend certain products to certain customers, on certain occasions or seasons, provide particular promotional offers in which the customer is interested in, etc. Further details about how the data is handled and how a solution is provided can be found in the upcoming sections.

Methodology

For the purpose of this project and as an example of how a solution can be provided for the above mentioned problems, we have considered a UK based online retailer data. For better understanding purpose let us first understand the background of the data being used. The data consists of all the transactions that took place between 01/12/2010 to 09/12/2011. This UK based online retailer is a non-store online retail. The main products of the company are occasional based gifts. They also have tie ups with wholesalers. There are 8 attributes associated with the data and they are as follows,

ATTRIBUTE NAME	DESCRIPTION	DATA TYPE
InvoiceNo	Invoice number, a 6-digit number uniquely assigned to each transaction.	Nominal
StockCode	Product code, a 5-digit integral number uniquely assigned to each distinct product.	Nominal
Description	Product name.	Nominal
Quantity	The quantity of each product per transaction.	Numeric
InvoiceDate	The date and time each transaction was generated.	Numeric
UnitPrice	Product price per unit in sterling.	Numeric
CustomerID	Customer number, a 5-digit integral number uniquely assigned to each customer.	Nominal
Country	Name of the country where each customer resides.	Nominal

First the data needs to be prepared in order to perform Recency, Frequency and Monetary model based clustering analysis. Once the data is prepared clustering can be applied on the dataset. The following procedures are followed to arrive at the end result:

- Customer Segmentation
- Dimensionality Reduction
- Principal Component Analysis
- Exploratory Analysis
- Cluster analysis
- Market basket analysis
- Machine Learning Predictions

Code, Results and Discussion

Customer Segmentation:

Applying Pareto Principle – This is commonly referred to as the 80-20 rule on our dataset by applying it to our RFM features. Pareto's rule says 80% of the results come from 20% of the causes. Similarly, 20% customers contribute to 80% of your total revenue. In our case 80 % of the revenue is generated by less than 20 % of the customer population.

Customer segmentation using RFM score involved using quartiles. We assign a score from 1 to 4 to Recency, Frequency and Monetary. Four is the best/highest value that can assigned to a customer. One is the lowest/worst value that can be assigned. A final RFM score is calculated simply by combining individual RFM score numbers. With more granularity, it becomes difficult as we have to deal more combinations of the RFM score. (Quintile -5 would lead to 555 possible combinations, which becomes a challenging task). Two different segmentations are used to generate Quartiles as High Recency is not good where as high monetary and frequency is a good aspect for the retailer with respect to revenue. Generated quartiles are combined to get RFM score. The following result is obtained:

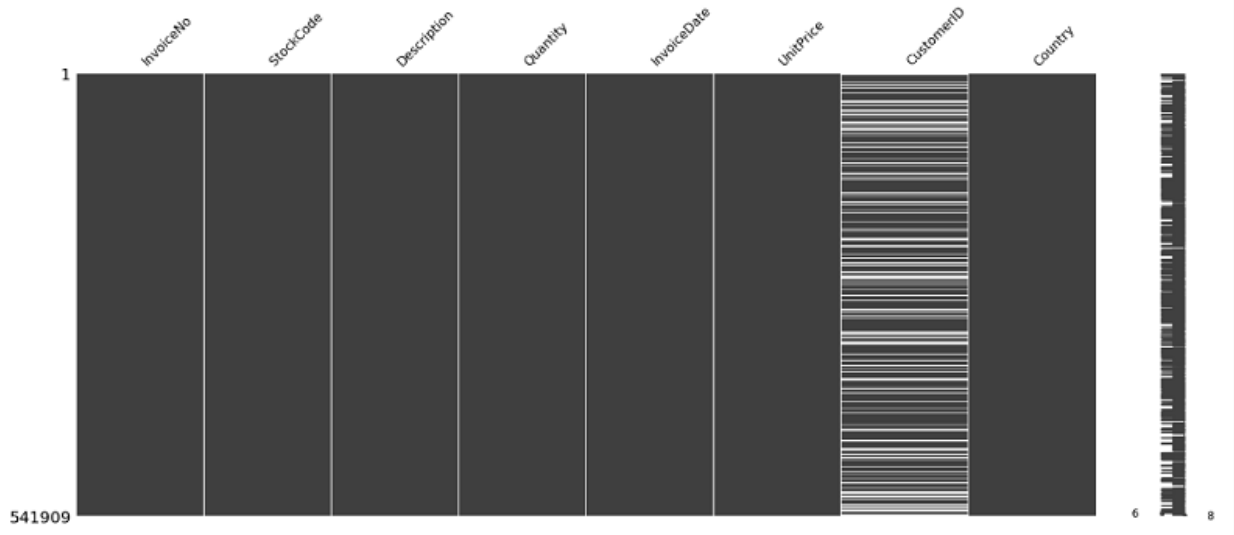
- Best Customers: 356
- Loyal Customers: 752
- Big Spenders: 966
- Almost Lost: 64
- Lost Customers: 9
- Lost Cheap Customers: 353

The above generated data can be used to come up with marketing strategies like make use of best customers for referral, identify the customers at risk and try to retail them by giving those additional offers apart from the ongoing seasonal offers, advertise for products by identifying the potential buyers, etc.

Data Exploratory Analysis:

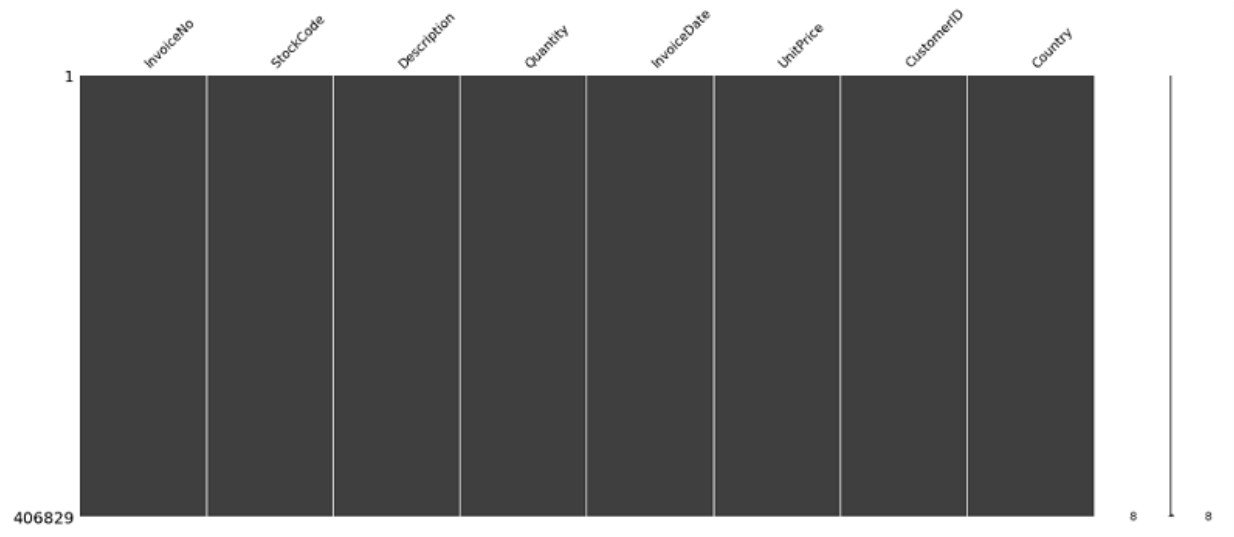
First we go over the data cleaning and feature engineering part. We check for any missing records in the dataset.

In our dataset there are missing records for features CustomerID.



Missing records

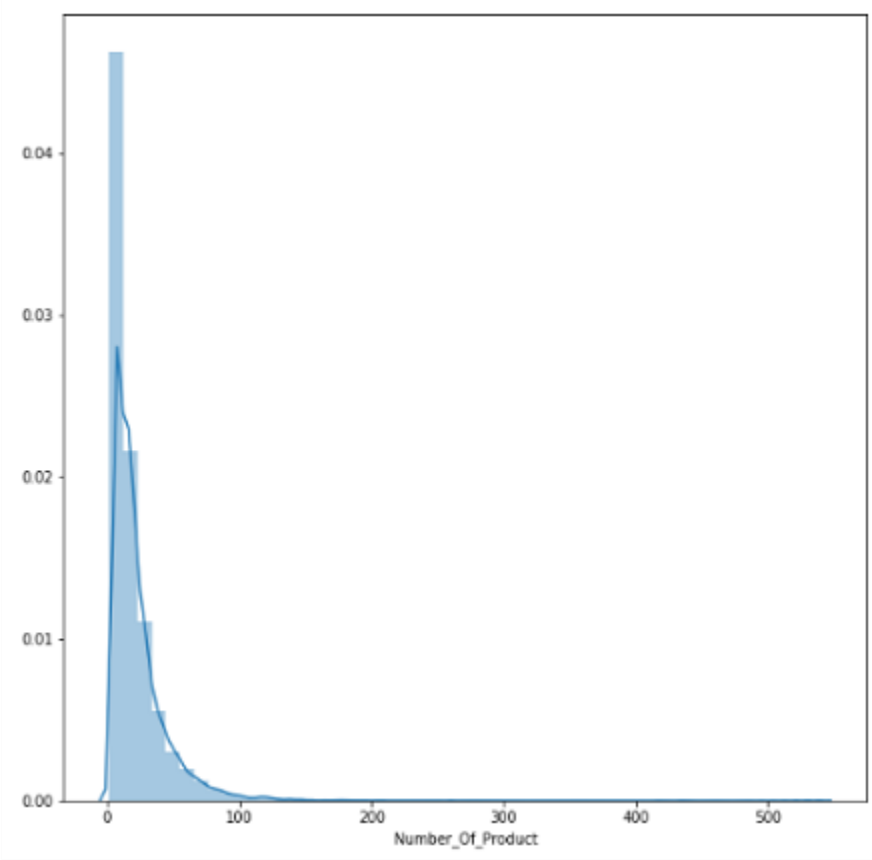
It is important to remove these records from the dataset.



Missing records removed

Once these missing records are cleaned the data is formatted according to the requirements. The following features are explored:

- Cancelled order
- Shopping basket



From the above plot we can make out that the number of products purchased by the customer in each transaction is less than 20.

- Data based on country

Dimensionality Reduction:

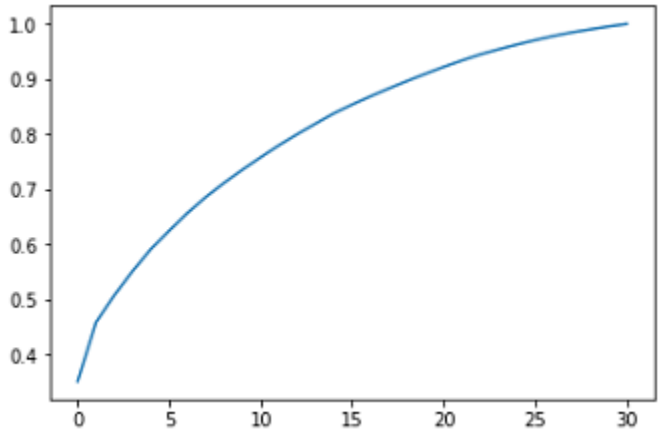
Once the dataset is cleaned and the purchase pattern is identified we move on to coming up with dimensions for the model which is specific to the products. We use a set of threshold to maintain feature count. This is achieved by:

- Calculating the total number of each product sold
- Sort them in descending order
- Select the first 30 features

This way dimensionality reduction is obtained and this helps in getting access to principal variables by filtering out the data that is not within the threshold set.

Principal Component Analysis:

One of the dimensionality reduction technique is principal component analysis. It uses an orthogonal transformation to convert observations into linearly uncorrelated variables called principal components. Here for dimensionality reduction we use PCA instead of thresholding which can be used by clustering algorithms. Standard Scaler is used to fit and transform product data. The obtained dataset which consists of all the principal features obtained by PCA analysis is saved into excel sheet.

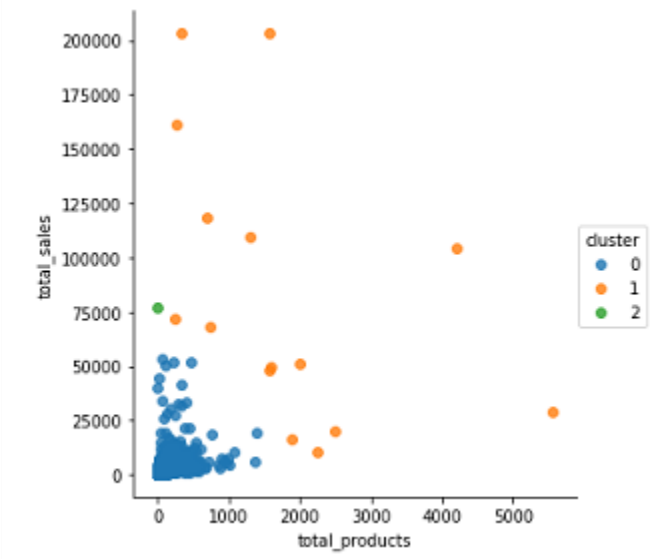


Cumulative explained variance

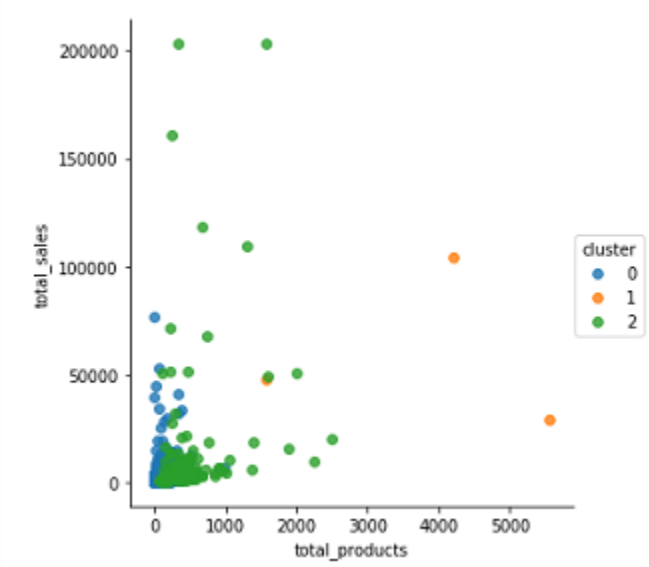
From the above plot we can notice that with the first 30 components can achieve 100 percent variance.

Cluster Analysis:

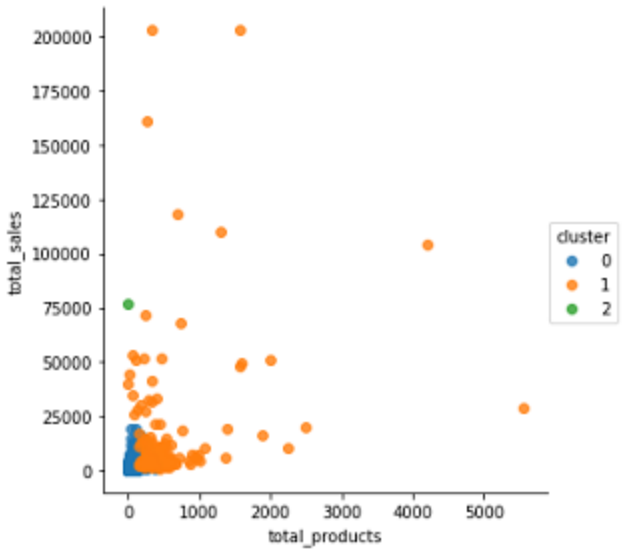
K-means clustering algorithm is used to classify the given dataset to n clusters. It classifies the dataset that has different features and compare their results.



Clustering appears to separate customers based on number of items and total amount spent for customer dataset which consists of transaction records.



For top 30 feature

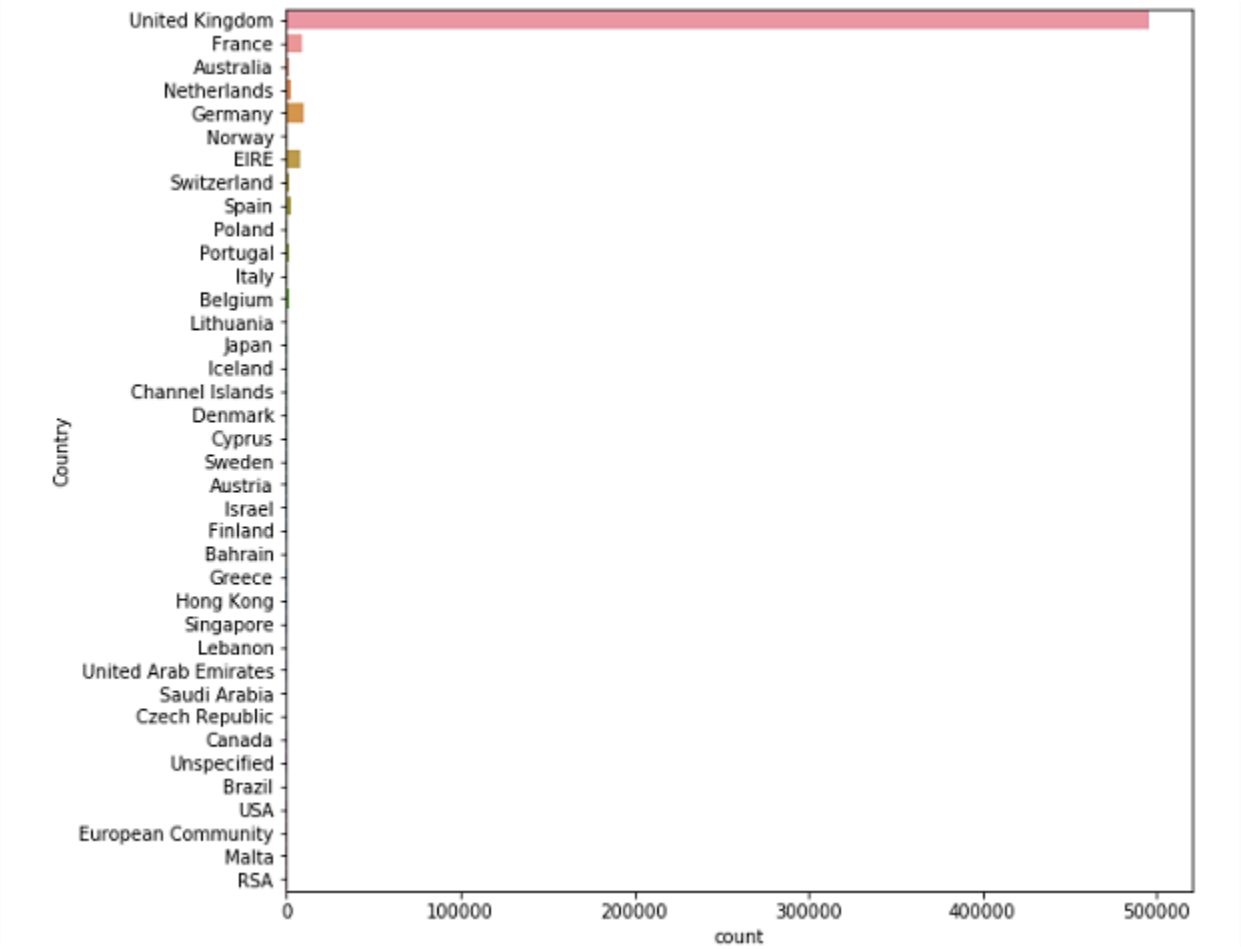


For PCA generated features

These two models are compared using rand index. Rand Index computes a similarity measure between two clustering by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clustering.

Market Basket Analysis:

Market basket analysis for transactions are applied country wise.



The above plot shows the number of purchases made by people residing in different countries. We can notice major market for this online retail store is United Kingdom.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE PINK)	0.096939	0.102041	0.073980	0.763158	7.478947	0.064088	3.791383
1	(ALARM CLOCK BAKELIKE PINK)	(ALARM CLOCK BAKELIKE GREEN)	0.102041	0.096939	0.073980	0.725000	7.478947	0.064088	3.283859
2	(ALARM CLOCK BAKELIKE RED)	(ALARM CLOCK BAKELIKE GREEN)	0.094388	0.096939	0.079082	0.837838	8.642959	0.069932	5.568878
3	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE RED)	0.096939	0.094388	0.079082	0.815789	8.642959	0.069932	4.916181
4	(ALARM CLOCK BAKELIKE RED)	(ALARM CLOCK BAKELIKE PINK)	0.094388	0.102041	0.073980	0.783784	7.681081	0.064348	4.153061

Market basket analysis for transactions that involve customers from France, the probability of product alarm clock bake like pink with alarm clock bake like green being purchased together is 7.3%. Since the lift score is greater than 1 for these two products, we can make use of alarm clock bake like pink to make predictions on alarm clock bake like green. This rule can be incorrect 3.28 times more often if this association rule was purely a random chance.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(PLASTERS IN TIN CIRCUS PARADE)	(PLASTERS IN TIN WOODLAND ANIMALS)	0.115974	0.137856	0.067834	0.584906	4.242887	0.051846	2.076984
1	(PLASTERS IN TIN WOODLAND ANIMALS)	(PLASTERS IN TIN CIRCUS PARADE)	0.137856	0.115974	0.067834	0.492063	4.242887	0.051846	1.740427
2	(PLASTERS IN TIN CIRCUS PARADE)	(ROUND SNACK BOXES SET OF 4 FRUITS)	0.115974	0.157549	0.050328	0.433962	2.754455	0.032057	1.488330
3	(ROUND SNACK BOXES SET OF 4 FRUITS)	(PLASTERS IN TIN CIRCUS PARADE)	0.157549	0.115974	0.050328	0.319444	2.754455	0.032057	1.298977
4	(PLASTERS IN TIN CIRCUS PARADE)	(ROUND SNACK BOXES SET OF4 WOODLAND)	0.115974	0.245077	0.056893	0.490566	2.001685	0.028470	1.481887

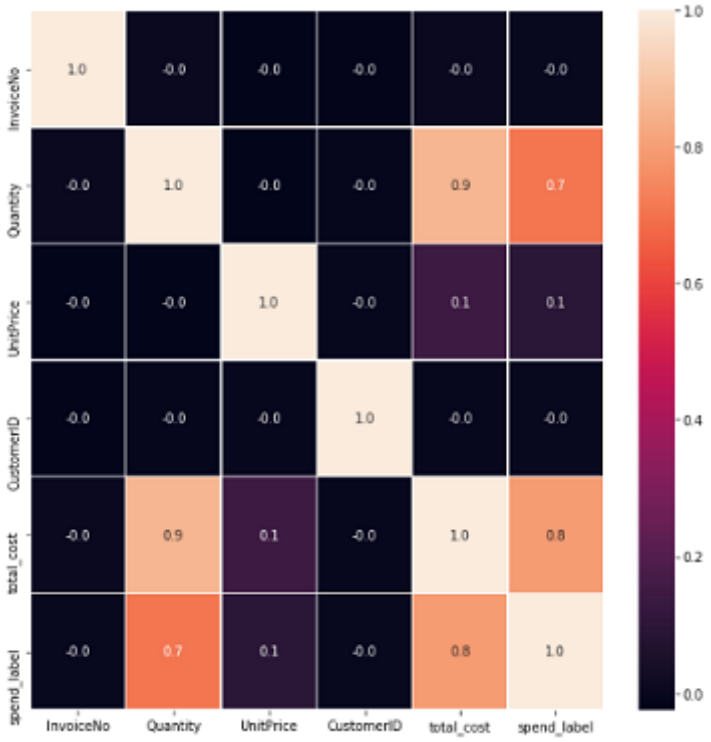
Market basket analysis for transactions that involve customers from Germany, the probability of product plasters in tin circus parade with plasters in tin woodland animals being purchased together is 6.7%. Since the lift score is greater than 1 for these two products, we can make use of plasters in tin circus parade to make predictions on plasters in tin woodland animals.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(ALARM CLOCK BAKELIKE RED)	(ALARM CLOCK BAKELIKE GREEN)	0.049821	0.046928	0.030160	0.605376	12.900183	0.027822	2.415142
1	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE RED)	0.046928	0.049821	0.030160	0.642694	12.900183	0.027822	2.659288
2	(PINK REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.037660	0.050035	0.030910	0.820768	16.403939	0.029026	5.300203
3	(GREEN REGENCY TEACUP AND SAUCER)	(PINK REGENCY TEACUP AND SAUCER)	0.050035	0.037660	0.030910	0.617773	16.403939	0.029026	2.517719
4	(ROSES REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.051267	0.050035	0.037553	0.732497	14.639752	0.034988	3.551237

Market basket analysis for transactions that involve customers from United Kingdom, the probability of product alarm clock bake like red with alarm clock bake like green being purchased together is 3.01%. Since the lift score is greater than 1 for these two products, we can make use of alarm clock bake like red to make predictions on alarm clock bake like green.

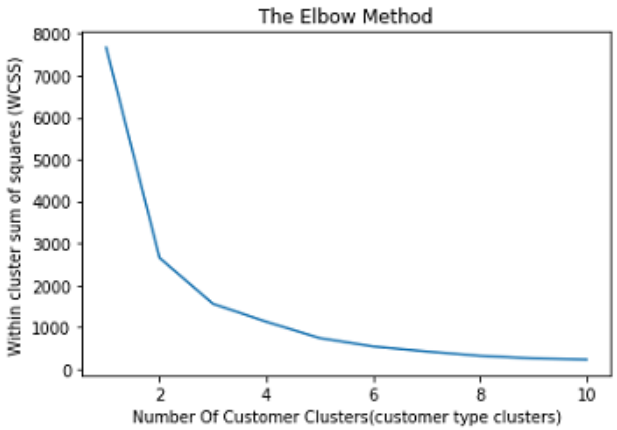
Machine Learning Predictions:

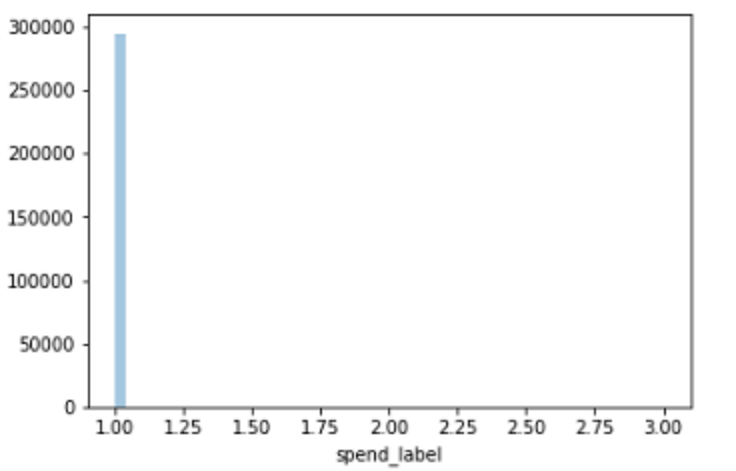
First the data is loaded and all non-applicable columns are removed. The customers are classified based on the amount they have spent. Along with this an additional feature 'spend_label' is added. We then look for correlation between the features that are under consideration.



Heat map to look for correlation

To make a better prediction, database with only features which have high correlative values is created and K-Means Clustering algorithm is applied. Elbow method is used to find the optimum number of clusters. Cluster that has a small sum of squares is more compact than a cluster that has a large sum of squares. We can notice that the value is smallest for n =8 or 10.





It is noticed from the above graph that the data is imbalanced, so only the transactions which involves lower values of the total amount spent by the customer is used to get rid of all outliers is used.

Dummy variables for descriptions and country are created. Also the columns that are not required are dropped and some are replaced by derived features. For better prediction data is normalized before model training.

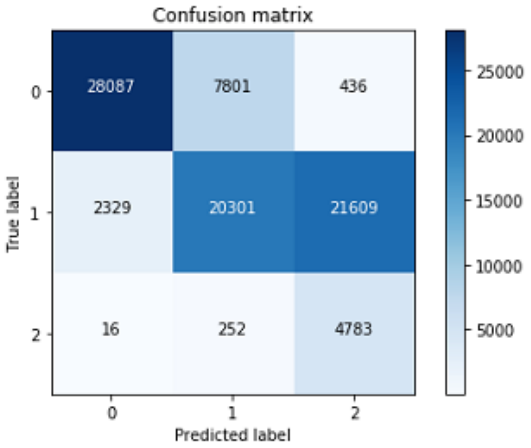
Machine learning - predicting customer spend:

Naive Bayes:

We get the following results,

- Accuracy: 62.11%
- Recall: 62.10549676454785 %
- Precision: 62.10549676454785 %

We can notice from the above values that there is scope for further improvement as model does not work as intended to a good extent.

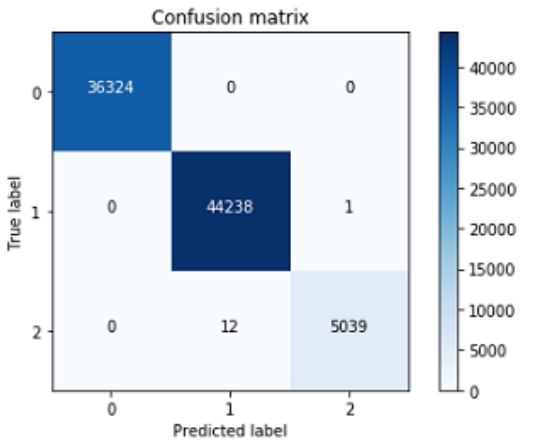


GradientBoost: LightGBM:

We get the following results,

- Accuracy: 99.98%
- Recall: 99.98481556754736 %
- Precision: 99.98481556754736 %

We can see that this model prediction capability is good.



Machine learning - predicting country of origin:

Naive Bayes:

We get the following results,

- Accuracy: 7.66%
- Recall: 7.663466255518957
- Precision: 7.663466255518957

Model does not work well.

GradientBoost: LightGBM:

We get the following results,

- Accuracy: 89.20%
- Recall: 89.1980283598477 %
- Precision: 89.1980283598477 %

This model gives better result.

RFM - Recency, Frequency, Monetary Analysis:

RFM analysis is used to understand customer behavior. Customers can be categorized by using the transaction data. RFM analysis provides insight on the customers and revenue.

- Discover market trends and Patterns
- Finding the customers at risk
- Identifying ways to retain customers
- Increase the revenue by identifying the target customers.

RFM analysis is used in combination with k-means clustering prediction algorithm. It is a customer segmentation analysis. It involves calculating

- Recency: Days since last purchase
- Frequency: Total number of purchases
- Monetary: Total amount spent by the customer

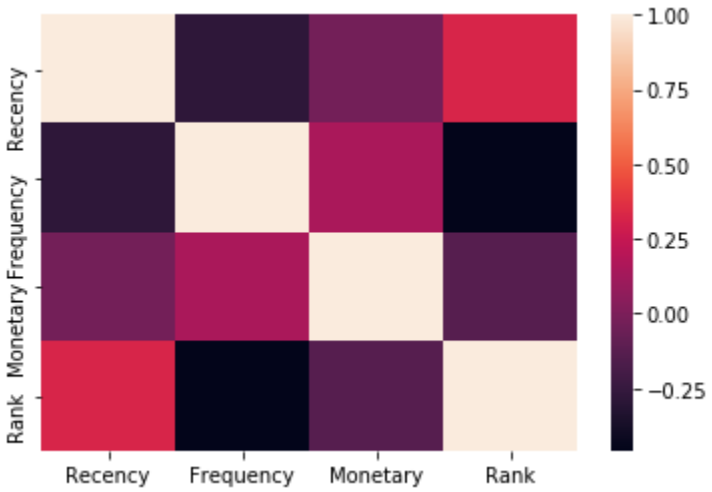
Studying this helps us identify the target customers who responds to promotional offers and the products that they might potentially buy.

RECENCY: total number of days since the customer made the last purchase. This is calculated with respect to the last day of the year that we have considered for RFM analysis.

FREQUENCY: The number of purchases made by the customer. This can be calculated by taking a look into the number of invoices in the transaction dataset.

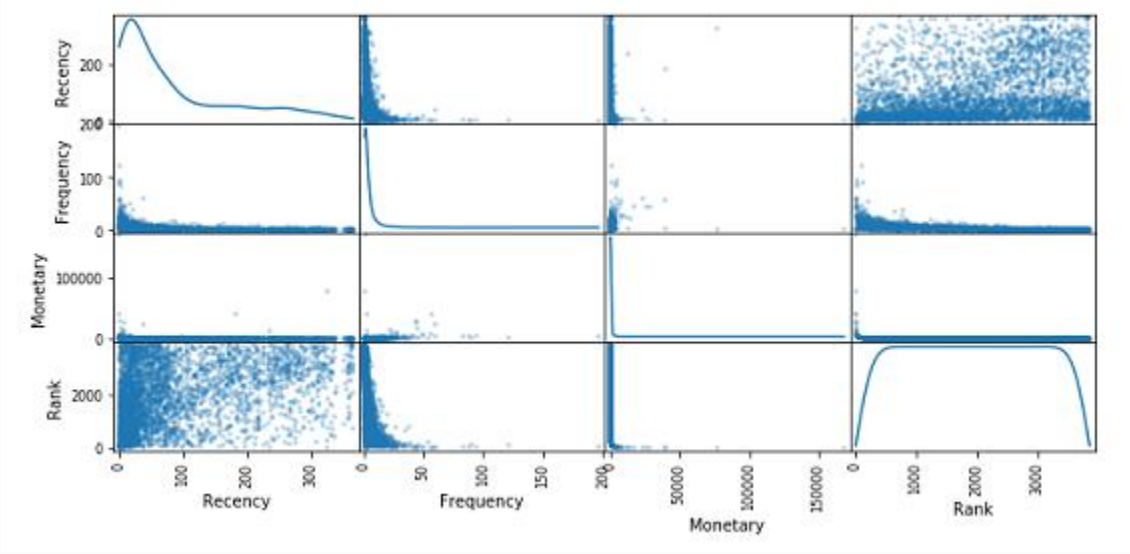
MONETARY: The amount of money spent by the customer by the year under consideration.

Once all these 3 values are computed it is added in the RFM table and verified for validity.



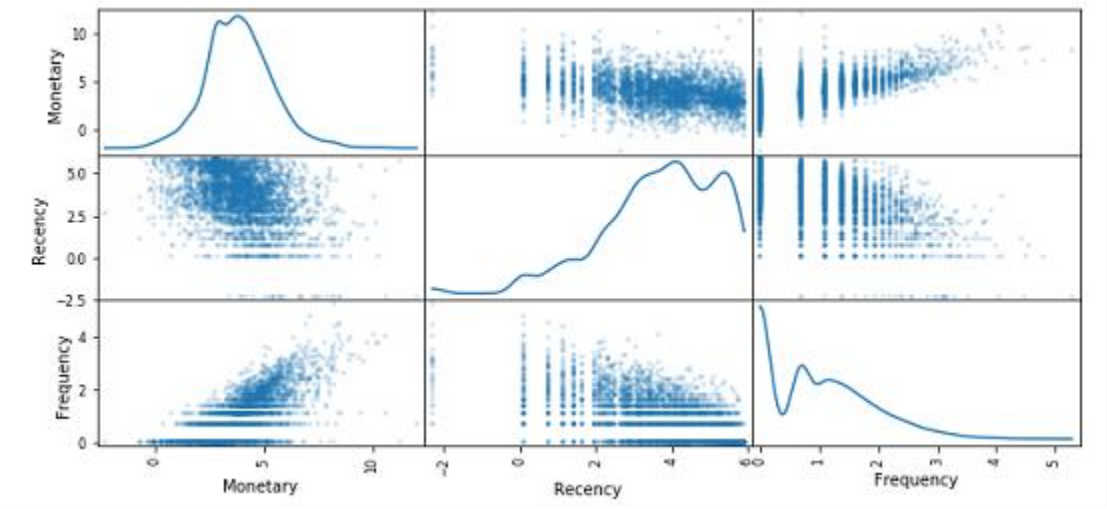
Correlation between Recency, Frequency and Monetary

K-means clustering is applied on RFM variables. It can be seen that correlation between Monetary and Frequency is a slightly higher but it is not too strong. Also for recency is too low and is more towards negative value.

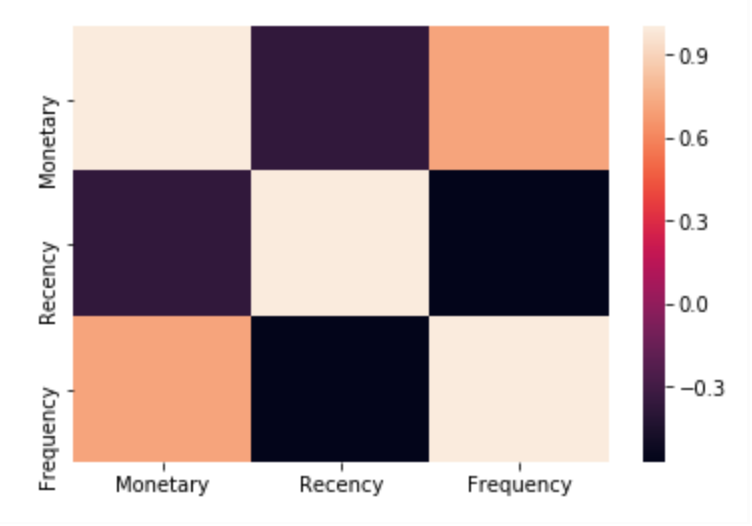


Distribution of Recency, Frequency and Monetary

Also distribution is skewed for all three cases, hence data can be normalized to overcome this. Once the data is normalized, we can see that the distribution is more normalized. But for frequency and Recency it is not showing much improvement.



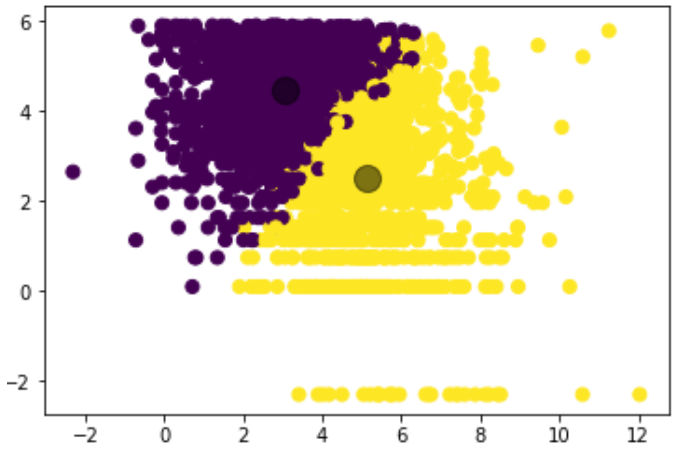
Distribution of Recency, Frequency and Monetary (after normalization)



Correlation between Recency, Frequency and Monetary (after normalization)

We can notice that correlation between Monetary and Frequency is strong in this case, but with recency it's still negative.

In this approach we are using Recency, Frequency and Monetary to form cluster of customers using k-means clustering algorithm. We can notice that there is no clear boundary of separation between the clusters. This could because of the presence of the outliers. The chances of data points being misclassified is high because of this reason.



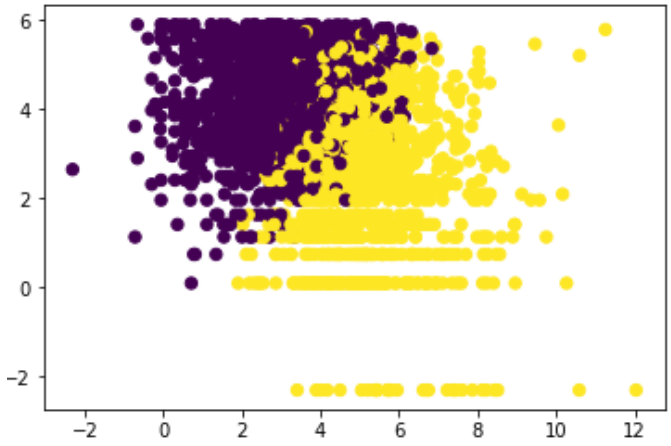
Visualizing a scatter plot

Drawbacks associated with this clustering are:

- The probability of a data point being in two clusters is high.
- This cannot be a guaranteed as a global solution.
- Does not work efficiently for models with limited feature sets.

So we tries working with Gaussian Mixture Model, but the results aren't that great when compared to the k-means clustering. Guassian Mixture model was used with the intension to exploit its flexibility to make use of unconstrained covariance. GMM involves probabilistic cluster assignment.

Gaussian Mixture Model Implementation:



GMM scatter plot

No improvement seen as there are no clear boundary of separation between the clusters.

Conclusion

Clustering and classification techniques are used to analyze customers in this project. Customers are categorized into various clusters based on how recently they have purchased a product, how often two or more products are purchased together, in what quantity a product is purchased, is a product being purchased on any particular occasion, how frequently a product is purchased, how much money has a customer spent over a period of time, etc. We can see there are distinct combinations of Recency, Frequency and Monetary values of customers in the clusters. These clusters can be used by the retailers to come up with appropriate marketing strategies that targets certain customer in particular. This data can also be used to stock up their warehouse and maintain good demand-supply chain.

Future Work

For the future work the project can be divided into 3 parts. Firstly, we can look for any latent variables that are present in the database. Then we can check whether these variables affects the amount of transaction that takes place on a given day at a given time. Next, we can play around with different clustering techniques and come up with a better quality of cluster. Finally, we can build better classification models that provide us better rules which can be used to categorize customers using more specific classification code. This model can also include various factors such as advertising, social media data, etc.

We have used a small set of data for our project. But this can be extended by using a larger dataset thus applying BigData analysis which will lead to better convincible result. Using larger dataset can give opportunity to explore appropriate BigData computing techniques in similar fields involving MapReduce/Hadoop system.

References

[1] https://archive.ics.uci.edu/ml/datasets/Online%20Retail

[2] https://link.springer.com/article/10.1057/dddmp.2013.20

[3]https://www.hbs.edu/faculty/Publication%20Files/kris%20Analytics%20for%20an%20Online%20Retailer_6ef5f3e6-48e7-4923-a2d4-607d3a3d943c.pdf