Application of Machine Learning in E-Commerce

Ashish Jha (asj49@pitt.edu)

Lakshmi Ravichandran (lar146@pitt.edu)

Abstract:

We have developed multiple classifiers based on demographic data and past usage pattern of Wireless network subscriber to predict and target user with the intention to switch of another network. In the competitive market with soaring competition and high customer accusation value the best strategy for Telecom companies with good market share is to retain the existing customers, especially the customer with high revenue. We have also found factors that are important in leading to a customer churning out of the current service provider. Finally, we have made a metric to target customer so as to maximize our revenue.

Problem description –

For the profit of any businesses let it be small or large, customer retention is a key feature. To retain customers, we should provide them with best and suitable discounts, rewarding the customers and provide personalized services. But, the main aim is to retain customers who are loyal and bring more revenue to the business. I this case, customer segmentation is important. Our problem is to clusters the customers for an online based gift retail shop based on different features and then based the customer segmentation, predict and classify new customers to improve retention rate. Provide insights on Customer purchase patterns and product popularity. Identity important features that could help ecommerce business in predicting future demands for products. Group customer into different segments based on purchasing pattern and other features.

Literature Survey –

1. https://tinyurl.com/Analytics-for-an-Online-Retail - Forecasts demands for new product based on features from similar product styles in an online event based business. Algorithm to set optimal price based on predicted demand to get the inventory empty.

2. https://tinyurl.com/A-decision-making-framework - This paper is closely related to the project. It presents data mining techniques to predict monthly supplies, group customers into different segments and target marketing strategies towards different customer categories to help reduce inventory.

3. https://ieeexplore.ieee.org/document/7069993 - Dynamic pricing in AWS spot instances: Amazon AWS offers compute resources to customers following two pricing models – (i) On-demand price where the price is fixed for a particular instance type throughout the life-time of the instance (grouped by number of cores, memory size etc.) (ii) Spot price where the price varies dynamically based on the demand for resources in

the spot market and available capacity. The spot pricing follows an auction-based model, where customers submit bid price for the instance type that they are requesting. If the bid price is above the current spot price and there is availability of resources, then their request would be satisfied. While spot instances provide cost savings compared to on-demand instances, the instance could be revoked with a 2 minute notification if need for additional capacity arises. Thus, the cost savings of spot instances should be weighed against the increased start-up delay and increased probability of failures. A system that predicts the current spot prices based on available data provided by the cloud vendor will help customers in many ways. First, customers could make informed decisions on whether to go for on-demand instances or spot instances. Further, price-aware applications may be developed in order to maximize the benefits of spot instances. B - Analytics for an Online Retailer: Demand Forecasting and Price Optimization- Paper Concern with pricing and predicting demand for products that it has never sold before, to maximize sales and revenue. The author uses machine learning techniques to estimate historical sales and predict future demand of new products. The author also considers the dependence of a product’s demand on the price of competing products, pose new challenges on translating the demand forecasts into a pricing policy.

Dataset 1

Dataset and features:

Data is a subset of one-year data of the telecom company. The dataset consists of customer demographic information mostly categorical features like Area, Age, ethnicity among others. there are several usage information as continuous features like total revenue, avg monthly revenue, average number of minutes of usage and categorical features like, Number of active subscriber. In total there are 81 distinct features and the number of observations is 66291. In the dataset Target feature is the column 'churn’ which is binary feature depicting if a user has moved out of the telecom service provider. Every observation has unique identifier as customer_id which can be leveraged to target customer from the model.

Data Cleaning and Feature engineering:

Dealing with null values:

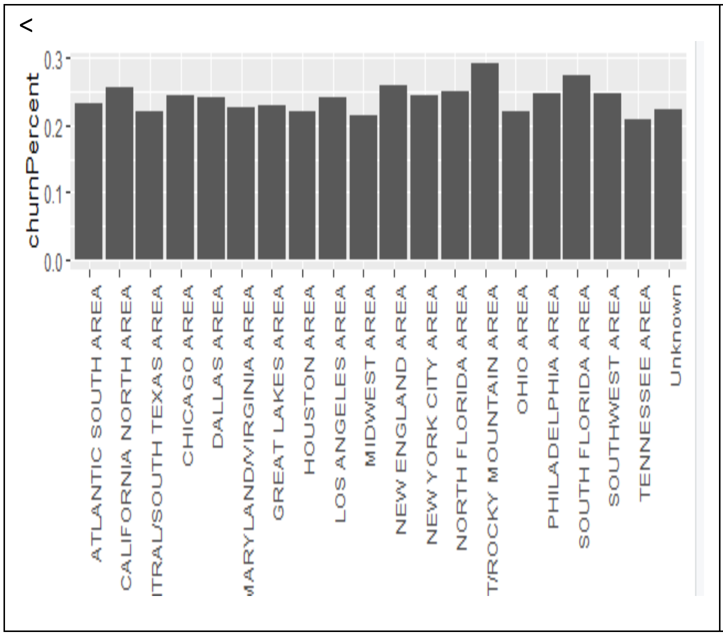
There were 14 features where the percentage of total features among total observation is more than 10%. We have removed these features from the dataset. we have calculated the event rate for all the categorical features which is just the percentage of churn in that feature category.

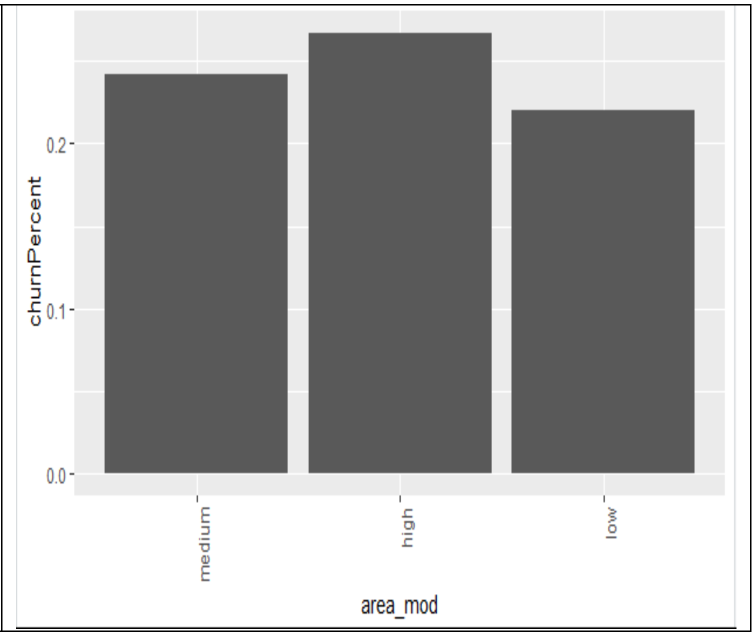
For categorical variables: Event rate for each level in a categorical variable is compute and NA factors are imputed with factors having similar rate of churn percent.

Dealing with features have several levels:

Ideally features should not have factors with more than 3 levels. We have encounter a lot of features with more than 10 levels.

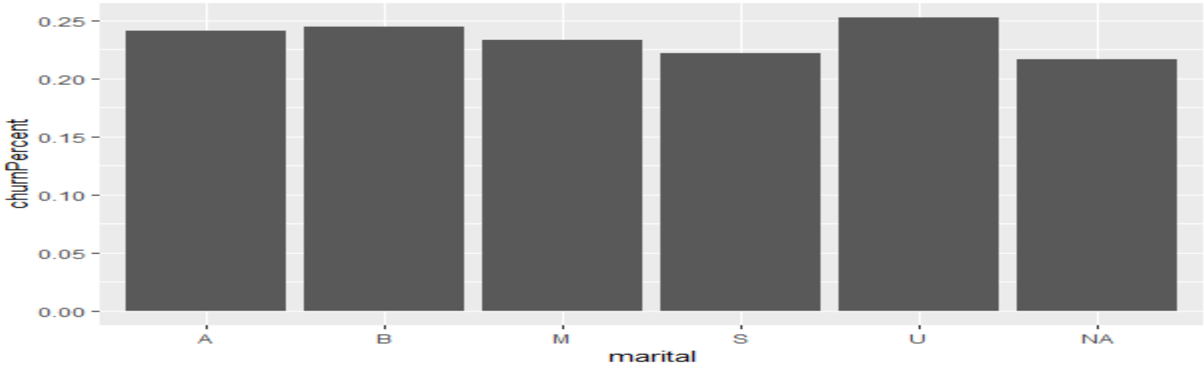
we have merged the feature in 3 levels combining levels with similar rate of churn. Like area is an nominal feature with more than 20+ feature instead of grouping them geographically we decided to merge area in the buckets having similar churn rate.





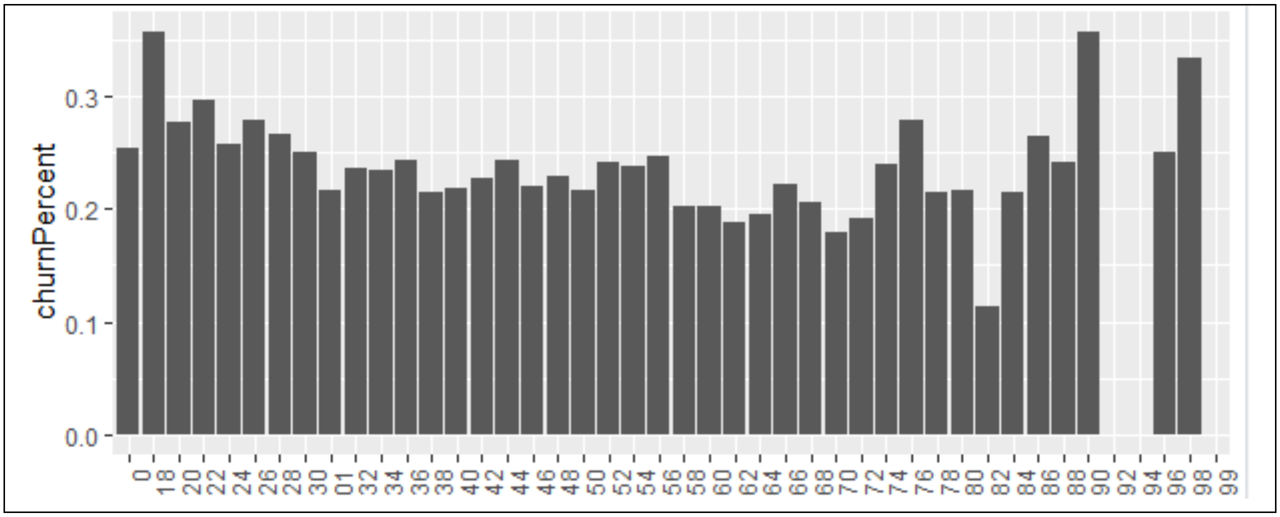
Removing features with similar rate:

Ideally there should be noticeable difference between the event rate in each level. In the due process we have removed features where found no noticeable difference between the various levels of features like for ‘marital status’ we have found that all the levels have similar rate of churn among thus we choose to omit it from our dataset to train the model.



Converting continuous features to factor:

Typically for age related columns which are usually continious features but we have decided to bucket the dates in several group of youth, late20's and others as to abe to understand the impact of the age group on the churn. Age group also leads to explainable and feasible models as categorical features.



Continuous Features

There were only a few of the continuous features having NA, thus we decided to simple mean imputation strategy for the data imputation.

All our categorical features in dataset are ordinal in nature. the For modelling purpose we have converted all the categorical features as dummy features using one hot dummy encoding.

Modelling

Logistic regression (LR): Our target variable is binary. This it we can easily use logistic regression as modelling algorithm. In our prediction, we want to measure the relationship between a dependent variable with binary classification (churn), and various independent variables, we chose to start with the LR model. LR model assume that the relationship between independent variable and the dependent variable is a linear function. It has become a standard classification method as it is easy to use and provides quick and robust results. In our application, the task is to build a classification model that estimates the probability of a customer churning out of the network provider base on the parameters we have generated. In logistic regression, hypothesis is defined as:

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$$

where g is a sigmoid function.

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left[-y^{(i)}\log(h_\theta(x^{(i)})) - (1-y^{(i)})\log(1-h_\theta(x^{(i)}))\right],$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}$$

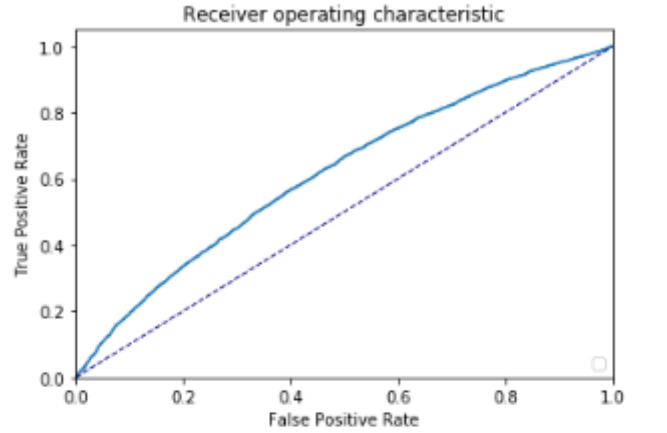
Associated cost function.

Result & Discussion:

The final processed dataset is divided into train and test dataset with train containing 75% of the data where as testset contains the rest 25%.

We have tried Logistic regression for modelling on the processed training data. We use the generated model predict the probability of churn in the test dataset.

We managed to get the accuracy of 76% on the test set with AUC score of 0.614



Comparing accuracy of various modelling algorithm:

	Modelling Algorithm	Accuracy	AUC
1	Logistic regression	76%	0.614
2	QDA	60%	NA
3	LDA	76%	NA
4	Random Forest	72%	0.580
5	Adaboost	76%	0.65
6	SVM	NA	NA

Discussion:

Our model with the dataset had the maximum accuracy of 75%, which is with Logistic regression and AdaBoost. Our Algorithm with SVM did not converge thus we were unable to get the accuracy from the SVM algorithm. As to solve the problem at hand we can use the probability prediction from the logistic regression to target customer. We have bucketed customer in segments of high, medium and low probability and total revenue. We can specifically target with either having high probability and high total revenue then we can more to customer with high probability and medium total revenue and thus moving to customer with medium probability and high total revenue.

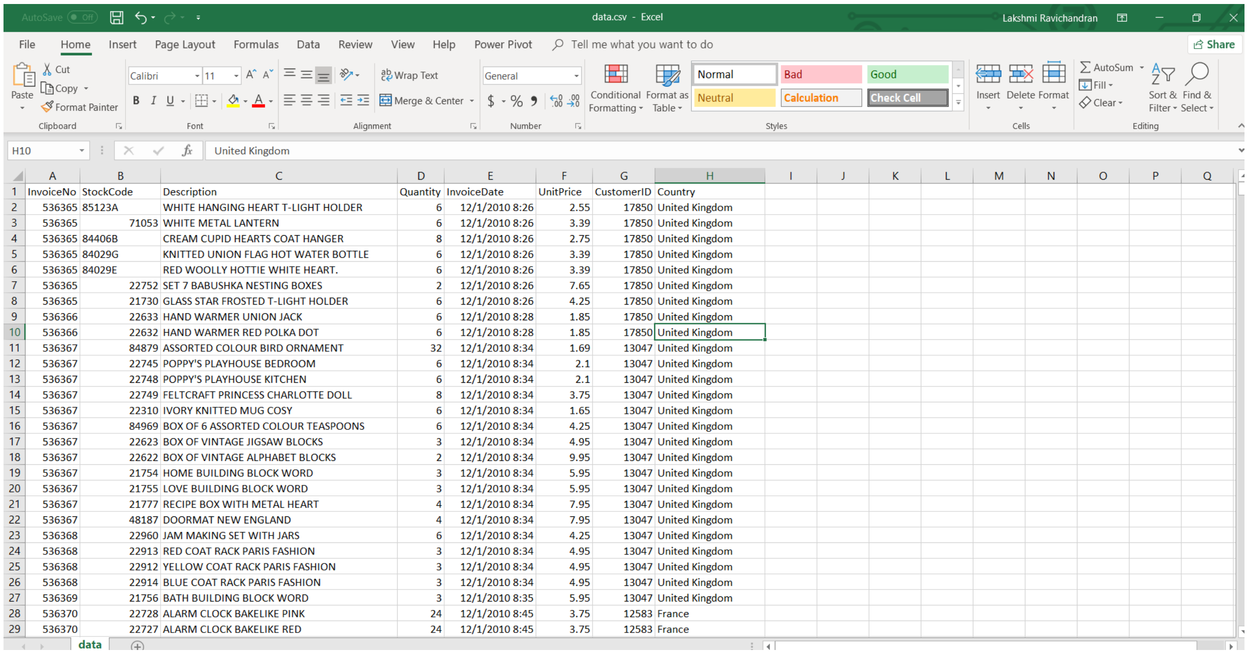
With the help of R and above strategy we were able to publish the customer ID of the potential customer to be targeted.

Having more data can also help improve the accuracy of the model. Since we did not have a lot of observation we couldn't use the neural network and deep learning models.

Data set 2 –

Transactional data which records data from 12/01/2010 and 12/09/2011 for a registered online retail store based in United Kingdom. This retail stored deals mainly in all-occasion gifts. The data set has 8 features – InvoiceNo., StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country. The total number of customer transaction observations are 541909 in the raw data. These are actual transaction data obtained from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/index.php) .

Sample raw data set before data cleaning and feature enginnering -

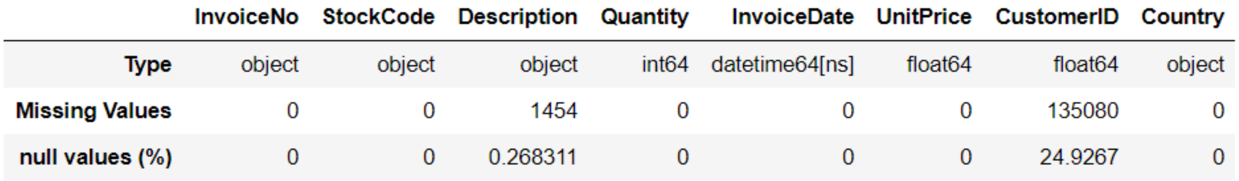


	A	B	C	D	E	F	G	H
1	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
2	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850	United Kingdom
3	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850	United Kingdom
4	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850	United Kingdom
5	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850	United Kingdom
6	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850	United Kingdom
7	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12/1/2010 8:26	7.65	17850	United Kingdom
8	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	12/1/2010 8:26	4.25	17850	United Kingdom
9	536366	22633	HAND WARMER UNION JACK	6	12/1/2010 8:28	1.85	17850	United Kingdom
10	536366	22632	HAND WARMER RED POLKA DOT	6	12/1/2010 8:28	1.85	17850	United Kingdom
11	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	12/1/2010 8:34	1.69	13047	United Kingdom
12	536367	22745	POPPY'S PLAYHOUSE BEDROOM	6	12/1/2010 8:34	2.1	13047	United Kingdom
13	536367	22748	POPPY'S PLAYHOUSE KITCHEN	6	12/1/2010 8:34	2.1	13047	United Kingdom
14	536367	22749	FELTCRAFT PRINCESS CHARLOTTE DOLL	8	12/1/2010 8:34	3.75	13047	United Kingdom
15	536367	22310	IVORY KNITTED MUG COSY	6	12/1/2010 8:34	1.65	13047	United Kingdom
16	536367	84969	BOX OF 6 ASSORTED COLOUR TEASPOONS	6	12/1/2010 8:34	4.25	13047	United Kingdom
17	536367	22623	BOX OF VINTAGE JIGSAW BLOCKS	3	12/1/2010 8:34	4.95	13047	United Kingdom
18	536367	22622	BOX OF VINTAGE ALPHABET BLOCKS	2	12/1/2010 8:34	9.95	13047	United Kingdom
19	536367	21754	HOME BUILDING BLOCK WORD	3	12/1/2010 8:34	5.95	13047	United Kingdom
20	536367	21755	LOVE BUILDING BLOCK WORD	3	12/1/2010 8:34	5.95	13047	United Kingdom
21	536367	21777	RECIPE BOX WITH METAL HEART	4	12/1/2010 8:34	7.95	13047	United Kingdom
22	536367	48187	DOORMAT NEW ENGLAND	4	12/1/2010 8:34	7.95	13047	United Kingdom
23	536368	22960	JAM MAKING SET WITH JARS	6	12/1/2010 8:34	4.25	13047	United Kingdom
24	536368	22913	RED COAT RACK PARIS FASHION	3	12/1/2010 8:34	4.95	13047	United Kingdom
25	536368	22912	YELLOW COAT RACK PARIS FASHION	3	12/1/2010 8:34	4.95	13047	United Kingdom
26	536368	22914	BLUE COAT RACK PARIS FASHION	3	12/1/2010 8:34	4.95	13047	United Kingdom
27	536369	21756	BATH BUILDING BLOCK WORD	3	12/1/2010 8:35	5.95	13047	United Kingdom
28	536370	22728	ALARM CLOCK BAKELIKE PINK	24	12/1/2010 8:45	3.75	12583	France
29	536370	22727	ALARM CLOCK BAKELIKE RED	24	12/1/2010 8:45	3.75	12583	France

Data cleaning –

Since the data set is taken from actual transactions. The chance of missing values, duplicates and error is more. These incorrect and missing data are not exceptions, and it is common in a large data set. The correction method could be data imputation, drop the rows with missing if the percentage of missing values are less. The corrective measure help increase the accuracy of the model.

Percentage of missing values in the data set across all columns –

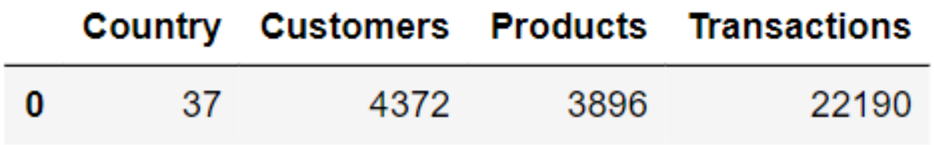


	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
Type	object	object	object	int64	datetime64[ns]	float64	float64	object
Missing Values	0	0	1454	0	0	0	135080	0
null values (%)	0	0	0.268311	0	0	0	24.9267	0

Since only 24% of CustomerID are missing, the missing value rows are dropped. That also cleaned the missing values in description column. Duplicate rows are also dropped which counts to 5225. Other cleaning done on data are, Invoice date column is changed to to_datetime format (format='%m/%d/%Y %H:%M')

After conversion - Sample data - 2010-12-01 08:26:00

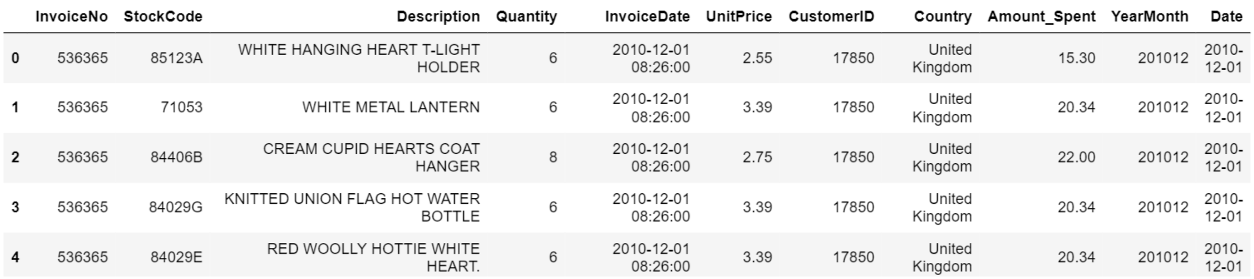
After data cleaning, the unique count values of customers, products and transactions are as follows –



	Country	Customers	Products	Transactions
0	37	4372	3896	22190

CustomerID is converted to integer and columns – amount spent (quantity * unit price), year month and date (split from invoice date column) are added.

After data cleaning – Sample dataset –



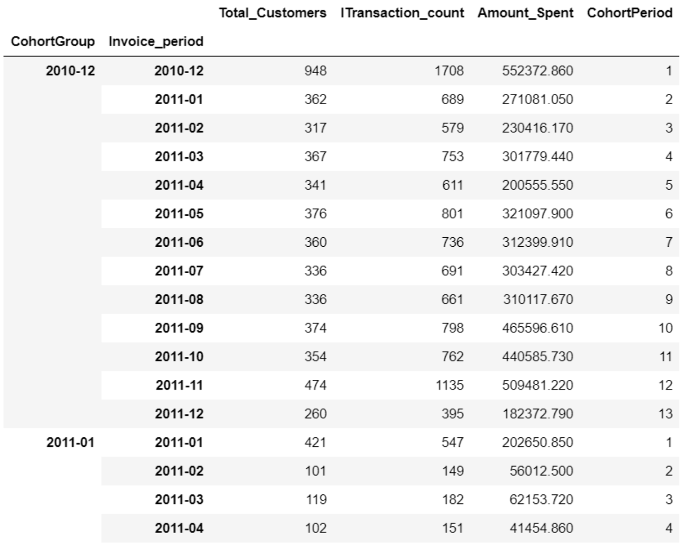
	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Amount_Spent	YearMonth	Date
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850	United Kingdom	15.30	201012	2010-12-01
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850	United Kingdom	20.34	201012	2010-12-01
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850	United Kingdom	22.00	201012	2010-12-01
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850	United Kingdom	20.34	201012	2010-12-01
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850	United Kingdom	20.34	201012	2010-12-01

Exploratory Data Analysis -

- **Cohort analysis –**

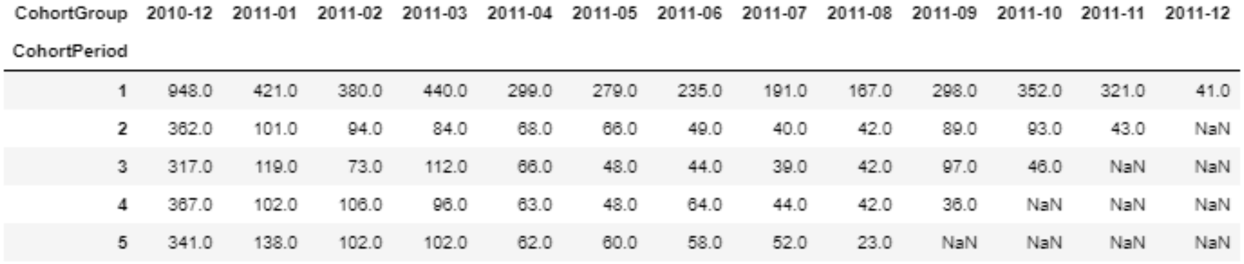
Cohort analysis is a data analytics technique that is suitable to conduct exploratory data analysis of large data sets. In this technique, we group data points into categories based on common features and then study changes in features across groups. Similar to clustering techniques in machine learning, this is a primitive analysis done using feature statistics. For example in our project, we group customers based on first month of purchase and then study retention rate among these groups. The cohort identifiers are Invoice period, cohort group and cohort period. These are required to group customers based on cohort analysis. – Invoice period (year – month combination). Cohort group is done based on month in which the customers did their first purchase and group them into buckets. Finally, we compute cohort period - which is the time gap since the first purchase. It is basically, stages of purchase of the customer. For instance, if a customer's first purchase was in March 2011 and the second purchase after 2 months i.e. May 2011. Here the first purchase will be in the March 2011 group and which is the first lifecycle stage where as May 2011 will be in the second life cycle stage. The difference between the first purchase and the purchase we want to know (Cohort period - Invoice Date) /30)+1 gives the cohort period. This helps us learn about customer purchase pattern and their impact on the revenue growth of the online retailer. In our problem, we used cohort analysis to find the retention rate of customer with respect to total transaction and total amount spent by each customer. Below is the sample data with the cohort identifiers and their corresponding total number of unique customers, total unique transactions and total amount spent in each group.

Customer retention rate -



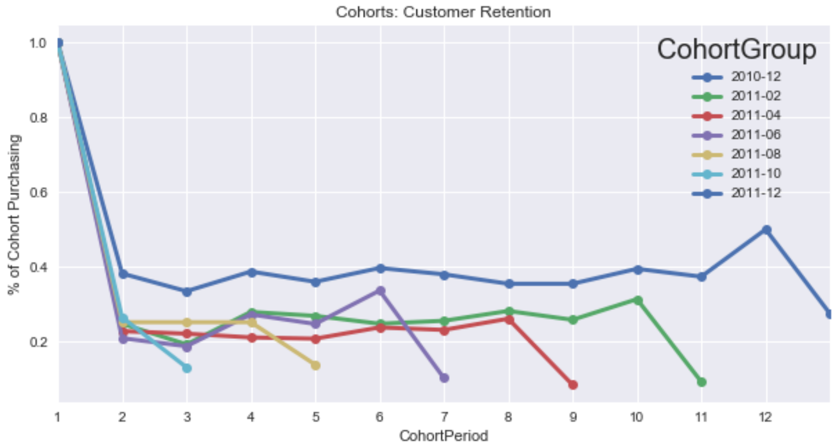
		Total_Customers	ITransaction_count	Amount_Spent	CohortPeriod
CohortGroup	Invoice_period				
2010-12	2010-12	948	1708	552372.860	1
	2011-01	362	689	271081.050	2
	2011-02	317	579	230416.170	3
	2011-03	367	753	301779.440	4
	2011-04	341	611	200555.550	5
	2011-05	376	801	321097.900	6
	2011-06	360	736	312399.910	7
	2011-07	336	691	303427.420	8
	2011-08	336	661	310117.670	9
	2011-09	374	798	465596.610	10
	2011-10	354	762	440585.730	11
	2011-11	474	1135	509481.220	12
	2011-12	260	395	182372.790	13
2011-01	2011-01	421	547	202650.850	1
	2011-02	101	149	56012.500	2
	2011-03	119	182	62153.720	3
	2011-04	102	151	41454.860	4

- In the below matrix, the first row has the number of unique customers in each cohort group and the first purchase months of the groups in first row is shown in the columns. In the other rows we compute the unique customers in each group that continued purchasing in the subsequent months and we normalize all rows with the count in the first row. Thus the first row is all 1.0. Also, across all columns we see customer retention rate is very low. For example if we take the first column, all customers whose first purchase month is 2010-12, only < 39% customers purchased again in subsequent months.

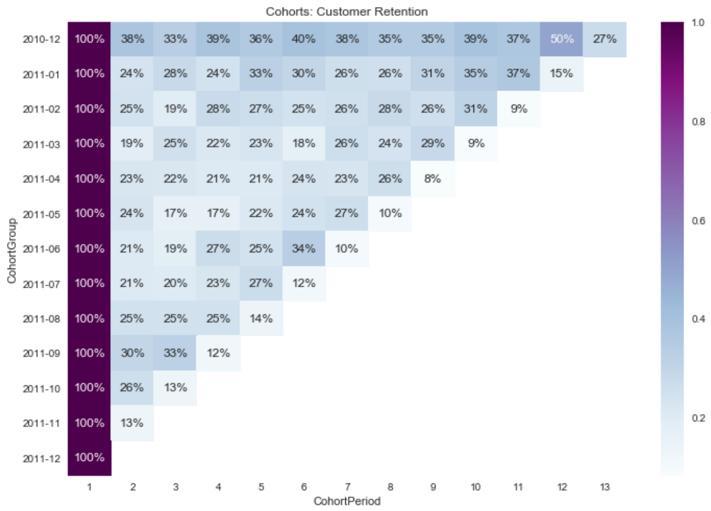


CohortGroup	2010-12	2011-01	2011-02	2011-03	2011-04	2011-05	2011-06	2011-07	2011-08	2011-09	2011-10	2011-11	2011-12
CohortPeriod													
1	948.0	421.0	380.0	440.0	299.0	279.0	235.0	191.0	167.0	298.0	352.0	321.0	41.0
2	362.0	101.0	94.0	84.0	68.0	66.0	49.0	40.0	42.0	89.0	93.0	43.0	NaN
3	317.0	119.0	73.0	112.0	66.0	48.0	44.0	39.0	42.0	97.0	46.0	NaN	NaN
4	367.0	102.0	106.0	96.0	63.0	48.0	64.0	44.0	42.0	36.0	NaN	NaN	NaN
5	341.0	138.0	102.0	102.0	62.0	60.0	58.0	52.0	23.0	NaN	NaN	NaN	NaN

Below plot shows that, low customer retention across all groups and first group higher retention than other groups.

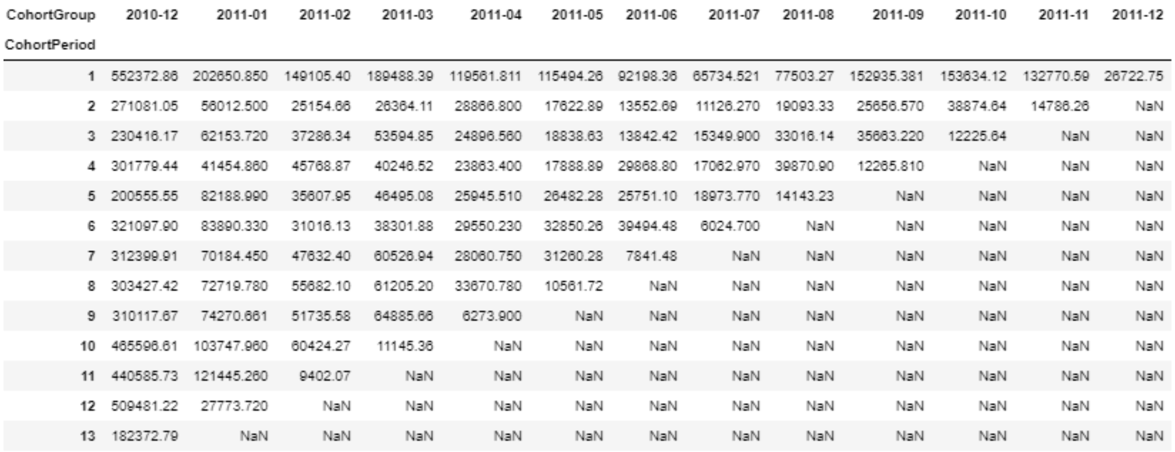


Above matric show in percentage with heat map for better understanding.

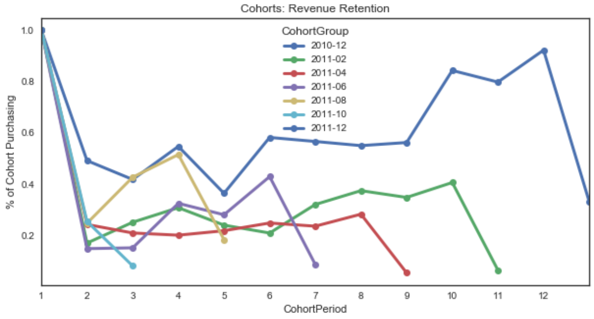


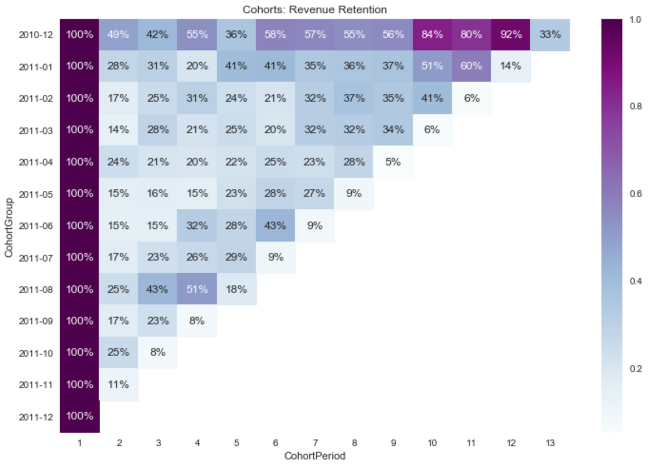
Revenue retention -

Similar analysis is done to get revenue retention rate with total amount spent in each cohort group. Below graphs show similar trends for revenue retention. But, with revenue we can find the seasonal increase in revenue especially in November. (Holiday season – gift retailer)

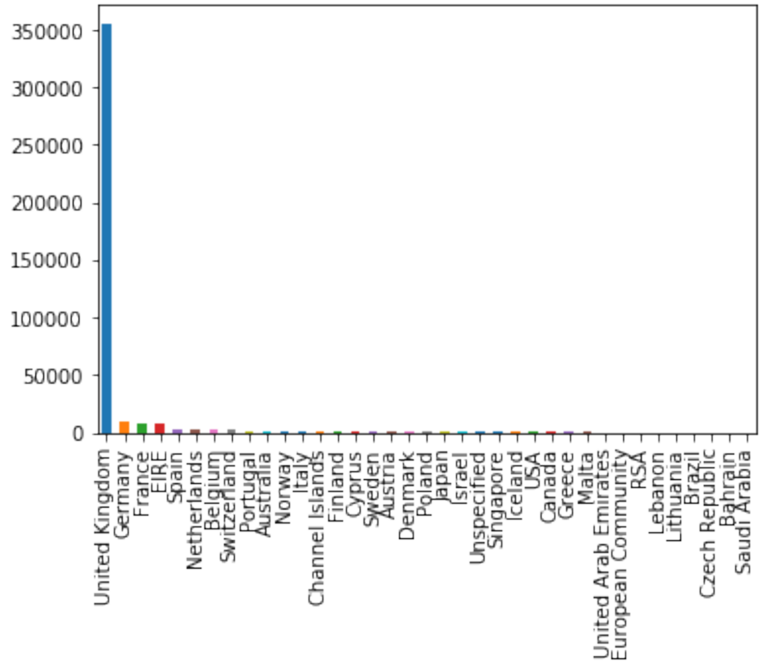


CohortGroup	2010-12	2011-01	2011-02	2011-03	2011-04	2011-05	2011-06	2011-07	2011-08	2011-09	2011-10	2011-11	2011-12
CohortPeriod													
1	552372.86	202650.850	149105.40	189488.39	119561.811	115494.26	92198.36	65734.521	77503.27	152935.381	153634.12	132770.59	26722.75
2	271081.05	56012.500	25154.66	26364.11	28866.800	17622.89	13552.69	11126.270	19093.33	25656.570	38874.64	14786.26	NaN
3	230416.17	62153.720	37286.34	53594.85	24896.560	18838.63	13842.42	15349.900	33016.14	35663.220	12225.64	NaN	NaN
4	301779.44	41454.860	45768.87	40246.52	23863.400	17888.89	29868.80	17062.970	39870.90	12265.810	NaN	NaN	NaN
5	200555.55	82188.990	35607.95	46495.08	25945.510	26482.28	25751.10	18973.770	14143.23	NaN	NaN	NaN	NaN
6	321097.90	83890.330	31016.13	38301.88	29550.230	32850.26	39494.48	6024.700	NaN	NaN	NaN	NaN	NaN
7	312399.91	70184.450	47632.40	60526.94	28060.750	31260.28	7841.48	NaN	NaN	NaN	NaN	NaN	NaN
8	303427.42	72719.780	55682.10	61205.20	33670.780	10561.72	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9	310117.67	74270.661	51735.58	64885.66	6273.900	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
10	465596.61	103747.960	60424.27	11145.36	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
11	440585.73	121445.260	9402.07	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
12	509481.22	27773.720	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
13	182372.79	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

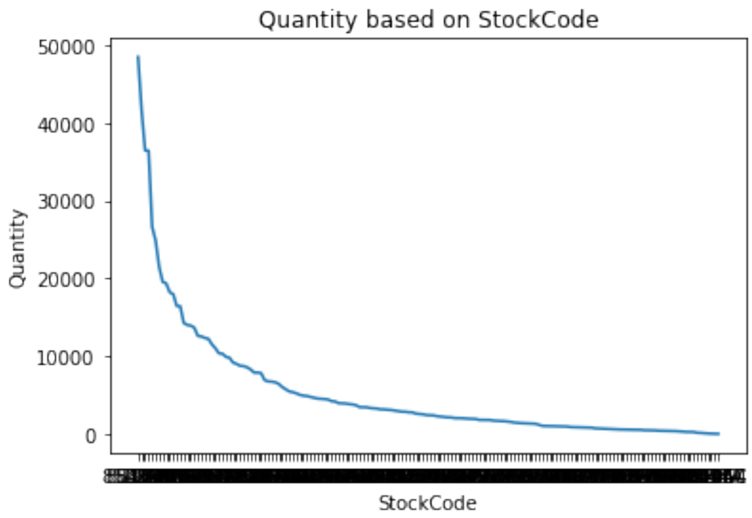




- The plot below shows that “Country” feature is not a significant one. Since it has unbalanced data, significantly large number of customers are from United Kingdom – reason could be the retailer is UK based. Using this feature could bias the model. Thus, not considered for clustering and prediction.

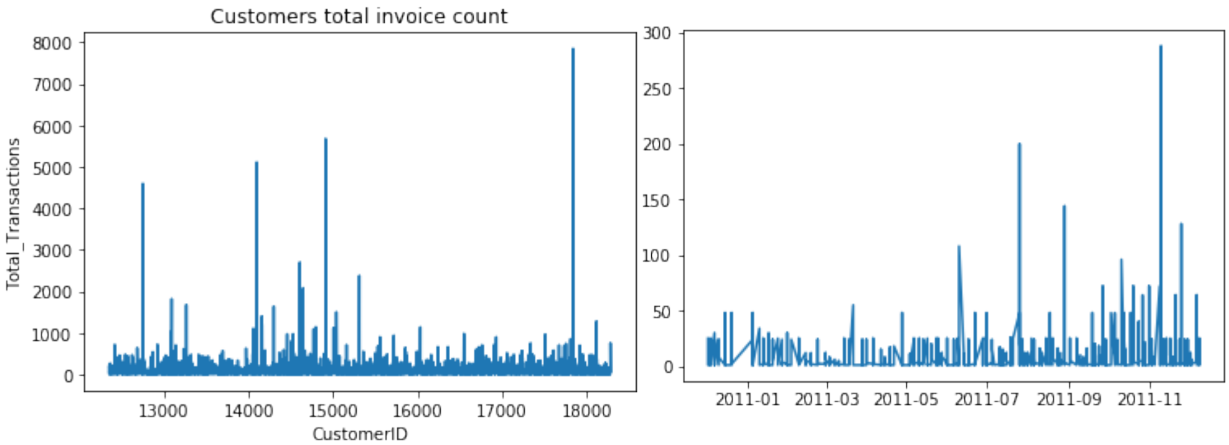


- Quantity based on stock code(product description)

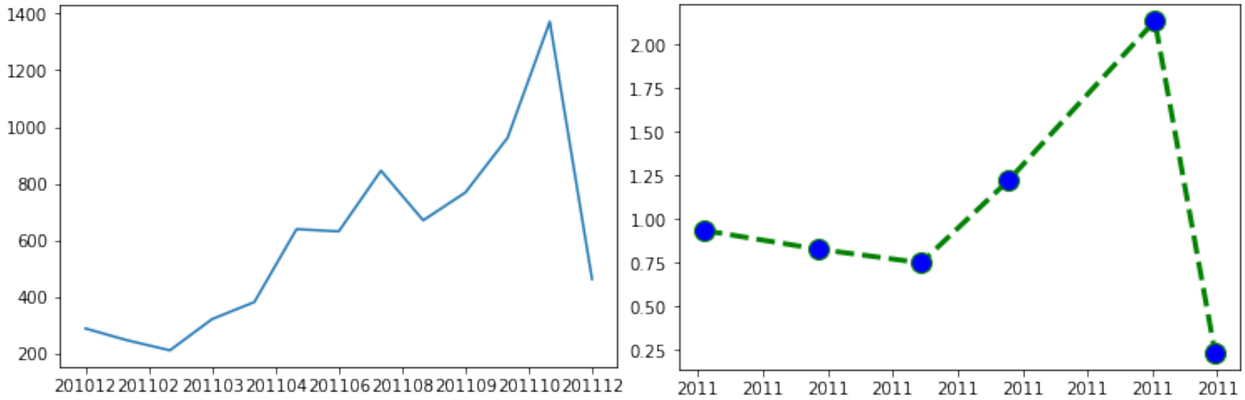


There is a skew in product popularity – few products are sold in large quantities than most products. This could help us categorize the product into different buckets instead of taking all the products individually.

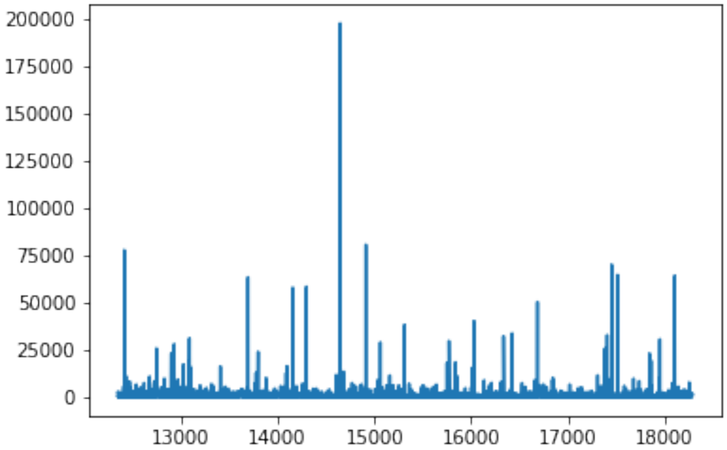
- From the below graph, there are clearly, 3 groups of customers – very frequent, frequent and less frequent customers. This can be understood more with time series graph. The figure on the right shows time series transaction count for the top customer. It can also be inferred that more transaction is done after November – holiday season. Similar plots are taken for top 15 customer (Refer python code). For the last figure, percentage of change of invoice count over year month is taken to learn about frequency and retention of the customer. But, this can be effectively done using cohort analysis – grouping customers based on a column and then find the retention rate.



-



- Total quantity of each customer – This graph also shows categories of customers based on the total quantity bought in the 13 month time frame. Some outstanding total quantity count shows that they could be wholesalers (insight)

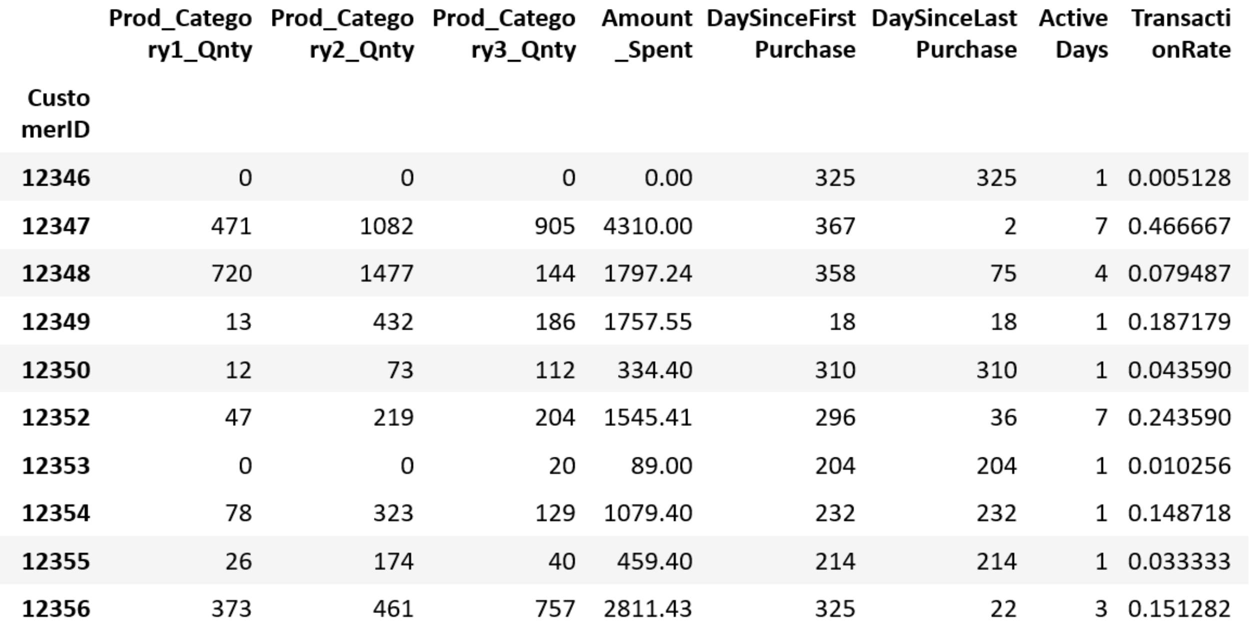


Feature Engineering –

The raw data consists of unprocessed and simple features like customer invoice data, product code etc. Using the features as such is not prudent since they do not represent the data well to clustering and classification algorithms. So we perform feature engineering to represent the data well to the algorithms there by getting good performance.

- For our problem, we choose features based on intuition. We choose 'first purchase date', 'last purchase date' as above to capture the customer entry and exit in a queue model.
- To represent purchase pattern of each customer, we choose the average rate of transaction and the number of active days. (Method used attached with jupyter book)
- Based on the insight from the EDA that products have a popularity distribution similar to a Zipf curve, we choose two thresholds of product quantity sold and categorize products into 3 groups - highly popular, medium popular and less popular then count the number of products in each group purchased by the customer.

Below is the final data with 8 features populated from the features from raw data for each customer. So, the number of observations is the total unique customers after data cleaning. This dataset provides us with significant features for customer segmentation – based on product preference, purchasing power (amount spent) and frequency of purchase (active days and transaction rate)



	Prod_Category1_Qnty	Prod_Category2_Qnty	Prod_Category3_Qnty	Amount_Spent	DaySinceFirstPurchase	DaySinceLastPurchase	ActiveDays	TransactionRate
CustomerID								
12346	0	0	0	0.00	325	325	1	0.005128
12347	471	1082	905	4310.00	367	2	7	0.466667
12348	720	1477	144	1797.24	358	75	4	0.079487
12349	13	432	186	1757.55	18	18	1	0.187179
12350	12	73	112	334.40	310	310	1	0.043590
12352	47	219	204	1545.41	296	36	7	0.243590
12353	0	0	20	89.00	204	204	1	0.010256
12354	78	323	129	1079.40	232	232	1	0.148718
12355	26	174	40	459.40	214	214	1	0.033333
12356	373	461	757	2811.43	325	22	3	0.151282

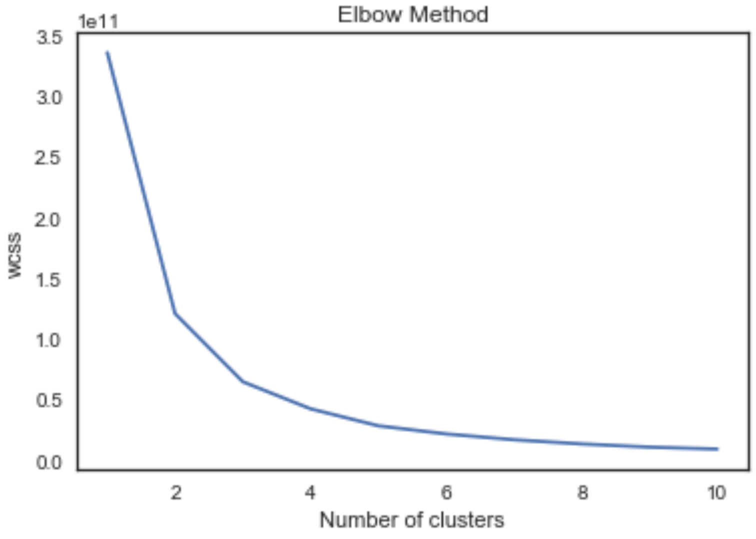
Model –

K means clustering for customer segmentation –

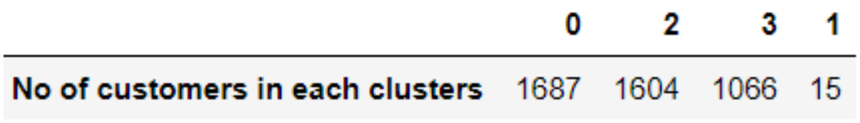
Based on the features we chose above, we perform customer segmentation. We choose the standard k-means clustering algorithm to cluster customers into multiple groups. The main parameter in k-means is the number of clusters to group the data set into. Further, the number of iterations to be done by the algorithm can also be tuned to provide better clustering efficiency.

Elbow plot to choose the number of clusters for k means clustering –

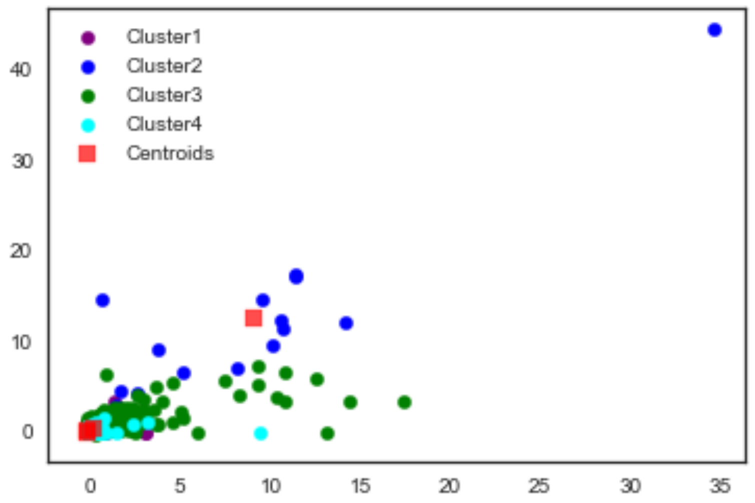
We conduct an analysis to find the best number of clusters for our dataset. We leverage elbow plot that shows how the sum of squares error of each point in a cluster from its means changes when we change the number of clusters. We observe from the plot that N=4 gives low error and beyond which increasing the number of clusters does not have much reduction. So we use N=4 as the parameter for the k-means clustering algorithm.



The below scatter plot shows 4 clusters of customer groups and the centroid point of each cluster. The number of customers in each cluster are

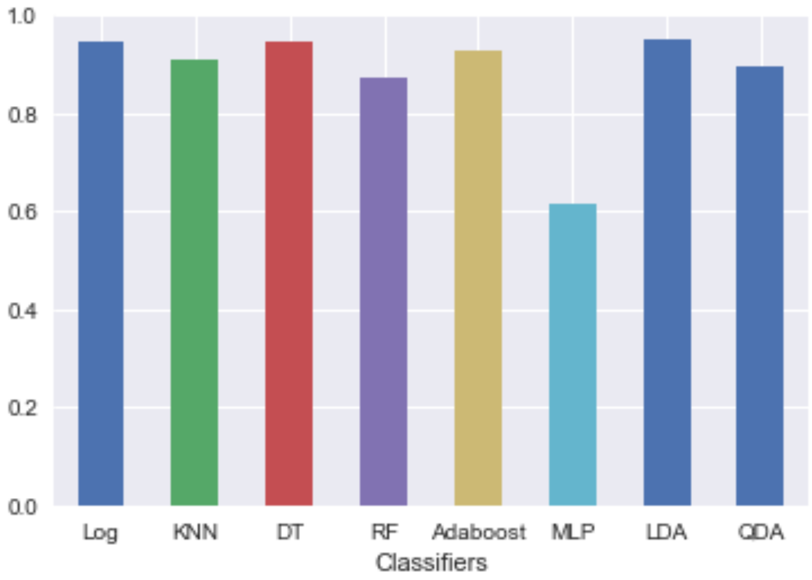


	0	2	3	1
No of customers in each clusters	1687	1604	1066	15



Classifying a new customer into one of the above 4 clusters –

These clusters generated by the k-means clustering algorithm above are used to label the dataset into 4 classes (0, 1, 2, and 3). We chose numeric classes to be generic and provide the flexibility to try many classification algorithms. Different classification models were implemented with train and test split of the data. Overall accuracy observed in most models were above 90% except for MLP. This could be of multiple causes. The main reason is the algorithms are suitable for binary classification whereas our problem is a multi class classification problem. MLP would have been a suitable algorithm for our problem, but the neural network has to be tuned so that the number of output layer neurons match the number of classes.



Future work and improvements –

We are currently working on implementing the multi class classification using MLP to get more insights and tune the results.

References –

[1] V. K. Singh and K. Dutta, "Dynamic Price Prediction for Amazon Spot Instances," 2015 48th Hawaii International Conference on System Sciences, Kauai, HI, 2015, pp. 1513-1520. doi: 10.1109/HICSS.2015.184.

[2] Cheng Wang, Qianlin Liang, and Bhuvan Urgaonkar. 2017. An Empirical Analysis of Amazon EC2 Spot Instance Features Affecting Cost-effective Resource Procurement. In Proceedings of the 8th ACM/SPEC on International Conference on Performance Engineering (ICPE '17). ACM, New York, NY, USA, 63-74. DOI: https://doi.org/10.1145/3030207.3030210

[3] https://www.bcg.com/documents/file13853.pdf

https://tinyurl.com/Analytics-for-an-Online-Retail

https://tinyurl.com/A-decision-making-framework

https://ieeexplore.ieee.org/document/7069993