

Term 3 Project- Customer Analytics

Amit Maity, Soumya Shukla
Date: 5th May 2019

Problem 1: Classification

- Overview of the project
- Features/Variables for model development
- Exploratory Data Analysis (EDA)
- Correlation
- Data Preparation
- Feature Importance
- Implementing ML Models
 - Logistics Regression
 - Random Forest
 - XG Boost
 - SVM
- Models Summary

Problem 2: Shipment Analysis w.r.t. Best customer

- Best 1000 Customer Analysis

Problem 3: Clustering

- Customer Segmentation with Clustering
 - Hierarchical Clustering
 - Kmeans Clustering

Problem 4: Sentiment Analysis

- Snapdeal, Flipkart, Amazon India

- Starting a new venture in the Business world requires extensive research of market and goods. It is important to know your target audience in order to convert them into loyal customers and improve your services.
- This project allows us to gain full visibility into how customers use their products. The better you understand and know your customers , the more accurately will you be able to draw predictions regarding their future buying behaviour patterns.
- The dataset provided by the client contained 10999 observations of 12 variables.
- From a Data Scientist viewpoint, we are to perform various logistic/ probabilistic techniques and derive the most accurate model

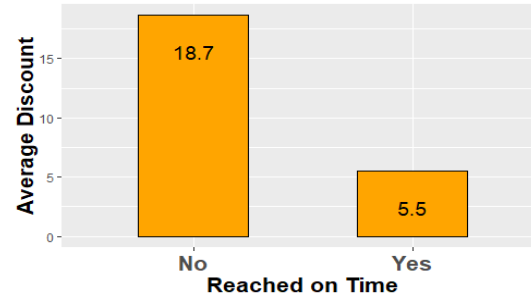
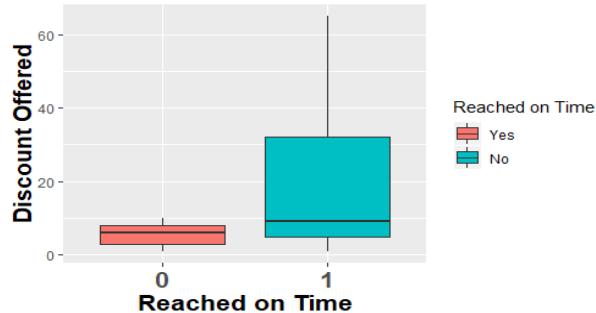
Features/Variables for model development

Feature Name	Feature Details
ID	Id number of the customer
Warehouse_block	Warehouses used for dispatching product (A,B,C,D,F)
Mode_of_Shipment	Mode of transport (Ship, Air and Road)
Customer_care_calls	Number of inquiry calls made by customer
Customer_rating	Customer rated on various parameters, 1 being the lowest (Worst), 5 being highest (Best)
Cost_of_the_Product	It is the cost of the product in USD
Prior_purchases	Number of prior purchases
Product_importance	Products categorized in the range of high, medium and low
Gender	Male or female
Discount_offered	Percentage of discount offered on that specific product.
Weight_in_gms	Product weight in grams
Reached.on.Time_Y.N	Reached On Time :- '0' and NOT Reached On Time :- '1'

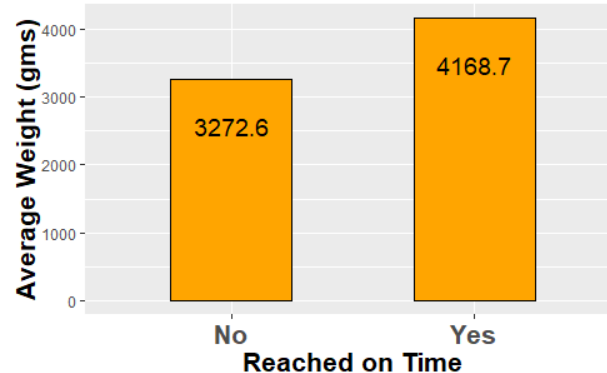
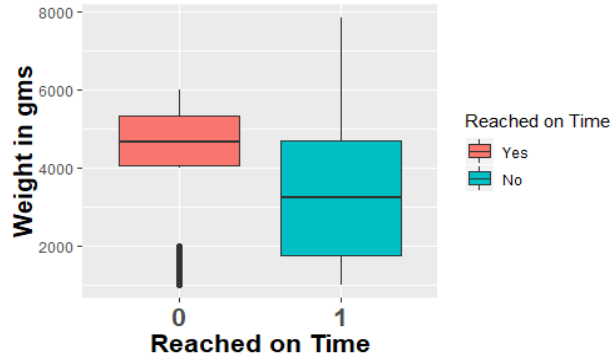
Key performance indicator (KPI) or response variable for model development is
Reached.on.Time_Y.N

Exploratory Data Analysis (EDA)

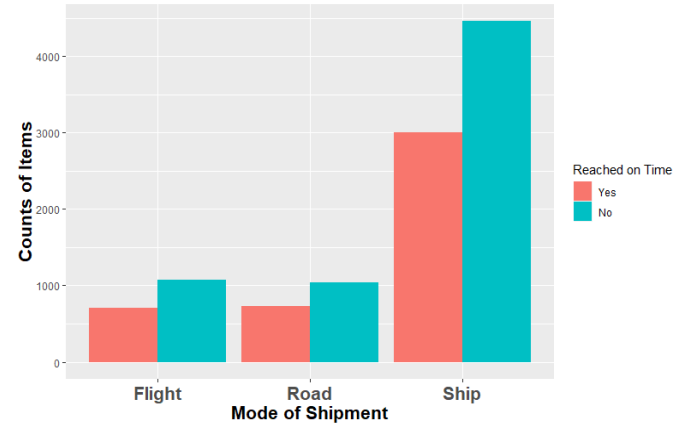
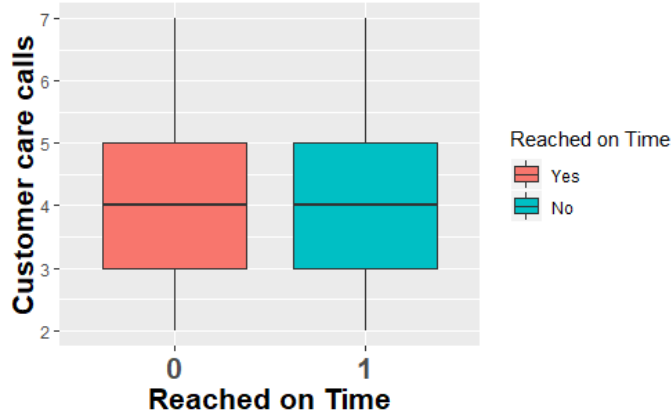
- The data was analysed using the R programming. Various insights were gathered using ggplot



Items with higher discount do not reach on time

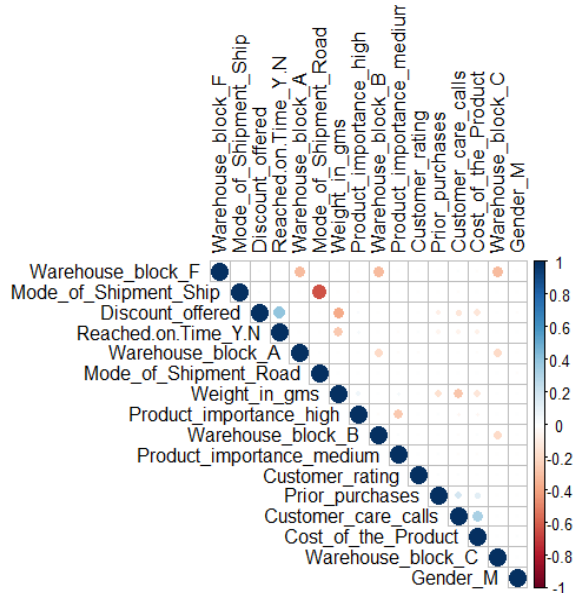


Heavier items reach earlier as compared to lighter ones



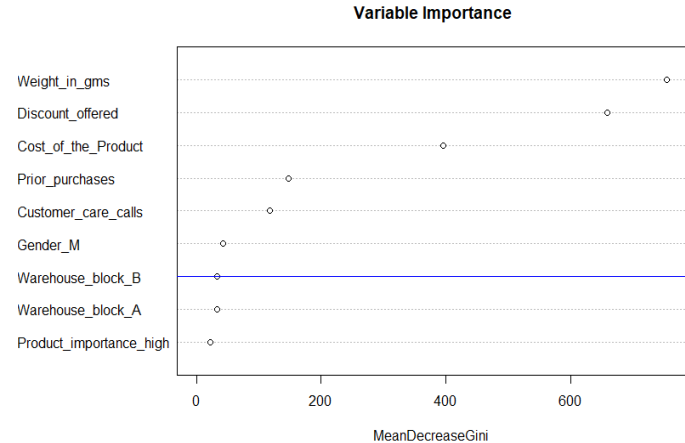
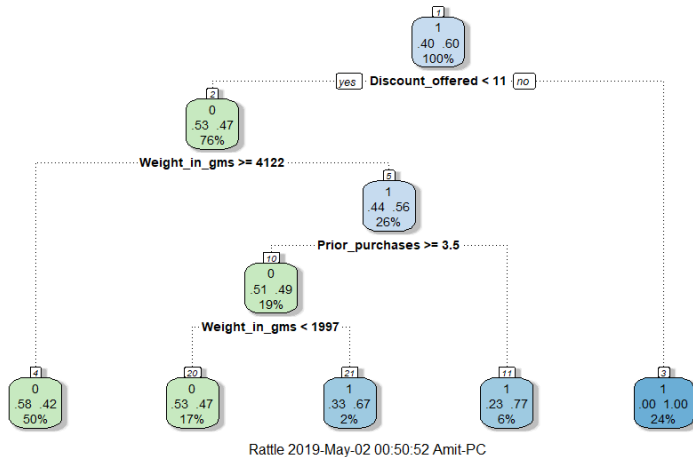
- There is no significant relation between customer care calls and delay of items.
- Ship cargo is the preferred method of order dispatch

Correlation Plot



- Positive correlation between discount offered and reached on time column
- Positive correlation between customer care calls and cost of the product column
- Strong negative correlation between shipment in Ship & Road

- Missing Value Imputation
 - There is no missing value in given dataset
- Feature Engineering
 - New features like customer_rating_per_calls, prior_purchase_per_calls, weight_per_discount, discount_per_weight are created from existing data.
 - As per observations new features does not create much impact on model.
- Train and Test split
 - First dataset splits into 70:30 ratio for Trainsplit and Test set
 - Further Trainsplit split into Train and Validation set in 70:30 ratio
 - Find the common features from Train and Validation set to fit the model



- As we discussed earlier: Weight_in_gms, Discount_offered are two most important features
- Cost_of_the_Product, Prior_purchases, customer_care_calls are the few less important features

Logistic regression model: A set of explanatory variables/independent variables is used to predict a logit transformation of the dependent/response variable

$$\text{logit}\{\Pr(Y = 1|\mathbf{x})\} = \log\left\{\frac{\Pr(Y = 1|\mathbf{x})}{1 - \Pr(Y = 1|\mathbf{x})}\right\} = \beta_0 + \mathbf{x}'\beta$$

$Y \rightarrow$ Response variable (binary=0 or 1)

$X \rightarrow$ Explanatory variables

$B_0 \rightarrow$ Intercept parameter

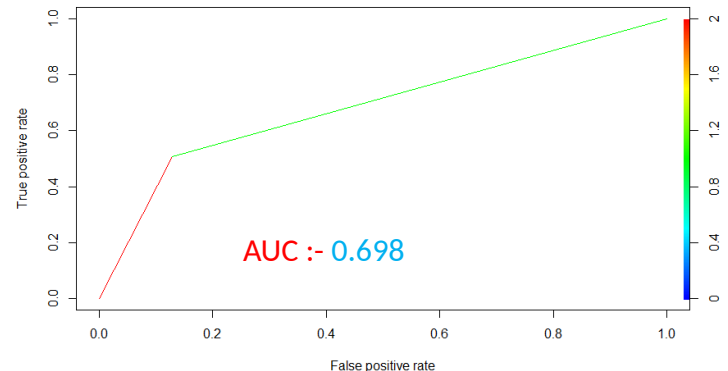
$\beta \rightarrow$ Slope parameters

Model Results

		Predicted		
Actual		Positive	Negative	
	Positive	TP = 1158	FN = 171	1329
	Negative	FP = 973	TN = 998	1971
		2131	1169	

T \rightarrow True; F \rightarrow False; P \rightarrow Positive; N \rightarrow Negative

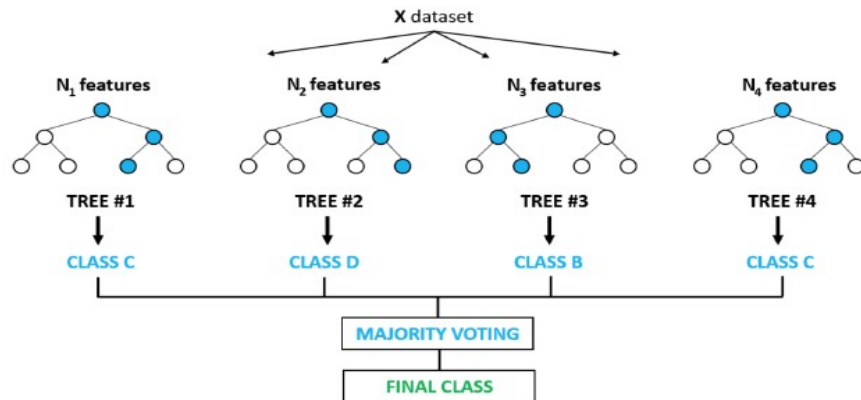
Accuracy : 65.3%
Sensitivity: 54.3%
Specificity: 85.3%



Random forest model: A top down decision tree (multiple trees) based algorithm

$$\text{Entropy} = -p \log_2(p) - q \log_2(q)$$

Entropy measures the homogeneity of the subset data

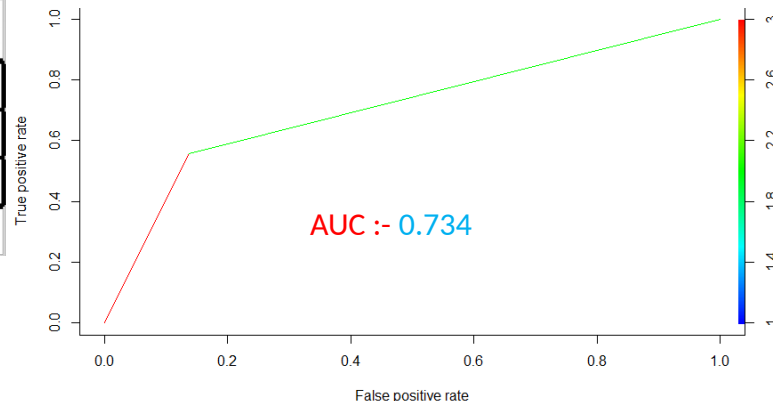


Model Results

Actual	Predicted		
	Positive	Negative	
	Positive	Negative	
Positive	TP = 1217	FN = 112	1329
Negative	FP = 932	TN = 1039	1971
	2149	1151	

T → True; F → False; P → Positive; N → Negative

Accuracy : 68.3%
Sensitivity: 56.6%
Specificity: 90.2%



Model Development Cont..

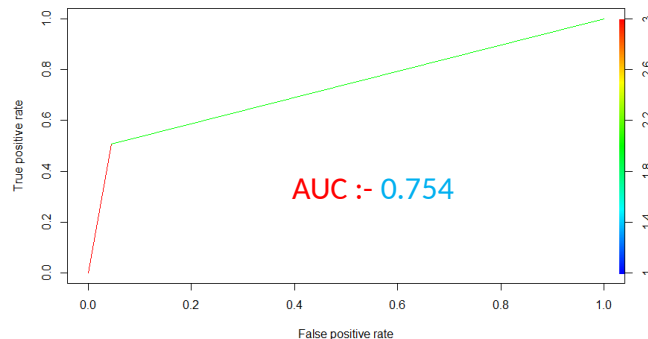
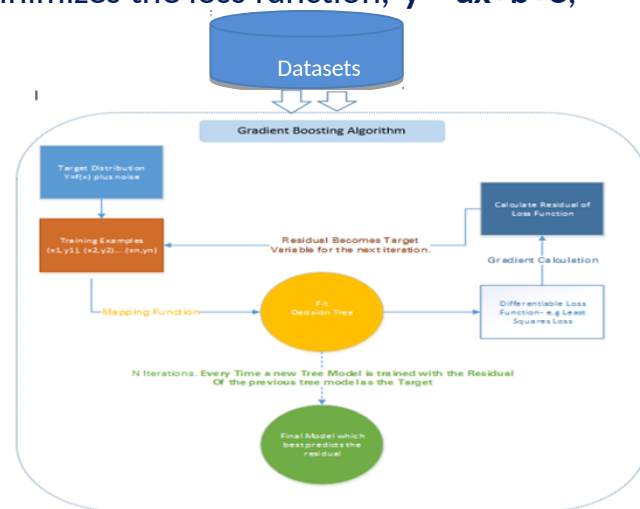
XG Boost model: In Extreme Gradient Boosting many models are trained sequentially. It is a numerical optimization algorithm where each model minimizes the loss function, $y = ax + b + e$, using the Gradient Descent Method

Model Results

		Predicted		
Actual		Positive	Negative	
	Positive	TP = 1269	FN = 60	1329
	Negative	FP = 973	TN = 998	1971
		2242	1058	

T → True; F → False; P → Positive; N → Negative

Accuracy : 68.7%
Sensitivity: 56.6%
Specificity: 94.3%



Model Development Cont..

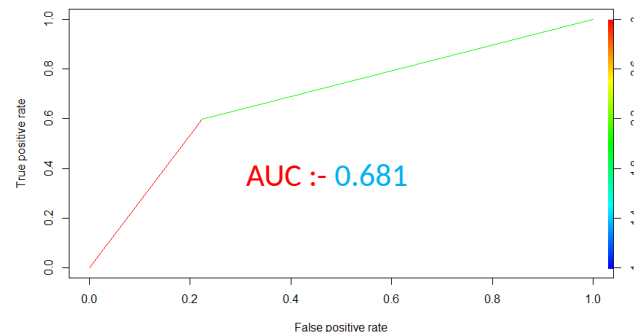
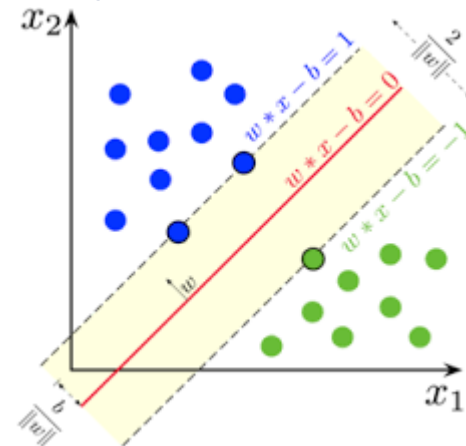
SVM model: It is a supervised learning **models** with associated learning algorithms that analyze data used for classification and regression analysis. The algorithm creates a line or a hyperplane which separates the data into classes.

Model Results

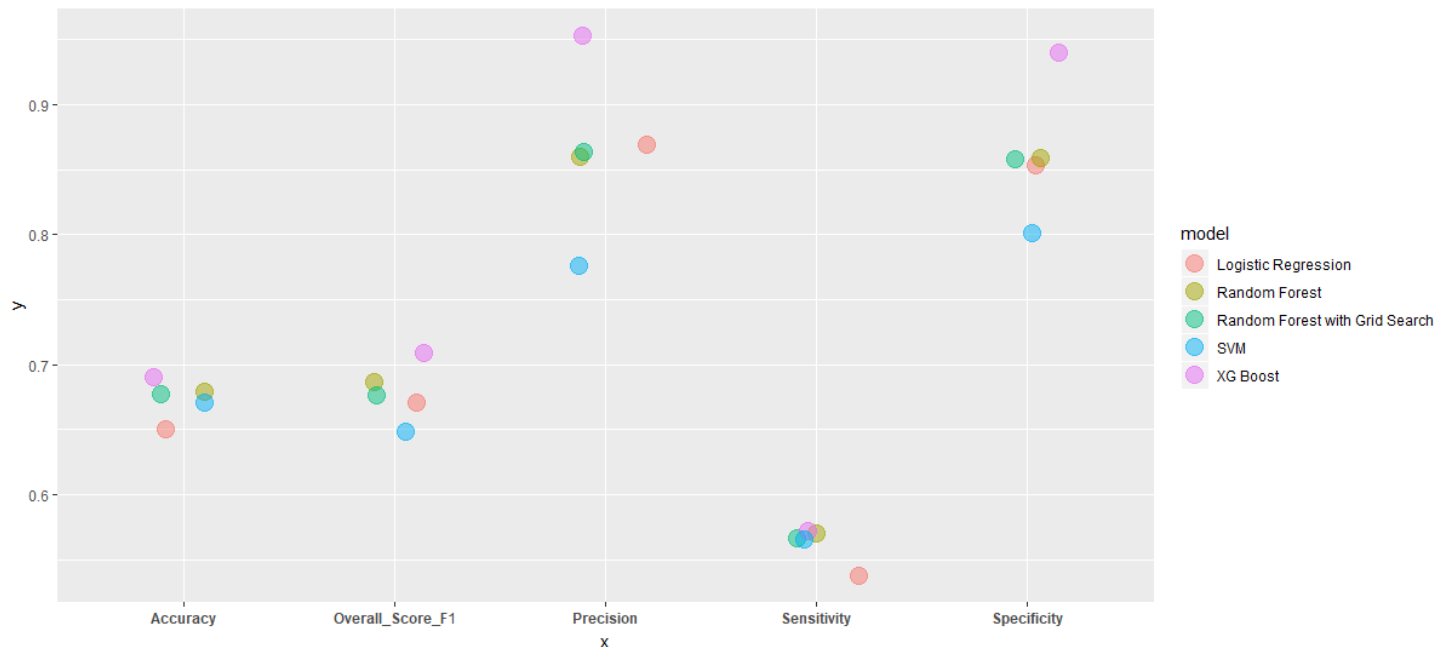
		Predicted		
Actual		Positive	Negative	
	Positive	TP = 1031	FN = 298	1329
	Negative	FP = 793	TN = 1178	1971
		1824	1476	

T → True; F→ False; P→ Positive; N→ Negative

Accuracy : 66.9%
Sensitivity: 56.5%
Specificity: 79.8%



Models Summary



Model	Accuracy	Sensitivity	Specificity	Precision	Overall_Score_F1
Logistic Regression	0.65	0.54	0.85	0.87	0.67
Random Forest	0.68	0.57	0.86	0.86	0.69
XG Boost	0.69	0.57	0.94	0.95	0.71
SVM	0.67	0.57	0.8	0.78	0.65

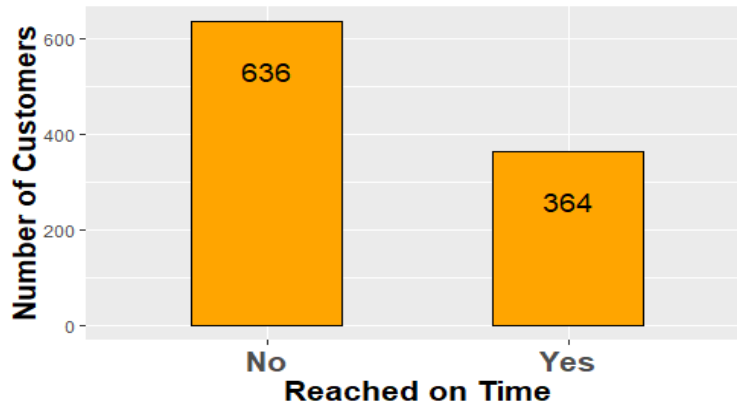
Shipment Analysis w.r.t. Best customer

Assumption

Rating Variables	1	2	3
Product_importance	High	Medium	Low
Customer_care_calls	≤ 3	$>3 \ \& \ \leq 5$	>5
Customer_rating	$==5$	$>=3 \ \& \ <5$	< 2
Prior_purchases	>5	$>2 \ \& \ \leq 5$	$==2$
Cost_of_the_Product	≥ 251	$>169 \ \& \ <251$	≤ 169

- Higher the product importance, prior purchases, customer rating and cost of the product is higher customer score
- Lower the customer care calls is the best customer
- We used binning technique and quartile approach to do segmentation

Shipment vs Best Customer Plots



Conclusion

- Shipment to top customers are not reaching on Time
- There are scope of improvements in services

Clustering is a critical aspect of customer segmentation that allows marketers to better tailor their marketing efforts to various audience subsets in terms of promotional, marketing and product development strategies etc.

- Clustering methods:
 - **Hierarchical Clustering:** Algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.
 - **Kmeans Clustering:** Aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean
- Data Preprocessing Techniques:
 - **Scaling:** Used to standardize the rang of independent variables of Data
 - **Euclidian Distance:** To measure the distance between two data points

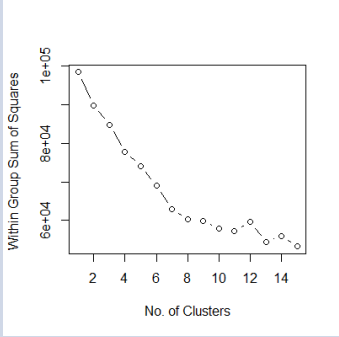
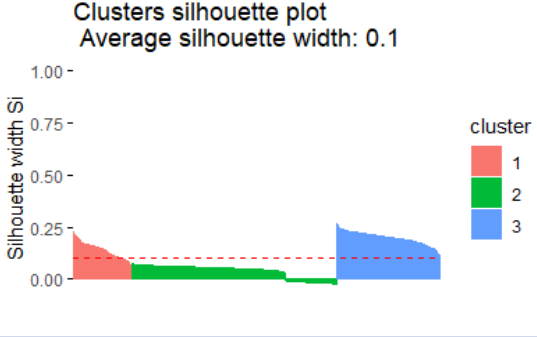
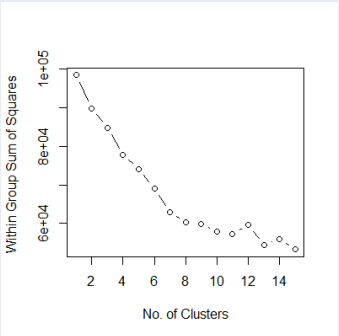
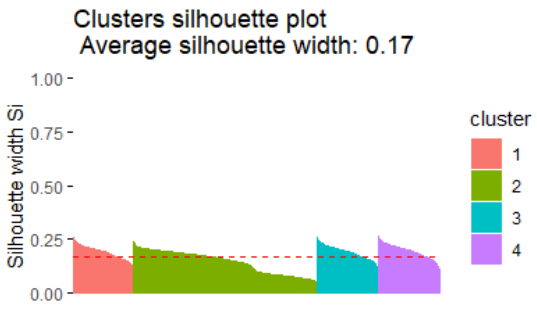
Methods Used:

Hierarchal clustering : 'Ward' function was carried out for clustering. Ward forms matrix of Euclidean distances and checks for variance from cluster to cluster.

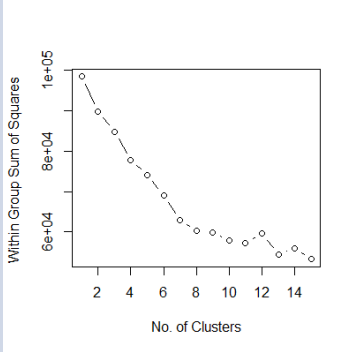
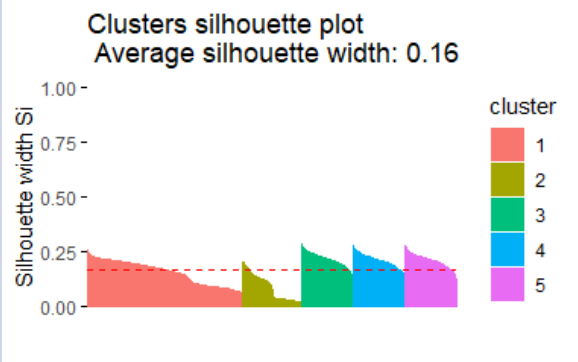
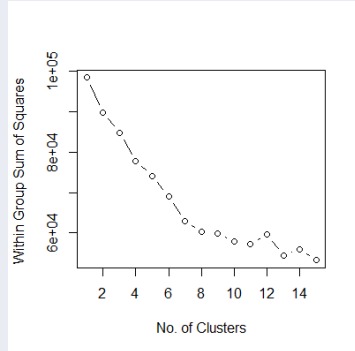

KMeans Clustering: Elbow curve and Silhouette coefficient used to measures how well our observation is clustered and it estimates the average distance between clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters.

Customer Segmentation with Clustering

Comparison between different K values : Based on elbow curve and silhouette coefficient, we find that our model fits best at k~6.

Value of K	Elbow curve	Silhouette plots
3		<p>Clusters silhouette plot Average silhouette width: 0.1</p> 
4		<p>Clusters silhouette plot Average silhouette width: 0.17</p> 

Customer Segmentation with Clustering

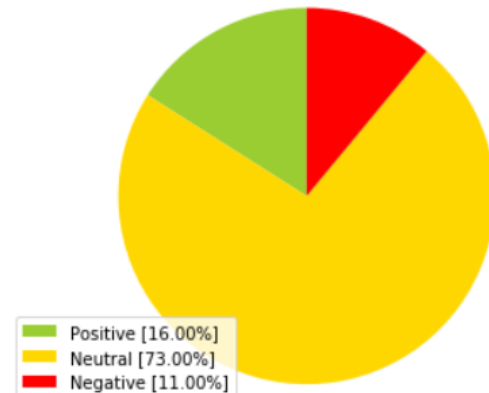
Value of K	Elbow curve	Silhouette plots
5	 <p>Within Group Sum of Squares</p> <p>No. of Clusters</p>	<p>Clusters silhouette plot</p> <p>Average silhouette width: 0.16</p>  <p>Silhouette width Si</p> <p>cluster</p> <p>1</p> <p>2</p> <p>3</p> <p>4</p> <p>5</p>
6	 <p>Within Group Sum of Squares</p> <p>No. of Clusters</p>	<p>Clusters silhouette plot</p> <p>Average silhouette width: 0.17</p>  <p>Silhouette width Si</p> <p>cluster</p> <p>1</p> <p>2</p> <p>3</p> <p>4</p> <p>5</p> <p>6</p>

Customer response across competitors is analysed and compared.

- Data Preparation:
 - **Text Cleansing:** Character case conversion, removal of alpha numeric characters and small length words
 - **Stop Words removal:** Removing commonly used words such as 'a, an, the, in' etc.
 - **Stemming:** Reducing inflected words to their base form
 - **Tokenisation:** Break up a sequence of strings into pieces such as words, keywords, phrases, symbols and other elements
- Sentiment Analysis:
 - **Wordcloud:** Visualisation format to highlight important textual data points
 - **Textblob:** Simplifying language processing by performing tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

As evident from wordcloud and pie chart, snapdeal has comparatively more negative customer response than positive. Words reflecting negative sentiment are- cancel ,worst ,harass ,apology etc.

0	[articl, whether, snapdeal, flipkart, paytm, hike , they, reportedli, want, creat, wechat clon, country , could, breakthrough, ultim, drop, idea let , wait, watch]
2	[snapdeal, custom, care, number 6289683655 8144291766 6289683655 8144291766]
3	[snapdeal, custom, care, number 6289683655 8144291766 6289683655 8144291766]
4	[snapdeal, custom, care, number , 6289683652 8144291766, 6289683652 8144291766, further]
6	[snapdeal, snapdeal, custom, care, number 6289683655 8144291766 6289683655 8144291766]

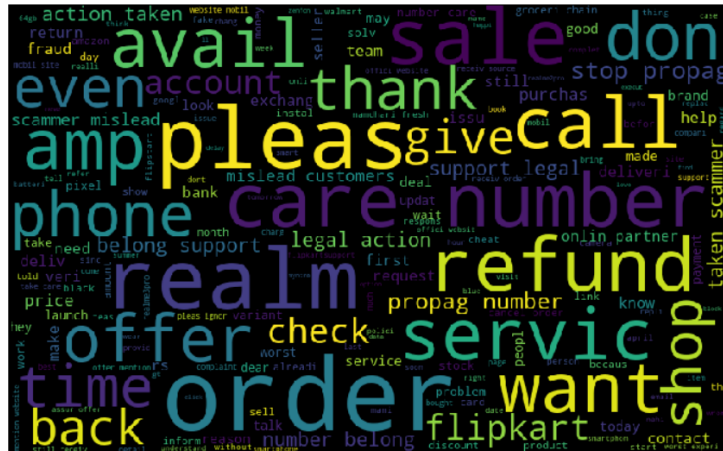


Comparison with competitive companies

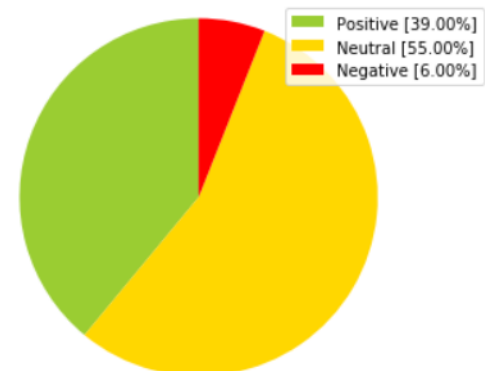
Competitor : Flipkart

As evident from wordcloud and pie chart, Flipkart has more positive customer response than negative. Words reflecting positive sentiment are- support ,care ,refund ,thank etc.

	clean_tokens
2	[anoth, trick, from, tomorrow , govern, amazon, india , walmart flipkart, they, bring]
3	[helplin, number, custom, care 6206090859]
4	[amazon, 1000 , better, they, mistak, they, compens, make, their, custom, happy so, that, custom, will, remain, with, them, next, purchase , flipkart, make, custom, vulner]
5	[stop, propag, number, that, belong, flipkart, custom, support , legal, action, will, taken, against, scammer, that, mislead, customers , arvind]
7	[face, issu, that, return, product, 19th, april, claim, that, seller, refund, amount, 19th]



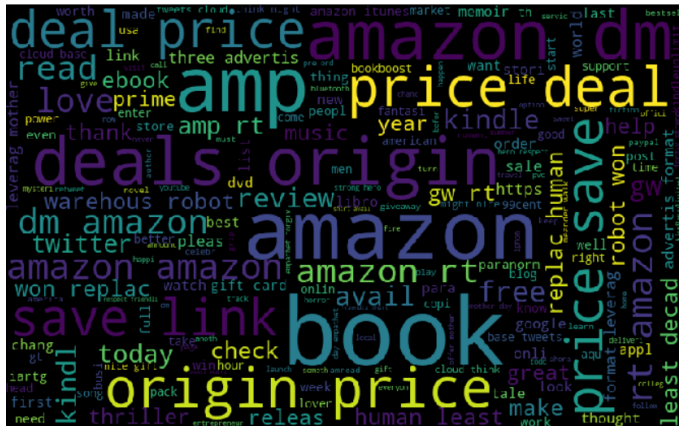
How people are reacting on flipkart by analyzing 1000 Tweets.



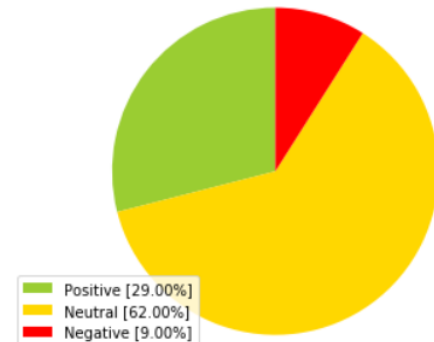
Competitor : Amazon

negative. Words reflecting positive sentiment are- price save ,care ,love ,free etc.

	clean_tokens
4	[amazon 1500 2 rt 5 1 , 21]
5	[free, amazon, gift, card , amazon, free, gift, card]
6	[5 2, amazon, , 10 , j , green, offici, band, score]
7	[, , hakod]
9	[amazon , ,]



How people are reacting on amazon by analyzing 1000 Tweets.



Thank You