# **Customer Profiling**

### **Objective**

To profile customers based on their shared characteristics or experiences within a time span, and then analyze those groups to gather targeted insights on their behaviors and actions.This information would be used by the customer service team to further refine their SOPs based on the behavior.

The use cases which are to be catered are as follows:

* Ability to handle the customer based on the various parameters.
* Route customers based on customer behavior to specific agents.
* Priority queue of customers based on his profile.

### **Variables Selection**

The variables which are used for profiling in the initial iteration are as follows:

| Business Buckets | Variable Name |
| --- | --- |
| **Basic** | asp,items\_per\_order |
| **Return,Refund and Canceled** | refund\_per\_sales,return\_items\_frac,contacts\_per\_order,qc\_fail\_per\_return,cancelled\_orders\_frac,exchange\_items\_frac |
| **Discount** | discount\_per\_sales |
| **Customer Satisfaction** | nps,avg\_csat |
| **Ordering Channel** | Online or offline |
| **Geography** | Geo (metro,non-metro) |

### **Variables Definitions and Formulas**

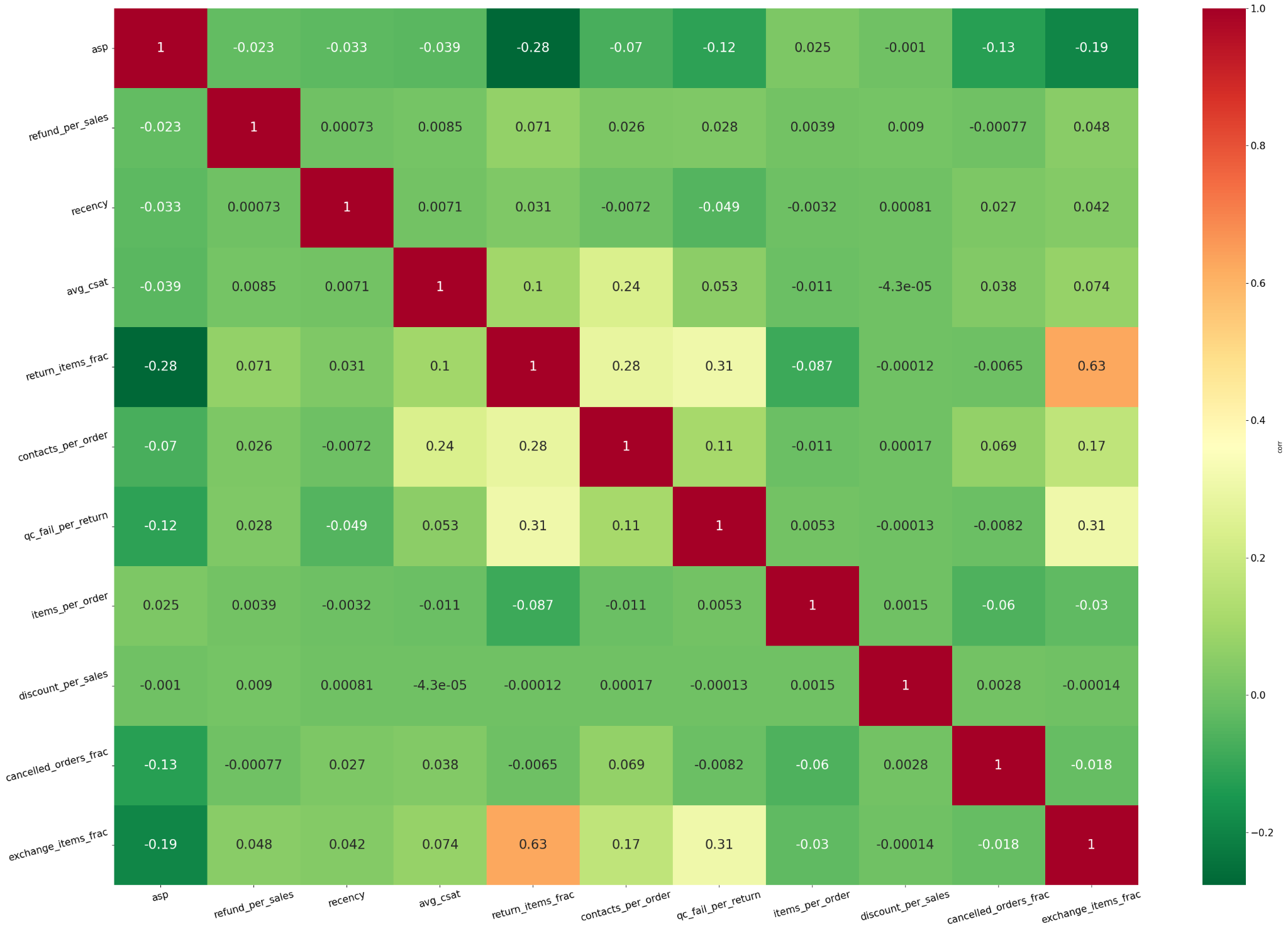
* **ASP** : **Average Selling Price** : (total sales - return refund amount - canceled amount) / (total items - refund items - canceled items)
* **Refund per sales** : **Refund to Sales Ratio** : refund amount / total sales
* **Qc fail per return** : **Ratio of items which are qc fail at the warehouse with respect to total returned items** : number of qc fail items / total returned items
* **Canceled orders frac** : **Canceled Orders to Total Orders Ratio** : number of canceled orders / total number of orders
* **Return items frac : Returned items to Total items** : number of items returned / total number of items
* Exchange items frac : Exchange items to Total items : number of items exchanged / total number of items
* **Discount per sales** : **Discount to Sales Ratio** : total discount amount (excluding bogo discount)/ total sales
* **Items per order** : **Average items bought per order** : total items / total number of orders
* **Contacts per order** : **Average number of times the customer has reached out to the customer service team per order** : total contacts / total number of orders

### **Descriptive Statistics of the Variables Used for Training**

|  | **asp** | **discount\_per\_sales** | **items\_per\_order** | **refund\_per\_sales** | **return\_items\_frac** | **contacts\_per\_order** | **avg\_csat** | **qc\_fail\_per\_return** | **cancelled\_orders\_frac** | **exchange\_items\_frac** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **count** | 6459555 | 6459555 | 6459555 | 6459555 | 6459555 | 6459555 | 6459555 | 6459555 | 6459555 | 6459555 |
| **mean** | 1341.143 | 1.633 | 1.864 | 0.146 | 0.07 | 0.095 | 0.008 | 0.035 | 0.026 | 0.031 |
| **std** | 1405.38 | 1340.862 | 0.815 | 5.255 | 0.202 | 0.428 | 0.086 | 0.176 | 0.131 | 0.13 |
| **min** | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **1%** | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **5%** | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **25%** | 477 | 0 | 1.14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **50%** | 1000 | 0.01 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **75%** | 1800 | 0.3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **90%** | 2764 | 1.27 | 2.67 | 0.64 | 0.27 | 0 | 0 | 0 | 0 | 0 |
| **95%** | 3700 | 1.63 | 3 | 1.05 | 0.5 | 0.75 | 0 | 0 | 0 | 0.25 |
| **99%** | 6591 | 2.33 | 4 | 2 | 1 | 2 | 0 | 1 | 1 | 0.67 |
| **max** | 34375 | 2968984 | 83.33 | 12332.33 | 1 | 117.5 | 1 | 1 | 1 | 1 |

### **Correlation of Variables**

With the help of a correlation heatmap, identification of patterns and interdependencies between two variables have been analyzed to study the effect of change in one parameter on other parameters. In a correlation heatmap, each variable is represented by a row and a column, and the cells show the correlation between them. We can understand the strength and the direction of correlation using the index provided at the right. Darker colors indicate stronger correlations, while lighter colors indicate weaker correlations, this can also be understood with the help of numeric values associated with them.



### **Data Modelling**

The algorithm used for profiling customers is PCA + K Means. We were able to explain the overall 98% variance in the dataset and segmented the data into 3 clusters based on above mentioned variables.

After getting the clusters the classification algorithm(decision trees) was tuned to capture the behavior of new unseen customers.

### **Data Set Summary**

All those customers who have ordered with us in the time period between 1 Feb 2024 and 2 Aug 2022 were a part of our customer base.

Overall Set - 92.28 lakhs

Training Set - 64.59 lakhs

Test Set - 27.68 lakhs

Clustering Algorithm was run on a training dataset.

**Observations Training Set:**

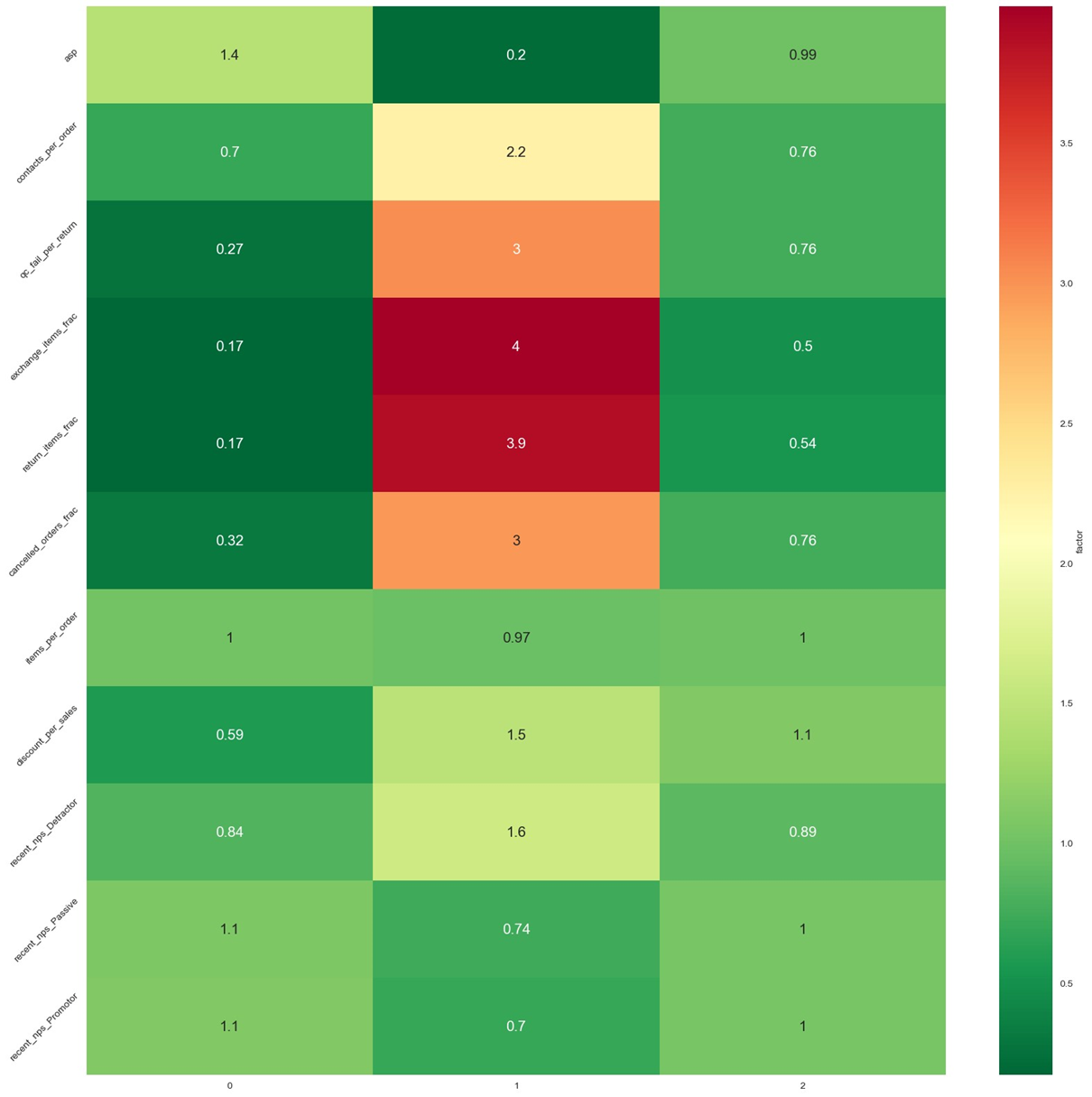
Characteristics of 3 clusters which have been created is described as below:

The major variables which were impacting the customer behavior

* **ASP**
* **Return Items Frac**
* **Exchange Items Frac**
* **Discount Per Sales**
* **Cancelled Orders Frac**
* **Qc Fail Per Return**
* **Contacts Per Order**

This is evident from the heatmap of ratio of cluster mean to population mean.

High variation of the ratio for every feature across the clusters indicates the heterogeneity i.e the importance of that feature in distinguishing the clusters.



|  | **Best** | **Good** | **Worst** |
| --- | --- | --- | --- |
| **Distribution** | **2065458 (31.98%)** | **3273029 (50.67%)** | **1121068 (17.36%)** |
|  | **Cluster 0** | **Cluster 2** | **Cluster 1** |
| **Asp** | **High** | **Medium** | **Low** |
| **Return Items Frac** | **Low** | **Medium** | **High** |
| **Exchange Items Frac** | **Low** | **Medium** | **High** |
| **Discount Per Sales** | **Low** | **High** | **High** |
| **Cancelled Orders Frac** | **Low** | **Medium** | **High** |
| **QC Fail Per Return** | **Low** | **Medium** | **High** |
| **Contacts Per Order** | **Low** | **Low** | **High** |

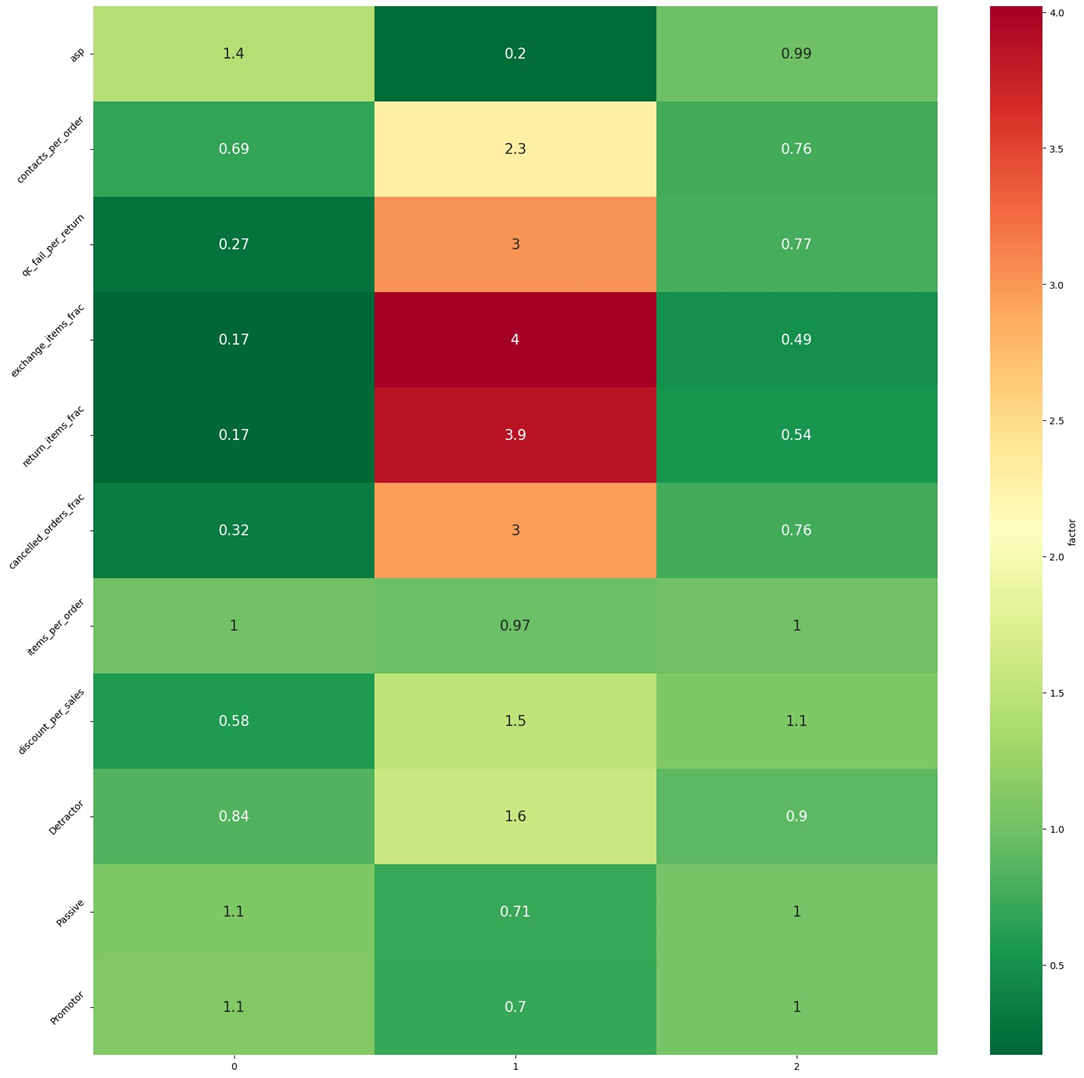
### **Descriptive Statistics of Variables Across Clusters**

|  | **label** | **0** | **1** | **2** |
| --- | --- | --- | --- | --- |
| **asp** | **count** | **2065458.0** | **1121068.0** | **3273029.0** |
| **mean** | **2779.8** | **38.8** | **879.3** |
| **std** | **1627.7** | **50.8** | **353.1** |
| **min** | **1525.0** | **0.0** | **175.0** |
| **25%** | **1850.0** | **0.0** | **600.0** |
| **50%** | **2234.0** | **0.0** | **850.0** |
| **75%** | **3098.0** | **84.0** | **1199.0** |
| **max** | **34375.0** | **185.0** | **1554.0** |
| **refund\_per\_sales** | **count** | **2065458.0** | **1121068.0** | **3273029.0** |
| **mean** | **0.0** | **0.7** | **0.0** |
| **std** | **0.1** | **12.6** | **0.2** |
| **min** | **0.0** | **0.0** | **0.0** |
| **25%** | **0.0** | **0.0** | **0.0** |
| **50%** | **0.0** | **0.5** | **0.0** |
| **75%** | **0.0** | **1.1** | **0.0** |
| **max** | **0.9** | **12332.3** | **1.0** |
| **avg\_csat** | **count** | **2065458.0** | **1121068.0** | **3273029.0** |
| **mean** | **0.0** | **0.0** | **0.0** |
| **std** | **0.1** | **0.1** | **0.1** |
| **min** | **0.0** | **0.0** | **0.0** |
| **25%** | **0.0** | **0.0** | **0.0** |
| **50%** | **0.0** | **0.0** | **0.0** |
| **75%** | **0.0** | **0.0** | **0.0** |
| **max** | **1.0** | **1.0** | **1.0** |
| **return\_items\_frac** | **count** | **2065458.0** | **1121068.0** | **3273029.0** |
| **mean** | **0.0** | **0.3** | **0.0** |
| **std** | **0.0** | **0.4** | **0.1** |
| **min** | **0.0** | **0.0** | **0.0** |
| **25%** | **0.0** | **0.0** | **0.0** |
| **50%** | **0.0** | **0.2** | **0.0** |
| **75%** | **0.0** | **0.5** | **0.0** |
| **max** | **0.9** | **1.0** | **0.9** |
| **exchange\_items\_frac** | **count** | **2065458.0** | **1121068.0** | **3273029.0** |
| **mean** | **0.0** | **0.1** | **0.0** |
| **std** | **0.0** | **0.3** | **0.1** |
| **min** | **0.0** | **0.0** | **0.0** |
| **25%** | **0.0** | **0.0** | **0.0** |
| **50%** | **0.0** | **0.0** | **0.0** |
| **75%** | **0.0** | **0.3** | **0.0** |
| **max** | **1.0** | **1.0** | **1.0** |
| **contacts\_per\_order** | **count** | **2065458.0** | **1121068.0** | **3273029.0** |
| **mean** | **0.1** | **0.3** | **0.1** |
| **std** | **0.3** | **0.8** | **0.3** |
| **min** | **0.0** | **0.0** | **0.0** |
| **25%** | **0.0** | **0.0** | **0.0** |
| **50%** | **0.0** | **0.0** | **0.0** |
| **75%** | **0.0** | **0.0** | **0.0** |
| **max** | **22.0** | **32.0** | **117.5** |
| **qc\_fail\_per\_return** | **count** | **2065458.0** | **1121068.0** | **3273029.0** |
| **mean** | **0.0** | **0.1** | **0.0** |
| **std** | **0.1** | **0.3** | **0.2** |
| **min** | **0.0** | **0.0** | **0.0** |
| **25%** | **0.0** | **0.0** | **0.0** |
| **50%** | **0.0** | **0.0** | **0.0** |
| **75%** | **0.0** | **0.0** | **0.0** |
| **max** | **1.0** | **1.0** | **1.0** |
| **items\_per\_order** | **count** | **2065458.0** | **1121068.0** | **3273029.0** |
| **mean** | **1.9** | **1.8** | **1.9** |
| **std** | **0.6** | **1.1** | **0.8** |
| **min** | **1.0** | **1.0** | **1.0** |
| **25%** | **1.7** | **1.0** | **1.0** |
| **50%** | **2.0** | **2.0** | **2.0** |
| **75%** | **2.0** | **2.0** | **2.0** |
| **max** | **20.0** | **71.0** | **83.3** |
| **discount\_per\_sales** | **count** | **2065458.0** | **1121068.0** | **3273029.0** |
| **mean** | **0.1** | **8.2** | **0.3** |
| **std** | **0.3** | **3218.6** | **0.8** |
| **min** | **0.0** | **0.0** | **0.0** |
| **25%** | **0.0** | **0.0** | **0.0** |
| **50%** | **0.0** | **0.2** | **0.0** |
| **75%** | **0.1** | **1.0** | **0.3** |
| **max** | **21.4** | **2968984.0** | **963.8** |
| **cancelled\_orders\_frac** | **count** | **2065458.0** | **1121068.0** | **3273029.0** |
| **mean** | **0.0** | **0.1** | **0.0** |
| **std** | **0.1** | **0.3** | **0.1** |
| **min** | **0.0** | **0.0** | **0.0** |
| **25%** | **0.0** | **0.0** | **0.0** |
| **50%** | **0.0** | **0.0** | **0.0** |
| **75%** | **0.0** | **0.0** | **0.0** |
| **max** | **0.9** | **1.0** | **0.9** |

### **Observations Test Set**

After performing the clustering on the training set the decision tree is trained for the classification of the test set. The heatmap of the ratio of cluster mean to population

mean is shown below.



|  | **Best** | **Good** | **Worst** |
| --- | --- | --- | --- |
| **Distribution** | **885310 (31.98%)** | **1402701 (50.67%)** | **480370 (17.36%)** |
|  | **Cluster 0** | **Cluster 2** | **Cluster 1** |
| **Asp** | **High** | **Medium** | **Low** |
| **Return Items Frac** | **Low** | **Medium** | **High** |
| **Exchange Items Frac** | **Low** | **Medium** | **High** |
| **Discount Per Sales** | **Low** | **High** | **High** |
| **Canceled Orders Frac** | **Low** | **Medium** | **High** |
| **QC Fail Per Return** | **Low** | **Medium** | **High** |
| **Contacts Per Order** | **Low** | **Low** | **High** |

### 

### **Order Distribution Across Clusters**

|  |  | **Training Dataset** | | **Testing Dataset** | |
| --- | --- | --- | --- | --- | --- |
| **label** | **order\_bucket** | **count** | **percent** | **count** | **percent** |
| **0** | **1** | **1467962** | **71.07%** | **628007** | **70.94%** |
| **2** | **384297** | **18.61%** | **165613** | **18.71%** |
| **3** | **127840** | **6.19%** | **54881** | **6.20%** |
| **4** | **48817** | **2.36%** | **21080** | **2.38%** |
| **5** | **19609** | **0.95%** | **8500** | **0.96%** |
| **6** | **8753** | **0.42%** | **3761** | **0.42%** |
| **7\_to\_10** | **7335** | **0.36%** | **3148** | **0.36%** |
| **more\_than\_10** | **845** | **0.04%** | **320** | **0.04%** |
|  |  |  |  |  |  |
| **1** | **1** | **780732** | **69.64%** | **334979** | **69.73%** |
| **2** | **194316** | **17.33%** | **82908** | **17.26%** |
| **3** | **69832** | **6.23%** | **29809** | **6.21%** |
| **4** | **31567** | **2.82%** | **13407** | **2.79%** |
| **5** | **16038** | **1.43%** | **6973** | **1.45%** |
| **6** | **9270** | **0.83%** | **3932** | **0.82%** |
| **7\_to\_10** | **13428** | **1.20%** | **5857** | **1.22%** |
| **more\_than\_10** | **5885** | **0.52%** | **2505** | **0.52%** |
|  |  |  |  |  |  |
| **2** | **1** | **1960091** | **59.89%** | **839738** | **59.87%** |
| **2** | **705452** | **21.55%** | **301710** | **21.51%** |
| **3** | **295036** | **9.01%** | **126312** | **9.00%** |
| **4** | **138372** | **4.23%** | **59866** | **4.27%** |
| **5** | **71177** | **2.17%** | **30680** | **2.19%** |
| **6** | **38485** | **1.18%** | **16768** | **1.20%** |
| **7\_to\_10** | **50104** | **1.53%** | **21447** | **1.53%** |
| **more\_than\_10** | **14312** | **0.44%** | **6180** | **0.44%** |

### 

### **Customer Scores**

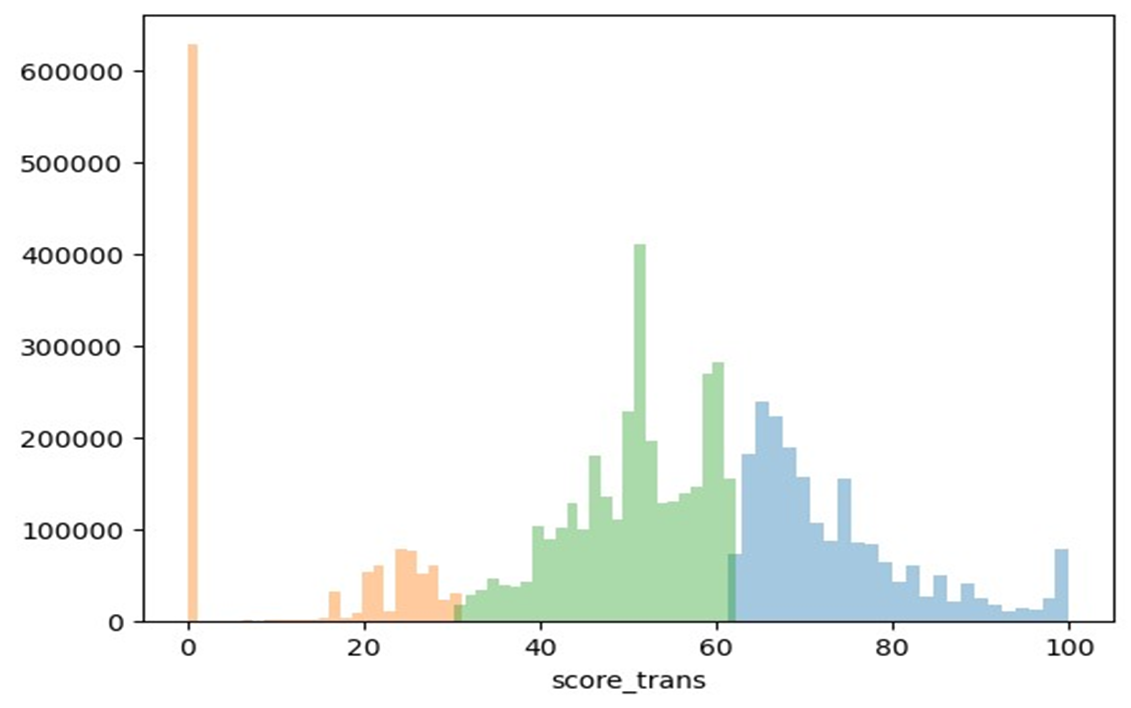
To create customer scores, weights were assigned to the variables on the basis of their gini index and nature of their impact (positive or negative) on the basis of the domain knowledge.

Weight of each variable and the nature of their impact on the score is shown below:

| **features** | **weight** | **nature** |
| --- | --- | --- |
| asp | 0.770 | positive |
| return\_items\_frac | 0.123 | negative |
| exchange\_items\_frac | 0.042 | negative |
| discount\_per\_sales | 0.034 | negative |
| cancelled\_orders\_frac | 0.025 | negative |
| qc\_fail\_per\_return | 0.004 | negative |
| contacts\_per\_order | 0.002 | negative |

Score Distribution Across Clusters is shown below :

| **cluster** | **count** | **mean** | **std** | **min** | **1%** | **5%** | **10%** | **25%** | **50%** | **75%** | **95%** | **99%** | **max** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 2065458 | 73.45 | 9.65 | 61.40 | 62.42 | 63.16 | 64.12 | 66.20 | 70.37 | 78.18 | 96.73 | 100.00 | 100.00 |
| 1 | 1121068 | 10.83 | 12.13 | 0.00 | 0.00 | 0.00 | 0.06 | 0.26 | 0.82 | 24.36 | 28.47 | 30.72 | 31.06 |
| 2 | 3273029 | 50.85 | 7.41 | 30.25 | 32.29 | 36.61 | 40.39 | 45.99 | 51.50 | 57.39 | 60.91 | 61.87 | 62.21 |



Observations from above histogram:

| Cluster | Characteristic | Score Range |
| --- | --- | --- |
| 0 (Best) | High ASP,Low Return,  Low Discount to Sales Ratio | 61.4 to 100 (31.98%) |
| 2 (Worst) | High ASP,Low Return,High Discount to Sales Ratio | 30.25 to 62.21 (17.36%) |
| 1 (Good) | Low ASP,High Return,  High Discount to Sales Ratio | 0 to 31.06 (50.67%) |

### **Distribution of Customers Across Ordering Channel,Satisfaction and Geography**

Assumption : Customer Satisfaction variable is defined as Detractor and Non-Detractor. Promoters,Passives and the customers for which nps response is missing are defined as Non-Detractors and others as Detractors.

On the basis of channel(online vs offline),geo(metro vs non-metro) and satisfaction (detractor vs non-detractor) cohorts are defined. And different customer handling strategies can be defined for each cohort.

Distribution of customers on above outlines is shown below:

| **cluster** | **channel** | **geo** | **satisfaction** | **overall** | **percentage** |
| --- | --- | --- | --- | --- | --- |
| **Best** | **offline** | **Metro** | **Detractor** | 18094 | 0.61% |
| **Non-Detractor** | 1495171 | 50.67% |
| **Non-Metro** | **Detractor** | 13624 | 0.46% |
| **Non-Detractor** | 1195337 | 40.51% |
| **online** | **Metro** | **Detractor** | 1551 | 0.05% |
| **Non-Detractor** | 78758 | 2.67% |
| **Non-Metro** | **Detractor** | 2311 | 0.08% |
| **Non-Detractor** | 145922 | 4.95% |
|  |  |  |  |  |  |
| **Worst** | **offline** | **Metro** | **Detractor** | 10694 | 0.67% |
| **Non-Detractor** | 438978 | 27.41% |
| **Non-Metro** | **Detractor** | 7967 | 0.50% |
| **Non-Detractor** | 359133 | 22.43% |
| **online** | **Metro** | **Detractor** | 5225 | 0.33% |
| **Non-Detractor** | 183682 | 11.47% |
| **Non-Metro** | **Detractor** | 12958 | 0.81% |
| **Non-Detractor** | 582801 | 36.39% |
|  |  |  |  |  |  |
| **Good** | **offline** | **Metro** | **Detractor** | 20151 | 0.43% |
| **Non-Detractor** | 1805590 | 38.62% |
| **Non-Metro** | **Detractor** | 16386 | 0.35% |
| **Non-Detractor** | 1543559 | 33.01% |
| **online** | **Metro** | **Detractor** | 8012 | 0.17% |
| **Non-Detractor** | 378877 | 8.10% |
| **Non-Metro** | **Detractor** | 15614 | 0.33% |
| **Non-Detractor** | 887541 | 18.98% |

### 

### **Production Working**

* Assign one time labels to all the transacting customers in the last 18 months through clustering.
* For D-1, new customers (which have placed first order with us) are assigned labels and scores everyday through classification and inserted into the database.
* After every three weeks clustering is again run on the entire dataset including repeat to relabel the entire customer base.

Schema fields that would be used to define the customer handling strategy are as follows:

| **Field** | **Description** |
| --- | --- |
| **Customer Phone** | Customer contact number |
| **Channel** | Online (1) or Offline(0) |
| **Location type** | Metro or Non-Metro |
| **Satisfaction** | Detractor or Non-Detractor |
| **Customer Profile** | 0 (Best),2 (Good),1(Worst) |
| **Customer Score** | Customer Score (Between 0 to 100) |

: