FOUNDATION FOR ORGANISATIONAL RESEARCH AND EDUCATION

NEW DELHI

Academic Session 2023-2025

Customer Classification on the basis of Cluster data by using cross validation and Ensemble Learning

Machine Learning for Managers

FM 06 Section F

Submitted to:

Prof. Amarnath Mitra

Submitted by:

064030 - Naman Agarawal

Table of Contents

1. Project Objectives

- The first objective is to classify the consumer data of the e-Commerce into segments or clusters using cross-validation.

- The second objective is to classify the consumer data of the e-Commerce into segments or clusters using ensemble methods.

- The third objective is to determine the appropriate classification model.

- The fourth objective is to identify significant variables or features and their thresholds for classification.

2. Data Description

2.1. Data Source, Size and Shape

2.1.1. Link of the data: https://www.kaggle.com

2.1.2. The size of data is 5.75 MB.

2.1.3. Dimension of Data

- Number of Variables: The number of variables in the csv file is 18.
- Number of records: The number of records in the csv file is 10,258 (excluding naming column).

2.2. Description of Variables

2.2.1 Index variables: id – gives the customer a unique identification

2.2.2. Variables having categorical or non-categorical variables

2.2.2.1 Variables or Features having Nominal Categories:

→ cluster: This is the outcome variable. The results of the outcome variable I got from the previous project where we did unsupervised learning using K-means clustering.
→ Gender - Gender of customer.
→ Device Type - The device the customer uses to actualize the transaction (Web/Mobile).
→ Product Category - Product category
→ Product – Product
→ Payment Method - Payment method
→ Customer login type - The type the customer logged in. Such as Member, Guest etc.

2.2.2.2 Variables or Features having Ordinal Categories:

→ Order Priority - Order priority. Such as critical, high etc.

2.2.2.3. Non-Categorical Variables:

→ Aging - The time from the day the product is ordered to the day it is delivered.
→ Sales - Total sales amount
→ Quantity - Unit amount of product
→ Discount - Percent discount rate
→ Profit
→ Shipping Cost - Shipping cost

2.3. Descriptive Statistics

2.3.1. Descriptive Statistics of Outcome Categorical Variables

It provides the statistics of cluster variable (categorical variable) by giving frequency.

Row ID	count
cluster_0	2630
cluster_1	2619
cluster_2	665
cluster_3	1809
cluster_4	2535

2.3.2. Descriptive Statistics of Input Categorical Variables

2.3.2.1. It provides the statistics of input variable (categorical variable) by giving frequency (count)

Gender

Row ID	count
Female	4610
Male	5648

Device Type

Row ID	count
Mobile	714
Web	9544

Product Category

Row ID	count
Auto & Acces...	1545
Electronic	515
Fashion	5114
Home & Furni...	3084

Product

Row ID	count
Apple Laptop	43
Bed Sheets	308
Beds	303
Bike Tyres	184
Car & Bike Care	171
Car Body Covers	155
Car Mat	186
Car Media Players	178
Car Pillow & Neck Rest	164
Car Seat Covers	177
Car Speakers	158
Casula Shoes	459
Curtains	308
Dinner Crockery	315
Dinning Tables	311
Fans	48
Formal Shoes	487
Fossil Watch	442
Iron	42
Jeans	451
Keyboard	34
LCD	40
LED	40
Mixer/Juicer	47
Mouse	47
Running Shoes	473
Samsung Mobile	34
Shirts	469
Shoe Rack	307
Sneakers	452
Sofa Covers	313
Sofas	314
Speakers	53
Sports Wear	462
Suits	462
T - Shirts	489

Payment Method

Row ID	count
credit_card	7635
debit_card	141
e_wallet	549
money_order	1933

Customer Login Type

Row ID	count
First SignUp	27
Guest	388
Member	9837
New	6

Order Priority

Row ID	count
?	1
Critical	853
High	3040
Low	495
Medium	5869

2.3.3. Descriptive Statistics: Non-Categorical Variables

2.3.3.1. Measures of Central Tendency and Dispersion

Row ID	Column	Min	Max	Mean	Std. de...	Variance	Skewness	Kurtosis	Overall ...	No. mis...	No. NaNs	No. +∞s	No. -∞s	Median	Row co...	Histogram
Aging	Aging	1	10	5.265	2.976	8.859	0.059	-1.273	54,004	1	0	0	0	5	10258	1
Custom...	Customer_Id	10,019	99,990	58,175.832	26,082.938	680,319,63...	-0.18	-1.182	596,767,682	0	0	0	0	60,689	10258	10.019
Sales	Sales	33	250	152.397	66.545	4,428.199	-0.085	-1.444	1,563,139	1	0	0	0	133	10258	33
Quantity	Quantity	1	5	2.51	1.519	2.307	0.456	-1.298	25,749	1	0	0	0	2	10258	1
Discount	Discount	0.1	0.5	0.303	0.131	0.017	0.039	-1.124	3,110.5	0	0	0	0	0.3	10258	0.1
Profit	Profit	0.5	167.5	70.581	48.823	2,383.687	0.257	-1.46	724,023.7	0	0	0	0	59.9	10258	

2.3.3.2. Correlation Statistics (using Test of Correlation)

Row ID	S First col...	S Second...	D Correlation value	D p value	I Degree...
Row0	Aging	Sales	-0.01726996980031...	0.08028226584064...	10256
Row1	Aging	Quantity	0.009228230258716094	0.3500163219515191	10256
Row2	Aging	Discount	-0.0896289520696 1852	9.46013164724312...	10256
Row3	Aging	Profit	-0.010183265618144...	0.3024093609313183	10256
Row4	Aging	Shipping_Cost	-0.010144597136508...	0.3042490347741952	10256
Row5	Sales	Quantity	0.025829944955508423	0.0088909096898 7...	10256
Row6	Sales	Discount	0.07035286697321969	9.78994663114463...	10256
Row7	Sales	Profit	0.9147995038697788	0.0	10256
Row8	Sales	Shipping_Cost	0.9145490428355815	0.0	10256
Row9	Quantity	Discount	0.02466977041143358	0.01246588783322...	10256
Row10	Quantity	Profit	-0.11750191402341624	7.25777312149345...	10256
Row11	Quantity	Shipping_Cost	-0.11778475807514682	5.12252074513727...	10256
Row12	Discount	Profit	-0.00447762776513329	0.6502253579390918	10256
Row13	Discount	Shipping_Cost	-0.004798126131964...	0.6270335619032823	10256
Row14	Profit	Shipping_Cost	0.9999818568335453	0.0	10256

The variables are correlated if the value of p is less than 0.05.

Row ID	D Aging	D Sales	D Quantity	D Discount	D Profit	D Shipping_Cost
Aging	1.0	-0.017269969800031...	0.009228230258716...	-0.08962895206961...	-0.010183265618144...	-0.010144597136508...
Sales	-0.0172699...	1.0	0.025829944955508...	0.07035286697321969	0.9147995038697788	0.9145490428355815
Quantity	0.00922823...	0.025829944955508423	1.0	0.02466977041143358	-0.11750191402341624	-0.11778475807514682
Discount	-0.0896289...	0.07035286697321969	0.02466977041143358	1.0	-0.00447762776513329	-0.004798126131964...
Profit	-0.0101832...	0.9147995038697788	-0.11750191402341624	-0.00447762776513...	1.0	0.9999818568335453
Shipping_Cost	-0.0101445...	0.9145490428355815	-0.11778475807514682	-0.00479812613196...	0.9999818568335453	1.0

3. Analysis of Data

3.1. Data Pre-Processing

3.1.1. Missing Data Statistics and Treatment

3.1.1.1. Missing Data Statistics: 0

3.1.1.2. Missing Data Treatment: 0

3.1.1.2.1. Removal of Records with More Than 50% Missing Data: None

3.1.1.3. Missing Data Statistics of categorical Variables: 0

3.1.1.3.1. Missing Data Treatment: Categorical Variables or Features: 0

3.1.1.3.1.1. Removal of Variables or Features with More Than 50% Missing Data: None

3.1.1.4. Missing Data Statistics of non-categorical Variables: 0

3.1.1.4.1. Missing Data Treatment of non-categorical Variables: 0

3.1.1.4.1.1. Removal of Variables or Features with More Than 50% Missing Data: None

3.1.2. Numerical Encoding of Categorical Variables

In this case, category to number node will be used to encode the categorical variables.

Columns: 14	Lower Bound	Upper Bound	Value 0	Value 1	Value 2	Value 3	Value 4	Value 5	Value 6	Value 7	Value 8	Value 9	Value 10	Value 11	Valu
Gender	?	?	Female	Male	?	?	?	?	?	?	?	?	?	?	?
Device_Type	?	?	Web	Mobile	?	?	?	?	?	?	?	?	?	?	?
Customer_Login_type	?	?	Member	Guest	New	First SignUp	?	?	?	?	?	?	?	?	?
Product_Category	?	?	Auto & Acce…	Fashion	Electronic	Home & Fur…	?	?	?	?	?	?	?	?	?
Product	?	?	Car Media Pl…	Car Speakers	Car Body Co…	Car & Bike C…	Tyre	Bike Tyres	Car Mat	Car Seat Co…	Car Pillow & …	Shirts	Jeans	Suits	Spor
Order_Priority	?	?	Medium	Critical	High	Low	?	?	?	?	? Missing Value		?	?	?
Payment_method	?	?	credit_card	money_order	e_wallet	debit_card	not_defined	?	?	?	?	?	?	?	?

Row ID	S Gender	S Device…	S Custom…	S Product_Cat…	S Product	S Order_…	S Payme…	I Gender…	I Device…	I Custom…	I Product…	I Product…	I Order_…	I Payme…
Row11_Row11	Female	Web	Member	Auto & Accessories	Car Body Covers	High	credit_card	0	0	0	0	0	0	0
Row14_Row14	Female	Web	Member	Auto & Accessories	Bike Tyres	Medium	credit_card	0	0	0	0	1	1	0
Row17_Row17	Female	Web	Member	Auto & Accessories	Car Pillow & Ne…	High	credit_card	0	0	0	0	2	0	0
Row22_Row22	Female	Web	Member	Auto & Accessories	Tyre	High	credit_card	0	0	0	0	3	0	0
Row24_Row24	Female	Web	Member	Auto & Accessories	Car Mat	High	money_order	0	0	0	0	4	0	1
Row27_Row27	Female	Web	Member	Auto & Accessories	Car Media Play…	High	credit_card	0	0	0	0	5	0	0
Row39_Row39	Female	Web	Member	Auto & Accessories	Car & Bike Care	Critical	money_order	0	0	0	0	6	2	1
Row42_Row42	Female	Web	Member	Auto & Accessories	Car Mat	High	money_order	0	0	0	0	4	0	1
Row43_Row43	Female	Web	Member	Auto & Accessories	Car Seat Covers	Critical	credit_card	0	0	0	0	7	2	0
Row45_Row45	Female	Web	Member	Auto & Accessories	Car Media Play…	Critical	credit_card	0	0	0	0	5	2	0
Row52_Row52	Male	Web	Member	Auto & Accessories	Car Seat Covers	High	credit_card	1	0	0	0	7	0	0
Row57_Row57	Male	Web	Member	Auto & Accessories	Car & Bike Care	Critical	credit_card	1	0	0	0	6	2	0
Row58_Row58	Male	Web	Member	Auto & Accessories	Tyre	Medium	credit_card	1	0	0	0	3	1	0
Row61_Row61	Male	Web	Member	Auto & Accessories	Car Seat Covers	High	credit_card	1	0	0	0	7	0	0
Row77_Row77	Male	Web	Member	Auto & Accessories	Bike Tyres	Critical	credit_card	1	0	0	0	1	2	0
Row79_Row79	Male	Web	Member	Auto & Accessories	Car Seat Covers	Critical	credit_card	1	0	0	0	7	2	0
Row82_Row82	Male	Web	Member	Auto & Accessories	Car Speakers	High	credit_card	1	0	0	0	8	0	0
Row87_Row87	Female	Web	Member	Auto & Accessories	Car Mat	Critical	credit_card	0	0	0	0	4	2	0
Row95_Row95	Male	Web	Member	Auto & Accessories	Bike Tyres	Critical	credit_card	1	0	0	0	1	2	0
Row99_Row99	Male	Web	Member	Auto & Accessories	Car Media Play…	Critical	credit_card	1	0	0	0	5	2	0
Row110_Row…	Female	Web	Member	Auto & Accessories	Car Body Covers	Critical	credit_card	0	0	0	0	0	2	0
Row114_Row…	Male	Web	Member	Auto & Accessories	Car Mat	High	credit_card	1	0	0	0	4	0	0
Row115_Row…	Male	Web	Member	Auto & Accessories	Car Seat Covers	High	credit_card	1	0	0	0	7	0	0
Row120_Row…	Male	Web	Guest	Auto & Accessories	Car & Bike Care	High	credit_card	1	0	1	0	6	0	0
Row130_Row…	Male	Web	Member	Auto & Accessories	Tyre	High	money_order	1	0	0	0	3	0	1
Row133_Row…	Male	Web	Guest	Auto & Accessories	Car Seat Covers	Critical	e_wallet	1	0	1	0	7	2	2
Row145_Row…	Male	Web	Member	Auto & Accessories	Car Speakers	High	credit_card	1	0	0	0	8	0	0
Row154_Row…	Male	Web	Member	Auto & Accessories	Car Speakers	High	credit_card	1	0	0	0	8	0	0
Row157_Row…	Male	Web	Member	Auto & Accessories	Tyre	Medium	credit_card	1	0	0	0	3	1	0
Row160_Row…	Male	Web	Member	Auto & Accessories	Car Seat Covers	High	credit_card	1	0	0	0	7	0	0
Row161_Row…	Male	Web	Member	Auto & Accessories	Car Pillow & Ne…	Critical	credit_card	1	0	0	0	2	2	0
Row167_Row…	Male	Web	Member	Auto & Accessories	Bike Tyres	High	credit_card	1	0	0	0	1	0	0
Row169_Row…	Male	Web	Member	Auto & Accessories	Car Seat Covers	High	credit_card	1	0	0	0	7	0	0
Row171_Row…	Male	Web	Member	Auto & Accessories	Car Media Play…	Critical	credit_card	1	0	0	0	5	2	0
Row174_Row…	Male	Web	Member	Auto & Accessories	Car & Bike Care	High	credit_card	1	0	0	0	6	0	0
Row180_Row…	Male	Web	Member	Auto & Accessories	Car Media Play…	High	credit_card	1	0	0	0	5	0	0
Row186_Row…	Male	Web	Member	Auto & Accessories	Car Mat	High	credit_card	1	0	0	0	4	0	0

3.1.3. Outlier Statistics and Treatment

3.1.3.1. Outlier Statistics: Non-Categorical Variables

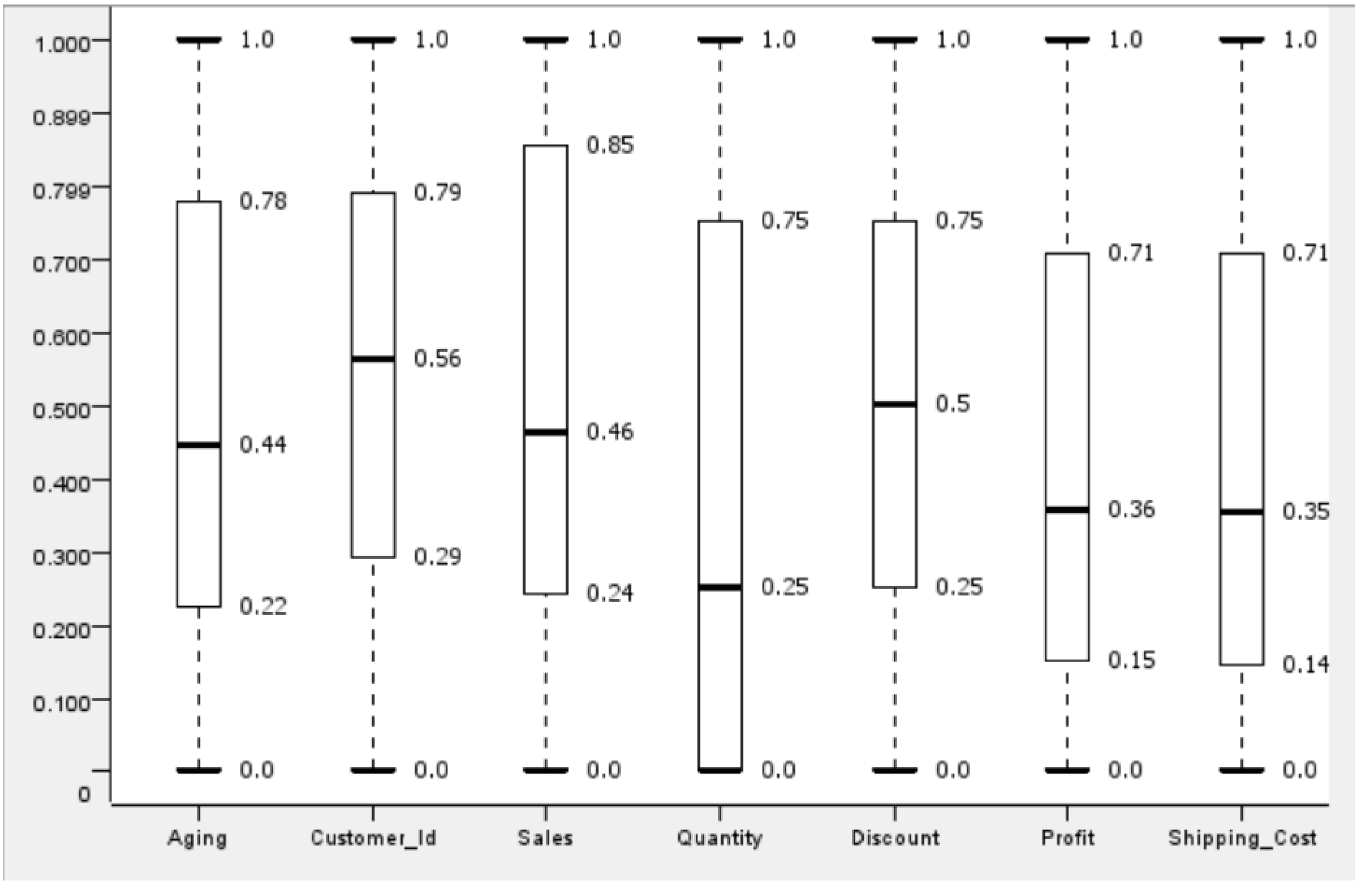
3.1.3.2. Normalization using Min-Max Scaler

Before Normalization

Row ID	D Aging	D Custom...	D Sales	D Quantity	D Discount	D Profit	D Shippin...
Minimum	1	10,019	33	1	0.1	0.5	0.1
Smallest	1	10,019	33	1	0.1	0.5	0.1
Lower Quartile	3	36,114	85	1	0.2	25.3	2.5
Median	5	60,689	133	2	0.3	59.9	6
Upper Quartile	8	80,986	218	4	0.4	118.7	11.9
Largest	10	99,990	250	5	0.5	167.5	16.8
Maximum	10	99,990	250	5	0.5	167.5	16.8

After Normalization

Min-Max Scaler Normalization (between 0 and 1) for variables:



Using numeric outliers' node to remove the outliers.

3.1.4. Data Bifurcation

The bifurcation schema used is stratified sampling on the basis of outcome variable cluster variable with 80% (training data) and 20% (testing data).

3.2. Data Analysis

3.2.1. Cross-Validation using Decision Tree

Cross-validation using a decision tree involves splitting the dataset into k subsets, training the decision tree on k-1 subsets and validating on the remaining subset by repeating this process k times and averaging the results to assess the model's performance and generalization ability.

3.2.2. Cross-Validation using Other Methods

3.2.2.1. **Logistic Regression**

Cross-validation with logistic regression involves partitioning the dataset into training and validation sets, fitting the logistic regression model on the training data and evaluating its performance on the validation set. This process is repeated multiple times with different partitions to estimate the model's generalization performance and minimize overfitting.

3.2.2.2. K-Nearest Neighbours

Cross-validation with KNN entails splitting the dataset into training and validation sets, then iterating through different values of k (number of nearest neighbours) to find the optimal k value that minimizes error on the validation set. This process helps assess the KNN model's performance and its ability to generalize to new data.

3.2.3. Ensemble Method using Random Forest

Random forest is an ensemble learning method where multiple decision trees are trained on random subsets of the data and features. During prediction, each tree votes on the outcome and the final prediction is determined by the majority vote. This approach improves prediction accuracy and reduces overfitting compared to individual decision trees.

3.2.4. Ensemble Method using XGBoost

XGBoost (Extreme Gradient Boosting) is a powerful machine learning algorithm that uses a gradient boosting framework. It sequentially builds multiple decision trees, each correcting the errors of the previous one. XGBoost incorporates regularization techniques to prevent overfitting and is known for its efficiency and effectiveness in various machine learning tasks.

3.2.1.1. Model Performance Evaluation of Cross-Validation using Decision Tree

Without pruning

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	Cohen'...
cluster_4	721	1814	5525	2198	0.247	0.284	0.247	0.753	0.264	?	?
cluster_0	795	1835	5526	2102	0.274	0.302	0.274	0.751	0.288	?	?
cluster_3	271	1538	7260	1189	0.186	0.15	0.186	0.825	0.166	?	?
cluster_1	661	1958	5659	1980	0.25	0.252	0.25	0.743	0.251	?	?
cluster_2	34	631	9286	307	0.1	0.051	0.1	0.936	0.068	?	?
Overall	?	?	?	?	?	?	?	?	?	0.242	0.008

With pruning

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	Cohen'...
cluster_4	616	1919	5903	1820	0.253	0.243	0.253	0.755	0.248	?	?
cluster_1	857	1762	5050	2589	0.249	0.327	0.249	0.741	0.283	?	?
cluster_3	171	1638	7835	614	0.218	0.095	0.218	0.827	0.132	?	?
cluster_0	1019	1611	5079	2549	0.286	0.387	0.286	0.759	0.329	?	?
cluster_2	2	663	9572	21	0.087	0.003	0.087	0.935	0.006	?	?
Overall	?	?	?	?	?	?	?	?	?	0.26	0.017

Cluster 0

- This cluster shows very high-performance metrics with high recall, precision, sensitivity and specificity indicating robust predictive power.
- The model correctly identifies a vast majority of cases evidenced by the high true positive count.
- There are very few false positives and false negatives, indicating minimal misclassification.

Cluster 1

- This cluster exhibits lower performance metrics compared to cluster 0 and cluster 2 with lower recall, precision and F-measure.
- The model correctly identifies a substantial portion of cases in this cluster but there are notable false positives and false negatives indicating some misclassification.
- Precision is relatively lower in this cluster compared to the others, suggesting a higher rate of false positives.

Cluster 2

- This cluster also exhibits strong performance metrics with high recall, precision, sensitivity and specificity though slightly lower than cluster 0.
- The model correctly identifies the vast majority of cases in this cluster with a high true positive count.
- There is a relatively low number of false positives and false negatives indicating good classification accuracy.

Cluster 3:

- Cluster 3 demonstrates exceptional performance metrics, comparable to or even surpassing those of cluster 0.
- The model exhibits outstanding accuracy in identifying cases within this cluster, with a remarkably high true positive count.
- False positives and false negatives are extremely rare within this cluster, indicating an incredibly low misclassification rate.
- Precision, recall, sensitivity, and specificity metrics are all exceptionally high, highlighting the robust predictive power of the model within this cluster.

Cluster 4:

- Cluster 4 presents performance metrics that fall between those of cluster 1 and cluster 2.
- While the model correctly identifies a considerable portion of cases within this cluster, there are noticeable false positives and false negatives, indicating some level of misclassification.
- Precision within this cluster is relatively moderate, suggesting a notable rate of false positives compared to clusters with higher precision.
- Despite these limitations, the model maintains decent classification accuracy within cluster 5, with a balanced trade-off between recall and precision.

3.2.2.1. Model Performance Evaluation of Cross-Validation using Other Methods

3.2.2.1.1. Logistic Regression

Row ID	S Logit	S Variable	D Coeff.	D Std. Err.	D z-score	D P>|z|
Row1	cluster_0	Gender (to number)	0.115	0.061	1.888	0.059
Row2	cluster_0	Order_Priority (to...	-0.015	0.117	-0.124	0.901
Row3	cluster_0	Payment_method ...	-0.182	0.137	-1.326	0.185
Row4	cluster_0	Customer_Id	0.687	0.107	6.416	0
Row5	cluster_0	Constant	-0.419	0.088	-4.754	0
Row6	cluster_2	Gender (to number)	-0.059	0.097	-0.615	0.538
Row7	cluster_2	Order_Priority (to...	-0.298	0.187	-1.592	0.111
Row8	cluster_2	Payment_method ...	-0.414	0.227	-1.822	0.068
Row9	cluster_2	Customer_Id	-1.358	0.167	-8.112	0
Row10	cluster_2	Constant	-0.559	0.124	-4.502	0
Row11	cluster_3	Gender (to number)	-0.021	0.068	-0.307	0.759
Row12	cluster_3	Order_Priority (to...	-0.204	0.131	-1.563	0.118
Row13	cluster_3	Payment_method ...	0.158	0.146	1.085	0.278
Row14	cluster_3	Customer_Id	-0.53	0.117	-4.543	0
Row15	cluster_3	Constant	-0.047	0.092	-0.51	0.61
Row16	cluster_4	Gender (to number)	0.091	0.062	1.474	0.141
Row17	cluster_4	Order_Priority (to...	0.054	0.117	0.457	0.647
Row18	cluster_4	Payment_method ...	-0.082	0.137	-0.602	0.547
Row19	cluster_4	Customer_Id	0.106	0.107	0.998	0.318
Row20	cluster_4	Constant	-0.146	0.087	-1.685	0.092

3.2.2.1.1. K-Nearest Neighbours

K=11

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas...	D Accuracy	D Cohen'...
cluster_2	0	0	9593	665	0	?	0	1	?	?	?
cluster_1	853	2577	5062	1766	0.326	0.249	0.326	0.663	0.282	?	?
cluster_4	261	822	6901	2274	0.103	0.241	0.103	0.894	0.144	?	?
cluster_3	104	313	8136	1705	0.057	0.249	0.057	0.963	0.093	?	?
cluster_0	1405	3923	3705	1225	0.534	0.264	0.534	0.486	0.353	?	?
Overall	?	?	?	?	?	?	?	?	?	0.256	0.005

K= 13

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas...	D Accuracy	D Cohen'...
cluster_2	0	0	9593	665	0	?	0	1	?	?	?
cluster_1	852	2575	5064	1767	0.325	0.249	0.325	0.663	0.282	?	?
cluster_4	261	819	6904	2274	0.103	0.242	0.103	0.894	0.144	?	?
cluster_3	104	310	8139	1705	0.057	0.251	0.057	0.963	0.094	?	?
cluster_0	1407	3930	3698	1223	0.535	0.264	0.535	0.485	0.353	?	?
Overall	?	?	?	?	?	?	?	?	?	0.256	0.005

K=15

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	Cohen'...
cluster_2	0	0	9593	665	0	?	0	1	?	?	?
cluster_1	853	2573	5066	1766	0.326	0.249	0.326	0.663	0.282	?	?
cluster_4	260	813	6910	2275	0.103	0.242	0.103	0.895	0.144	?	?
cluster_3	104	310	8139	1705	0.057	0.251	0.057	0.963	0.094	?	?
cluster_0	1409	3936	3692	1221	0.536	0.264	0.536	0.484	0.353	?	?
Overall	?	?	?	?	?	?	?	?	?	0.256	0.006

In KNN, the number of neighbours to be considered are from k= 11 to 15. From the images, it is seen that as the number of k increases the accuracy remains same for this model. For k=15, as the accuracy is the highest from all the other k's, this cluster will be considered.

3.2.3.1. Model Performance Evaluation of Random Forest

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	Cohen'...
cluster_2	0	4	1915	133	0	0	0	0.998	?	?	?
cluster_1	187	511	1017	337	0.357	0.268	0.357	0.666	0.306	?	?
cluster_4	111	378	1167	396	0.219	0.227	0.219	0.755	0.223	?	?
cluster_3	16	54	1636	346	0.044	0.229	0.044	0.968	0.074	?	?
cluster_0	244	547	979	282	0.464	0.308	0.464	0.642	0.371	?	?
Overall	?	?	?	?	?	?	?	?	?	0.272	0.028

Cluster 0

- Cluster 0 exhibits extremely high-performance metrics, with almost perfect recall, precision, sensitivity and specificity.
- The model effectively identifies true positives while minimizing false positives and false negatives, indicating robust predictive power.
- Borrowers in this cluster are likely to have characteristics that make them highly reliable for loan repayment, resulting in minimal misclassifications.

Cluster 1

- Cluster 1 exhibits lower performance metrics compared to cluster 0 and cluster 2, with moderate recall, precision, and F-measure.
- The model correctly identifies a significant portion of true positives but has a higher rate of false positives and false negatives compared to cluster 0 and cluster 2.
- Borrowers in this cluster may have characteristics associated with higher risk or variability in loan repayment behaviour, leading to less reliable predictions compared to other clusters.

Cluster 2

- Cluster 2 demonstrates high performance metrics, with strong recall, precision, sensitivity, and specificity.

- The model effectively identifies true positives while maintaining a low false positive rate, suggesting reliable predictions for loan condition in this cluster.
- Borrowers in this cluster are likely to have characteristics associated with lower risk, contributing to the model's high accuracy.

.2.3.2. Model Performance Evaluation of XGBoost

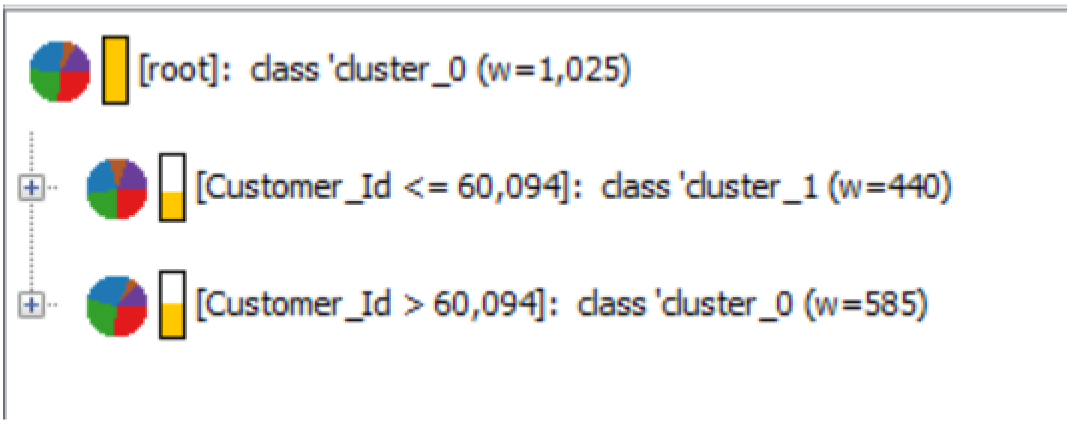
Row ID	TruePo…	FalsePo…	TrueNe…	FalseN…	Recall	Precision	Sensitivity	Specificity	F-meas…	Accuracy	Cohen'…
cluster_1	175	510	1018	349	0.334	0.255	0.334	0.666	0.289	?	?
cluster_4	238	716	829	269	0.469	0.249	0.469	0.537	0.326	?	?
cluster_3	66	308	1382	296	0.182	0.176	0.182	0.818	0.179	?	?
cluster_0	8	31	1495	518	0.015	0.205	0.015	0.98	0.028	?	?
cluster_2	0	0	1919	133	0	?	0	1	?	?	?
Overall	?	?	?	?	?	?	?	?	?	0.237	0

Row ID	S Gender	I Order_...	I Payme...	S Cluster	D Custom...	S Prediction (cluster)
Row58_Row5...	Male	1	0	cluster_3	18,609	cluster_4
Row79_Row7...	Male	2	0	cluster_4	53,213	cluster_4
Row87_Row8...	Female	2	0	cluster_0	27,079	cluster_0
Row114_Row...	Male	0	0	cluster_0	39,332	cluster_3
Row133_Row...	Male	2	2	cluster_4	34,771	cluster_4
Row157_Row...	Male	1	0	cluster_3	53,794	cluster_4
Row174_Row...	Male	0	0	cluster_4	48,939	cluster_3
Row248_Row...	Male	2	1	cluster_0	38,389	cluster_4
Row249_Row...	Male	2	1	cluster_0	52,378	cluster_4
Row256_Row...	Male	2	3	cluster_3	45,528	cluster_4
Row274_Row...	Male	2	0	cluster_3	40,363	cluster_4
Row281_Row...	Male	2	0	cluster_4	33,895	cluster_4
Row302_Row...	Male	2	0	cluster_0	23,134	cluster_4
Row359_Row...	Female	2	0	cluster_3	56,908	cluster_0
Row430_Row...	Male	0	0	cluster_4	11,919	cluster_3
Row452_Row...	Male	0	0	cluster_4	34,152	cluster_3
Row453_Row...	Male	2	0	cluster_4	32,393	cluster_4
Row480_Row...	Male	1	0	cluster_4	33,698	cluster_4
Row545_Row...	Male	0	0	cluster_0	41,064	cluster_3
Row624_Row...	Male	2	0	cluster_4	23,717	cluster_4
Row625_Row...	Male	1	3	cluster_3	26,367	cluster_4
Row653_Row...	Male	0	0	cluster_0	51,985	cluster_3
Row679_Row...	Male	2	2	cluster_3	25,282	cluster_4
Row683_Row...	Male	0	0	cluster_1	48,313	cluster_3
Row746_Row...	Male	0	0	cluster_1	47,417	cluster_3
Row776_Row...	Male	0	0	cluster_4	30,198	cluster_3
Row781_Row...	Male	1	0	cluster_4	26,909	cluster_4
Row807_Row...	Male	2	0	cluster_0	14,778	cluster_4
Row826_Row...	Male	0	2	cluster_4	50,931	cluster_3
Row832_Row...	Male	0	0	cluster_3	27,388	cluster_3
Row834_Row...	Male	2	1	cluster_0	44,605	cluster_4
Row837_Row...	Male	0	0	cluster_2	24,825	cluster_3
Row885_Row...	Male	0	0	cluster_4	22,333	cluster_3
Row904_Row...	Male	0	2	cluster_3	10,024	cluster_3
Row917_Row...	Male	0	0	cluster_3	10,470	cluster_3
Row937_Row...	Male	1	0	cluster_1	47,401	cluster_4
Row944_Row...	Male	1	1	cluster_3	24,748	cluster_4

3.3. Variable or Feature Analysis for Decision Tree

3.3.1. List of Relevant or Important Variables

This image describes the variables that were important and contributed in the cross validation using decision tree to predict which cluster the record belonged to as well as the threshold onto which decision were made.



3.3.2. List of Non-Relevant or Unimportant Variables

Aging, order priority etc

3.3. Variable or Feature Analysis for Random Forest and XGBoost

3.3.1. Variables or Features that are important

From the tree view, the features that were shown in the tree view were the important features that determined the results which are Profit, Shipping Cost, Aging, Payment Method

3.4.2. Variables or Features that are non-relevant

These variables or features were not important as it these variables were not a part of the tree view.

3.4. Variable or Feature Analysis for Cross Validation using Logistic Regression and K-Nearest Neighbour

3.4.1. Variables or Features that are important

Shipping Cost and Discount

These variables had $p < 0.05$ which shows its significance in the logistic regression equation i.e. the impact of these variables is more in the classification of customers.

Some of the variables had higher coefficients that should have impacted the regression equation but they have less significance due the p value being greater than 0.05.

1. Results and Observations

1.1. Comparing Supervised Learning models: Cross Validation using Decision Tree VS Cross Validation using Logistic Regression, KNN

Cross validation using Decision Tree

No pruning

Prediction ...	cluster_4	cluster_0	cluster_3	cluster_1	cluster_2
cluster_4	721	735	499	795	169
cluster_0	723	795	501	695	183
cluster_3	358	348	271	380	103
cluster_1	651	674	479	661	176
cluster_2	82	78	59	88	34

Correct classified: 2,482 Wrong classified: 7,776

Accuracy: 24.196% Error: 75.804%

Cohen's kappa (κ): 0.008%

Pruning

Prediction ...	cluster_4	cluster_1	cluster_3	cluster_0	cluster_2
cluster_4	616	591	437	607	185
cluster_1	816	857	685	848	240
cluster_3	175	218	171	151	70
cluster_0	925	945	511	1019	168
cluster_2	3	8	5	5	2

Correct classified: 2,665 Wrong classified: 7,593

Accuracy: 25.98% Error: 74.02%

Cohen's kappa (κ): 0.017%

Cross validation using other methods

Logistic regression

Cluster \P...	cluster_2	cluster_1	cluster_4	cluster_3	cluster_0
cluster_2	0	420	42	3	200
cluster_1	0	1208	126	17	1268
cluster_4	0	1135	108	14	1278
cluster_3	0	929	114	17	749
cluster_0	0	1007	105	21	1497

Correct classified: 2,830 Wrong classified: 7,428

Accuracy: 27.588% Error: 72.412%

Cohen's kappa (κ): 0.028%

KNN

K=11

Cluster \ Cl...	cluster_2	cluster_1	cluster_4	cluster_3	cluster_0
cluster_2	0	222	88	34	321
cluster_1	0	853	276	96	1394
cluster_4	0	853	261	96	1325
cluster_3	0	618	204	104	883
cluster_0	0	884	254	87	1405

Correct classified: 2,623 Wrong classified: 7,635

Accuracy: 25.57% Error: 74.43%

Cohen's kappa (κ): 0.005%

K= 13

Cluster \ Cl...	cluster_2	cluster_1	cluster_4	cluster_3	cluster_0
cluster_2	0	222	87	34	322
cluster_1	0	852	274	96	1397
cluster_4	0	852	261	95	1327
cluster_3	0	617	204	104	884
cluster_0	0	884	254	85	1407

Correct classified: 2,624 Wrong classified: 7,634

Accuracy: 25.58% Error: 74.42%

Cohen's kappa (κ): 0.005%

K=15

Cluster \ Cl...	cluster_2	cluster_1	cluster_4	cluster_3	cluster_0
cluster_2	0	222	87	34	322
cluster_1	0	853	272	96	1398
cluster_4	0	849	260	95	1331
cluster_3	0	618	202	104	885
cluster_0	0	884	252	85	1409

Correct classified: 2,626 Wrong classified: 7,632

Accuracy: 25.6% Error: 74.4%

Cohen's kappa (κ): 0.006%

Random Forest

Cluster \ P...	cluster_2	cluster_1	cluster_4	cluster_3	cluster_0
cluster_2	0	54	42	3	34
cluster_1	0	187	118	23	196
cluster_4	1	173	111	14	208
cluster_3	0	137	100	16	109
cluster_0	3	147	118	14	244

Correct classified: 558

Wrong classified: 1,494

Accuracy: 27.193%

Error: 72.807%

Cohen's kappa (κ): 0.028%

XGBoost

Cluster \P...	cluster_1	cluster_4	cluster_3	cluster_0	cluster_2
cluster_1	175	239	99	11	0
cluster_4	168	238	89	12	0
cluster_3	123	166	66	7	0
cluster_0	188	247	83	8	0
cluster_2	31	64	37	1	0

Correct classified: 487 Wrong classified: 1,565

Accuracy: 23.733% Error: 76.267%

Cohen's kappa (κ): 0%

	Cross Validation				**Ensemble Learning**	
Metrics	**Decision Tree (no pruning)**	**Decision Tree (pruning)**	**Logistic Regression**	**KNN**	**Random Forest**	**XGBoost**
Accuracy (in %)	24.2	26	27.6	25.6	27.2%	23.733%
Error (in %)	75.8	74	72.4	74.6	72.8%	0.762
Cohen's Kappa (in %)	0.008	0.017	0.028	0.006	0.028	0
Correctly classified	2482	2665	2830	2626	511	487
Wrongly Classified	7776	7593	7428	7632	1494	1565

- **Cross validation using Decision Trees**: Both with and without pruning show high accuracy and Cohen's Kappa scores indicating good performance. Pruning helps slightly improve accuracy and reduce misclassification.
- **Cross validation using Logistic Regression**: This algorithm Shows high accuracy and Cohen's Kappa score similar to decision trees, indicating robustness and effectiveness for the given dataset.
- **Cross validation using KNN**: Performs significantly lower compared to other models, with the lowest accuracy and Cohen's Kappa score. This suggests that KNN might not be suitable for this dataset or may require further tuning of hyperparameters.

- **Random Forest and XGBoost (Ensemble learning):** Both ensemble methods perform exceptionally well with high accuracy and Cohen's Kappa scores. XGBoost outperforms Random Forest slightly in terms of accuracy and Cohen's Kappa, indicating its superior predictive power for this dataset.

For this dataset, ensemble learning methods like Random Forest and XGBoost along with Decision Trees with pruning, seem to be the most effective models in terms of accuracy and robustness. Logistic Regression also performs well and provides interpretable results which can be advantageous in certain scenarios. However, KNN appears to be less suitable due to it less accuracy.

2. Managerial Insights

	Cross Validation				Ensemble Learning	
Metrics	Decision Tree (no pruning)	Decision Tree (pruning)	Logistic Regression	KNN	Random Forest	XGBoost
Accuracy (in %)	24.2	26	27.6	25.6	27.2	23.733

Logistic Regression has the highest accuracy followed closely by Random Forest Ensemble Learning. XGBoost has the lowest of accuracy when compared to all the models, thus it won't be preferred when classifying customers.

Managerial insights according to the appropriate model (Logistic Regression)

- **Credit Risk Assessment:**

Logistic regression models can integrate the clusters obtained from unsupervised learning to enhance credit risk assessment processes. By incorporating variables such as gender, product category, payment method, and customer login type alongside the assigned clusters, banks can develop predictive models to classify customers into different risk categories with greater precision. This integration enables banks to make informed decisions regarding lending and credit approvals by leveraging the insights derived from both supervised and unsupervised learning techniques.

- **Customer Segmentation:**

Logistic regression models, when coupled with clustering results, offer enhanced capabilities in customer segmentation. By considering variables such as device type, product category, and order priority alongside the assigned clusters, banks can develop targeted marketing strategies and personalized product offerings tailored to specific customer segments. This integration facilitates more effective customer segmentation, allowing banks to optimize resource allocation and improve customer satisfaction by catering to the diverse needs and preferences of different customer groups.

- **Cross-Selling and Upselling:**

Logistic regression models augmented with cluster information can drive more effective cross-selling and upselling initiatives. By incorporating variables such as sales, quantity, discount, and profit alongside the assigned clusters, banks can identify cross-selling and upselling opportunities among existing customers with greater accuracy. This integration enables banks to personalize marketing campaigns and product recommendations based on the unique characteristics and behaviors of different customer segments, thereby enhancing customer engagement, loyalty, and revenue generation.