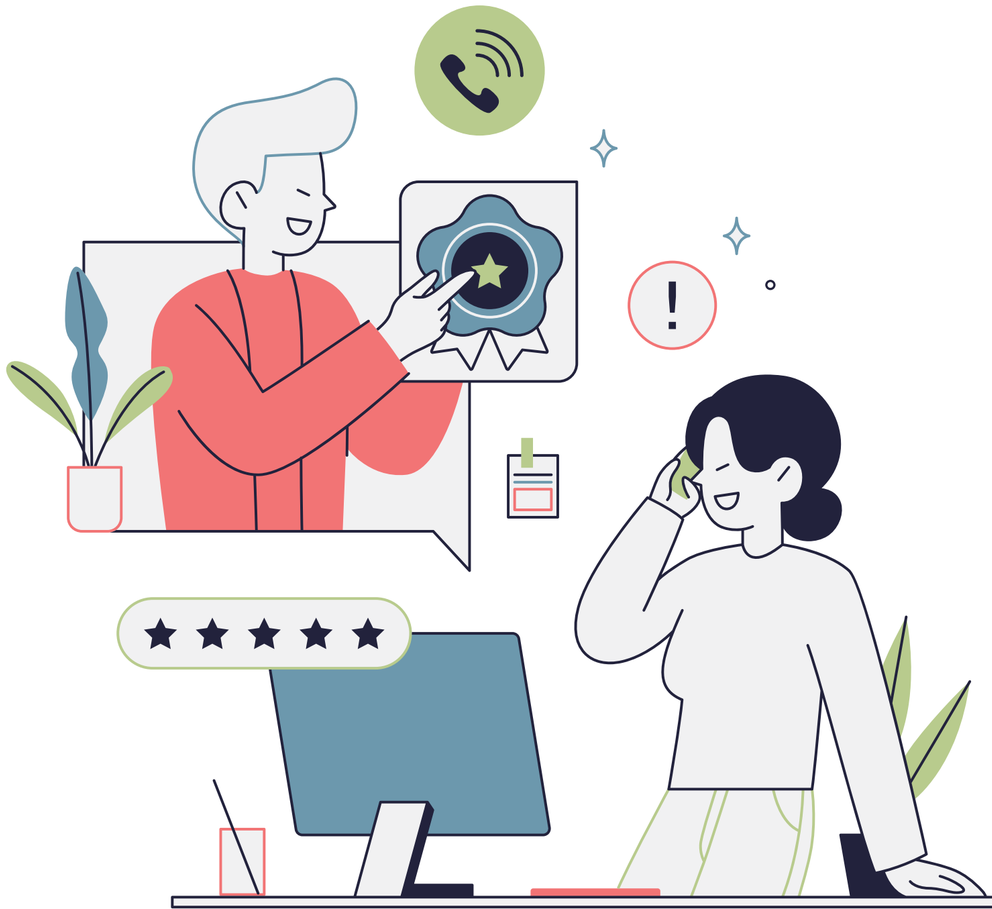
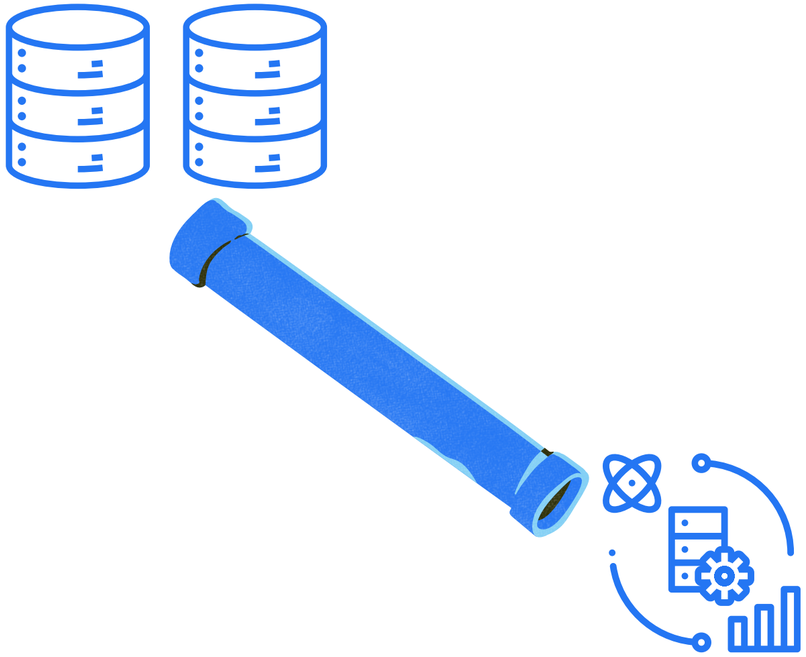
Shop Customer Clustering & Prediction



A deep understanding of customers is important for marketing strategies, business decision-making, and improved customer retention.

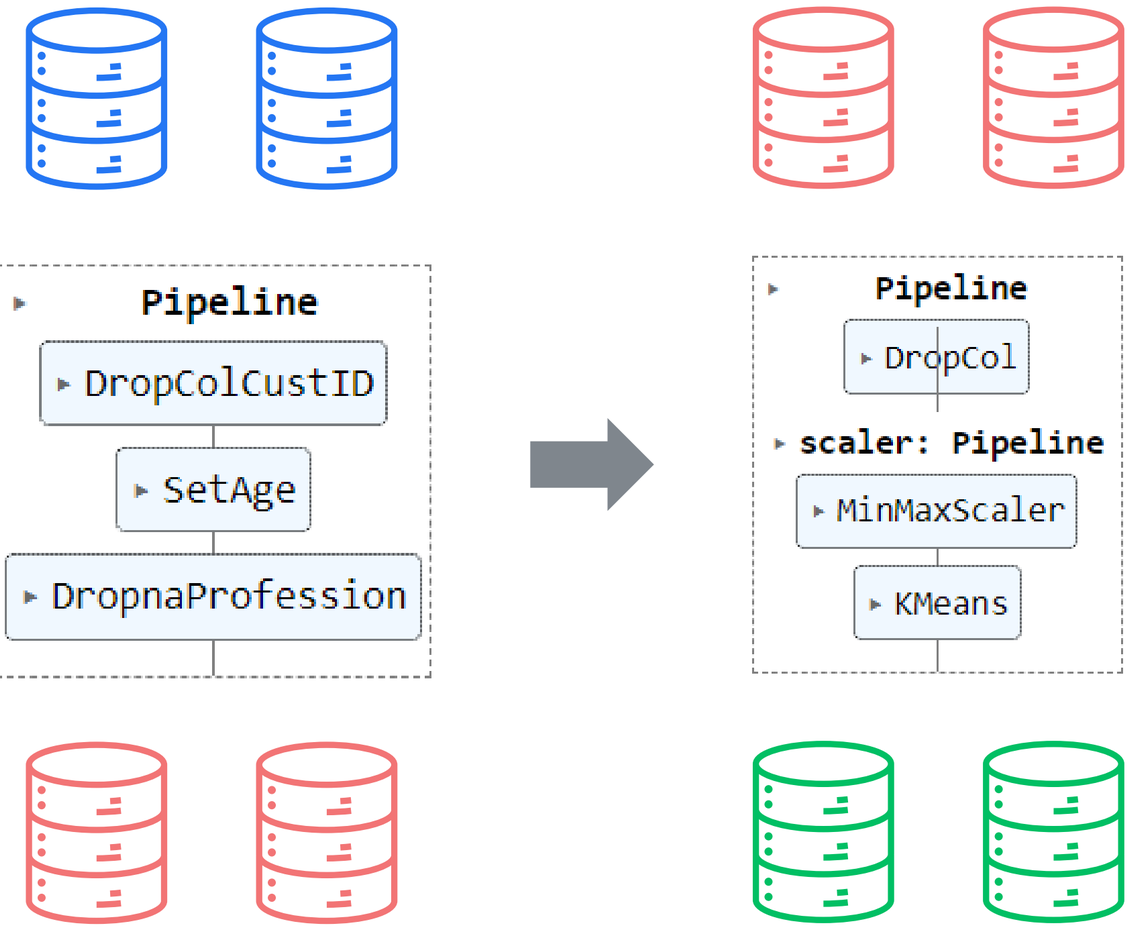


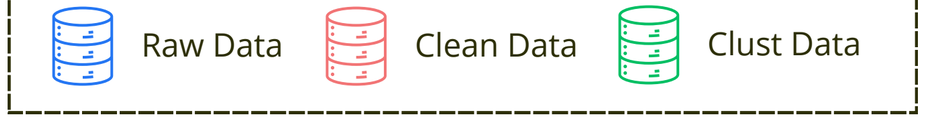
This project will try to cluster customer data and then predict customer clusters using K-Means + Logistic Regression using sklearn Pipeline

Data Snap!

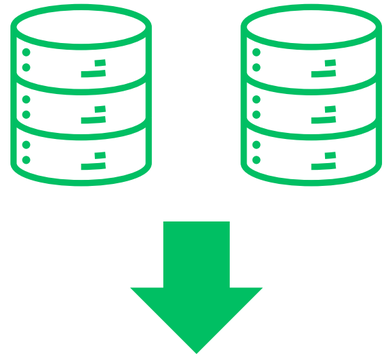
	CustomerID	Gender	Age	Annual Income ($)	Spending Score (1-100)	Profession	Work Experience	Family Size
0	1	Male	19	15000	39	Healthcare	1	4
1	2	Male	21	35000	81	Engineer	3	3
2	3	Female	20	86000	6	Engineer	1	1
3	4	Female	23	59000	77	Lawyer	0	2
4	5	Female	31	38000	40	Entertainment	2	6
...	...	...	...	...	...	...	...	...
1995	1996	Female	71	184387	40	Artist	8	7
1996	1997	Female	91	73158	32	Doctor	7	7
1997	1998	Male	87	90961	14	Healthcare	9	2
1998	1999	Male	77	182109	4	Executive	7	2
1999	2000	Male	90	110610	52	Entertainment	5	2

Cleaning & Clustering Pipe



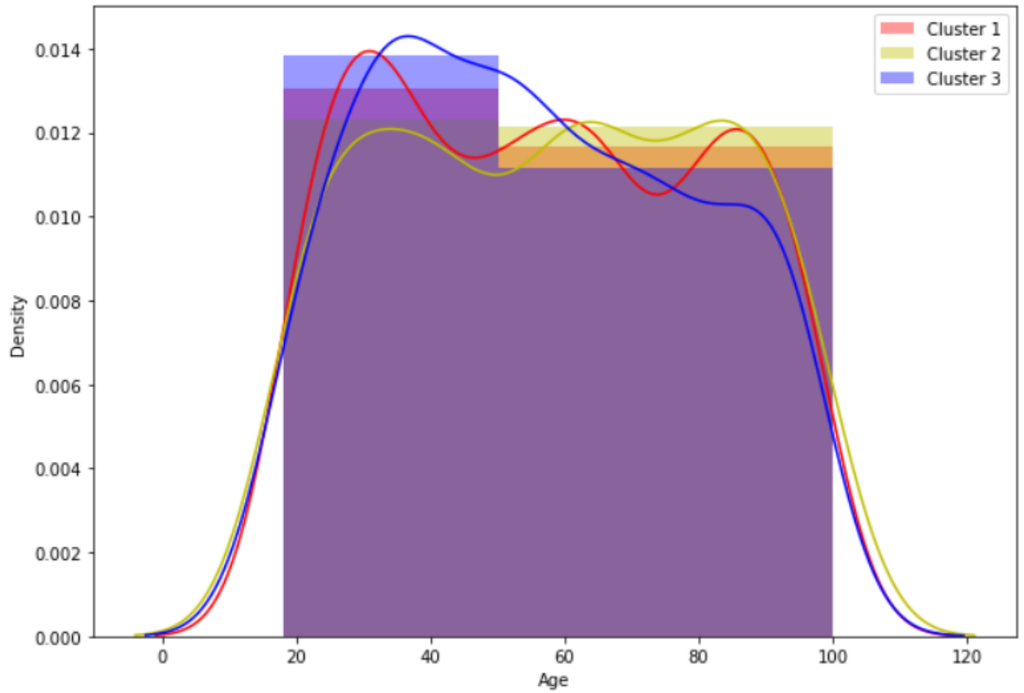


Clustering Result



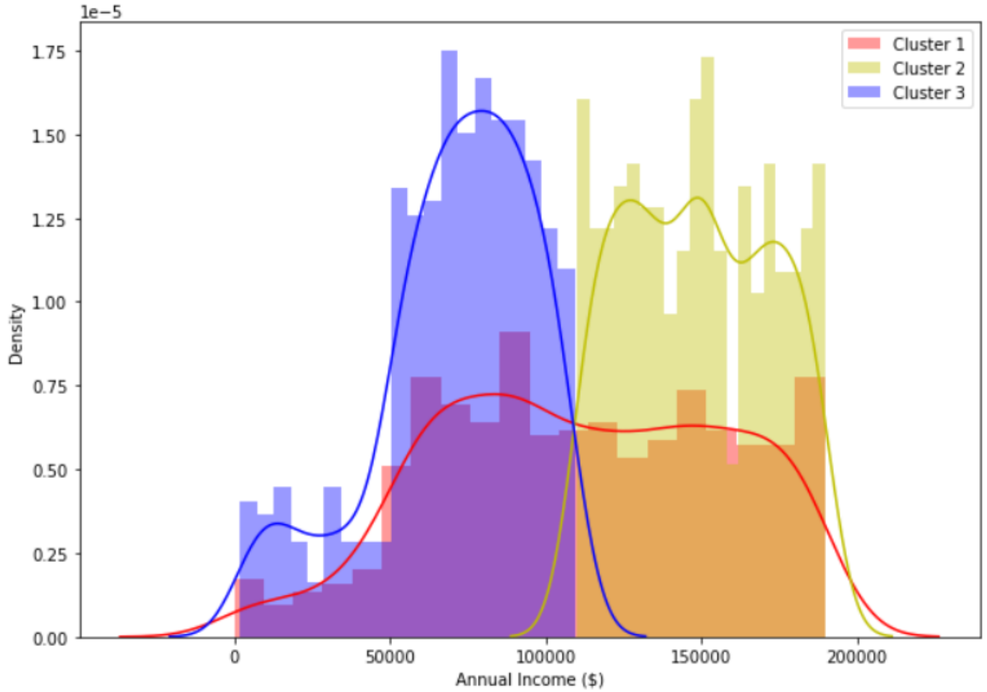
Cluster 1	**789**
Cluster 2	**458**
Cluster 3	**391**

Age vs Cluster



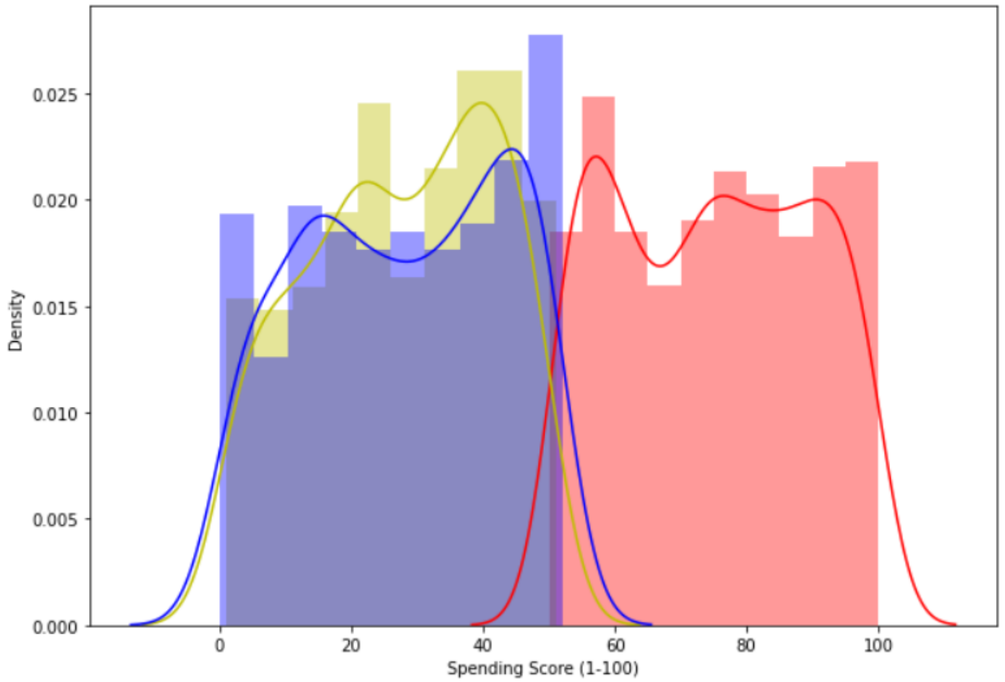
Cluster distribution based on age appears even across all ages. There is no specific age range that dominates a particular cluster.

Income vs Cluster



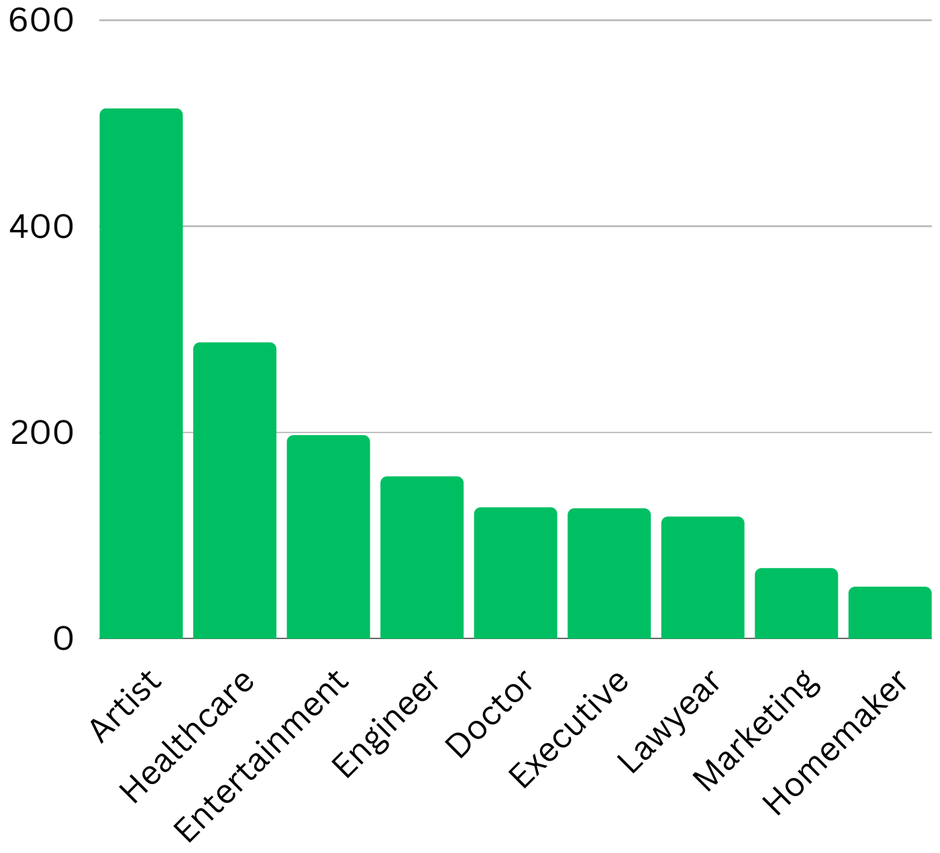
Annual income range 50000-100000 mostly falls into cluster 3, annual income range 100000-200000 mostly falls into cluster 2, while some individuals with annual income range 50000-170000 are placed in cluster 1.

Spending Score vs Cluster

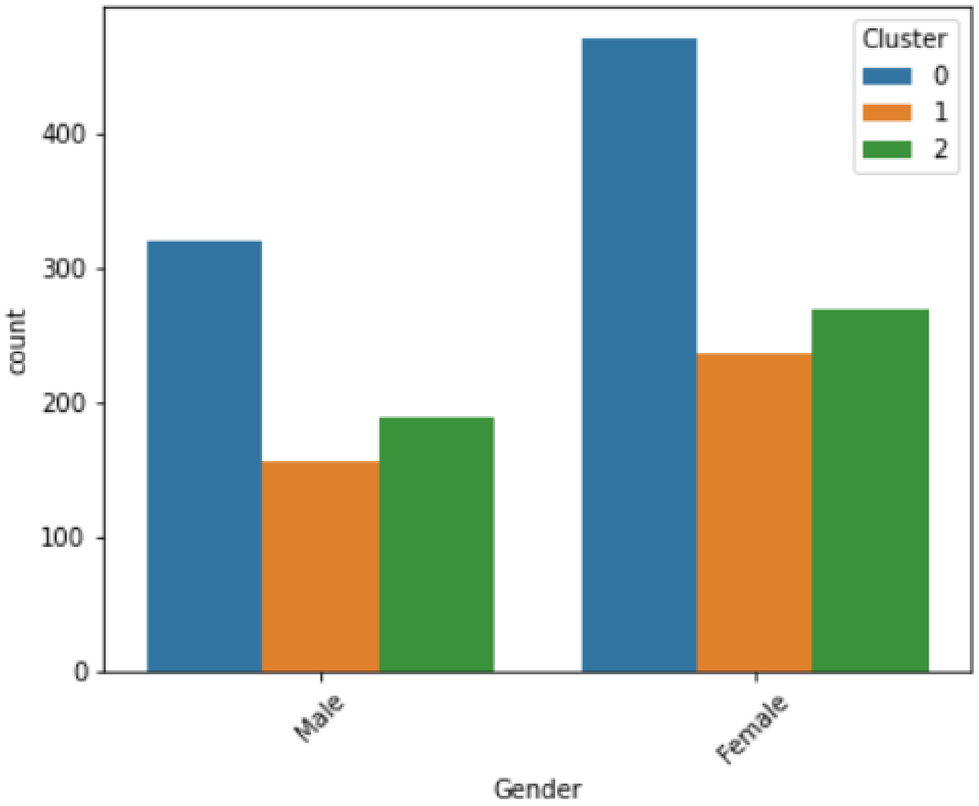


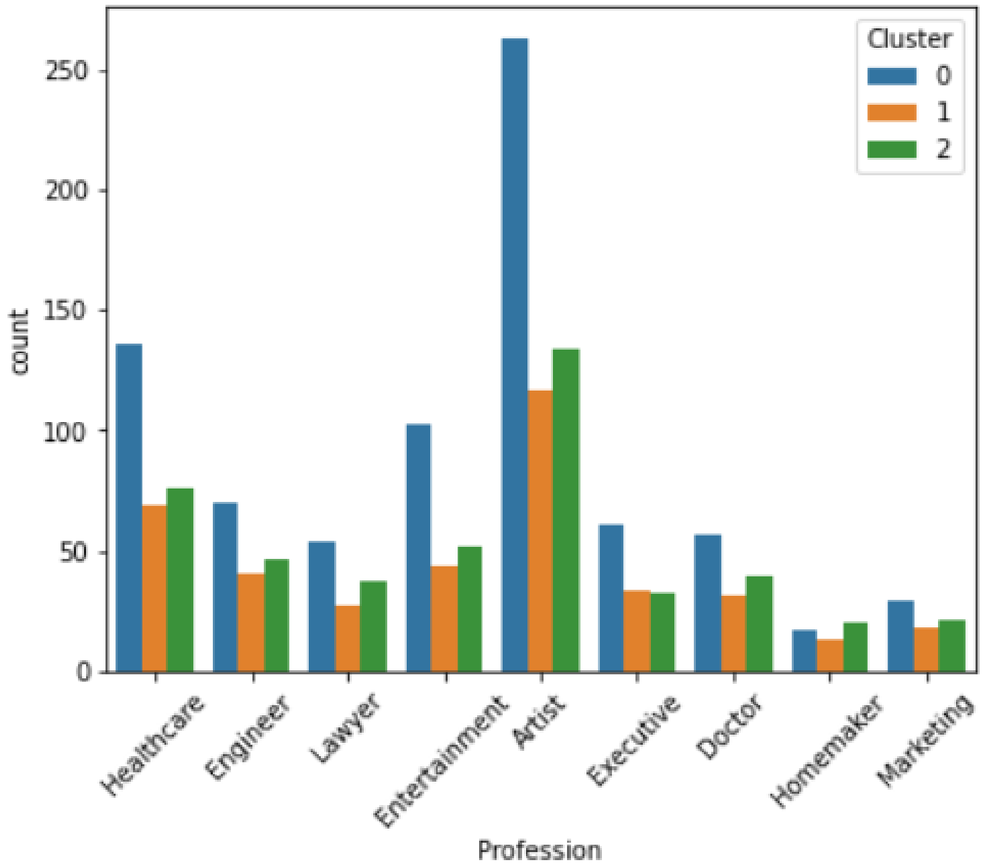
Spending score >50 tends to be present in cluster 1, while spending score <50 is evenly distributed between cluster 2 and cluster 3.

Profession Distribution



Categorical vs Cluster





Both females and males are more prevalent in cluster 1, and there is no significant correlation observed with the profession plot.

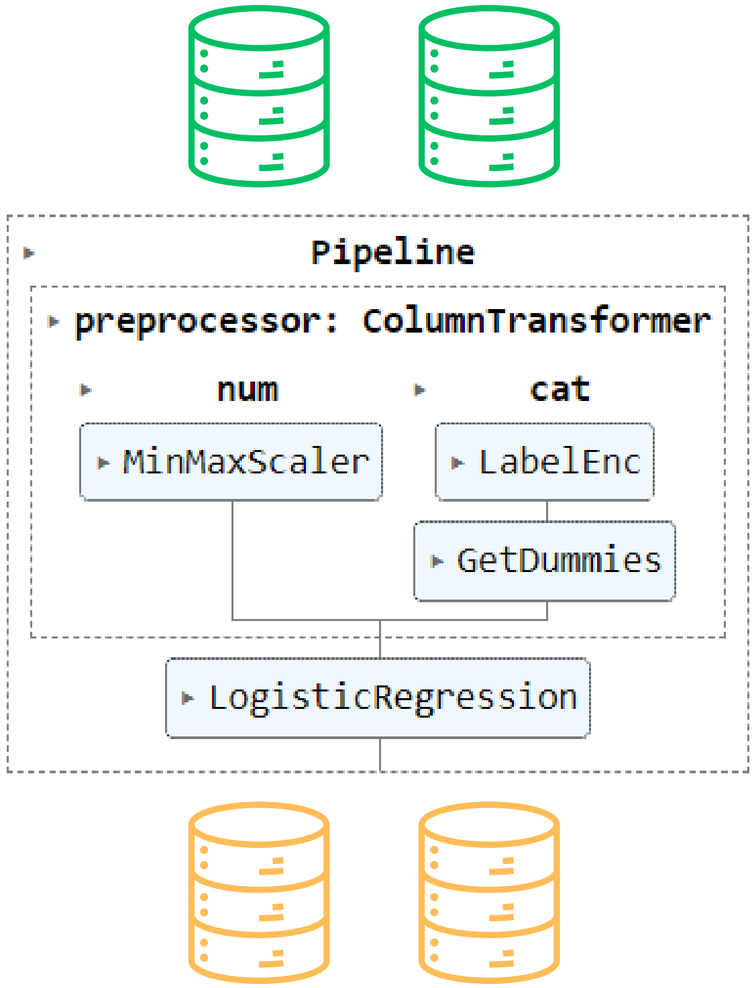
Split Data

X = ['Gender', 'Age', 'Annual Income ($)', 'Spending Score (1-100)', 'Profession', 'Work Experience', 'Family Size']
y = 'Cluster'

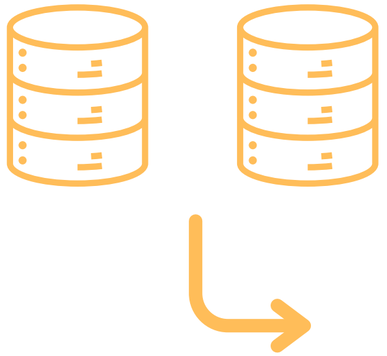
Split Data 70 : 30, stratify = y, random_state = 42

X_train	(1146, 7)
y_train	(1146,)
X_test	(492, 7)
y_test	(492,)

Prediction Pipe



Clustering Result



Accuracy = 0.9756

Confussion Matrix

		Predict		
Actual		250	5	2
		1	116	0
		3	1	134

Result

Sklearn Pipeline is very helpful in automating and simplifying the machine learning flow. It prevents data leakage, is efficient, and can be used for cross validation or hyperparameter tuning.

The prediction results using the logistic regression pipeline produce an accuracy value of 0.9756, indicating that the model can predict customer groups based on the given features quite accurately.

Confussion Matrix

	Predict		
Actual	230	5	2
	1	116	0
	3	1	134

- **Cluster 1** has 230 correct predictions, 5 incorrect predictions as Class 1, and 2 incorrect predictions as Class 2.
- **Cluster 2** has 116 correct predictions, 1 incorrect prediction as Class 0, and no incorrect predictions as Class 2.
- **Cluster 3** has 134 correct predictions, 1 incorrect prediction as Class 0, and 3 incorrect predictions as Class 1.

Thank You!

More Info :

Dataset : https://www.kaggle.com/datasets/datascientistanna/customers-dataset

Code Project : https://github.com/dewaadji/shop-cust-clustering-prediction/


https://www.linkedin.com/in/dewa-adji/


dewaadji12@gmail.com