

Capstone Project-4

Customer Segmentation

By-Prabir Debnath

Points of Discussion

1. The Problem Statement
2. Concept of ML Model
3. Summary of the Experience Set
4. Exploration and Pre-processing of Data
5. Building of Model-1
6. Building of Model-2
7. Building of Model-3
8. Final Conclusion

The Problem Statement

We are provided with an unlabeled dataset on the transaction details of online retail customers. All the transactions are occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers. The task is to explore and analyze the data and to build a clustering model for customer segmentation.

Concept of ML Model

The performance of a machine learning model depends on three factors:

i. Quality of Data

(cleaner experiences for better learning)

ii. Quantity of Data

(more experiences for better learning)

iii. Quality of Model

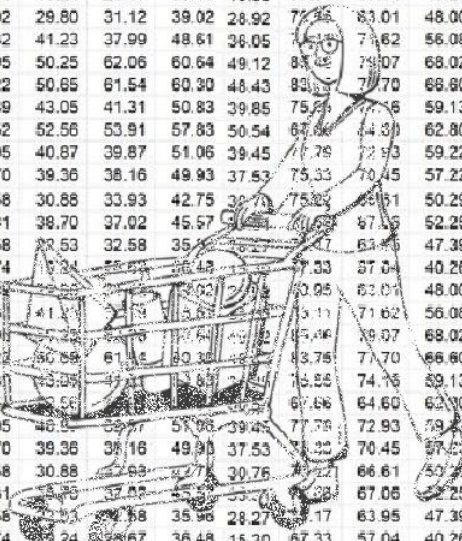
(proper model and right hyperparameters for better learning)

Summary of the Experience Set

Here, the dataset has 5,41,909 rows, which means 5,41,909 experiences about transaction of online retail customers and

It has 8 columns, which means each experience is observed along 8 features or dimensions.

Let's Decode the Experiences !



49.61	21.58	30.88	33.93	42.75	30.76	75.22	66.61	50.29	36.73	30.76
52.25	28.31	38.70	37.02	45.57	35.31	76.66	67.06	52.25	41.20	35.31
45.73	19.58	22.53	32.58	35.96	28.27	71.17	63.95	47.39	30.16	28.27
40.51	12.74	19.24	28.67	36.48	15.30	67.33	57.04	40.26	27.52	15.30
46.63	19.02	29.80	31.12	39.02	28.92	70.95	63.01	48.00	34.47	28.92
53.76	27.32	41.23	37.99	48.61	36.05	73.11	71.62	56.08	40.62	36.05
66.67	41.95	50.25	62.06	60.64	49.12	85.46	79.07	68.02	52.81	49.12
65.41	42.22	50.65	61.54	60.30	48.43	83.75	77.70	66.60	51.60	48.43
56.33	29.39	43.05	41.31	50.83	39.85	75.65	74.16	59.13	42.93	39.85
59.66	52.52	52.56	53.91	57.83	50.54	67.66	64.60	62.80	56.05	50.54
56.27	31.05	40.87	39.87	51.06	39.45	77.79	72.93	59.22	43.56	39.45
54.32	29.70	39.36	38.16	49.93	37.53	75.33	70.45	57.22	41.40	37.53
49.61	21.58	30.88	33.93	42.75	30.76	75.22	66.61	50.29	36.73	30.76
52.25	28.31	38.70	37.02	45.57	35.31	76.66	67.06	52.25	41.20	35.31
45.73	19.58	22.53	32.58	35.96	28.27	71.17	63.95	47.39	30.16	28.27
40.51	12.74	19.24	28.67	36.48	15.30	67.33	57.04	40.26	27.52	15.30
46.63	19.02	29.80	31.12	39.02	28.92	70.95	63.01	48.00	34.47	28.92
53.76	27.32	41.23	37.99	48.61	36.05	73.11	71.62	56.08	40.62	36.05
66.67	41.95	50.25	62.06	60.64	49.12	85.46	79.07	68.02	52.81	49.12
65.41	42.22	50.65	61.54	60.30	48.43	83.75	77.70	66.60	51.60	48.43
56.33	29.39	43.05	41.31	50.83	39.85	75.65	74.16	59.13	42.93	39.85
59.66	52.52	52.56	53.91	57.83	50.54	67.66	64.60	62.80	56.05	50.54
56.27	31.05	40.87	39.87	51.06	39.45	77.79	72.93	59.22	43.56	39.45
54.32	29.70	39.36	38.16	49.93	37.53	75.33	70.45	57.22	41.40	37.53
49.61	21.58	30.88	33.93	42.75	30.76	75.22	66.61	50.29	36.73	30.76
52.25	28.31	38.70	37.02	45.57	35.31	76.66	67.06	52.25	41.20	35.31
45.73	19.58	22.53	32.58	35.96	28.27	71.17	63.95	47.39	30.16	28.27
40.51	12.74	19.24	28.67	36.48	15.30	67.33	57.04	40.26	27.52	15.30
46.63	19.02	29.80	31.12	39.02	28.92	70.95	63.01	48.00	34.47	28.92
53.76	27.32	41.23	37.99	48.61	36.05	73.11	71.62	56.08	40.62	36.05
66.67	41.95	50.25	62.06	60.64	49.12	85.46	79.07	68.02	52.81	49.12
65.41	42.22	50.65	61.54	60.30	48.43	83.75	77.70	66.60	51.60	48.43
56.33	29.39	43.05	41.31	50.83	39.85	75.65	74.16	59.13	42.93	39.85
59.66	52.52	52.56	53.91	57.83	50.54	67.66	64.60	62.80	56.05	50.54
56.27	31.05	40.87	39.87	51.06	39.45	77.79	72.93	59.22	43.56	39.45

Exploration and Pre-processing of Data

We have done the exploration and pre-processing in seven steps to transform raw data into quality data for our ml model.

1. Connection with the Data
2. First Feelings of the Data
3. Deeper Understanding of the Data
4. Cleaning the Data
5. Treating Anomalies in the Data
6. Final Feature Selection from the Data
7. Preparation of Input Data

Cleaning the Data

- > We have removed 5,268 duplicate experiences from our dataset
- > We have dropped 1,35,037 null values in 'CustomerID' column as we cannot impute them
- > There were negative values in 'Quantity' column. We have removed the experiences with negative values in 'Quantity'
- > There were zero values in 'UnitPrice' column. We have removed the experiences with zero values in 'UnitPrice'
- > We have created 'Mean_UnitPrice' and 'Sum_Quantity' feature for each 'CustomerID' because from the average unit price, we can understand what they buy and from the sum of quantity , we can understand how much they buy.

Expected Segments of Customers

The segments of customers may be

Target-1: Buying costly product high quantity

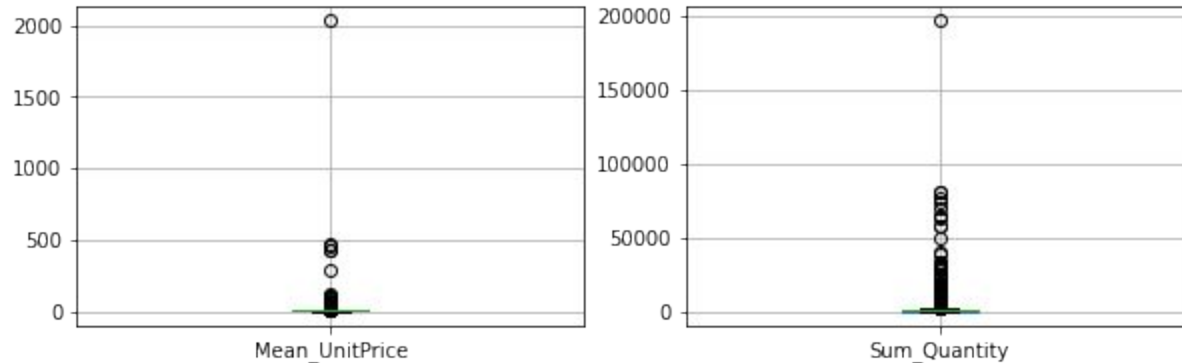
Target-2: Buying less costly product high quantity

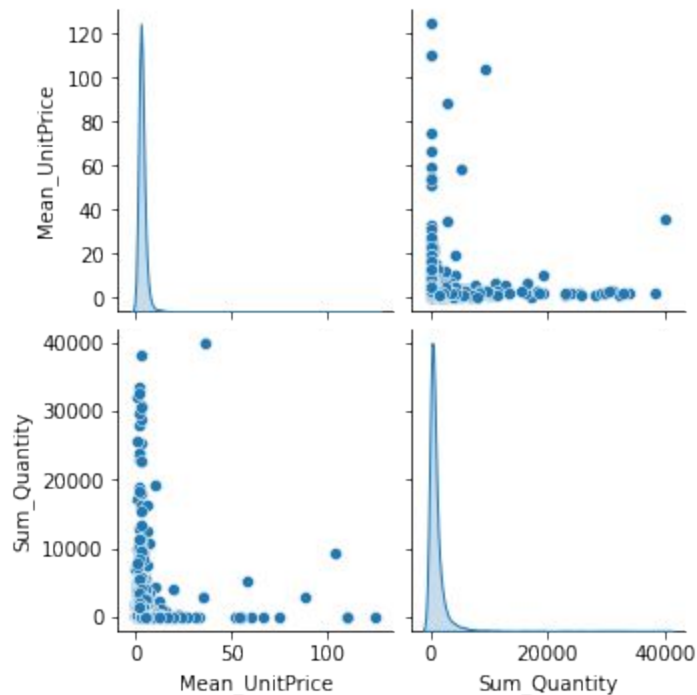
Target-3: Buying costly products less quantity

Target-4: Buying less costly product less quantity

Treating Anomalies in the Data

There are outliers for $\text{Mean_UnitPrice} > 250$ and $\text{Quantity} > 10000$ and thus we have removed these experiences from our dataset





Overall Feature Understanding

Here, the distribution of 'Mean_UnitPrice' and 'Sum_Quantity' are positively skewed. Thus we have done log transformation on these features to normalize their distribution.



Looking for Truly Independent Features

Here, all our input variables are truly independent as all the VIF values are below 10

	variables	VIF
0	Mean_UnitPrice	1.058642
1	Sum_Quantity	1.058642

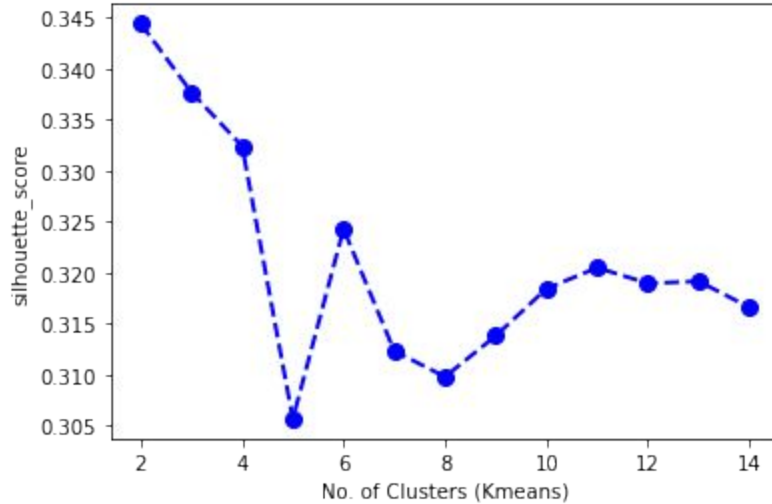
Let's Train the Models !



Building of Model-1

Final KMeans Model:

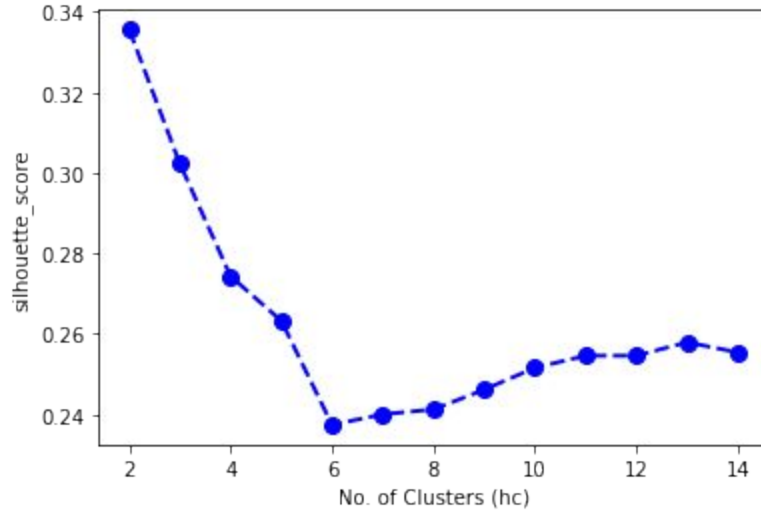
The best no. of clusters= 3
according to silhouette score



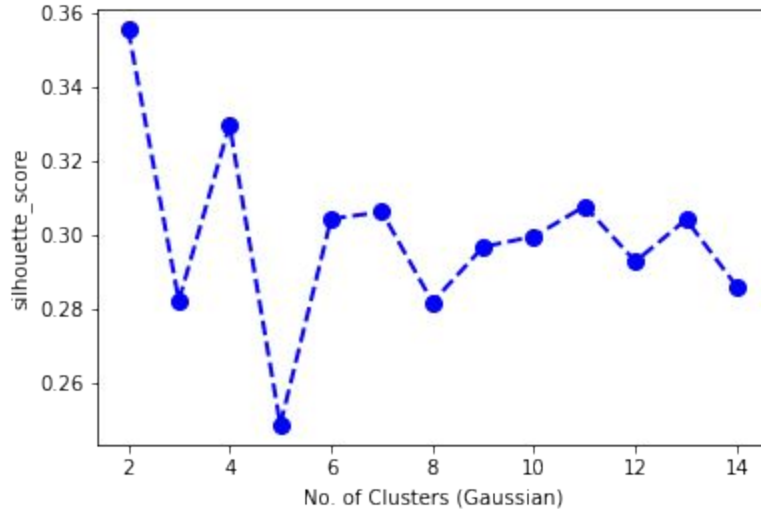
Building of Model-2

Final AgglomerativeClustering Model:

The best no. of clusters= 3 according to silhouette score

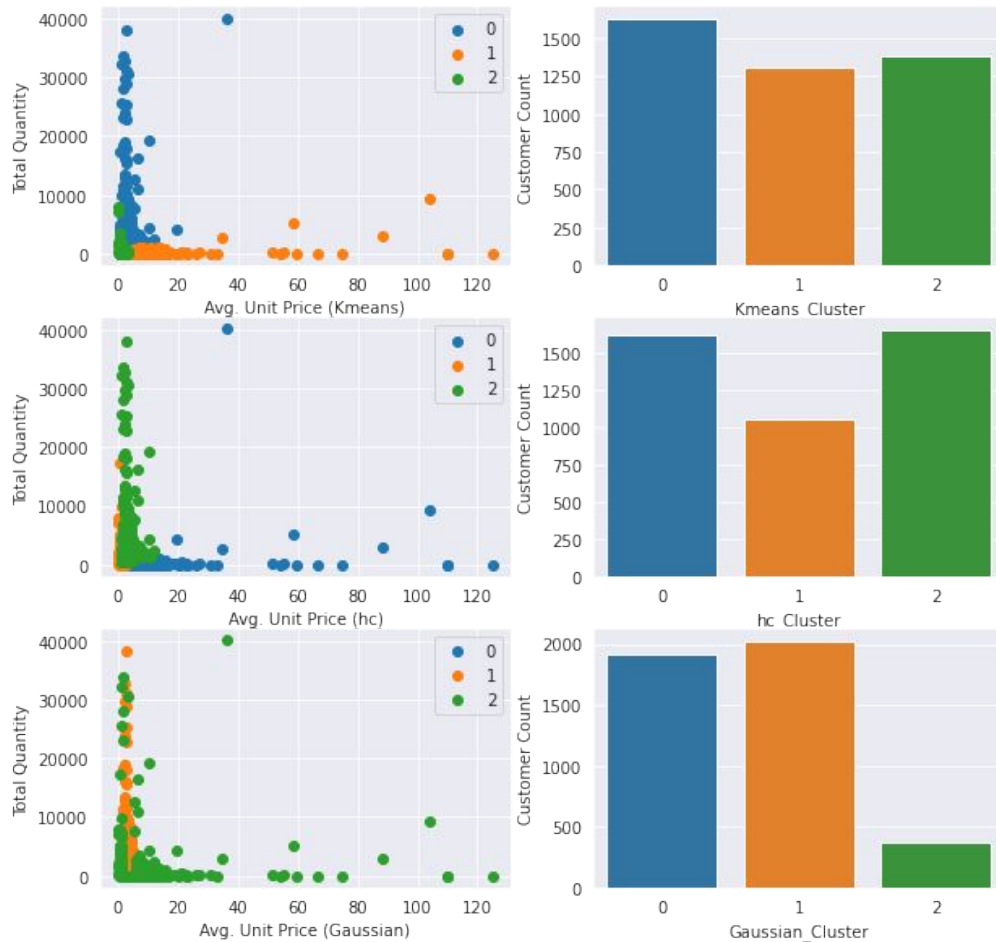


Building of Model-3



Final GaussianMixture Model:

The best no. of clusters= 4 according to silhouette score, but we are selecting 3 for comparison with other models



Final Conclusion

On the basis of the performance study of our three models, we are selecting **KMeans** model (*the best warrior*) for online retail customer segmentation, as it best fits our expected customer segmentation with minimum overlap among all three models



Thank you !