Customer Segmentation using Machine Learning

CS-725 Foundation of Machine Learning
Project Report

by

Gourav Tilwankar - 203070051
B Vasant - 203070085
Samay Pritam Singh - 20307R002

under the guidance of

Prof. Dr. Preethi Jyothi,
Dept. of CSE, IIT Bombay

Indian Institute of Technology Bombay
Powai, Mumbai, 400076

November 27, 2021

Declaration

I declare that this written submission represents my ideas in my own words and where other ideas or words or diagrams have been included from books/papers/electronic media, I have adequately cited and referenced the original sources.

I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will result in disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken.

Gourav Tilwankar - 203070051
B Vasant - 203070085
Samay Pritam Singh - 20307R002

Acknowledgement

I'd like to thank my guide, **Dr. Preethi Jyothi**, for giving me the opportunity to work on the topic of "Customer Segmentation using Machine Learning." I'd want to express my appreciation to all of the members of the CS725 Course members for their helpful ideas and kind advice during project duration.

Gourav Tilwankar, B Vasant, Samay Pritam Singh
Electrical Engineering Department
IIT Bombay

Abstract

Many e-commerce giants give bonuses to consumers just to make an account with them, as the making of this account enables them to understand the customer's purchasing pattern to derive the maximum profits from them. The goal of customer segmentation is to identify the most profitable customer segment, and improve marketing by personalizing to that niche. Customer segmentation helps understand the consumer base, and helps anticipate the needs of consumers. K-means clustering is one such machine learning algorithm that is widely used in customer segmentation. This project includes exploration of K-Means Clustering Algorithm on a dataset having numerical values features and clustering/grouping them according to similar attributes.

Contents

List of Figures

1 Introduction

This project uses a dataset which contains customer shopping data from a mall, which is pre-processed to merge the data of each customer from their multiple visits to the mall. This data is then processed again to remove null data, and is used to run a K-means clustering algorithm to find and visualize the most profitable customer segment based on it's main features.

2 Dataset

Dataset : https://github.com/cereniyim/Customer-Segmentation-Unsupervised-ML-Model

This dataset contains customers data in which the features are numeric, alphabetical as well as alphanumeric. It contains data points of 70000+ non-unique customers for 16 features. Although, we have used a pre-processed data available at same github repository in which the data points are as follows:

- 24874 customers (all unique)
- 4 features namely **Products Ordered, Average return Rate, Total Spending and Gender**
- Total Dataset: (24874 x 6)

3 Exploratory Data Analysis (EDA)

We visualize how the number of customers (density) varies with the three main features : Number of products ordered, Total spending, and Average return rate.

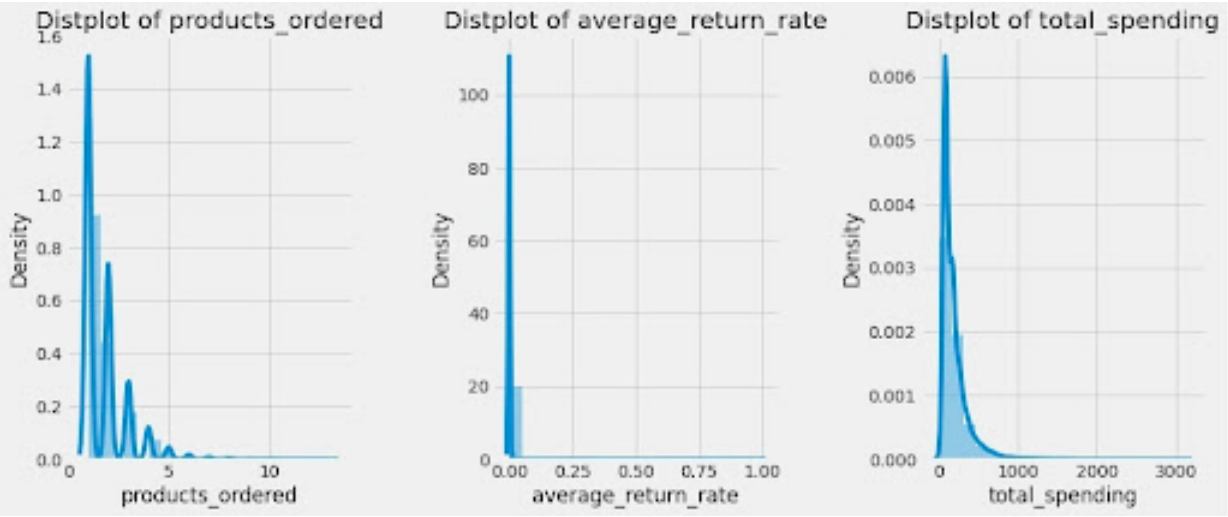

Figure 3.1: Feature Visualization

The 4^{th} feature, Gender does not contribute more on distribution due to equally likely values as shown in figure 3.2.

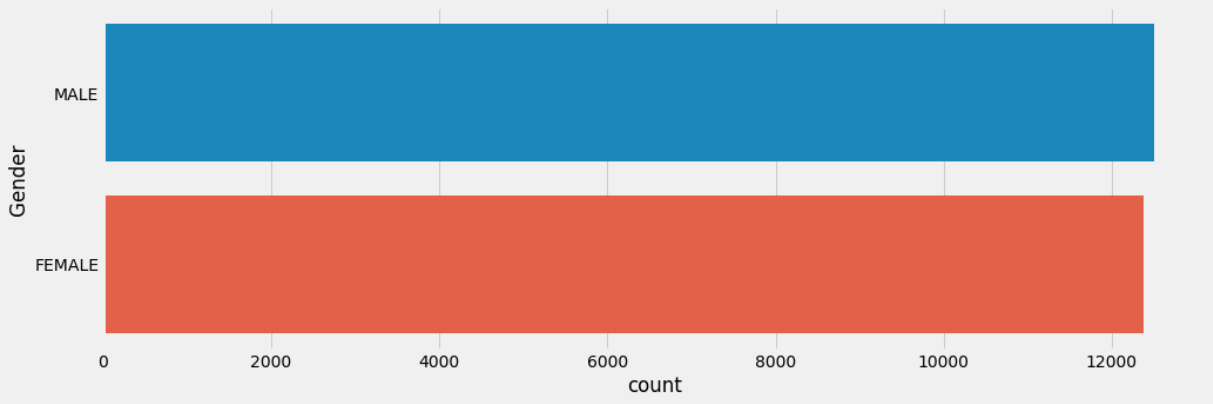

Figure 3.2: Number of data points w.r.t. Gender

Now, we visualize the correlation between three main features : Number of products ordered, Total spending, and Average return rate.

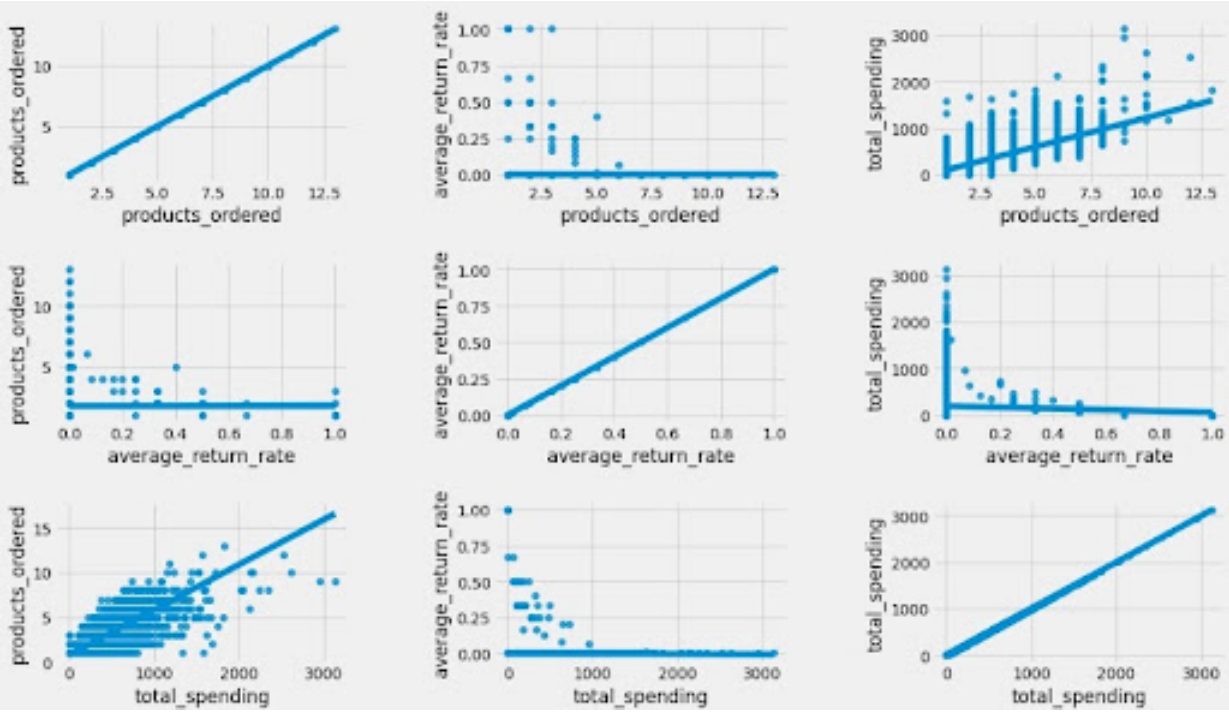

Figure 3.3: Feature Co-relation with every other Feature

We observe that taking the log of the features gives us a more evenly spread distribution.

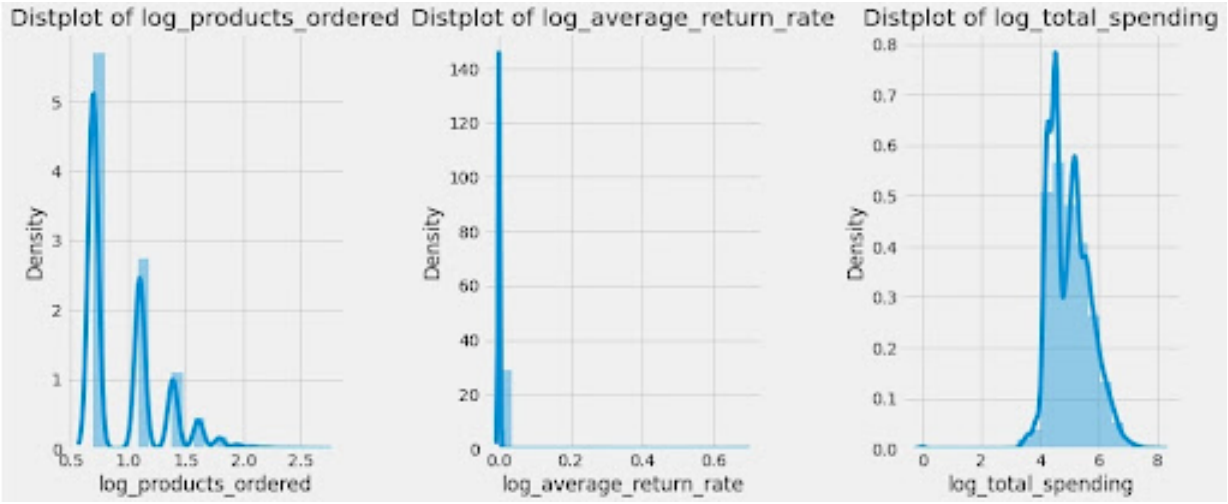

Figure 3.4: Feature Visualization after Log1p Transformation

4 K-Means Clustering

K-means clustering is an unsupervised learning algorithm which is simple, yet effective, thus making it a prime choice for customer segmentation. K-means clustering is an iterative algorithm which estimates 'K' centroids and then iteratively updates the position of the centroids that minimize the square of the distance to the points in each set. One of the most popular ways in which the hyperparameter 'K' is estimated, is by plotting the mean distance to centroid for discrete positive integer values of 'K', and choosing 'K' from the elbow point.

The K-means clustering algorithm includes the following steps:

- Select a dataset, and run manual/automatic pre-processing to clean the data, by removing nulls, removing outliers, etc.
- Estimate for K, and run the K-means clustering algorithm
  - Perform data assignment to the pre-processed data
  - Estimate for K
  - Initialize centroids randomly (using the K++ algorithm)
  - Run iterative centroid updates
  - Review Results, and optionally repeat for other possible values of K

Applying the K-means clustering algorithm to our dataset of customer segmentation includes the following steps:

1. Input dataset
2. Data pre-processing
3. Data visualization

4. K-means clustering for input features taken two at a time
5. K-means clustering for input features taken all together
6. Output clusters visualization

4.1 Hyperparameters and Tuning

The Hyper-parameters considered in this method and their ranges were taken as follows:

1. Number of Clusters - $1 \leq N \leq 15$
2. Cluster Centroid Initialization - k-means++ and random
3. Convergence Algorithm - Standard and Elkan
4. Number of Iterations - 200,300,400 and 500
5. Tolerance - $0.0001 \leq tol \leq 0.002$

4.2 Within-Cluster Sum of Squares method

In the Elbow method, we are actually varying the number of clusters (K) from 1 – 10. For each value of K, we are calculating WCSS (Within-Cluster Sum of Square). WCSS is the sum of squared distance between each point and the centroid in a cluster. When we plot the WCSS with the K value, the plot looks like an Elbow. As the number of clusters increases, the WCSS value will start to decrease. WCSS value is largest when K = 1. When we analyze the graph we can see that the graph will rapidly change at a point and thus creating an elbow shape. From this point, the graph starts to move almost parallel to the X-axis. The K value corresponding to this point is the optimal K value or an optimal number of clusters.

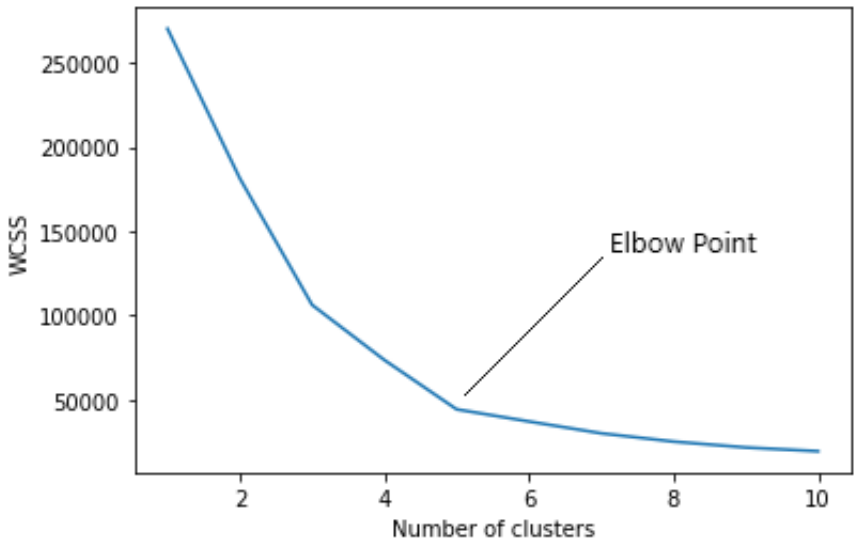

Figure 4.1: Elbow Plot

4.3 k-means++ Technique

The **k-means++** algorithm for random centroids initialization: It is an algorithm for choosing the initial values of centroids for the k-means clustering algorithm. The intuition behind this approach is that spreading out the k initial cluster centers is a good thing. The first cluster center is chosen uniformly at random from the data points that are being clustered, after which each subsequent cluster center is chosen from the remaining data points with probability proportional to its squared distance from the point's closest existing cluster center.

4.4 Elkan's Algorithm

Elkan's algorithm accelerates k-means by avoiding redundant distance calculations. This is done by using triangle inequality to keep track of lower and upper bounds for distances between points and centers.

4.5 2-variable Analysis

The **Elbow Plot** method was applied on two features, namely **Log of Products Ordered** and **Log of Total Spending**. The figure 4.2 shows large decrease in value of WCSS till number of clusters = 3.

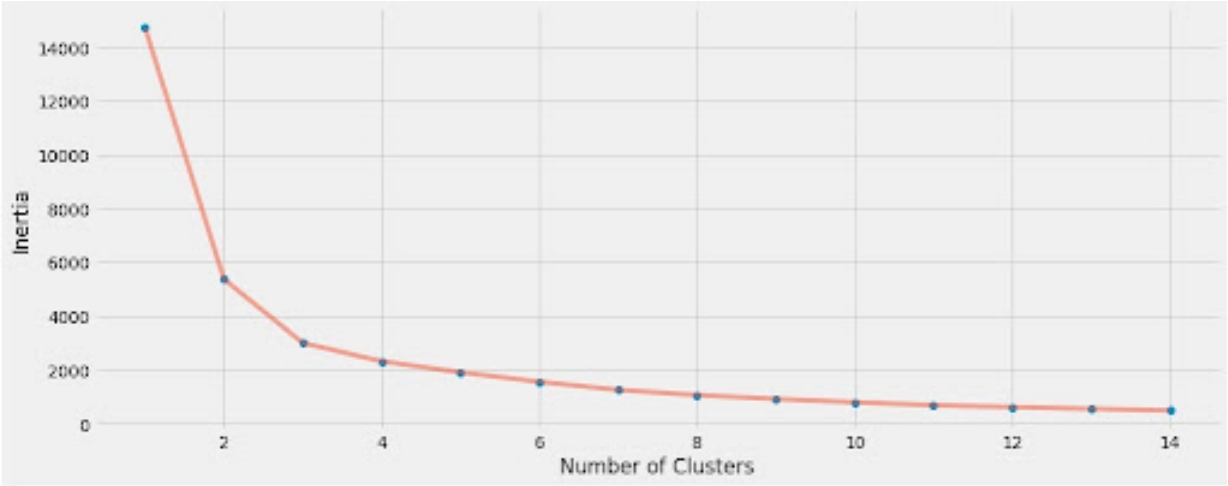

Figure 4.2: Elbow Plot for 2 features

4.6 3-variable Analysis

The **Elbow Plot** method was applied on three features, namely **Log of Products Ordered**, **Log of Average Return Rate** and **Log of Total Spending**. The figure 4.3 shows large decrease in value of WCSS till number of clusters = 4.

Figure 4.3: Elbow Plot for 3 features

5 Results

We plot the elbow plot for two of the features (Number of products ordered, and Total spending), and observe that the elbow point is at about K = 3 or 4, and we choose K = 3, and run a K-means clustering algorithm to find the following results shown in 5.1 Where violet points represent the above average customers, yellow points represent the average customers, and cyan points represent the below average customers. The pink data points represent the centroids of each of the three clusters.

5.1 2-variable Analysis

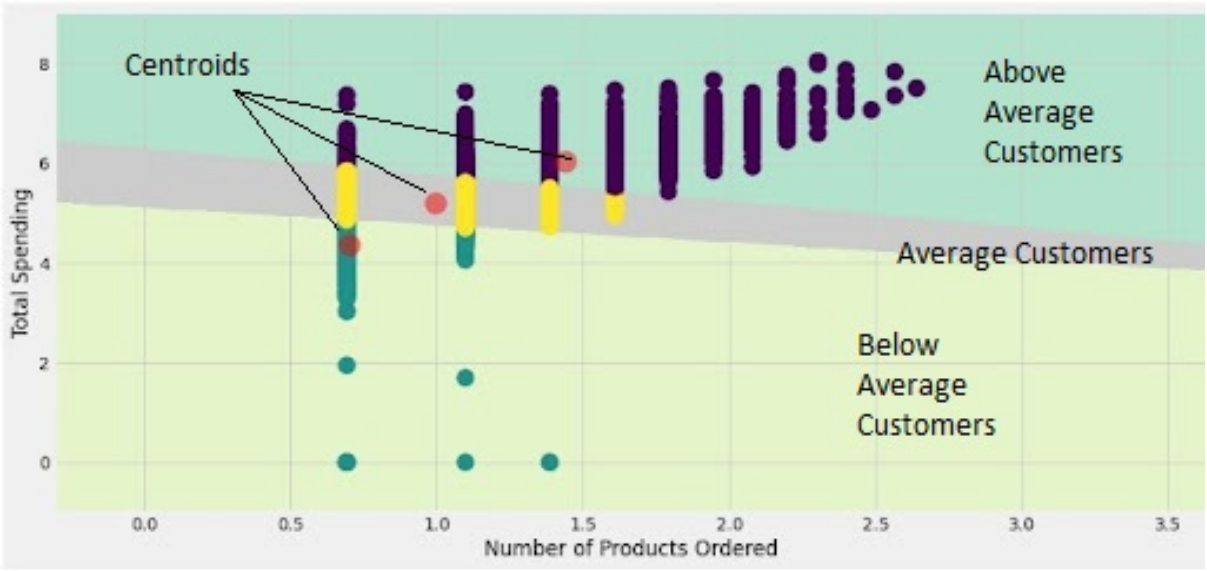

Figure 5.1: Clusters for 2 features

5.1.1 Centroids

Centroid\Feature	Number of Products Ordered	Total Spending
C2_1	1.43778251	6.05264533
C2_2	0.70305819	4.37369063
C2_3	0.99666264	5.21658115

Figure 5.2: Centroids for 2 features

5.1.2 Distance between Clusters

Distance between Cluster 1 and Cluster 2 is : 1.832678014729091
Distance between Cluster 2 and Cluster 3 is : 0.892562608962016

5.2 3-variable Analysis

Now, we plot the elbow plot for three of the features (Number of products ordered, Total spending, and Average return rate), and observe that the elbow point is at about K = 3 or 4, and we choose K = 4, and run a K-means clustering algorithm to find the following results. Where yellow points represent the Best customers, red points represent the above average customers, green points represent the below average customers, and blue points represent the customers from the least profitable segment.

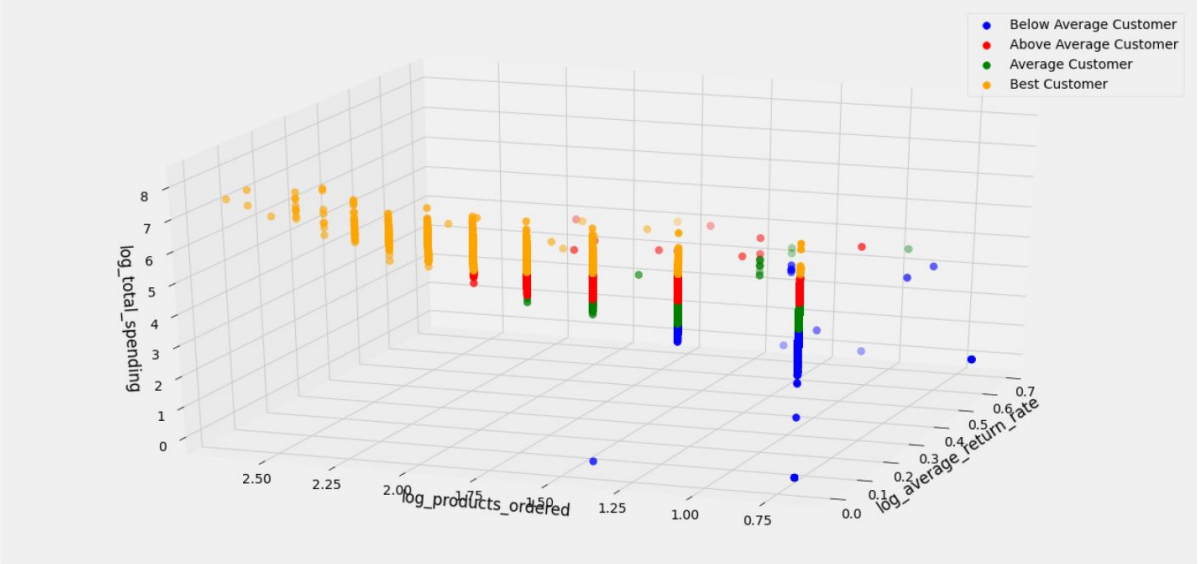

Figure 5.3: Clusters for 3 features

5.2.1 Centroids

Centroid\Feature	Log of Total Spending	Log of Average Return Rate	Log of No. of Products Ordered
C3_1	7.00601007e-01	1.15218211e-03	4.35093589e+00
C3_2	1.22324947e+00	5.20044802e-04	5.65169596e+00
C3_3	9.26994039e-01	5.47295737e-04	5.08068731e+00
C3_4	1.59846518e+00	5.55536094e-04	6.36573629e+00

Figure 5.4: Centroids for 3 features

5.2.2 Distance between Clusters

Distance between Cluster 1 and Cluster 2 is : 1.401834002087483
Distance between Cluster 2 and Cluster 3 is : 0.6432870011390299
Distance between Cluster 3 and Cluster 4 is : 1.4499049546075127

6 Conclusion

Customer segmentation using K-means clustering was applied on a dataset of customer purchases from a shopping mall, after data pre-processing. We visualize the data for each feature, understand the correlation between features, and then select two, and three features which are strongly correlated. We then use the elbow point method to estimate a value for K, and then run a K-means clustering algorithm on the data to obtain the mentioned results, where we visualize the clusters. These results can be used to find the most profitable customer segment and personalize their shopping experience to further increase profits.

7 Appendix

The GitHub link to project: https://github.com/GouravT8962/CS725_FML.git

References

[1] Krishnaveni K P,Blessy Paul P, "CUSTOMER SEGMENTATION USING MACHINE LEARNING", International Journal of Creative Research Thoughts (IJCRT), ISSN:2320-2882, Volume.9, Issue 7, pp.a236-a240, July 2021, Available at : http://www.ijcrt.org/papers/IJCRT2107034.pdf

[2] https://github.com/cereniyim/Customer-Segmentation-Unsupervised-ML-Model

[3] https://towardsdatascience.com/customer-segmentation-with-machine-learning-a0ac8c3d4d84

[4] https://blogs.oracle.com/ai-and-datascience/post/introduction-to-k-means-clustering

[5] https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/