MSc Project - Reflective Essay

Project Title:	Customer Segmentation Using Unsupervised Machine Learning Techniques
Student Name:	Tanishq Verma
Student Number:	210642458
Supervisor Name:	Dr Lin Wang
Programme of Study:	MSc Project ECS751P

The above project work on applying customer segmentation to the dataset using unsupervised machine learning techniques. The project applies five different unsupervised machine learning techniques which includes Hierarchical Clustering, K-means, DBSCAN, Gaussian Mixture Model and BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies). The project determines the best technique that could be used based on the average silhouette score and time taken. In addition, it also compares the similarity of the cluster generated by using adjusted rand score. The reflective essay illustrates the strength and weakness of the project including giving highlights on the possible future work. It also includes the relationship between theory and practical aspects. Finally, it gives details of sustainability and ethics followed during the implementation of the project.

Introduction

For this study, we have used Customer Segmentation using Unsupervised Machine Learning Techniques. We used real life open-source Online Retail Dataset which includes all the essential attributes to apply customer segmentation. We are using the RFM model which stands for Recency, Frequency and Monetary. RFM model is one of the most famous models for computing customer segmentation. It gives equal weight to each RFM attribute and using it, applies unsupervised machine learning techniques which is used to cluster customers. Before applying any techniques, since it's a real life open-source dataset, we applied data cleaning and preprocessing techniques. In addition, we had to generate each individual RFM attribute and add it to data frame. Before applying any techniques, we applied standard scaler which helps in quick and easy computation of the machine learning algorithm. We further visualizer the clustered generated, calculate the silhouette score and predict the cluster similarity for each of the following techniques.

Analysis of Strength/Weakness

This section of the essay will highlight on the strength and weakness of the project and further individual pros and cons of the five techniques applies in the study. We have visualized and made the relevant comparison wherever possible and suitable.

The project has various strength and weakness. To begin with, the primary advantage of the model and the study is that it can work on multiple unsupervised machine learning techniques and according to the data predict which techniques is the most efficient. Further, clustering of each techniques gives a 3d visualization of the data set cluster giving a better understanding of the dataset. Analysis and comparison using different metrics including the silhouette and adjusted rand score will simplify work to determine the best technique. However, on the other hand, customer segmentation is a vast and broad category and not all data set applied to the model may give expected result. Further, the techniques may differ according to the dataset and there may be a case for some dataset that any of techniques applied may not give good clustering results.

Discussing the strength and weakness of the techniques, to begin with, Agglomerative Hierarchical Clustering is famous clustering techniques. We implemented Agglomerative Hierarchical Clustering using different methods which includes complete linkage, single linkage, and Wards method. The main strength of this techniques is showing each step when clustering occurs using a diagram called Dendrogram. It depicts how clustering is done at each step and using different method and visualization dendrogram for each method, we can determine which method is the best. Further, we can determine the threshold value of sum of square distance, and it can develop number of clusters according to its own. However, the main disadvantage of this techniques is that for a large dataset, it takes a lot of computation time and memory when compared to other techniques. Overall, when we generate the silhouette score for the technique, the average silhouette score generates was approximately 0.432 which depicts that the clusters generated are good and not overlapping.

For K-means, we first applied Elbow method to determine the value of k or the number of clusters to be formed. We have visualized the graph and determine the value of k = 3. Further, we have applied k-means clustering for k=3 and visualized the cluster on 3d plot based on RFM. The strength of k-means is that we can find the effective value of k by using elbow method. Further, Clustering on large dataset is fast when compared to Hierarchical Clustering. The average value of Silhouette Score is 0.5799 for k=3 which depicts that clustering by k-means is better than Hierarchical Clustering. However, the main disadvantage of k-means is that for higher value of k, the sum of squared distance gets significant less difference which makes it harder to determine the actual value of k. As the value of k increase, we need advance version to determine the hyper parameters. Further, for dataset where data points have weight distribution i.e., some data points have more significance than other than it requires further techniques.

For the next technique, DBSCAN or Density Based Spatial Clustering with Noise, unlike the above techniques, it does not require to define the number of clusters to be formed but need to define specific variables which are used to determine the number of clusters to be generate. DBSCAN is effective for dataset to identify noise data segregating high density cluster from low density cluster. It can identify different sizes and shapes clusters from the dataset. However, the main disadvantage of DBSCAN is that it finds difficult to identify clusters of varied density. The average silhouette score generated after clustering for DBSCAN is 0.5656 which clearly shows that it makes good clusters.

For Gaussian Mixture Model which uses Gaussian distribution to cluster dataset into groups. The advantages of gaussian mixture model are that it works well on large and non-linear dataset. Gaussian mixture also has soft and hard clustering which can be used. Soft clustering can have datapoint lying between two cluster part of the cluster where weight of one cluster may be more compared to other depending upon the distribution. Hard clustering on the other hand defines that the data point shall only be part of single cluster even if it is part of both clusters. Further, the main deciding factor for defining which cluster shall the data point be part of is using the Gaussian distribution. The disadvantage of Gaussian mixture is that we need to specify how many clusters to be created for the dataset and there is no specific technique to identify the value of hyper parameter. The average silhouette score for our dataset when k=3 is 0.34 which is less compared to other techniques.

For BIRCH, as the name suggests, it uses hierarchical clustering where it takes a large dataset and starts creating a dataset summary. It takes small data points and cluster these points/clusters until the desire number of cluster are formed. It takes some parameters based on which the clustering is done. These includes threshold which states the maximum number of data points a sub cluster can hold while n cluster which is the number of clusters to be generated. The main advantage of BIRCH is that for implementing Hierarchical clustering it is better to use BIRCH which is faster compared to Agglomerate Hierarchical Clustering. In addition, it also creates a CF tree which is like the dendrogram generate while implementing

Agglomerative Hierarchical Clustering. The disadvantage of BIRCH is that parameter including the number of clusters to be generate i.e., the value of k needs to be defined unlike DBSCAN or K-means which uses elbow method to determine the value of K. The average silhouette score for our model is 0.71 which is the best average silhouette score generated when compared to each technique.

In general, each machine learning techniques that we have applied have their own strengths and weakness. Customer Segmentation being a vast and varied depending on what market segmentation is to be done, selecting the right clustering technique is vital as it can change the model generated and may not get the expected result. However, on the other hand, proper research, understanding the need of the market and criteria and using it to discover which technique to apply can give great result which will be advantageous for the business.

Future Work

As discussed above, although the project has a lot of strength still there are some weaknesses in the project. These weaknesses involve applying segmentation using unsupervised machine learning techniques with limited techniques. Customer Segmentation is a broad category where segmentation depends on the need of the business and the criteria on which clustering is to be performed. For some industries, the need may be different compared to other industries and so effective modification needs to be done by understanding the needs of the industry. Since customer segmentation is a broad category, based on this assumption, some possible future work includes using an industry specific dataset and understanding the requirement of the industry. Further, understanding the dataset and applying machine learning to it. For unsupervised, unlabeled data is suitable however for applying supervised machine learning, labeling data is critical and correct labelling is even more important.

In addition, exploring the different techniques for both supervised as well as unsupervised machine learning. Unsupervised machine learning includes Fuzzy k-means, OPTICS, mini-batch k-means and other while supervised includes linear regression, support vector machine and others. Apart from this, based on the requirement, different types of techniques to identify the accurate hyper-parameters for the different techniques shall also be explored and identified based on the technique.

Given more time for this project, the more work that could have been done would be implemented more unsupervised machine learning techniques and give an idea of how each technique is better than the other and exploring the different types of techniques to get the accurate hyperparameter. Further, additional implementation would involve a couple of famous supervised machine learning clustering algorithms and compare the supervised with the unsupervised. This will give an idea of which technique to be used based on business requirement. Since supervised machine learning requires to label the data, granting more time would be used by analyzing the best unsupervised machine learning, labelling a particular amount of data using the unsupervised the machine learning considering it as training data and applying the supervised machine learning based on this labeled data. Calculating the most appropriate clustering techniques using different parameters and not just using the average silhouette score.

According to [1], the study uses k-means, DBSCAN, and hierarchical clustering to applying customer segmentation. The study uses silhouette score but also uses CH (Calinski Harabasz Score) to analyze the quality of the structure created. [4] uses an extended version of RFM called RFMC proposed by classic RFM has introduced a new feature called proposed feature category. Using [2], it determines the value of k for k-means by elbow but also using Calinski Method. According to a different study [3], which uses Renyi and Shannon Entropy to calculates the randomness and uncertainty in the dataset can also be used for future work.

Studying [5] based on similar category where instead of customer segmentation uses risk analysis segmentation and uses different techniques involving k-nearest neighbor, random forest and other which could be used for future study. [6] uses k-means to apply customer segmentation and uses auto-encoder on predicted sales values to find optimal number of clusters.

Customer Segmentation has a vast domain and given more time could have explored the different domains. Working for specific industry and applying customer segmentation could have made business grow by understanding the customer behavior and developing business strategies including marketing strategies according to the customers.

Relationship between theory and practical work

During this study, the model that was created can implement different machine learning techniques on the dataset. Given a dataset, the model can use the dataset and apply five unsupervised machine learning techniques to it and cluster using RFM. However, since the data will be different, we have is applying data cleaning and pre-processing to it. Further, we also must get the RFM attributes so that clustering can be done based on the data frame.

Using this working machine learning clustering model, we can cluster customers into different cluster based on their behavior for this case it consists of Recency, Frequency and Monetary. As for the theory of the study, which includes segmenting customers based on their similarities so that business can study the behavior of the customer and develop new business strategies to keep up with the competitors in the industry. We implemented the machine learning model which can use the theory explained and can implement the clustering provided we get a dataset and the attributes which will be used for clustering.

Using the literature review and understanding in-dept about the market and industry, we did research on what factors the clustering needs to be performed. Using it, further, determine what key benefits it provides to the business. The implementation provided using this study can benefit any organization, business or MNCs to make appropriate change to it and further according to its own parameters of determining cluster can help them understanding the behavior of the customer. The model can help business to understand the needs of the customer, what products are in demand, and they can develop new business strategies using it.

Further study in this field for any specific industry by determining the key factors responsible for clustering can give an insight to these industry to segment customers. Customer Segmentation can have many benefits provided it is applied on the depending key factors.

Awareness of Legal, Social Ethical Issues and Sustainability

While conducting and implementing the study, we have taken into consideration all the social ethics and legal values and principal. To illustrate, the data set that is taken is from an open-source machine learning repository owned by University of California, Irvine. Being an open-source repository, it can be accessed and used by everyone. Further, while conducting literature and background review, we have studied various research published on different publication. We used the Queen Mary university registration to get the access to the papers. However, in some publication, the journal was paid and so did not get access to it. In addition, we have used local machine to implement and run the model. All the python packages that were required for the implementation of the model were open-source and can be accessed by anyone. Overall, during the implementation of the model and writing research, we have followed the social and legal ethics.

Conclusion

When analyzing the result, we can say that Hierarchical Cluster works well for small data set. K-means and DBSCAN generated good silhouette score and took significant less time. Gaussian Mixture Model, although did took very less time for computation, the cluster generated by calculating the average silhouette score was poor compared to other techniques. Finally, for BIRCH, which is the most effective technique generating the highest average silhouette score and low computation time.

From the above we can say that BIRCH performed the best for the dataset followed by k-means and DBSCAN. Hierarchical is better for small dataset and not viable for large. Gaussian Mixture Model generated the worst clusters although exploring it further could give better results.

References

1. J. Thomas and P. N, "Customer Segmentation in the Field of Marketing," 2021 4th International Conference on Recent Trends in Computer Science and Technology (ICRTCST), 2022, pp. 401-405, doi: 10.1109/ICRTCST54752.2022.9781964.
2. E. Umuhoza, D. Ntirushwamaboko, J. Awuah and B. Birir, "Using Unsupervised Machine Learning Techniques for Behavioral-based Credit Card Users Segmentation in Africa," in SAIEE Africa Research Journal, vol. 111, no. 3, pp. 95-101, Sept. 2020, doi: 10.23919/SAIEE.2020.9142602.
3. K. Bhade, V. Gulalkari, N. Harwani and S. N. Dhage, "A Systematic Approach to Customer Segmentation and Buyer Targeting for Profit Maximization," 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2018, pp. 1-6, doi: 10.1109/ICCCNT.2018.8494019.
4. S. Allegue, T. Abdellatif and K. Bannour, "RFMC: a spending-category segmentation," 2020 IEEE 29th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 2020, pp. 165-170, doi: 10.1109/WETICE49692.2020.00040.
5. S. Bhatia, "Pragmatic segmentation-based credit risk management using Machine Learning," 2022 International Conference on Communication, Computing and Internet of Things (IC3IoT), 2022, pp. 1-5, doi: 10.1109/IC3IOT53935.2022.9768006.
6. U. Sharma, G. Aditi, N. R. Roy and S. N. Singh, "Analysis of Customer Segmentation Clustering Techniques," 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2022, pp. 374-379, doi: 10.1109/Confluence52989.2022.9734147.