Customer Segmentation Using Machine Learning

Abstract

A customer is a critical aspect in the success of any business. Building a better relationship with customer help the organizations in increasing profit and customers satisfaction. The potential value of a customer to a company is a core ingredient in decision-making about marketing strategies. Customer Segmentation is one such strategy, which helps in identifying groups of similar customers based on their interactions with the products and then effectively implementing various marketing strategies for the suitable customers. In this work I implemented five clustering algorithms for the dataset to compare the performance. The algorithms implemented are K-Means, K-Means++, Manshift, Agglomerative, Spectral. For our case Agglomerative showed the best performance measured by Silhouette scores. I implemented six classification algorithms namely Logistic Regression, SVC, Decision Trees, Random Forest, KNN and Ridge Classifier and three use cases.

Goal:

The purpose of this project is to explore the use of different clustering algorithm for identifying and understanding user behaviour on mall dataset to increase customer satisfaction and increase profits for the business owner. The aim is to develop an application that can group and classify users based on their behavioural segmentation factors where in the segmentation is based on customer's behaviour pattern with a particular business or a website. Example Spending habits, purchasing habits, browsing habits, interaction with the brand, loyalty to a brand which then could be used to improve the mall's marketing strategies.

Proposed Idea:

Our proposed idea is shown in the Figure 1. I selected a data set of a mall which have details of a purchased product. After doing pre-processing and feature extraction, I will select features for clustering. The features will be used to group customers using 5 clustering algorithms. After making inferences from the cluster, I will implement 6 classification models to classify new customers. Based on the cluster of a customer and most purchased and correlated product, I will perform various use cases and send various offers via email.

Methodology:

Data Collection:

The dataset I am using for this project is from a shopping mall in London. It was collected by using various kinds of IOT (Internet of Things)evices used in Malls like POS machines, scanners, checkout systems. The data from all these devices is merged to form one dataset for analysis. Figure 2 shows a view of the dataset I am using. The dataset has total 8 attributes namely Invoice No, StockCode, Description, Quantity,InvoiceDate, UnitPrice, CustomerId, Country.

Prepared by –Chandu Pavan Budda
Id:1001977117

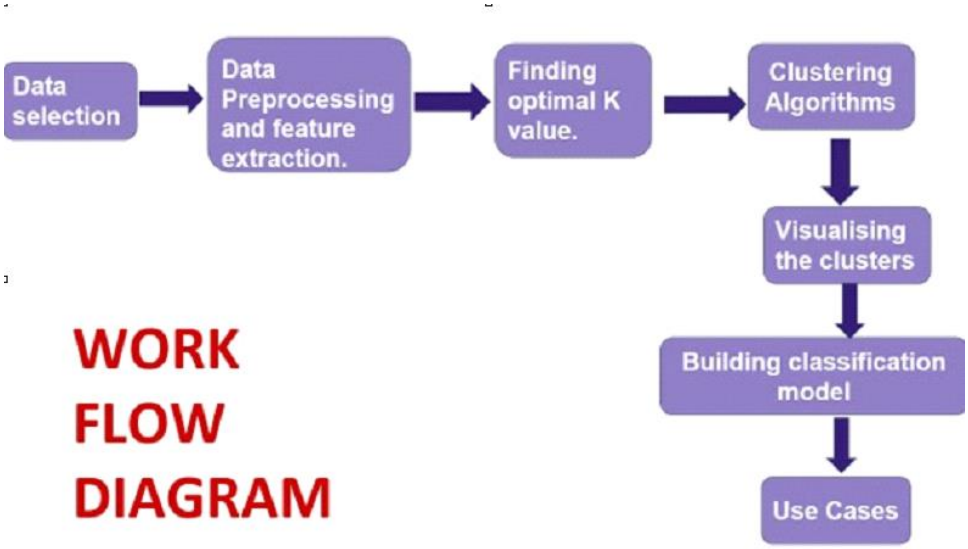

Figure: Workflow Diagram

```
In [5]: #top rows of data

        data.head()
```

Out[5]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

Figure: Data Set

Data Pre-processing:

Data pre-processing is the process of converting the raw data into useful and meaningful data to be used for further analytics and algorithms. It involves dealing with missing values, normalization of data, selection of relevant features etc. I have started with removing the null values and then selected the features best suited.

Data Cleaning:

This step involves dealing with missing and irrelevant parts in our dataset. For missing values, I can do two things: either remove the tuple from the data completely or fill those missing places with the most probable value. Most probable value can be mean or mode. In our dataset the attribute" Description" had 1454 and attribute "CustomerID" had 135080 missing values. I chose the former approach and ignored the rows containing missing values.

Feature Extraction:

I derived new features from the inbuilt features. The features are known as RFM (Recency,Frequency, Monetary).

1. Recency - how recently the customer did his purchase?
2. Frequency – How often do they purchase?
3. Monetary Value – How much revenue they generate?
4.Mean Days -is the average number of days customer has waited for his next purchase.
5. Also include top 5 products of each cluster as a feature.

Anova Analysis:

Out of these 9 features we intended to select the most relevant ones. For this task we used Analysis of Variance (ANOVA). It is statistical tool which consist of various models and their procedures. I used F-test of ANOVA to determine the top 5 relevant attributes which helps in forming clusters in our data. It was done using f classify and SelectKBest from sklearn.featureselection. This analysis gave the result that attributes"AvgSpend"," frequency","meanDays"and"recency" are more relevant than others. The work was continued using these attributes only

Clustering:

It is an unsupervised ML (Machine Learning) algorithm which divides the dataset into different clusters. There are various clustering algorithms available, but I started with most popular one i.e., K-Means Clustering. For that I need to find optimal K value first which was done using Elbow method. We implemented another version of K-means we call it K-Means ++

Elbow Method:

For clustering we need to find the optimal K value first. One of the ways to decide K value is to use Elbow method. We can do that by computing the cost of K-Means model with different values of K. It can be observed that the cost decreases as the value of K increases. But at a certain value of K the cost stops decreasing drastically. That value can be considered as the optimal K value. I have used elbow method for two datasets, one with "Avgspend"," frequency," and "recency" values only and another with added "meanDays" values. Plotted graphs for first and second datasets are Figure 3 and Figure 4, respectively. The graph of K v/s cost forms a shape of an elbow as in the figures. From both the graphs you can infer that optimal K value should 3. But there is a significant difference in cost value for both the datasets. Figure 5 shows the cost comparisons for RFM features and RFM + MeanDays features. Cost of having 3 clusters using RFM is 49.9 and cost of 3 clusters with RFM+meandays is 135.63. That is why I discarded MeanDays feature. Including meanDays also affects our accuracy.

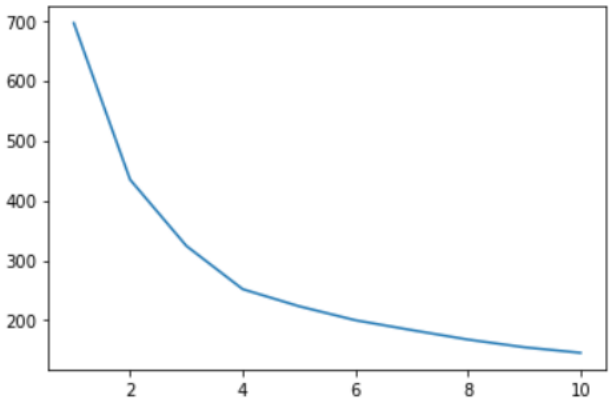

Figure 3

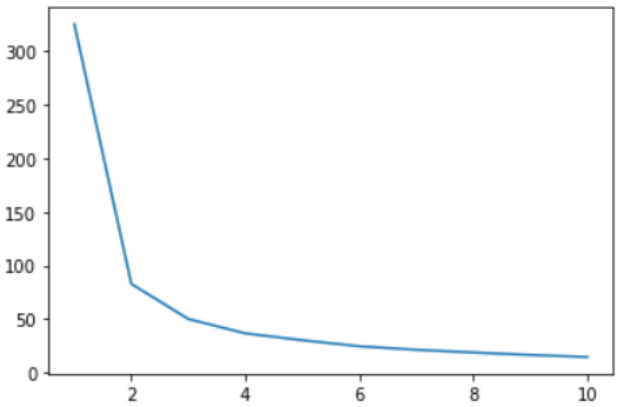

Figure 4

```
In [35]: #in both using mean days and without using mean days optimal k is 3
         print(cost1[2],cost2[2])
49.93351925902856 135.63892428443623
```

Figure 5

Data After:

Clustering Figure 6 shows how the clustered data looks like. Classification models will be built on the clustered data. The clustered data has one extra feature ”clno” which is the cluster number the customer id belongs to.

Visualizing The Clusters and Making Inferences:

I have used seaborn.pairplot to visualize the different scatter plots between the attributes ”freq”,” recency” and ”avgspend” shown in the Figure 7. It can be observed that most of the customers have frequency close to 0. The few portions of customers who are frequent have low average spend value and high recency. The recency value clearly divides the customers into clusters. High recency value tells you are in cluster 0 whereas low recency value says you are in cluster 1. The ones in cluster 2 have average recency value. If we consider the average spend value, you can see that most customers belong to low to medium category. The only ones whose average spend is high are less frequent and belong to cluster 0. We can infer that they have made a single purchase but huge one. Figure 8 shows the average RFM values for each cluster.

	id	freq	rec	avgspend	clno
0	12347.0	0.024291	0.994638	0.099186	0
1	12348.0	0.012146	0.798928	0.072380	0
2	12349.0	0.000000	0.951743	0.283126	0
3	12350.0	0.000000	0.168901	0.053869	1
4	12352.0	0.040486	0.903485	0.022632	0

Figure 6

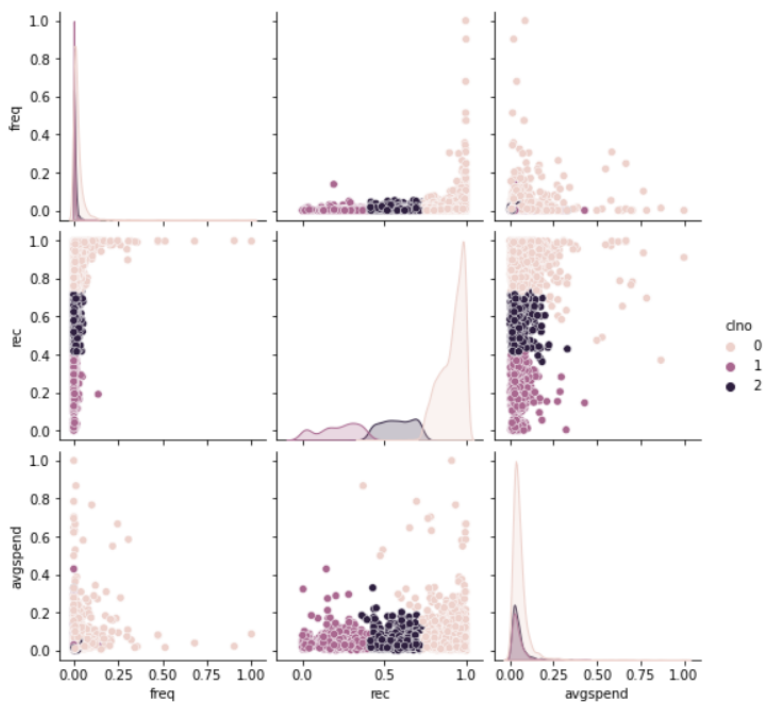

Figure 7

Classification Models:

For the use cases we need to create a classification model. I have used different classification techniques namely Logistic Regression, SVC, Decision Trees, Random Forest, KNN and Ridge Classifier. I compared accuracy scores of all these algorithms and found apart from Ridge classifier all were giving satisfactory results.

Model Training and Validation (Hyperparameter Tuning):

For classification task we need to define our independent variables and dependent variables(output). The independent variables were the attributes ”avgspend”,”rec” and ”freq” and the output label was the cluster number namely column ”clno”. Then I randomly split the data into training set (70%) and testing set (30%). Then the classification was done on our dataset using each algorithm one by one. For better results we also used K-Fold cross validation with number of folds as 6. The accuracy was calculated on each fold and average was returned. Each algorithm has its own set of hyperparameters, for example we have K in KNN and max depth in Decision trees. I checked the accuracy levels for different values of hyperparameters and chose the best one. I started with Random Forest classifier which has hyperparameters” max depth” and” number of estimators.”I checked 100 different combinations of both these values and for each combination cross validation was performed. The combination which gave highest accuracy was” max depth” =9 and” number of estimators” =70. Then again, I repeated the same process with Decision Tree classifier with” max depth” as hyperparameter. I checked for 10 different values of max depth and fount the optimal value as” max depth” =4. For KNN the hyperparameter is K (number of neighbours to be considered). After checking for 10 different values, I found the optimal value as 5. In case of SVC there are three hyperparameters I considered: C (Regularization parameter), gamma (kernel parameter) and kernel. We had to try out 160 different combinations to find the best one and it turned out to be” C” =600,” gamma” =1 and” kernel”=rbf. For Logistic Regression we have C (Regularization parameter) as hyperparameter and tried 8 values. Optimal” C” was 300. At last, for Ridge classifier hyperparameter was” alpha” whose 5 different values were tested, and optimal value was 1.

Model Evaluation:

For evaluating the different models, I calculated individual accuracy scores on our test data points. Classification report was also generated for each model which includes Precision, Recall, F1-score, and Support. The hyperparameters of models were set to their best value. On comparing the accuracy score of all these algorithms, we found that apart from Ridge classifier all were giving satisfactory results. I can say that K-Nearest Neighbour (KNN) works exceptionally well among the others but the problem with it is it is a lazy algorithm. It predicts most accurately but takes time to do so. Second best algorithm in terms of accuracy is Random Forest. Its accuracy level is almost equal to KNN and works faster. So, for classification I decided the Random Forest as final classification model.

Use Cases:

I have implemented three use cases for this project.

Product Suggestions:

In the first use-case, for a particular customer, I find the cluster to which it belongs and the most purchased product in that cluster. If the customer has not purchased any product from the top product in that cluster, then we suggest that product to him.

Market Basket Analysis:

Market Basket Analysis is used to find an association between the products. It finds the product which is most frequently brought together by the customers. I have used aprior algorithm to find these associations. I apply market basket to the clusters first and then we can suggest the offers.

Combo Offers:

In this use case, we find the most purchased product by the customer and the least sold product in the store. And we have provided a combined offer on both products.

Results:

Silhouette:
Coefficient It is a method of interpretation and validation of consistency within clusters of data. It is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from -1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

Clustering Results:
We have implemented five different clustering algorithms, K-Means,K-Means++, Mean shift, Spectral clustering, and Agglomerative clustering on RFM features of customers. We have calculated the silhouette coefficient for these algorithms. Table 1 shows the results. Based on the silhouette score, we have selected the Agglomerative Clustering algorithm.

Classification Results:
We have implemented six algorithms: Random Forest, Decision Tree, KNN, support vector classifier, Logistic Regression, and Ridge Classifier for classifications. Table 2 shows the results. The result from the clustering algorithm is used as the training set. Based on the accuracy of these algorithms, best

classification model is K Nearest Neighbours, but we do not take it since it is a lazy algorithm and after that the best is random forest.

Algorithm	Silhouette Score
K-Means	0.609185
K-means++	0.610002
Mean-shift	0.606873
Spectral	0.503549
Agglomerative	0.610704

Classification Method	Test-Accuracy	Train-accuracy
Random Forest	0.997708	1.000000
Decision Tree	0.996528	1.000000
K Nearest Neighbours	0.998450	0.999210
Support Vector	0.996528	0.996411
Logistic Regression	0.995045	0.989286
Ridge Classifier	0.666667	0.666508

Conclusion:

The purpose of this project was to explore the different clustering algorithms as a tool for identifying user behaviour on a customer dataset. To achieve the purpose of the project, six classification models have been trained to categorize users into different user personas. The performance of the classification models has been analysed using several evaluation metrics. The results of this study showed that Agglomerative clustering algorithm performed better than rest of the clustering algorithms and for classification models Random Forest performed better. We choose many features for clustering but Recency, Frequency, Monetary (RFM) gave best results for our dataset. For the clustered customers we implemented three use-cases. It is worth noting that the result of the project, like most other classification problems, is tightly connected to the data available. Should a couple of different decisions have been made during the project, the result would have been different, better, or worse. But in the end, the result of the project satisfied our goals from the project and the purpose have been fulfilled