Credit Card Default Prediction

Rahul Chauhan

Data Science Enthusiast

Almabetter, Banglore.

Content

TITLE PAGE

1. Abstract 0-1
2. Problem Statement 1-3

3. Introduction 3-4

4. What is customer segmentation 6-7

5. Steps Involved 7-9

6. Types of model used 9-13

9. Evaluation Metrics 13-14

10. Conclusion 14-15

1.Abstract:

let us think about how a company should think in order to make effective marketing policies. For a small company, the customer base is usually quite small and individually targetable. But, as a business grows in size, it will not be possible for the business to have an intuition about each and every customer. At such a stage, human judgments about which customers to pursue will not work and the business will have to use a data-driven approach to build a proper strategy.

For a medium to large size retail store, it is also imperative that they invest not only in acquiring new customers but also in customer retention. Many businesses get most of their revenue from their ‘best’ or high-valued customers. Since the resources that a company has, are limited, it is crucial to find these customers and target them. It is equally important to find the customers who are dormant/are at high risk of churning to address their concerns. For this purpose, companies use the technique of customer segmentation.

*Keywords: Data cleaning, Data analysis, Feature Engineering, Cluster model, Evaluation metrics*

2.Problem Statement

The **problem statement** is to identify major customer segments on a transnational data set in registered non-store online retail. We have to analyse the customer based on the history of their transaction which occour between **01/12/2010 and 09/12/2011.** The model should have a good clustering segmentation for the customer, any of the store will not get away from the customer.

The data description for the customer segmentation is:

### InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.

### StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.

### Description: Product (item) name. Nominal.

### Quantity: The quantities of each product (item) per transaction. Numeric.

### InvoiceDate: Invice Date and time. Numeric, the day and time when each transaction was generated.

### UnitPrice: Unit price. Numeric, Product price per unit in sterling.

### CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.

### Country: Country name. Nominal, the name of the country where each customer resides.

3.Introduction:

Bussiness is now a days major source to earn the money, so most of the people move towards the business. In short we can say that business is the key point to earn the money. So to build a good business we have to consider a customer behaviour with our company as in small business it does not need to cluster the segmentation of the customer but as the business increase we need to cluster the data of customer so we have to use Elbow method to find the optimal number of cluster for the segmentation of the customer. By using this evaluation metrics we came to know that in how many group we can group the customer into the forms so which forms the customer will be group we have to decide.

4.What is customer segmentation?

**customer segmentation** are used to develop tailor-made marketing campaigns and for designing overall marketing strategy and planning.

A key consideration for a company would be whether or not to segment its customers and how to do the process of segmentation. This would depend upon the **company philosophy** and the type of product or services it offers. The type of segmentation criterion followed would create a big difference in the way the business operates and formulates its strategy.

5.Steps involved:

* **Fill Missing value**

After loading the dataset we performed this method by replacing

nan values with zero. This procedure giving us a approach to tackle with the null value and making a foundation strong.

* **Data cleaning**

In this process we convert all string contained data to numeric data. As instead of this we will unable to do the EDA on the data of customer segmentation.

* **Exploratory data analysis**

In this procedure we simultaneously work with each features from our dataset and clearly visualize each and every point of aspects. This gives a graphically representation of entire dataset.

* **Extraction of un-useful data**

In entire data some features has un-useful for the data analysis. So it’s better to extract this kind of features from our dataset.

* **Feature Engineering**

In this part we done some feature extraction like month, date, day and year and with the help of this we explore our dataset and come up with more comfortable data to deal with the issues.

* **Import Baseline model**

As we discussed earlier we have to train model such as Recency, Frequency & Monetary. As it gives us a help to decide the better model for segmentation of the customer.

* **Use Evaluation Metrics**

The evaluation metrics such as silhouette score, DBSCAN(Density Based Spatial Clustring of Application With Noise) and Elbow method to calculate the score for better cluster from the model means by which model we can cluster the dataset of the customer segmentation well cleared. The elbow method used to find the optmal number of the cluster by which we can group the dataset into the various forms of the behaviour of the customer.

* **Choose the best model**

The final predictor model is the last point for our model then after we are confirm to use the model for clustering the dataset and make the final decision that in how many groups we can cluster the data.

6. More about RFM

RFM stands for Recency, Frequency, and Monetary. RFM analysis is a commonly used technique to generate and assign a score to each customer based on how recent their last transaction was (Recency), how many transactions they have made in the last year (Frequency), and what the monetary value of their transaction was (Monetary).

7. How RFM works:

* Customers are assigned a recency score based on date of most recent purchase or time interval since most recent purchase. This score is based on a simple ranking of recency values into a small number of categories. For example, if you use five categories, the customers with the most recent purchase dates receive a recency ranking of 5, and those with purchase dates furthest in the past receive a recency ranking of 1.
* In a similar fashion, customers are then assigned a frequency ranking, with higher values representing a higher frequency of purchases. For example, in a five category ranking scheme, customers who purchase most often receive a frequency ranking of 5.Finally, customers are ranked by monetary value, with the highest monetary values receiving the highest ranking. Continuing the five-category example, customers who have spent the most would receive a monetary ranking of 5. The result is four scores for each customer: recency, frequency, monetary, and combined RFM score, which is simply the three individual scores concatenated into a single value. The "best" customers (those most likely to respond to an offer) are those with the highest combined RFM scores. For example, in a five-category ranking, there is a total of 125 possible combined RFM scores, **and the highest combined RFM score is 555.**

7. Evaluation Metrics

1. Silhouette score:

**What is silhouette score in clustering ?**

Selecting the number of clusters with silhouette analysis on k-means clustering. The silhouette plot display a measure of how close each point in one cluster is to points in the neighbouring cluster.

**How does the silhouette method work?**

The silhouette method computes silhouette coefficients of each point that measures how much a point is similar to its own cluster compared to other clusters, by providing a succint graphical representation of how well each object has been classified.

DBSCAN:-

 Density Based Spatial Clustering Of Application With Noise.

There are two parameters:

**Epsilon Point:**

The distance taken to draw a circle considering as a radius of the

circle.

**Min\_Points:-**

 There must be present atleast this points. If we consider the

min\_points as 3 so there should be three points in the circle.

**Core Points:-**

The points which have atleast min\_points present in the circle are

called as core point.

**Boundary Point:**

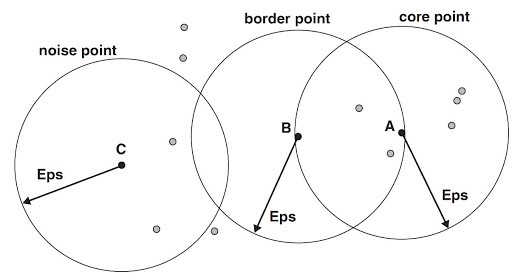
The point which are neighbour to the core point are called as

boundary point.

**Noise Point:-**

 The point which have neither neighbour point nor core point are

called as Noise Point.



Elbow method

This method is used to find the optimal number for the clustering of the dataset.

8.Conclusion

We have a dataset of the customer for non-store online retailer wholesaler, in which we analyze the transaction history as per Country. The behaviour of the each and every customer is easily analyze with the EDA. Then we done some preprocessing of the dataset and we come up with the feature engineering in which we explore the data and get some more view in the dataset. We build a RFM model in which we analyze the customer as according with this model and come up with the RFM score. Then we deal with some evaluation metrics such as silhouette score, Elbow method to find the optimal number of cluster and plot the graph.Then we come up with the final conclusion that we can cluster the customer based on the frequency and monetary so this is the good model to cluster and the optimal number for the cluster of customer data is 3. Hence we can go ahead with this number of cluster and it giving us a good segmentation for the customer and we can easily solve the business problem with this cluster model.