Advances in Economics, Business and Management Research, volume 203
Proceedings of the 2021 3rd International Conference on Economic Management and Cultural Industry (ICEMCI 2021)

CVM Model of Customer Purchasing Behavior Based on Clustering Analysis

Rui Zhao[1,*]

[1]*Department of Finance, National Taipei University of business,Tapei, Taiwan,China*
**Corresponding author. Email:952570041@qq.com*

ABSTRACT
At this time, companies progressively gain revenue from their long-term relationship customers. Many online retailers are earger to practice data analysis and CVM in businesses. However, many companies lack expertises to do so. This paper will present how to use data analysis in business cases. The main idea is to teach many executives how to have a better understand for their customer and therefore in their future progress they can be more producive to make decisions. On account of model built in this paper, customers were segmented into various meaningful groups using the k-means clustering algorithm. The main traits of these consumers in each part have been pointed out. Consequently, a set of recommendations is further provided to the business on consumer marketing.

Keywords: *Customer segmentation, Customer value measures, Data mining, k-means clustering*

1. INTRODUCTION

With the development of modern technology, people have witnessed a steady and vigorous increase in online business. Based, on this, data from customer behaviour not only can accurate gauge what customers need but also can have a deep understand on customer preference.

This paper aims at producing a predictive model to anticipate the purchases from a new customer. The first step is to classify the type of product. Therefore, this paper performed a classification of the customers' behaviour. The methods to segment the customer to distinct group is the k-meansclustering algorithm and decision tree.

On this stage, Ultimately trained several classifiers is done. The classifier is based on variables which are:

mean: amount of the basket of the current purchase categ_N with: percentage spent in product category with index

Then, the data were processed in two steps: first, all the data were assumed to define the category to each client belongs, and the classifications were compared with this category groups. One then found that 75% of clients are awarded the right classes. The performance of the classifier, therefore, seems accurate because it gave the potential shortcomings of the present-day model.

2. LITERATURE REVIEW

We next review the relevant literature about the segmentation model, Customer value measures, and classification models.

2.1 Segmentation of customers

The increasing calculate ability for data storage, which coupled with advanced data mining techniques, increased the possibility to grab information from online business. Data mining help to extract hidden or predictive information from sizable databases, enabling companies to find valuable customers, therefore companies can predict cutomers' purchase behavior and make proactive and knowledge-based decisions.

For classification methods such as neural networks, linear discriminant analysis, decision-tree induction, the main conception is to establish a model that allot each label to a new observation group. When one label can not find its observation group, unsupervised clustering methods can infer the class information to the distribution of observations.

2.2 Customer value measures (CVM)

Customer value measures are a method to identify the value of each customer.. It can use to analyze the value of each customers.

CVM base value = profit from each customer = revenue – cost

Then, adding in expert factors as multiplicative enhancements:

CVM = base value * f1*f2*f3*...

Then the formula for expert factors:

$$f_1(\text{field}) = 1 + \frac{1.5 - 1}{1 + \mathrm{e}^{-(\text{field} - \text{field}_{\text{mid}})/c}} \quad (1)$$

(This formula gives up to a 50% boost depending on the value of "field".)

Customer Value Measures ("CVM") is widely used among several industries including financial services. However, a single company find that sometomes this measure diversed from the true condition and was incompatible with applications from departments. In addition, methods focused on spot values without considering the fluctuations.

2.3 Cluster analysis

The goal of cluster analysis is to organize the test subjects into groups, according to their similarities. Cluster method is considered one of the most vital unsupervised learning methods. Like every other unsupervised method, the cluster method does not use previous classifications to reveal the underlying structure in the collection of data.

3. METHODOLOGY

The ultimate intention of this paper is to anticipate the purchase behavior that each customer has and to anticipate the amount of visits that customers will make during a year. Firstly, we need to clean up the data by deleting specified data which are came from a particular customer because these entries are useless for running the model.

3.1 Data for customer segmentation

The data can be categorized into products and customers. This data is composed of approximately 4000 clients' data details on purchases made on an E-commerce platform over one year. Each entry in the dataset includes the product bought from each customer.

3.2 Data pre-processing

To exploit a model-based clustering analysis, researchers need to pre-process the original dataset. Theoretically, customer segmentation methods can be segmented into distinct geographic, psychographic, behavior, demographic groups. For example, geographic segmentation is based on where customers live. Companies can use the data from their website to find out the characteristics of people in each regions[2].

This paper needs to divide customer segmentation methods into past purchases, consumption of each customer and so on. The data frame contains eight fields that are related to:

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.

3.3 Clustering of customers

Clustering analysis is widely used to conduct studies in terms of objects' variables. Hruschka (1996), Ozer (2001), and Weber (1996) use clustering techniques to segment customers and markets. The K-means clustering algorithm and the Kohonen Self-organizing map (1982) are the two most popular clustering techniques[4][5][6].

The K-means clustering algorithm randomly selects K initial cluster centers from N observations and assigns the remaining N-K observations to the nearest cluster based on the Euclidean distance. Then, the center of each cluster is recalculated according to the observation value assigned to each cluster. The next step is to reassign the observations to the closest cluster and recalculate the center of each cluster again until no observations are reassigned to the new cluster.

The best method of discovering the possibility to segment customers into accurate value clusters is using the K-means clustering algorithm based on the suitable database. From the dataframe, every products have a accurate stock code. Then, researchers can divide different kinds of products into categories. To measure the distance, one uses the cosine similarity in binary-coded matrices. The method also can use to reveal K-means clustering. Then, this paper employed silhouette score to gauge the quality of segmentation. Because the unstability of K-means method, this paper included 100 iterations on 3-8 clusters to run[3].

The results are:

- For n_clusters = 3 The average silhouette_score is : 0.25673884714979833
- For n_clusters = 4 The average silhouette_score is : 0.32045152822154677
- For n_clusters = 5 The average silhouette_score is : 0.3406799276763692
- For n_clusters = 6 The average silhouette_score is : 0.38970683534465045
- For n_clusters = 7 The average silhouette_score is : 0.3687191429652772
- For n_clusters = 8 The average silhouette_score is : 0.35286912641983464

According to the above data, the calculation revealed that when n-cluster=6, it has the highest average silhouette score in all iterations. silhouette refers to the interpretation and verification method of consistency within a data cluster. This technology provides a succinct optical representation that shows the classification of each object. The silhouette value is an index that measures the degree of similarity between an object and its cluster compared to other clusters. Therefore, if the silhouette maintains a high value, it means that the object matches well with its cluster, but it matches poorly with neighboring clusters. If most objects have high values, the cluster configuration is appropriate.

From the clear blueprint from this tech, one can see charasterics of each subject. To make this tech better, reaserachers must combine the original data with the related low-deimensinal database. In order to operate sizable variables of initial matrix, PCA method can help to deal with principal components and make changes to data. Researchers can explicate client clusters in a standardized matrix and select a amount of clusters by silhouette score because the only few foremost components will be used.

Accordingly, one can find that the biggest silhouette score in the clusters.

3.4 Customer value measurement

Customer Value Measurement is to measure the value of each customer to an enterprise's business. CVM helps companies to find out people who bring what value. Also, it can use for diagnosis. Generally, Customer Value Measures are a measurement formula, combining known data values with models. The basic formula of CVM is:

CVM= profit from each customer = revenue – cost

Then, start with a base value, one needs to measure revenue and cost at the customer level

CVM = (revenue – cost) x f1 x f2 x f3…

$$f_1(\text{field}) = Y_{\text{low}} + \frac{Y_{\text{high}} - Y_{\text{low}}}{1 + \text{e}^{-(\text{field}-\text{field}_{\text{mid}})/c}} \quad (2)$$

Yellow = boost amount, c = Range of “field” for the smoothing effect

Fieldmid = value of ”field” where there is a halfway boost

The CVM model can help reaserachers to find customer purchase behavior by using the recency and frequency data. In a long-term analysis, K-means method ultimately is to cluster customers. This procedure can help to grasp sizable data of customers and cluster each customer group of the same sort. Then reseachers can discover that 75% customers have been given proper courses. As a consequence, one can know the limited data can not infer the accurate data. Based on this deviation, it would be better to have data covering a longer period of time.

4. CONCLUSION

This paper used a case to reveal the method of providing online retail business with cusumer business intelligence. The distinct customer groups dislplay clearly every customer’s consumption level and preference. Therefore, company can select different tactics for customers. From analysis, there are two foremost processes: forming a prepared databases and modeling. From the data, it can reveal each customer’s preference and frenquency in consumption. Then, in line with distinct cusumption level, company can have a better tactics on sales by predicting the preference of each customer.

ACKNOWLEDGMENTS

This paper was written to have a better understanding of my Business Analytics study. Although the research took place online, thank Professor Stephen Coggshell for permitting me to commence this paper in this instance, do the necessary research work, and use departmental data.

I am deeply indebted to Professor Stephen Coggshel, whose help, stimulating suggestions, and encouragement helped me in all the research for and writing of this thesis. Furthermore, I would like to give my special thanks to my ta YuHui Li, whose patient and enthusiasm enabled me to complete this work. Finally, I would like to thank my teacher Min Han, who looked closely at the final version of the thesis for English style and grammar, correcting both and offering suggestions for improvement.

REFERENCES

[1] Roung-Shiunn Wu, Po-Hsuan Chou, Customer segmentation of multiple category data in e-commerce using a soft-clustering approach, Elsevier, 2010, 332-333

[2] uni Nurma Sari, Lukito Edi Nugroho, Ridi Ferdiana, P. Insap Santosa, Review on Customer Segmentation Technique on Ecommerce, American Scientific Publishers, 2011, 2-4

[3] Daqing Chen, Sai Laing Sain, Kuo Guo, Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining, Macmillan Publishers, 2012, 200-201

[4] Hruschka, H. (1996). Market definition and segmentation using fuzzy clustering methods. International Journal of Research in Marketing, 3(2), 117–134.

[5] Ozer, M. (2001). User segmentation of online music service using fuzzy clustering. Omega, 29(2), 193–206.

[6] Weber, R. (1996). Customer segmentation for banks and insurance groups with fuzzy clustering techniques. In J. F. Baldwin (Ed.), Fuzzy Logic (pp. 187–196). New York, NY: Wiley.