International Journal of Information Technology and Computer Science Applications (IJITCSA)
p-ISSN 2964-3139 e-ISSN (on Process)
Vol. 1, No. 1, page 1 - 10
Submitted 30/12/2022; Accepted 31/12/2022; Published 14/01/2023

Data Analysis Using Cluster and Logistic Regression Analysis (A Case Study)

[1*]Puspa Byanjankar, [2]Kabindra Marhatta, and [3]Yushma Himanshu
[1,2,3]Department of Computer Science and Engineering, Kathmandu University, Nepal
e-mail : [1]layarsn140@gmail.com, [2]ramesh.napit@hotmail.com, [3]zadhikari@gyalzen.com

Corresponding Autor: layarsn140@gmail.com

Abstract
Customer loyalty has been a concern to C&M. C&M implements logistic regression and cluster analysis to tackle customer churn on consulting services and products. Logistic regression analysis predicts whether chemical manufacturers and small personal services will purchase consulting services and training products with discount reduction in 18 months. Their pur-chase choices every 18 months are influenced by discounts or non-discount. Cluster analysis groups purchase power based on the age group. It forecasts business client's transaction through purchase duration and frequent purchase on consulting services and items. Thus, C&M builds a long-term relationship with chemical manufacturers and small personal ser-vices by creating customer satisfaction on our consulting services and products.

Keywords— C&M, business clients, logistic regression, cluster analysis, consulting services and training product, purchase.

1. Introduction
C&M Company provides consulting and tailored teaching services [1]. It was founded by Adam Lau in 2005. It offers services such as training materials, training class, both in house or training centers and consulting services [1]. Their target business clients are small personal services industries and large chemical manu-facturers [1]. Small personal services include tuition centres, food delivery, lock-smith shops, retail and travel agencies rarely purchase products and services from C&M [1]. Chemical manufacturers produce petrochemical materials to create plastic, soaps, detergents, fertilizers, rubbers, paints and so on [2]. They do not focus on buying solely on C&M's services and products over the one to two years [1]. Thus, C&M discovers that price competition, customer retention, profit loss and digital transformation influence existing customer base on their services and product purchases.

2. Business Issues

2.1 Price Competition
Intense competition has gradually arisen in the consulting industry. Consulting firms with similar services create price wars to win over their competitors while they propose affordable pricing to their target business clients. Business clients who are price-sensitive tend to switch services to competitors who suit their budget. They attract business clients with these tactics to enhance sales performance on consulting services and training products. They desire to expand their market and company size through this act. Thus, C&M is under pressure to lower their service charges.

C&M company segments purchase behavior such as purchase duration, habits, and demographics from previous months. Purchase behavior would impact business clients purchasing decisions from C&M. business clients re-munerate consulting services and training products due to excellent customer service, reasonable price, and an effective solution for satisfying outcomes. It analyzes different product categories purchased by small

personal services and chemical manufacturers. It can provide packages for the same price as the competitors to attract business clients with different needs and demands towards sales training methods. Distinctive resellers or dis-tributes of personal services provide their services based on location. This could satisfy their customers with requirements or pricing on the product or services.

Competitors constantly introduce new packages to attract business clients. Various choices are produced for business clients for product selection. Business clients who are price sensitivity would substitute C&M with the cheapest consulting services and training products to maximize their profit returns. Logistic regression analyses whether business clients repeat products to purchase based on price. C&M provides a discount reduction for continuous purchase from business clients. For example, C&M offers discount packages to chemical manufacturers and small personal services for each purchase consists of 2 product categories. This beats competition through stimu-lating C&M's sales.

2.1.1 Profit Loss

Covid-19 has affected worldwide economic growth. Companies minimize costs to sustain their business operations. Employers in personal services industries implement salary cuts to reduce their spending on consulting services. This assists companies to overcome losses. Whereas the pandemic had granted essential businesses such as food delivery, postal, tech, telecommunications, e-commerce, medical, hotel accommodations, banking and healthcare industries to operate and restrict entertainment, chemical manufacturers, logistics, clothing, beauty and food retail business. Chemical manufacturers had reduced chemical production to lower the risk of non-profit. Their profit margin could not achieve breakeven on their sales. Hence, cautions spending of business clients is forced to shrink their demand for petrochemical related products. On the other hand, strong income business clients have purchased consulting services and products for the last 24 months.

Clustering analysis monitors sales performance for good payback. They analyze a company's credit score to determine the financial ability of the company. This shrinks the risk of revenue loss. It clusters business clients based on their financials to detect turnover on products. It discovers business clients' frequent purchases of products based on their income. High-income business clients could afford to purchase additional products in the long run. If a product has low sales, C&M delivers the best deals to business clients. Therefore, this acts as a review of customer fulfilment on C&M's services and consultants.

Logistic regression forecast whether chemical manufacturer and small personal services will cancel C&M's services and training items. To build long term relationships with chemical manufacturers and small personal services, C&M implements customer lifetime value (CLTV) to generate revenue from chemical manufacturers and small personal services. C&M target business clients with strong purchase power through attractive promotions. It en-sures business clients repeat purchase from C&M's consulting services and training products. This reduces the chances of profit loss and gains revenue for C&M.

2.1.2 Customer Retention

Existing business clients had churn on C&M's products. They had zero purchases in six (6) months, 12 months, or 18 months. C&M's product had low competence with their rivals. As C&M focused on retaining new business clients, they had failed to maintain relationships with its current customer base. Business clients are not satisfied with the services provided by C&M. This would impact C&M's profit without its customer base. Therefore, C&M reinvent their services to keep up to date with chemical manufacturers and personal services needs and requirements on their business operation.

Loyalty programs create good relationships with business clients that would bring return or new business clients for lifetime value. It determines the influence of return purchase based on the age group. Different age groups have distinct desires and needs on products purchase [3]. Young business clients are interested in buying new products introduced to them. They are willing to experience innovations invented by C&M. Middle-aged business clients have a higher urge for product purchase. They stick to familiar products that have been used for a long time. The elderly have low interest in purchasing products from C&M as they have a stable lifestyle through purchasing similar products constantly. Furthermore, C&M provides training products, but consultants should train business clients' staff with real-life use cases to address their problems. This reduces customer retention towards C&M's services.

C&M predicts whether small personal services and chemical manufacturers would purchase consulting services and products on long term subscription. This detects the highest possibility of churn on business client's behavior [4]. Business clients determine their willingness to purchase based on the product's value in terms of monetary

value and worth on usage. Customer satisfaction affects customer retention on company's products and services [4]. Business client's behavior act based on their experience with C&M's quality services.

2.1.3 Digital Transformation

Moreover, C&M lack of consultants to assist business clients for solutions while customer expectations evolve from time to time. For example, chemical manufacturers transform their operations towards environment friendly and safety on developing chemicals or medicines. This impacts C&M to gradually shift their traditional business model to the tech business model. C&M are forced to incorporate their system with technologies. They invest heavily in technology to create innovative alternatives to meet customer needs. Hence, the style of business operations changes due to pandemic.

C&M classified business clients purchase decision through online or brick-and-mortar. Spending of business clients is influenced by online research on consulting packages [5]. They consider price and products functional by comparing with various options in the online market. Regular business clients who have a high purchase rate will receive returns such as free training products delivery from C&M. However, the offline market creates business client's customer experience on direct contact with a sales representative [5].

Logistic regression analyzes whether business clients are satisfied with C&M based on their return of products. Each customer's profile consists of name, demographics, previous product transactions and keywords search on C&M's website. They analyze online transactions to customize customer experience in C&M's system. They would constantly purchase online with C&M if C&M manage to capture business clients' requirements. This prevents the chances of returned products to C&M with speedy purchase process. Therefore, it provides personalized recom-mendations such as strategies to push infrequent products according to business clients' determined budget and queries to the consultants.

2.2 Gap and limitations of the results

Pandemic has changed the traditional way of operating businesses. Consultants could not physically reach out to business clients. Chemical manufacturers and small personal services face difficulty carrying out the project with C&M. Chemical manufacturer encounter covid-19 disruption and eco-friendly challenges to operate the factory. Nevertheless, small personal services transform their selling products or business to adapt to market changes due to COVID-19. Thus, C&M had discovered business clients had low urge on purchasing C&M's consulting services and training products. This affects C&M's profit returns and customer base.

3. Methodology

3.1. Logistic Regression Analysis

This study aims to predict factors that would influence the discount buyer in the last 18 months. Similarly with regression [6], Logistic regression is a model to forecast the dependent variable's relationship with independent variables.

Table 1: The dependent and independent variable

DISBUY	BUY18
1	1
0	0
0	1
1	0
0	0
0	0

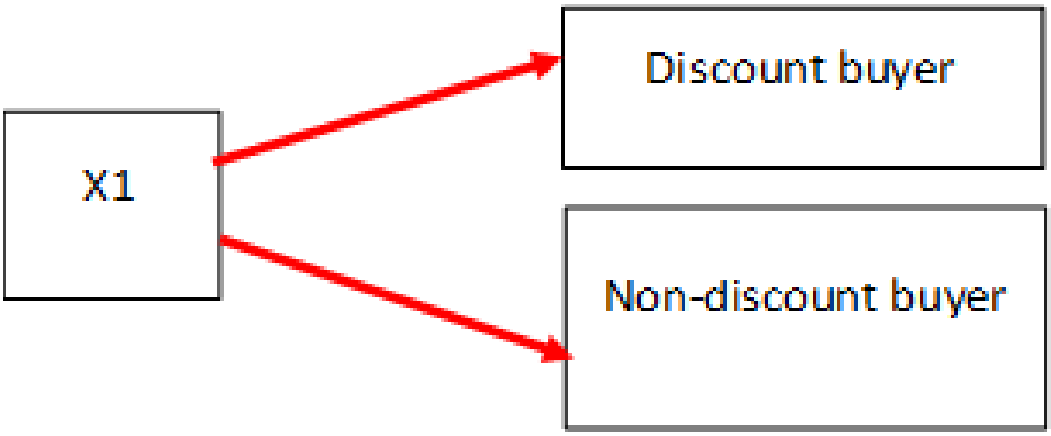

Figure 1. Logistic regression model.

The dependent variable (Y) represents DISBUY and independent variable (X) represents BUY18. BUSINESS CLIENTS had discount buyers in the past 18 months serves as 1 whereas BUSINESS CLIENTS who are non-discount buyers in the past 18 months serves as 0. This indicates the outcome of regression analysis act as binary for the dependent variable (see Figure 1). Figure 1 shows logistic regression model BUY18 that has relation with DISCBUY. It produces outcome as discount buyer and non-discount buyer. Logit function [4]:

$$logit(\pi) = \log\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \beta_1 x_1 \quad (1)$$

π indicates the probability of the event (Osborne, 2015). Probability of the event:

$$\pi(y) = \frac{e\,(\beta_0 + \beta_1 x_1)}{1 + e(\beta_0 + \beta_1 x_1)} \quad (2)$$

Table 2: Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood		Likelihood Ratio Chi-Square	DF	Pr > ChiSq
Intercept only	Intercept & Covariates			
5829.340	5826.038	3.3021	1	0.0692

3.3021 likelihood ratio chi-square indicates C&M data fits perfectly in logistic regression model. Thus, C&M do not reject business clients had discount buyers in the past 18 months.

Table 3: Analysis of Maximum Likelihood Estimates

	Intercept	BUY18
Estimate	-1.0305	0.1001
Standard Error	0.0375	0.0548
Wald Chi-Square	754.65	3.34
Pr > ChiSq	< 0.0001	0.0676
Standardized Estimate	-	0.357
Exp (Est)	0.357	1.105

β_o represents intercept and β_1 represents BUY18. BUY18 is the most significant to DISCBUY. Therefore, $P(DISCBUY = 1) = -1.0305$ + 0.1001 * BUY18. Probability of DISBUY:

$$P(DISBUY = 1) = \frac{e^{-1.0305+0.1001*(1)}}{(1 + e^{-1.0305+0.1001*(1)})} \quad (3)$$

1.105 odds ratio indicates that BUY18 changes when DISBUY increases by each unit. Hence, High purchase from business clients when C&M provides discount to business clients.

3.2. Cluster Analysis

This study aims to predict consumer churn from C&M's existing customer base. Clustering analysis segment business clients into distinct purchase power groups. These groups are group into high, medium, and low purchase power group to discover business clients purchase pattern. K Medoids Clustering Equation (X is BUY18 and Y is VALUE24):

$$\sqrt{(x1 - x2)^2 + (y1 - y2)^2} \quad (4)$$

Table 4: Root-Mean-Square Standard Deviation

Segment Id	Root-Mean-Square Standard deviation
1	0.69608
2	0.349967
3	0.775552

Table. 5: K-Medoids

Transformed: # of purchases 12 months	Transformed: # of purchases 18 months	Transformed: # of purchases 6 months	Transformed: Total value of purchases last 24 months
1.445625	1.4924213	1.424442	19.79299
1	1.004233	1	13.52331
1.177347	1.436313	1	19.64115

Table. 6: Distance of K-Medoids

Distance to Nearest Cluster	Nearest Cluster
3.375466	3
2.590178	3
2.590178	2

K-medoids calculates the distance between clusters and medoids [7]. It selects random medoids for each cluster. It compares the distance to achieve the smallest distance. The nearest cluster from distance is subdivided into groups of clusters.

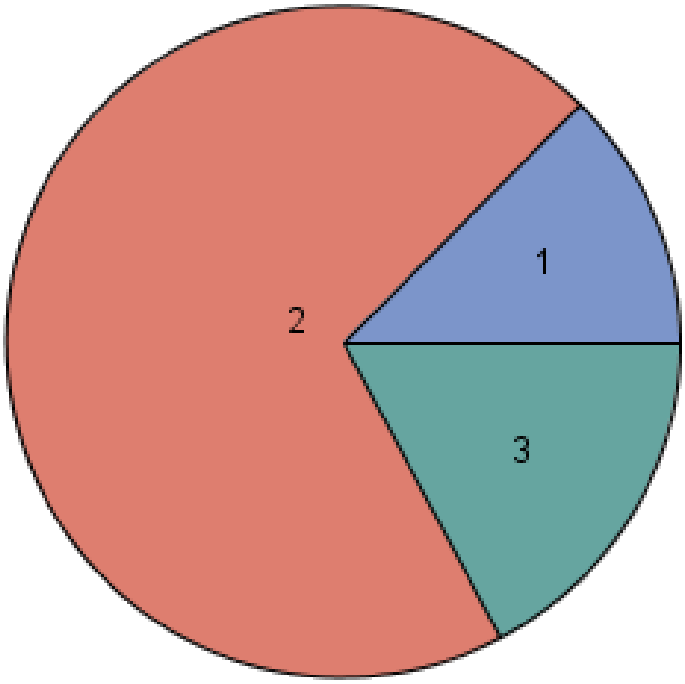

Figure 2. Segment for each cluster

4. Results and Discussion

4.1. Results

DISCBUY has 0.1001 relations with BUY18. Discount purchase influences 18 months purchase. Business clients select the most cost-effective consulting packages with discounts. They consume consulting services and training products in bulk to save company's cost. They purchase consulting services and training products online or offline. Websites provide sufficient and detailed information for online business clients to compare price on C&M's consulting packages. Purchases from business clients are not price sensitive offline. Hence, C&M could differentiate themselves from the competitors to prevent business clients churn towards competitors services. This assists C&M to rebrand their position on the consumer's market (see Figure 3).

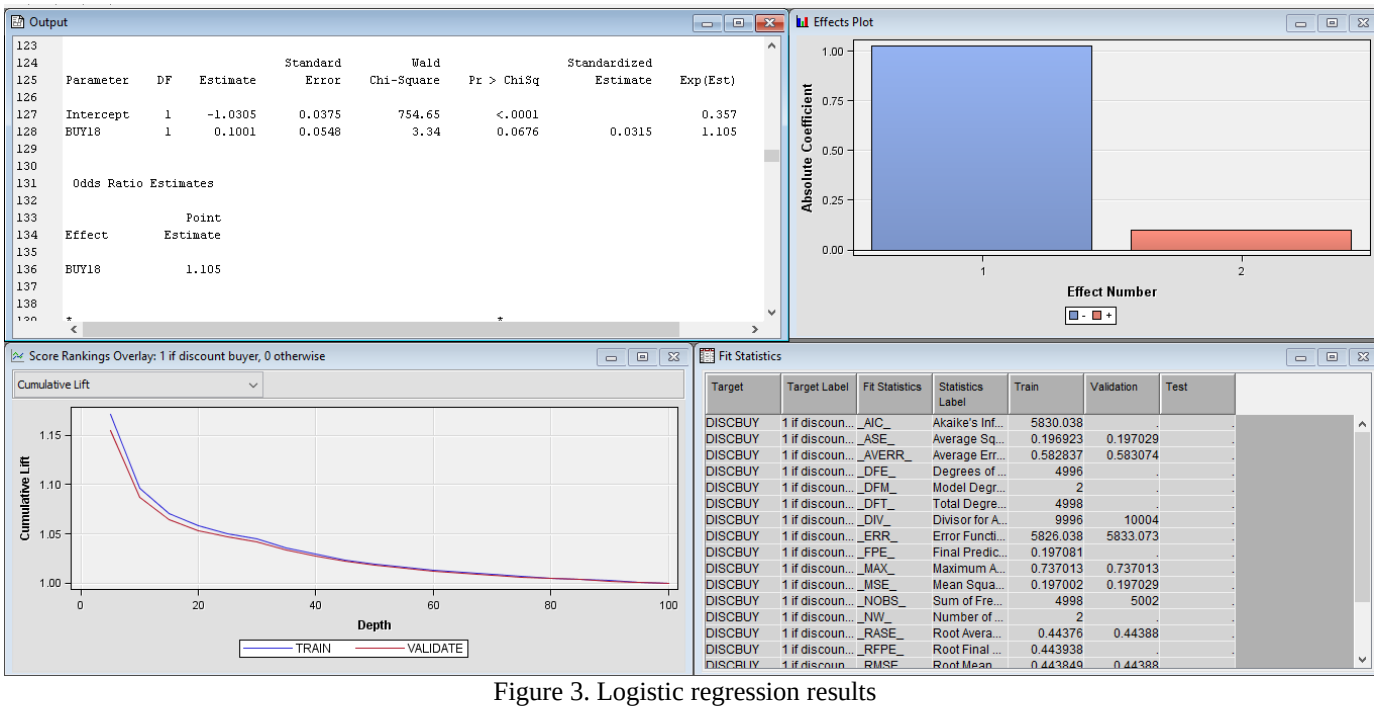

Figure 3. Logistic regression results

Figure 4 shows graph to determine income and product return from business clients based on residence location Business clients in location B have higher income ($63,521) than location A ($26,452). This indicates business clients in location B have high purchase power to buy desire product or services. High purchase power provides a wide range of products for decision making. It benefits the company to buy the best product or services and price based on their needs and budget. It can purchase the product with the lowest price provided to them. It reduces the chances of product return as it has met their requirements.

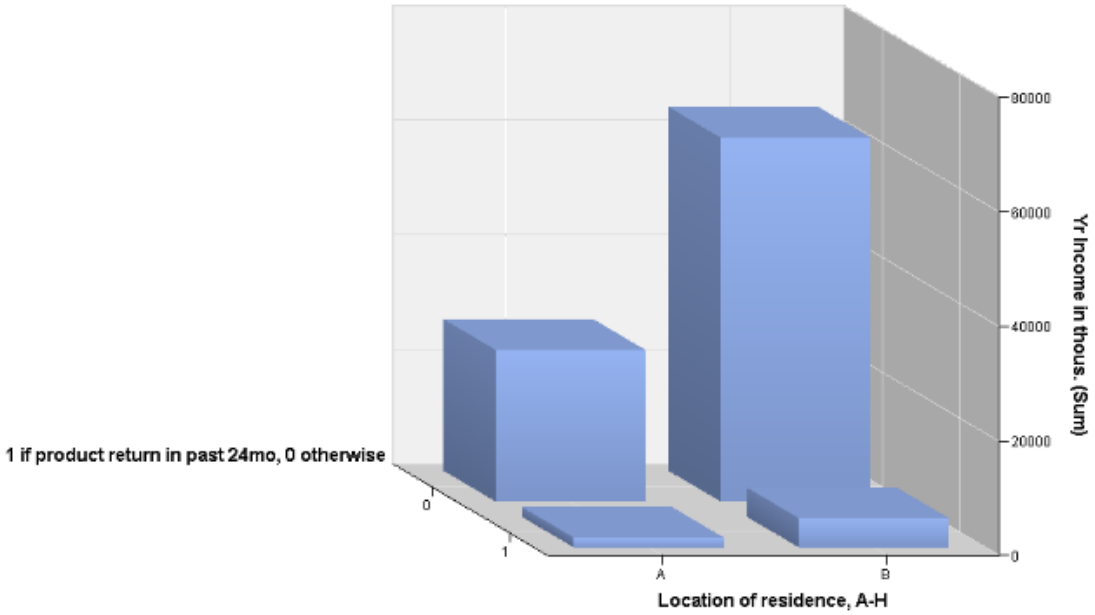

Figure 4. Graph to determine income and product return from business clients

Location A has the lowest income ($1,717) compare to location B. Low-income limits company's choices towards the product or services when there is limited budget. This will affect their satisfaction towards the product or service selections. However, company in location A sells products solely to domestic market as a small group of business clients and business purchase products from them. This has shrunk the income of the company in location A (see Figure 5).

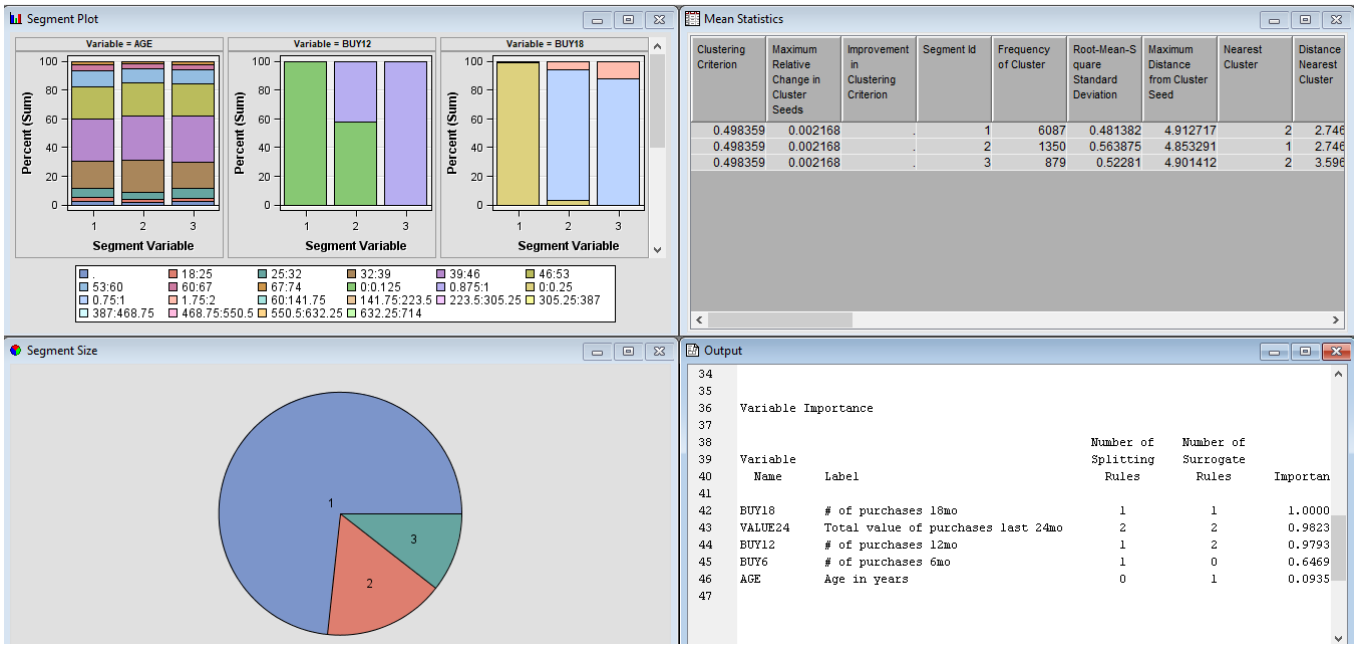

Figure 5. C&M's cluster results

C&M data are clustered into 3 segments. Segment 1 represents medium purchase power group, segment 2 represents high purchase power group and segment 3 represents low purchase power group.

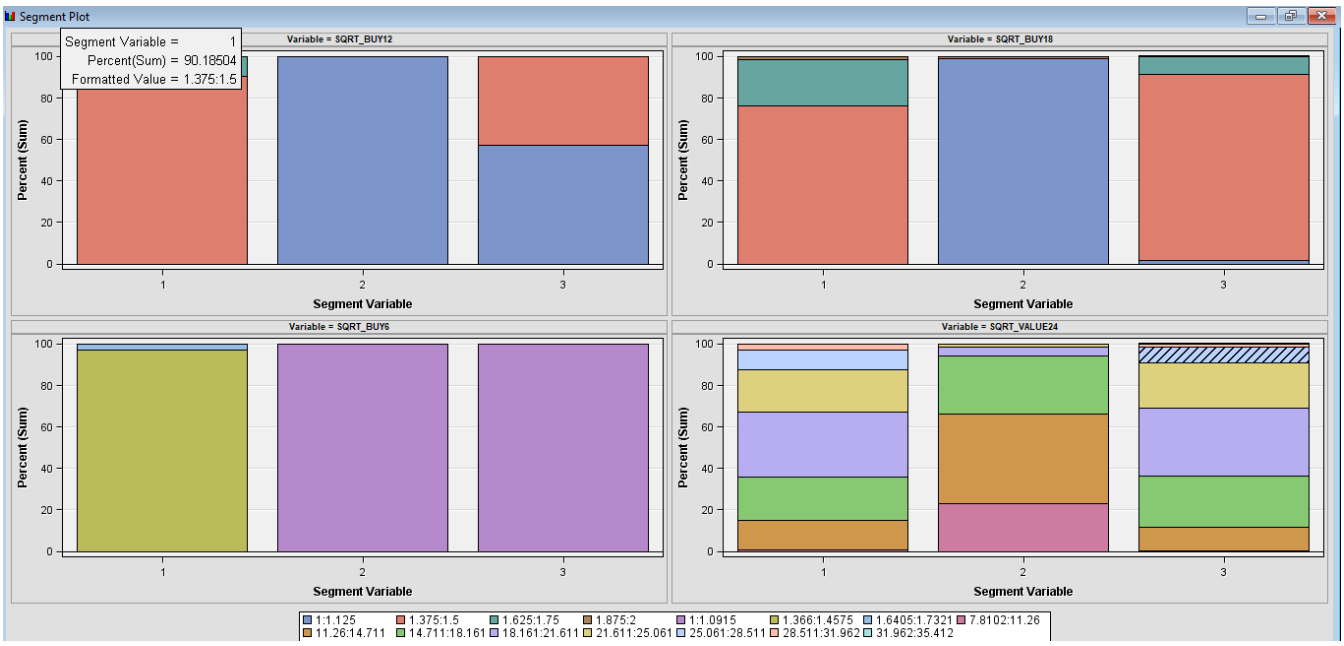

Figure 6. Segment plot of clustering C&M's data

Three (3) segments contain BUY6, BUY12, BUY18 and VALUE24 variables. C&M analyzes purchase demand for 6 months, 12 months and 18 months based on total value of purchase in 24 months. BUY18 has the highest worth in segment 1 and 2. business clients purchase the highest number of products throughout 18 months. C&M offers best deals on price to attract customers. Therefore, product sales turn into profits for C&M.

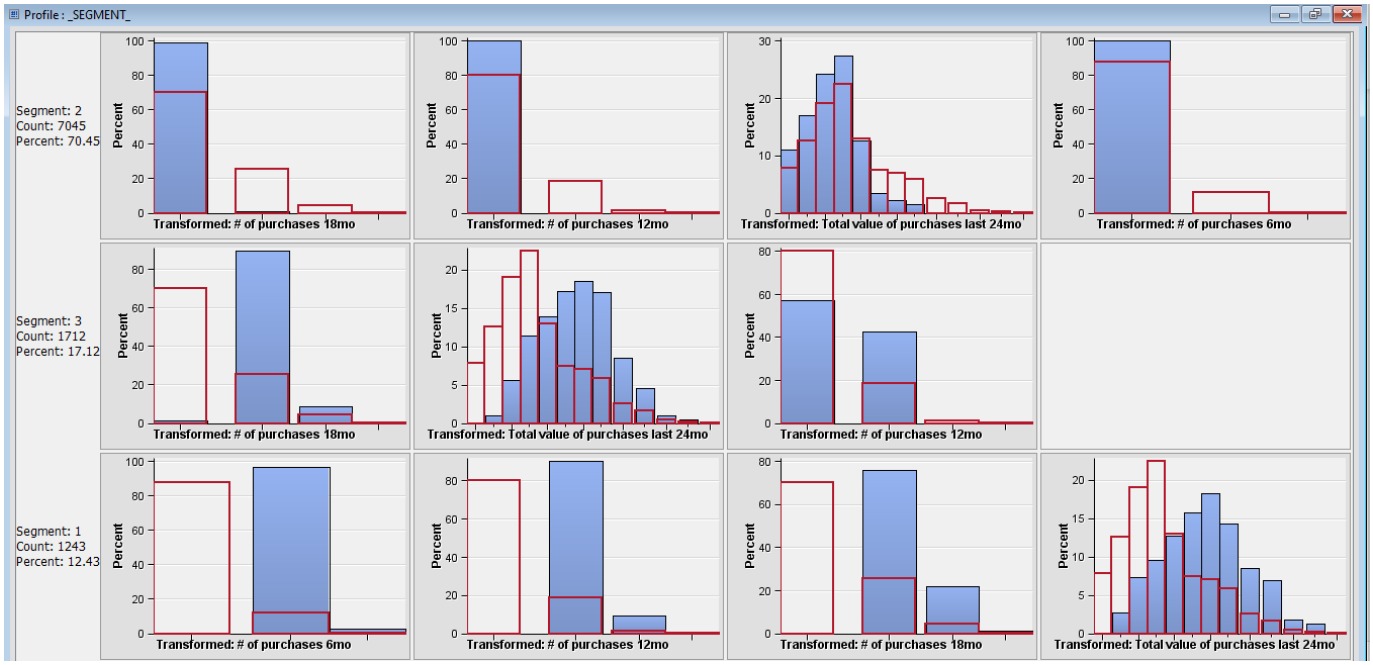

Figure 7. Segment profile of C&M data

Segment 1 known as medium purchase power group. It has highest purchase power in 6 months while high purchases in 12 and average purchase in 18 months. business clients purchase relatively low purchase in 24 months. Hence, chemical manufacturers purchase products constantly to load their stock. This creates a direct supply chain to pharmacies. Segment 2 is known as high purchase power group. It has the highest 6 months, 12 months, and 18 months purchase in C&M. Young business clients of segment 2 have strong satisfaction on C&M's innovation services for long term subscription. Segment 2 has high constant purchase power and attraction for consulting services and training products. Therefore, the highest purchase for last 24 months among all segments. Segment 3 known as low purchase power group. Small personal services have no purchase in 6 months. They have medium purchase for 18 months. Total value of purchases lasts 24 months is low. business clients of segment 3 prefer to stock up products in every one year and six months to save cost for a personal services company. They control their budget to resell products to their consumers. Hence, reasonable pricing draws customer's attention and business to the company.

Different age group purchases different quantity and category of products. Innovate services interest the location coverage of youngsters. Technologies open youngsters' minds for innovations around the world. On the contrary, business clients in the middle age stick to familiar products with high purchase of products. Elderly have low interest in products purchase. They purchase similar products consistently as they are emotionally attached to C&M's goods. Hence, C&M gathers transactions in each age group to detect trends on purchase behaviour. These transactions indicate that frequent purchases with their income. It will sustain existing business clients with returns through loyalty program.

4.2. Discussion

Researcher face issues to identify dependent variable in logistic regression. They "force quantitative data into a dichotomous variable" [8]. This creates different effects with inaccurate dependent variables. Thus, there is no occurrence of odds ratio estimates in logistic regression. Logistic regression is determined by eliminating unusable data. It limits the number of variables as inputs [9]. It rejects unusable variables and connects applicable variables to C&M. Furthermore, it must not have missing data to train and validate the model. This affects the significance of the effects. Data imputation aims to filter missing data in C&M's dataset. Hence, strong correlation influences on dependent and independent variable [10].

C&M created a consulting platform for consultants to connect and engage with business clients. The consulting platform track number of visitors when they visit the website. C&M detects and analyze business client's behavior through viewing customer database. It monitors website traffic flow through conducting live stream with customer engagement [11]. Live stream allows customers to access anywhere and anytime. Neural network recommends consulting services and packages based on customer's preferences on C&M's website. C&M classified business clients into on peak and off-peak session to detect future 1-month sales trend in C&M's website. They detect profitable sales based on location to produce effective strategies from C&M's consultant. It enhances productive sales to promote consulting services and training products online [12].

5. Conclusion

Pandemic has affected C&M's sales due to the offline approach to chemical manufacturers and small personal services. They had low purchase on 6, 12 and 18 months. C&M provide consulting platform to track timely sales trend with neural network. The customer database is transformed into a dashboard to present insights on customer purchase behaviour. In order to survive through pandemic, C&M decides to expand their services overseas. It generates extra income to support daily operations in C&M. Moreover, the researcher implements logistic regression for C&M. It occurs problems such as quantitative data as dependent variable and missing data. Therefore, reject unusable data, restrict number of input variables and data imputation solves the problem.

BIBLIOGRAPHY

[1] . R. S. Collica, Customer segmentation and clustering using SAS enterprise miner. SAS Institute Inc., 2017.

[2] . D. C. Y. Foo, "The Malaysian chemicals industry: From commodities to manufacturing," AIChE, 09-Nov-2015. [Online]. Available: https://www.aiche.org/resources/publications/cep/2015/november/malaysian-chemicals-industry-commodities-manufacturing. [Accessed: 20-Aug-2022].

[3] . R. Helm and S. Landschulze, "How does consumer age affect the desire for new products and brands? A multi-group causal analysis," Review of Managerial Science, vol. 7, no. 1, pp. 29–59, 2011.

[4] . A. Anand and G. Bansal, “Predicting customer’s satisfaction (dissatisfaction) using logistic regression,” International Journal of Mathematical, Engineering and Management Sciences, vol. 1, no. 2, pp. 77–88, 2016.

[5] . R. Archacki, K. Protextor, D. Ratajczak, and N. Rich, “Capturing the offline impact of online marketing in B2B,” BCG Global, 13-Apr-2022. [Online]. Available: https://www.bcg.com/publications/2019/capturing-offline-impact-online-marketing-b2b. [Accessed: 20-Aug-2022].

[6] . H. Surbakti, “Risk perception in the correlation between the tendency of using internet and customers’ willingness to use online payment system,” Journal of Management Information System & E-commerce, vol. 1, no. 2, 2014.

[7] . M. Z. Hossain, M. N. Akhtar, R. B. Ahmad, and M. Rahman, “A dynamic K-means clustering for data mining,” Indonesian Journal of Electrical Engineering and Computer Science, vol. 13, no. 2, p. 521, 2019.

[8] . B. E. Huitema, The analysis of covariance and alternatives: Statistical methods for experiments, quasi-experiments, and single-case studies. Hoboken, NJ: Wiley, 2011.

[9] . T. G. Nick and K. M. Campbell, “Logistic regression,” Topics in Biostatistics, pp. 273–301, 2007.

[10] . P. Ranganathan and R. Aggarwal, “Common pitfalls in statistical analysis: Linear Regression Analysis,” Perspectives in Clinical Research, vol. 8, no. 2, p. 100, 2017.

[11] . M. M. Jeon, S. (A. Lee, and M. Jeong, “E-social influence and customers’ behavioral intentions on a bed and breakfast website,” Journal of Hospitality Marketing & Management, vol. 27, no. 3, pp. 366–385, 2017.

[12] . A. Cobham and P. Janský, “Measuring misalignment: The location of US multinationals’ economic activity versus the location of their profits,” Development Policy Review, vol. 37, no. 1, pp. 91–110, 2018