Project Report

on

CUSTOMER SEGMENTATION USING MACHINE LEARNING

Submitted to

Shri Ramdeobaba College of Engineering & Management, Nagpur

(An Autonomous Institute Affiliated to Rashtrasant Tukdoji Maharaj Nagpur University)

for partial fulfillment of the degree in

Bachelor of Engineering

(Information Technology)

Eighth Semester

by

ABHISHEK VERMA

DEEPANSHU SAHU

LAXMIKANT KABRA

Under the Guidance of

Prof. Purshottam Assudani

Department of Information Technology

Shri Ramdeobaba College of Engineering & Management, Nagpur-13

2021-22

CERTIFICATE

This is to certify that the Project Report on

"CUSTOMER SEGMENTATION USING MACHINE LEARNING"

is a bonafide work and it is submitted to

Shri Ramdeobaba College of Engineering & Management, Nagpur

(An Autonomous Institute Affiliated to Rashtrasant Tukdoji Maharaj Nagpur University)

by

ABHISHEK VERMA

DEEPANSHU SAHU

LAXMIKANT KABRA

For partial fulfillment of the degree in

Bachelor of Engineering in Information Technology,

Eighth Semester

during the academic year 2021-22
under the guidance of

Prof. Purshottam Assudani
Assistant Professor
Department of Information Technology
RCOEM, Nagpur

Dr. P. D. Adane
Head, Department of Information Technology,
RCOEM, Nagpur

Dr. R. S. Pande
Principal
RCOEM, Nagpur

RCOEM
Shri Ramdeobaba College of Engineering and Management, Nagpur

Department of Information Technology

Shri Ramdeobaba College of Engineering & Management,

Nagpur-13

2021-22

DECLARATION

We hereby declare that the thesis titled **"Customer Segmentation using Machine Learning"** submitted here in, has been carried out in the Department of Information Technology of Shri Ramdeobaba College of Engineering & Management, Nagpur. The work is original and has not been submitted earlier as a whole or part for the award of any degree/diploma at this or any other Institute/University.

Date: 05/05/2022

Place: Nagpur

Abhishek Verma
Roll Number: 26

Deepanshu Sahu
Roll Number: 34

Laxmikant Kabra
Roll Number: 47

Approval Sheet

This thesis entitled **"Customer Segmentation using Machine Learning"** by Abhishek Verma, Deepanshu Sahu, Laxmikant Kabra is approved for the degree of Bachelor of Engineering in Information Technology.

Name & signature of Supervisor(s)

Name & signature External Examiner(s)

Name & signature RRC Members

Name & signature of HOD

Date: 07/05/2022

Place: Nagpur

Acknowledgements

Presentation inspiration and motivation have always played a key role in the success of any venture. It goes without saying that we are indebted to several people who have extended their cooperation and help in completing this project. We present our sincerest gratitude to all of them. First, we are thankful to **Dr. R. S. Pande,** Principal, Shri Ramdeobaba College of Engineering and Management, Nagpur, for providing the requisite facilities for the project completion no word can express our grateful thanks to our guide **Prof. P. J. Assudani** for his constant guidance, supervision and invaluable cooperation in every step of the progress of this project. We take great pleasure in acknowledging **Prof. Madhuri Tayal,** Assistant Prof. Department of Information Technology, Shri Ramdeobaba College of Engineering and Management, for their support. And at last but not least, we acknowledge our parents and all our faculty member for being such a nice source of encouragement & moral support that helped us tremendously in this project. It is our pleasure to submit this project report entitled, "CUSTOMER SEGMENTATION USING MACHINE LEARNING" in partial fulfillment of the requirement for the award of the degree of bachelor's in information technology.

Name of the Projectees

Abhishek Verma

Deepanshu Sahu

Laxmikant Kabra

CONTENTS

ABSTRACT

As the legal cannabis industry emerges from its nascent stages, there is increasing motivation for retailers to look for data or strategies that can help them segment or describe their customers in a succinct, but informative manner. While many cannabis operators view the state-mandated traceability as a necessary burden, it provides a goldmine for internal customer analysis. Traditionally, segmentation analysis focuses on demographic or RFM (recency frequency monetary) segmentation. Yet, neither of these methods has the capacity to provide insight into a customer's purchasing behavior. With the help of Front Ventures, a battle-tested multinational cannabis operator, this report focuses on segmenting customers using cannabis-specific data (such as flower and concentrate consumption) and machine learning methods (K-Means and Agglomerative Hierarchical Clustering) to generate newfound ways to explore a dispensary's consumer base. The findings are that there are roughly five or six clusters of customers with each cluster having unique purchasing traits that define them. Although the results are meaningful, this report could benefit with exploring more clustering algorithms, comparing results across dispensaries within the same state, or investigating segmentations in other state markets

LIST OF FIGURES

CHAPTER - 1
INTRODUCTION

1. INTRODUCTION

In this section we will look briefly at what can be solved for the business around which the whole project works. We would investigate details of customer segmentation and LTV into later chapters. This section would also inform data acquisition and scope for the analysis that is under consideration here.

1.1 The Business Problem

Any organization in retail, regardless of the enterprise, finally ends up gathering, creating, and manipulating 1 record over the direction in their lifespan. These records are produced and recorded in a lot of contexts, maximum drastically withinside the shape of shipments, tickets, worker logs, and virtual interactions. Each of those times of records describes a small piece of ways the organization operates, for higher or for worse. The extra get entry to records that one has, the higher the photograph that the records can delineate. With a clean photograph crafted from records, information formerly unseen start to emerge that spur new insights and innovations. The sheer length and complex nature of records withinside the actual international make the above mission a great deal simpler stated than accomplished, aleven though. The upward push of overall performance metrics and interactive dashboards have ushered in a brand-new generation of searching at records. Many times, the records protected in dashboards are on the superficial stage: How a great deal did save X make at some stage in December? What is our pinnacle five merchandise? What is our month-to-month COGS (Cost of Goods Sold)?. While dashboards deliver records that regularly have crucial significance in deliver chain control and operations, they're restricted withinside the experience that they pass over records and insights that require better stage of records mining and evaluation. Companies that make use of right records technological know-how and records mining practices permit themselves to dig in addition into their personal working techniques, which in flip lets in them to optimize their business practices. As a result, there are growing motivations for investigating phenomena and records that can't be absolutely answered: Why is product B bought extra on the primary Saturday of each month as compared to different weekends? If a patron sold product B, will they prefer product C? What are the defining tendencies of our clients? Can we be expecting what clients will need to buy? It is the latter 1/2 of the closing query so that it will be the extensive attention of this paper. In particular, this paper discusses the outcomes of a patron segmentation evaluation mission accomplished alongside 4Front Ventures. 4Front Ventures (hereby called 4Front) is a consulting and control company withinside the criminal hashish enterprise that operates numerous cultivations, manufacturing, and retail web sites

throughout the country. As 4Front maintains increasing into new markets, it's miles important for them to have a experience of who their clients are. Not simply the goods they prefer to buy, however whilst they prefer to buy them, how regularly they need to buy them, and what their lifetime price can be to the 1 In general, records manipulation isn't malicious nor consists of. malintent in its nature. It is the easy procedure of changing records from one layout right into a extra usable, beneficial one. four organization. While a number of those questions are extra uncomplicated than others, it's miles clean that all of them require records munging, evaluation, and presentation that contain talents and strategies past what's required of a conventional analyst. By integrating gadget learning2 practices and traditional commercial enterprise understandings, the trails to answering those questions have become extra intertwined with that of a comparable query: What segments or businesses of clients can we have? After analyzing clustering and analyzing approximately it in several different contexts, it has become clean that clustering 4Front's retail clients have become one manner to research the shopping styles and behaviors of its clients.

1.2 Acquisition of Data

Finding readied, usable records for evaluation in a commercial enterprise context is a rarity. As such, it's miles vital to gather as a great deal record as feasible, however additionally in a layout that meets a huge form of financial, moral, and computational issues. But earlier than discussing those, it's miles first crucial to explain the approaches wherein the applicable retail records are saved and applied throughout the organization. Without delving into personal information, the extensive concept is that most retail records are saved in numerous SQL databases. Because of emphasis on seed-to-sale traceability, numerous country guidelines, and absence of opposition withinside the software program marketplace, maximum groups are required to combine their complete commercial enterprise up to at least one factor of sale (POS) gadget this is regular throughout the organization. If the organization is vertically incorporated, the POS extends to their cultivation and manufacturing software program. Some software program providers, including Bio Track, Green bits, Viridian, have flourished withinside the enterprise through offering absolutely incorporated software program referred to as seed-to-sale systems. In the backend, servers save their records in SQL databases constructed to conform with country guidelines and standards. On the front end, they supply applicable records or perception thru interactive dashboards, reporting modules, or easy visuals to retail managers or analysts. As a right of way result of 4Front's a success growth into new and growing markets, they've incurred unexpected demanding situations with records managing

and garage. Even alevin though the methods and techniques that 4Front makes use of to promote and marketplace their merchandise are, for the maximum part, regular throughout country markets, their records garage and records accessibility is contingent upon their markets and get entry to third-birthday birthday celebration software program. Certain software program, whilst bearing in mind first-rate 2Generally, gadget getting to know is the technological know-how of making and the usage of fashions and algorithms that are expecting or organization records in a statistically significant manner. It is a subset of synthetic intelligence (AI). five reviews and key visuals do now no longer have any constructed-in backend capability for outlets to get entry to the uncooked records. Luckily, one of the software programs utilized in more than one 4Front's working markets lets in for a backend SQL editor that lets in for direct queries, alevin though there may be very restricted documentation at the database shape is supplied through the software program organization. Nonetheless, it's miles feasible for patron records to be amassed for clients who exist withinside the country markets with the precise seed-to-sale software program. However, simply getting access to the records/understanding wherein it's miles a small step withinside the universal records amassing procedure. Roughly speaking, it's miles feasible to categories the numerous records acquisition approaches into 3 wonderful categories. First, it become vital to set up any moral issues or constraints to using records. When first-time clients input a dispensary, they're supplied with a shape that asks for verifiable demographic facts including their call, age, and address. In addition, they're additionally requested in the event that they consent to the organization the usage of their records for evaluation and advertising functions. Each patron's solution to the preceding query is one-warm encoded into the database: zero for "no" and 1 for "yes". The clients protected on this evaluation, thus, are simplest the clients who answered "yes" to the query and feature a 1 for the price for the precise feature. Furthermore, to defend the anonymity of every of the clients, it's also vital to prune away all touchy facts from every patron. In different words, the simplest demographic/touchy facts of every patron that the evaluation will use is the age of the patron. The sex, address, call, and different touchy or private facts is indifferent from the patron at some stage in evaluation. Each patron is uniquely diagnosed with an ID that lets in for regular evaluation, however the IDs are generated internally, because of this that that the patron has no expertise in their ID. Essentially, whilst there may be a manner for this system to hold song of a specific patron's purchases, it isn't feasible for this system to consist of clients who do now no longer consent to the usage of their records for this purpose, or for this system to tie the purchases to a specific call or address3 . Second, gathering the records in a green way closely is based on a robust expertise of the shape of the database. Without revealing too many information, there had been 4 crucial records tables withinside the

database that contained applicable facts. • The clients desk consists of the patron id, variety of visits, overall quantity spent, whether or not they consent to us the usage of their records, and three It is feasible, alevin though, to tie these facts outdoor of the context of this system. Namely, through having access to the database in a one-of-a-kind manner without regard to the above issues. 6 a long time of every patron. These records are wished for figuring out precise clients and offering the beginnings of a number of the records utilized in clustering. • The tickets desk contained all facts concerning tickets four, including the price tag ID, time of transaction, overall quantity spent, the patron ID worried withinside the price tag, and which worker finished the price tag. The time of transaction, overall quantity spent, and related patron ID are applicable for this evaluation. • The income desk hosts records associated with every character sale (i.e., every character product sold). This includes a income ID, the price tag ID that the sale is related to, the fee of the sale (fee of the object), and the product ID related to the sale. This desk consists of many IDs and different records that intersect with different tables which might be crucial for this evaluation. From this desk, it's miles feasible to acquire nearly all of the applicable records for every price tag/patron. • The merchandise desk consists of the vital facts approximately every product the shop has, including its ID, whilst it become brought to the gadget, and which product class five it belongs to. This desk is often used for debugging functions and for offering a few contexts that makes it simpler to become aware of and classify merchandise. Lastly, there had been sure computational issues to do not forget whilst gathering records as well. Though the database is installation to deal with lacking values already, there had been numerous columns in numerous tables that had malformed or lacking values that required extra attention. Incorrect self-pronounced dates, voided tickets, and tickets with $zero in income had to be pruned from the dataset. In addition, any applicable discipline with a lacking or bad price had to be pruned or corrected from the dataset. Though the variety of affected times is small, it become important to deal with those malformed times due to the fact they avoided clean evaluation later. After taking the above approaches and issues into account, it become feasible to gather the applicable records in a unmarried question the usage of the software program’s SQL editor. The records become then outputted right into a CSV file (with round 250,000 rows) for smooth viewing, importing, and evaluation. four in retail jargon, a price tag is largely a receipt. It is an evidence of transaction. five It might be found out later those the preliminary product class assignments are incomplete/tough to parse. It is vital to gather those now to give you a better manner to categories merchandise withinside the latter elements of this mission.

1.3 Scope of Analysis

In general, the techniques used to acquire the records for this mission can effortlessly be prolonged into different applicable contexts/analyses. While there may be clean price in the usage of the equal records to research shopping styles or to construct an object primarily based totally collaborative filtering recommender gadget, neither of those is the point of interest for this paper. The scope of the paper is restricted to the subsequent 4 intertwined goals:

1. To cluster clients primarily based totally on not unusual place shopping behaviors for destiny operations/advertising initiatives

2. To comprise first-class mathematical, visual, programming, and commercial enterprise practices right into a considerate evaluation this is understood throughout a lot of contexts and disciplines

3. To check out how comparable records and algorithms will be utilized in destiny records mining initiatives.

4. To create an expertise and notion of ways records technological know-how may be used to resolve actual-international troubles Before delving into the information of the mission and its implications, the following bankruptcy discusses what patron segmentation evaluation is and the motives for its importance.

CHAPTER - 2

LITERATURE REVIEW

2. LIRERATURE REVIEW

A literature review is an overview of the works that have been published previously regarding a specific topic. This helps you to get an idea of the solution that other people have given so far and how you can improve it and take it to the next level.

2.1 Review of Customer Segmentation Technique on Ecommerce

Ecommerce transactions are no longer a new thing. Many people shop with ecommerce and many companies use ecommerce to promote and to sell their products. Because of that, overloading information appears on the customers' side. Overloading information occurs when customers get too much information about a product then feel confused. Personalization will become a solution to overloading problem. In marketing, personalization technique can be used to get potential customers in a case to boost sales. The potential customer is obtained from customer segmentation or market segmentation. This paper will review customer segmentation using data, methods and process from a customer segmentation research. The data for customer segmentation were divided into internal data and external data. Customer profile and purchase history were treated as the internal data while server log, cookies, and survey data were as the external data. These data can be processed using one of several methods: Business Rule, Magento, Customer Profiling, Quantile Membership, RFM Cell Classification Grouping, Supervised Clustering, Customer Likeness Clustering, Purchase Affinity Clustering and Unsupervised Clustering. In this paper, those methods were classified into Simple technique, RFM technique, Target technique, and Unsupervised technique and the process was generalized in determining the business objective, collecting data, data preparation, variable analysis, data processing, and performance evaluation. Customer behavior in accessing ecommerce when viewing a product on ecommerce was recorded in server log with time. Duration when seeing the product can be used as customer interest in the product so that it can be used as a variable in customer segmentation.

Basically, they used RFM technique to depict a segmentation model. But only theoretical approach has been given and no practical implementation has been conducted or proposed.

2.2 Customer Segmentation using Machine Learning

The emergence of many competitors and entrepreneurs has caused a lot of tension among competing businesses to find new buyers and keep the old ones. As a result of the predecessor, the need for exceptional customer service becomes appropriate regardless of the size of the business. Furthermore, the ability of any business to understand the needs of each of its customers will provide greater customer support in providing targeted customer services and developing customized customer service plans. This understanding is possible through structured customer service. Each segment has customers who share the same market features. Big data ideas and machine learning have promoted greater acceptance of automated customer segmentation approaches in favor of traditional market analytics that often do not work when the customer base is very large. In this paper, the k-means clustering algorithm is used for this purpose. The Sklearn library was developed for the k-Means algorithm (found in the Appendix) and the program is trained using a 100-pattern two-factor dataset derived from the retail trade. Characteristics of average number of customer purchases and average number of monthly customers.

Customer segmentation is done using K-Means and elbow criterion method. Imbalanced dataset was used. Result are not documented properly.

2.3 A Systematic Approach to Customer Segmentation and Buyer Targeting for Profit Maximization

Nowadays, maintaining customer loyalty and attention span of the customers are major challenges faced by the retail industry. This leads to the need for reinforcement of marketing strategies from time to time. This paper proposes a systematic approach for targeting customers and providing maximum profit to the organizations. An important initial step is to analyze the data of sales acquired from the purchase history and determine the parameters that have the maximum correlation. Based on respective clusters, proper resources can be channeled towards profitable customers using machine learning algorithms. K-Means clustering is used for customer segmentation and Singular Value Decomposition is used for providing appropriate recommendations to the customers. This paper also deals with the drawbacks of the recommender system like cold start problem, sparsity, etc and how they can be overcome.

Used K-Means clustering for customer segmentation and Singular Value
Decomposition is used for providing appropriate recommendations to the customers

2.4 RFM ranking–An effective approach to customer segmentation

The efficient segmentation of customers of an enterprise is categorized into groups of similar behavior based on the RFM (Recency, Frequency and Monetary) values of the customers. The transactional data of a company over is analyzed over a specific period. Segmentation gives a good understanding of the need of the customers and helps in identifying the potential customers of the company. Dividing the customers into segments also increases the revenue of the company. It is believed that retaining the customers is more important than finding new customers. For instance, the company can deploy marketing strategies that are specific to an individual segment to retain the customers. This study initially performs an RFM analysis on the transactional data and then extends to cluster the same using traditional K-means and Fuzzy C- Means algorithms. In this paper, a novel idea for choosing the initial centroids in K- Means is proposed. The results obtained from the methodologies are compared with one another by their iterations, cluster compactness and execution time.

RFM analysis technique was used for segmentation. K-means and Fuzzy C-Means were implemented and concluded K-Means as a better choice of algorithm.

2.5 How to strengthen customer loyalty, using customer segmentation

Do you provide exceptional customer service?", "Is the customer service in your company extraordinary?" "How to convert satisfied customers into loyal customers?" - are the most frequent questions of today's managers and have driven the research on this article to getting the answer to a highly important marketing topic "How to strengthen customer loyalty using customer segmentation?". Anyone who has bought a product, or a service has probably suffered at least once from a company's apparent indifference to what should be its first concern: the customer experiences. If this is the case, the company is in a wrong direction, since loyalty is the most powerful tool in today competitive market. To strengthen the bonds with these high-profit customers, innovative companies are deploying enterprise wide strategies built on consumer segmentation.

This paper provides a prerequisite for customer segmentation dataset based on customer satisfaction, recommendation and future expectation.

2.6 Customer Segmentation using Machine Learning

In recent years, every e-commerce enterprise focuses on Customer Relationship Management(CRM) to provide the better services to the customer as compared to their competitors. Building a better relationship with customer help the enterprises in increasing profit and customers retention and satisfaction. It is necessary for enterprises to identify the potential customers in the market by mining the customer data to gain profitable insight. One of the efficient ways to identify the different customer characteristics is by applying clustering analysis. In this paper, different clustering approach has been presented in order to segment the customer and apply the different marketing strategies accordingly. The possibility of hybrid combination of clustering algorithm can outperform individual model has also been discussed.

This paper talks about dividing a market into different buyers with different behaviors by usage of CRM as part of the organizations business strategy for enhancing customer service satisfaction by using different types of clustering algorithm like Affinity Propagation Algorithm, Density Based Clustering, etc.

CHAPTER - 3

METHODOLOGY

3. METHODOLOGY

This section deals with the steps/process/methodologies used in making of Customer Segmentation. It also discusses about the technologies used in each step. It starts from data collection, data cleaning, analysis on data and the final representation of the result of the various analysis done on the given customer data.

3.1 DATA COLLECTION

As we know, machines initially learn from the data that we give them. It is of the utmost importance to collect reliable data so that your machine learning model can find the correct patterns. The quality of the data that you feed to the machine will determine how accurate your model is. If you have incorrect or outdated data, you will have wrong outcomes or predictions which are not relevant.

To make sure we use data from a reliable source, as it will directly affect the outcome of your model. Good data is relevant, contains very few missing and repeated values, and has good representation of the various subcategories/classes present.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850	United Kingdom
5	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12/1/2010 8:26	7.65	17850	United Kingdom
6	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	12/1/2010 8:26	4.25	17850	United Kingdom
7	536366	22633	HAND WARMER UNION JACK	6	12/1/2010 8:28	1.85	17850	United Kingdom
8	536366	22632	HAND WARMER RED POLKA DOT	6	12/1/2010 8:28	1.85	17850	United Kingdom
9	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	12/1/2010 8:34	1.69	13047	United Kingdom

Figure 3.1 Snapshot of data

3.2 DATA PREPROCESSING

After we have your data, we must prepare it. We can do this by:

Putting together all the data you have and randomizing it. This helps make sure that data is evenly distributed, and the ordering does not affect the learning process.

Cleaning the data to remove unwanted data, missing values, rows, and columns, duplicate values, data type conversion, etc. You might even have to restructure the dataset and change the rows and columns or index of rows and columns.

Visualize the data to understand how it is structured and understand the relationship between various variables and classes present.
Splitting the cleaned data into two sets - a training set and a testing set. The training set is the set your model learns from. A testing set is used to check the accuracy of your model after training.

3.3 DATA ANALYSIS

Data analysis is a process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, and is used in different business, science, and social science domains. In today's business world, data analysis plays a role in making decisions more scientific and helping businesses operate more effectively.

3.3.1 BRIEF INTRODUCTION

For a retailer, understanding the components of their consumer base is key to maximizing their potential in a market; the retailer that attracts the most customers will acquire the most market share, citrus paribus. In fact, the high costs of gaining a new customer or getting back an old customer force retailer to seriously consider how to allocate resources to optimize not just volume of customers, but the retention of them as well 7. Additionally, it is a common understanding in the retail industry that the Pareto Principle—more likely than not—applies to the company: 80% of profits come from 20% of the customers. One crucial reason why this principle holds is because retail businesses thrive on repeat purchases. As a consequence, a net change of one customer can significantly impact a business' profit in the long run. Therefore, it is generally in the best interest of the retailer to devote efforts to retaining customers by understanding them on as deep of a level as necessary. However, examining the intricate, rich relationships between a retailer and their consumer base involves understanding how different components of the base behave. Namely, how different segments of customers act similarly or differently from other segments10. One method of approaching customer understanding is through the lens of customer segmentation. In short, customer segmentation analysis is the process of grouping customers in such a way that customers within one particular group are similar to each other but different from customers in other

groups. In general, there are two paths of segmentation: a priori and post hoc. A priori analysis involves creating the segments beforehand and then, after examining data, placing each customer within the segments11. Rather than having the customer data dictate the types of segments formed, certain outside knowledge or structure would dictate the preferred segmentations. As such, the key unit of analysis here are the created segments, not necessarily the customers themselves. On the other hand, post hoc analysis leverages the data to form the segments, rather than the other way around. In a sense, post hoc analysis is a direct consequence of advancements in data collection and reliability whereas a prior analysis arose to prominence several years before such beneficial advancements. Regardless of the context, advancing technology has opened doors for post hoc analysis to succeed as a segmentation method in the retail industry. So, modern retailers and data scientists tend to perform customer segmentation using techniques residing under the post hoc umbrella, which will be the focus of the remainder of the paper. While the goal of customer segmentation analysis has been consistent among retailers for many years, approaches in the past relied on much weaker analytical techniques than available today. It is nonsensical to blame companies in the past who failed to utilize their data properly; the technology and data infrastructure simply were not ubiquitous or cheap enough to allow for companies to collect massive amounts of data as they do today. Yet, many companies still found rudimentary methods to attempt to understand their customers, the most traditional involving purely demographic analysis. Demographic analysis is segmenting customers solely on demographic features, such as age, sex, race, or income. It is built upon the assumption that retail behavior is defined by the demographics of the surrounding neighborhood of a store's consumer base. The distillation of customers to only a few well-understood and categorized demographic features meant it was easier for retailers to collect and utilize data from their customers, since it was relatively easy to take a limited number of specific characteristics and generate reasonable predefined categories. Furthermore, demographic analysis also thrived because it became a quick, cheap, and easy model to predict how new customers would interact. So, demographic segmentation allowed for retailers to collect only relevant data—which in turn requires minimal labor and thus cost—that kept analysis and communication of the analysis at a common level. Despite the success of many popular marketing firms, the increasing accessibility of retail technology revealed that demographic segmentation had no capacity to produce insight with consumer purchase histories. Once retailers and marketing researchers began to tinker with different methods of segmentation, it became clearer sooner rather than later that deeper behavioral

segmentation would quickly supersede purely demographic segmentation. Instead of attempting to divide customers based on their demographics, retailers began segmenting their customers based on their purchasing patterns, mostly using a technique known as the Recency-Frequency Monetary (RFM) method

3.4 MODULE 1: RFM ANALYSIS

A standard implementation of the RFM model is cheap and simple: once each of the components are defined in a way that makes them easy to collect, it is a relatively menial task for a retailer to visualize the results, which makes interpretation easy as well. Usually, the results of an RFM analysis would include three plots—one for each combination of two variables (e.g., Recency and Frequency)—with the inferred segments and their defining characteristics. RFM analysis became a staple of modern marketing for its simplicity and its cheap cost to implement as well to communicate efficiently16. In a way, the visualization aspect alone gave utility to the RFM model, allowing managers to effectively glean insights from the analysis. Yet, as the retail industry evolved in parallel with the technology boom, it became dramatically easier for retailers to collect data at a larger scale, which also meant it became easier to mine at a larger scale. In the case of the cannabis industry, the mandate that each operator must have a secure and sound traceability system allows operators— who know how to access their data— virtually unlimited potential in performing higher level analysis. While RFM modelling is based on only three features, modern customer segmentation can involve several hundred or even several thousand features. As a result, the segments of the analysis become much finer, much richer to allow retailers to understand their customers at levels simply unattainable from RFM or demographic analysis. One of the more popular ways retailers have been able to acquire such specific data regarding their customers is through a loyalty program 18. In a loyalty program, the customer benefits by receiving certain discounts, but the natural by-product19 of the loyalty card is the data that the retailer can mine to better serve their customers and boost profits 20. By using this data, retailers can create specific marketing campaigns, target certain customer segments with uniquely tailored discounts, or even invite old customers back into the store. This data allows for retailers to conduct ultra-specific marketing strategies that has transformed the way retailers compete in the age of Big Data. To perform customer segmentation analysis at a high level, retailers have begun to incorporate aspects of machine learning into the analysis of

their customers. More specifically, retailers are utilizing unsupervised machine learning tools such as clustering and dimensionality reduction to approach

analysis in ways that cannot be matched without machine learning. Instead of focusing on only a few features or customers at a time, it is possible to write programs and implement algorithms that can consider several more features or several more instances than traditional spreadsheets can hold or process. Because of this massive potential, retailers across all industries are attempting to leverage clustering algorithms such as K-Means or hierarchical clustering to segment their customers more accurately and quickly. The faster and better retailers can cluster their customers, the quicker they can market to them and thus acquire market share.

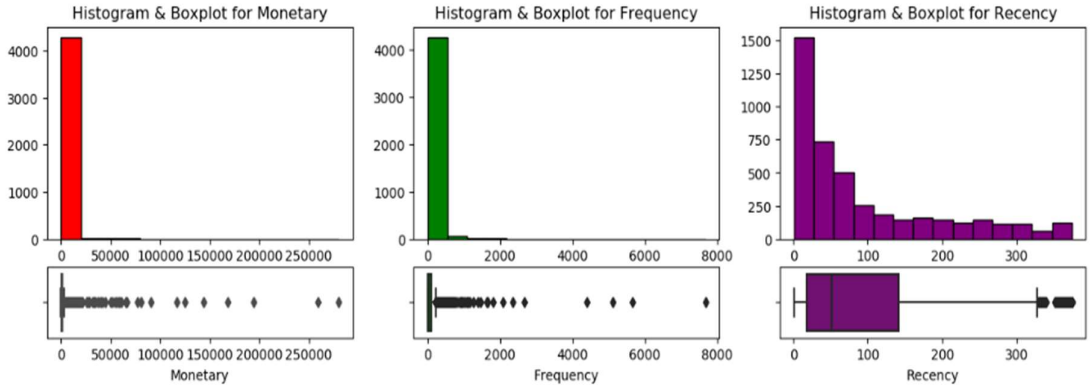

Figure 3.2 RFM data distribution

3.4.1 CHALLENGES OF PERFORMING ANALYSIS

The benefits of customer segmentation analysis are clear. By having a stronger understanding of their consumer base, retailers can properly allocate resources to collect and mine relevant information to boost profits. However, getting to the point of performing high-level customer segmentation analysis is more difficult than originally thought for many retailers. Many retailers may have the rights to the necessary data to perform the analysis, but do not have either the ability to access it in a user-friendly manner or have an employee that has the skillset to work with it. The lack of proper personnel or equipment to handle the necessary volume of data is perhaps the biggest hindrance to smaller firms being able to perform such analysis. The popularity of opensource programming software such as R or Python has certainly helped make this type of analysis more accessible, but it still would require retailers having someone on their team who can code in either of those languages. Additionally, some retailers are simply unaware of either the extent of their data collection or are not yet inspired to dig

into it. Nevertheless, retailers that have not fully adopted customer segmentation analysis are likely not doing so simply because they cannot afford to spend the time, money, or labor to perform the analysis. Therefore, it is an aim of this paper to show that this rich analysis can be performed cheaply and efficiently. However, there is a far subtler but still consequential reason why retailers do not implement customer segmentation analysis: it is too complicated to understand. When compared to traditional demographic segmentation or RFM analysis, high-level customer segmentation analysis requires far more precise knowledge of machine learning and the mathematics that describe how the algorithms work. In addition, traditional marketing analysts are not equipped with the math or programming skills necessary to successfully implement customer segmentation analysis with machine learning methods 21; similarly, programmers and data analysts are not well-suited to handle marketing tasks. This poses another conundrum as it involves transforming a typical marketing assignment—segmenting customers based on purchasing behaviors— into a purely programming one, which means the marketing team does not have the skills to code it up themselves, but the programming team does not have the marketing skills to interpret the results. Hence, there is a necessity for a hydride role that involves knowledge of the business, programming, and marketing. In modern workspaces, this role is dubbed the data scientist or information specialist. In sum, customer segmentation analysis is the process of trying to understand a consumer base by splitting it up into segments. While traditional analysts found some success with demographic or RFM analysis, these models simply do not have the technological capabilities to provide rich insight into more specific details regarding the customers. On the other hand, customer segmentation analysis that is combined with machine learning methods can transform the way a retailer thinks about their data. As such, retailers are trying to find cheap, easy ways to implement and communicate how clustering can be used to segment their customers. Now that there has been plenty of introduction into customer segmentation analysis, it is time to look under the hood of some clustering algorithms before finally engaging in discussion of the analysis.

3.4.2 INTERQUARTILE RANGE

In statistics, a trimmed estimator is an estimator derived from another estimator by excluding some of the extreme values, a process called truncation. This is generally done to obtain a more robust statistic, and the extreme values are considered outliers. Trimmed estimators also often have higher efficiency for mixture distributions and heavy-tailed

distributions than the corresponding untrimmed estimator, at the cost of lower efficiency for other distributions, such as the normal distribution.

Given an estimator, the n% trimmed version is obtained by discarding the n% lowest and highest observations: it is a statistic on the middle of the data. For instance, the 5% trimmed mean is obtained by taking the mean of the 5% to 95% range. In some cases, a trimmed estimator discards a fixed number of points (such as maximum and minimum) instead of a percentage.

The median is the most trimmed statistic (nominally 50%), as it discards all but the most central data, and equals the fully trimmed mean – or indeed fully trimmed mid-range, or (for odd-size data sets) the fully trimmed maximum or minimum. Likewise, no degree of trimming has any effect on the median – a trimmed median is the median – because trimming always excludes an equal number of the lowest and highest values.

Most often, trimmed estimators are used for parameter estimation of the same parameter as the untrimmed estimator. In some cases the estimator can be used directly, while in other cases it must be adjusted to yield an unbiased consistent estimator.

For example, when estimating a location parameter for a symmetric distribution, a trimmed estimator will be unbiased (assuming the original estimator was unbiased), as it removes the same amount above and below. However, if the distribution has skew, trimmed estimators will generally be biased and require adjustment. For example, in a skewed distribution, the nonparametric skew (and Pearson's skewness coefficients) measure the bias of the median as an estimator of the mean.

When estimating a scale parameter, using a trimmed estimator as a robust measure of scale, such as to estimate the population variance or population standard deviation, one generally must multiply by a scale factor to make it an unbiased consistent estimator.

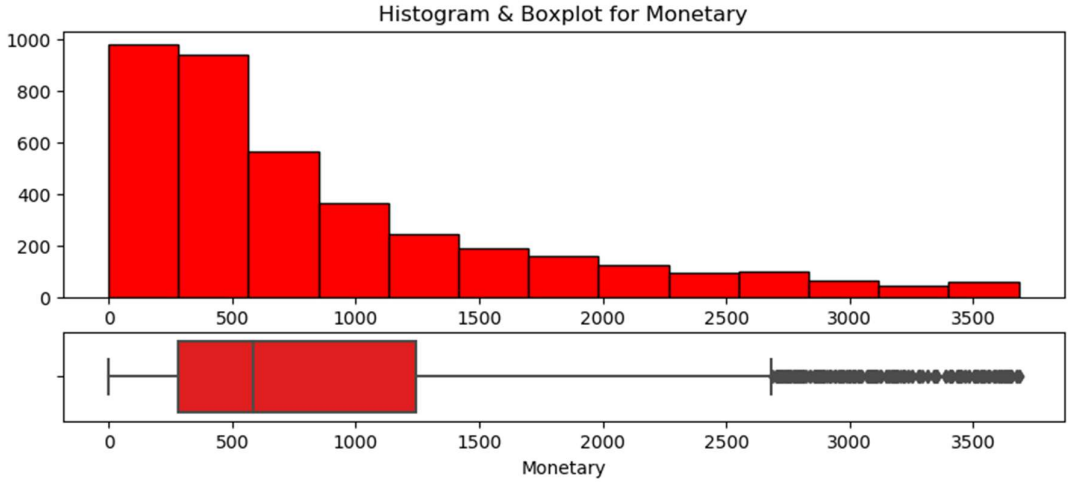

Figure 3.3 Monetary

3.4.3 HOPKINS STATISTIC

The Hopkins statistic (introduced by Brian Hopkins and John Gordon Skellam) is a way of measuring the cluster tendency of a data set. It belongs to the family of sparse sampling tests. It acts as a statistical hypothesis test where the null hypothesis is that the data is generated by a Poisson point process and are thus uniformly randomly distributed. A value close to 1 tends to indicate the data is highly clustered, random data will tend to result in values around 0.5, and uniformly distributed data will tend to result in values close to 0.

$$H = \frac{\sum_{i=1}^{m} u_i^d}{\sum_{i=1}^{m} u_i^d + \sum_{i=1}^{m} w_i^d}$$

ui - nearest neighbour distances from uniformly generated sample points to sample data from given dataset,

wi - nearest neighbour distances within sample data from given dataset.

We found our Hopkins value nearly equal to 0.932 that means our data is highly cluster able.

3.4.4 K-MEANS ALGORITHM

K-means is a technique for data clustering that may be used for unsupervised machine learning. It can classify unlabeled data into a predetermined number of clusters based on similarities (k).

The K-means clustering algorithm computes centroids and repeats until the optimal centroid is found. It is presumptively known how many clusters there are. It is also known as the flat clustering algorithm. The number of clusters found from data by the method is denoted by the letter ‘K’ in K-means.

In this method, data points are assigned to clusters in such a way that the sum of the squared distances between the data points and the centroid is as small as possible. It is essential to note that reduced diversity within clusters leads to more identical data points within the same cluster.

3.4.5 SILHOUTTE ANALYSIS

Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1].

Silhouette coefficients (as these values are referred to as) near +1 indicate that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster.

We ran a loop from 2 to 15 and applied silhouette analysis for each point and taken the best among them.

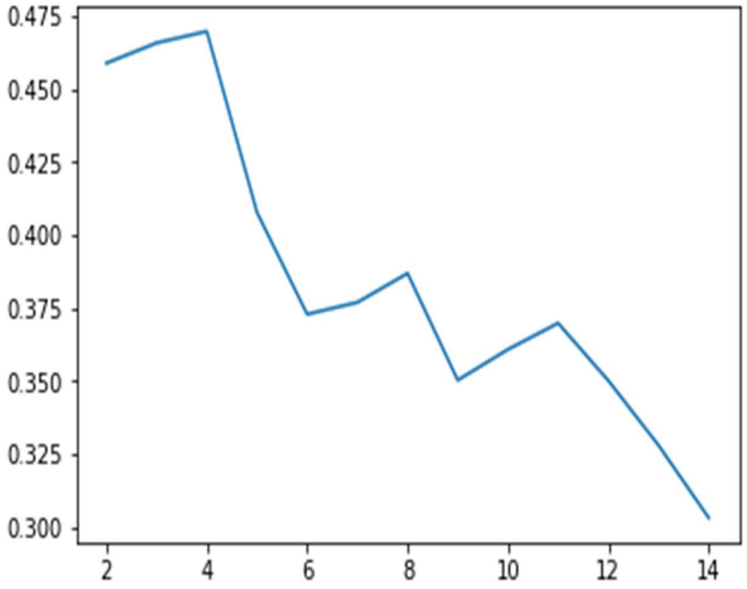

Figure 3.4 Silhouette analysis result

3.4.6 K-MEANS IMPLEMENTATION

The approach behind this simple algorithm is just about some iterations and updating clusters as per distance measures that are computed repeatedly. k is the number of clusters that are to be formed. This content represents the implementation of K-Means algorithm from the scratch using numpy, pandas and plotly.

One notable thing about this algorithm is that, it changes the cluster groups whenever it is re-evaluated. Hence, a data point may become a part of different cluster after the program is executed again.

Initially, two dataframes are created that will be used as set of current means and previous means. Before the iteration, one dataframe from these two contains randomly selected means from the targeted data set itself.

```
1     2049
0      986
2      757
3      122
dtype: int64
```

Figure 3.5 Cluster results

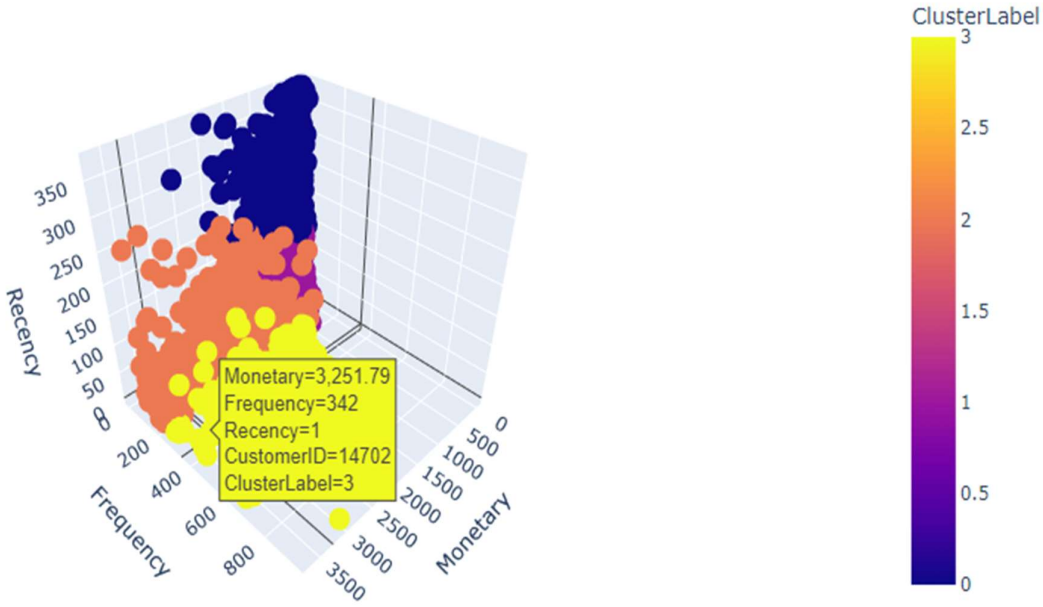

Figure 3.6 3-D result graph

3.5 MODULE 2

We invest in customers (acquisition costs, offline ads, promotions, discounts & etc.) to generate revenue and be profitable. Naturally, these actions make some customers super valuable in terms of lifetime value but there are always some customers who pull down the profitability. We need to identify these behavior patterns, segment customers and act accordingly.

Calculating Lifetime Value is the easy part. First, we need to select a time window. It can be anything like 3, 6, 12, 24 months.

- Define an appropriate time frame for Customer Lifetime Value calculation
- Identify the features we are going to use to predict future and create them
- Calculate lifetime value (LTV) for training the machine learning model
- Build and run the machine learning model
- Check if the model is useful

Deciding the time frame really depends on your industry, business model, strategy and more. For some industries, 1 year is a very long period while for the others it is very short. In our example, we will go ahead with 6 months.

3.5.1 RFM SCORE CALCULATION

First, we calculate recency cluster for each customer by giving the Recency value as input to K-Means algorithm and use the cluster labels as the Recency score. We would wish to have lower cluster labels for customers with higher Recency value, but this is not guaranteed by K-Means algorithm, hence we pass our dataset (that contains the cluster labels) to a user defined function order_cluster() to get this done. The figure below shows the assigned cluster labels (recency cluster) to the given recency values.

	CustomerID	Recency	RecencyCluster
0	14620	12	3
1	14740	4	3
2	17068	11	3
3	12971	4	3
4	15194	6	3

Figure 3.7 Recency cluster results

In the similar fashion, the value of frequency cluster and revenue cluster was calculated. The only change was customers with higher frequency/revenue values were assigned higher cluster labels (in the form of frequency and revenue cluster). The figures below show the same.

	CustomerID	Recency	RecencyCluster	Frequency	FrequencyCluster	Revenue	\
0	14620	12	3	30	0	393.28	
1	15194	6	3	64	0	1439.02	
2	18044	5	3	57	0	808.96	
3	18075	12	3	35	0	638.12	
4	15241	0	3	64	0	947.55	

	RevenueCluster
0	0
1	0
2	0
3	0
4	0

Figure 3.8 RFM cluster results

3.5.2 CUSTOMER LIFETIME VALUE

Now, the values of recency cluster, frequency cluster and revenue cluster are added together to get the overall RFM score for each customer. Then, we calculate the revenue of the next six months of each customer using our dataset and plot a graph of RFM score against the next six months revenue to distribute the customers in low-value, mid-value and high-value segments.

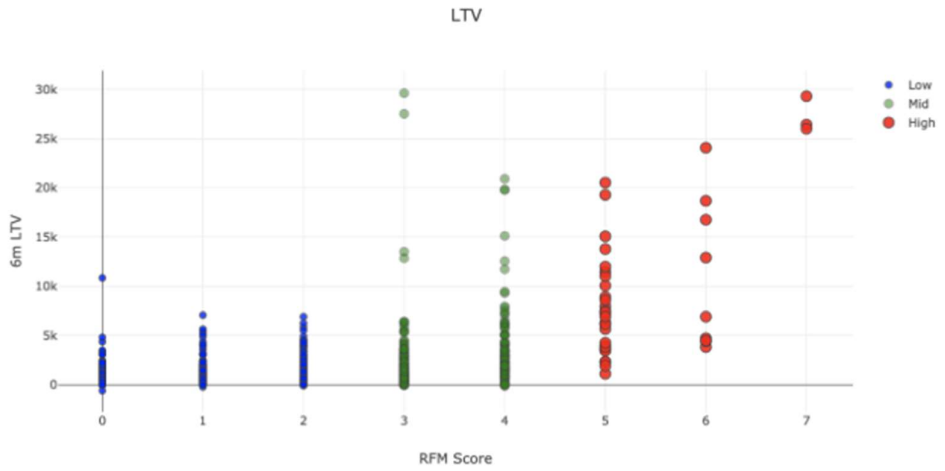

Figure 3.9 RFM score

Before building the machine learning model, we need to identify what is the type of this machine learning problem. LTV itself is a regression problem. A machine learning model can predict the $ value of the LTV. But here, we want LTV segments. Because it makes it more actionable and easier to communicate with other people. By applying K-means clustering, we can identify our existing LTV groups and build segments on top of it.
Considering business part of this analysis, we need to treat customers differently based on their predicted LTV. For this example, we will apply clustering and have 3 segments (number of segments really depends on your business dynamics and goals):

- Low LTV
- Mid LTV
- High LTV

3.5.3 FEATURE ENGINEERING

Need to do some feature engineering. We should convert categorical column “segment” into numerical columns. The three columns that are formed after this step are Segment_High-Value, Segment_Low-Value and Segment_Mid-Value.

We checked the correlation of features against our label, LTV clusters. We will split our feature set and label (LTV) as x and y. We created training and test dataset. Training set was used for building the machine learning model. We applied our model to test set to see its real performance.

3.5.4 XGBOOST CLASSIFIER

XGBoost is an ensemble learning method. Sometimes, it may not be enough to rely upon the results of just one machine learning model. Ensemble learning offers a systematic solution to combine the predictive power of multiple learners. The resultant is a single model which gives the aggregated output from several models.

The models that form the ensemble, also known as base learners, could be either from the same learning algorithm or different learning algorithms. Bagging and boosting are two widely used ensemble learners. Though these two techniques can be used with several statistical models, the most predominant usage has been with decision trees.

It was observed that 3 months Revenue, Frequency and RFM scores were helpful for our machine learning models. A quite strong ML library called XGBoost was used to do the classification. It has become a multi classification model since we had 3 groups (clusters). Following are the results:-

```
Accuracy of XGB classifier on training set: 0.93
Accuracy of XGB classifier on test set: 0.78
             precision    recall  f1-score   support

          0       0.85      0.92      0.88       414
          1       0.48      0.35      0.41       113
          2       0.64      0.45      0.53        20

avg / total       0.76      0.78      0.77       547
```

Figure 3.10 Final result

CONCLUSION

With enough records, specific area knowledge, and a background in machine learning, developing a set of scripts to cluster the uncooked records was viable. After engineering applicable capabilities and reformatting the records, it was viable to carry out consumer segmentation evaluation through K-Means clustering algorithm.

We clustered our customers into four segments using the RFM technique on the given dataset. After that, we used XGBoost Classifier to predict the life-time value segment (low-value, mid-value and high-value) of a particular customer for the next six-months. The classifier made the prediction with an accuracy of 78 percent on the test-set.

REFERENCES

[1] Sari, Juni Nurma, et al. "Review on customer segmentation technique on ecommerce." *Advanced Science Letters* 22.10 (2016): 3018-3022.

[2] A. Banduni, Ilavendhan. "Customer Segmentation using Machine Learning" International Journal of Creative Research Thoughts (IJCRT) Volume 7, Issue 2:(116-112)

[3] Bhade, Kalyani, et al. "A Systematic Approach to Customer Segmentation and Buyer Targeting for Profit Maximization." 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, 2018.

[4] Christy, A. Joy, et al. "RFM ranking–An effective approach to customer segmentation." Journal of King Saud University-Computer and Information Sciences 33.10 (2021): 1251-1257.

[5] Bulletin of the Transilvania University of BraşovSeries V: Economic Sciences • Vol. 9 (58) No. 2 – 2016

[6] Monil, Patel. "Customer Segmentation Using Machine Learning." International Journal for Research in Applied Science and Engineering Technology 8.6 2104–2108. Web.

[7] Marcus, Claudio. "A practical yet meaningful approach to customer segmentation." Journal of consumer marketing (1998).

[8] Hwang, Hyunseok, Taesoo Jung, and Euiho Suh. "An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry." Expert systems with applications 26.2 (2004): 181-188.

[9] Kim, Su-Yeon, et al. "Customer segmentation and strategy development based on customer lifetime value: A case study." Expert systems with applications 31.1 (2006): 101-107.

[10] Cooil, Bruce, Lerzan Aksoy, and Timothy L. Keiningham. "Approaches to customer segmentation." Journal of Relationship Marketing 6.3-4 (2008): 9-39.