

**MINOR PROJECT
ON
CUSTOMER ANALYTICS AND BEHAVIOUR
PREDICTION**

By:

Priyanka Parashar 17803005

Aradhya Mathur 17803011

Mitushi Agarwal 17803016

Under Supervision – Dr. Adwitiya Sinha



May 2020

**For the partial fulfillment of the Degree of Bachelor of
Technology in Computer Science Engineering**

**DEPARTMENT OF COMPUTER SCIENCE ENGINEERING & INFORMATION TECHNOLOGY
JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY, NOIDA**

TABLE OF CONTENTS

S No.	CHAPTER	PAGE NUMBER
1	DECLARATION	3
2	ACKNOWLEDGEMENT	4
3	INTRODUCTION	5-6
4	PROBLEM STATEMENT	7
5	RELATED WORK	8-17
6	WORK PLAN	18-26
7	EXPERIMENTAL WORK	27-28
8	RESULT AND ANALYSIS	29-32
9	CONCLUSION	33
10	FUTURE SCOPE	34
11	REFERENCES	35-36

DECLARATION

We hereby declare that this submission is our own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Place: Jaypee Institute of Information Technology, Noida

Date: 15/05/2020

(Mitushi Agarwal)

(Priyanka Parashar)

(Aradhya Mathur)

ACKNOWLEDGEMENT

We wish to express our sincere gratitude to *Dr. Adwitiya Sinha* for mentoring us in our minor project. We are grateful for her valuable guidance and encouragement in carrying out this project efficiently within the given time constraints.

INTRODUCTION

A huge amount of data is generated every second- based on every action, every purchase, every single click by the user that is now documented by the corporations to enhance their already existing algorithms- therefore there is a lot of data available, we can work upon and further used to generate various results. However, a lot of this data is unstructured, sparse and redundant, from an analytical point of view, unless processed. It rarely provides any conclusive insights to user reactions by itself, and thus, a proper framework is required to actually make sense of this seemingly random data and reach a point of comprehension. E-commerce has become a crucial platform that consists a large database of products with billions number of retailers and consumers.

According to a study it costs you five times more to acquire new customers than it does to retain current customers. And the existing customers are more likely to spend on a new product of the company as compared to the new customers, this is where customer loyalty comes into action. Some customers spend more time and money with the brands they're loyal to. These customers also tell their friends and colleagues about those brands, too which drives referral traffic and word-of-mouth marketing. Customer loyalty is a valuable concept because it allows us to take the risk of predicting the actions and behavior of people we trust. It is also a very integral part of a business and helps in creating a brand image. This eventually helps a brand to retain more customers and improve customer loyalty as well as brand loyalty. In an era of strong customer relationship management (CRM) emphasis, firms strive to build valuable relationships with their existing customer base.

Therefore, once you understand the importance of customer value, you will want to get your hands on all the data you can about your customers, to help you decide how best to serve them. However, not all customer data is equal. Balancing the data of your relevance or importance is the first task.

The Customer Loyalty Prediction is a Model that was formulated to understand the most important aspects of a customer's lifestyle. It is a model that acts as an intermediate between the customers and the merchants. The main target of this model is to find a balance between both the involved associates that is the customers and the merchants by

predicting the loyalty score of the customers. Customer retention increases your customers' lifetime value and boosts your revenue. While acquiring new customers is essential, organizations must lay emphasis on retaining existing ones and creating loyal customers who will ensure stable business operations. The predicted score of the customers categorizes the customers into various segments which further help us to analyze the purchasing behavior (Products a customer shall purchase based on his/her previous purchases), of the best and the loyal group of customers, which are the main target of this model.

In today's hypercompetitive business environment, marketing teams of the business stakeholders need to know their customers inside out and maximize the return on every dollar spent. A marketer is expected to get the most out of their budgets, growing the top line through new customer acquisition while keeping expenses down and boosting the bottom line through effective advertising. Predictive analytics can bring critical improvements to this process, increasing the efficiency of marketing efforts through segmentation.

The customer behaviour analysis is an observational measure of how customers interact with your company. Customer Segmentation is the first step in which the customers are first segmented based on their common characteristics and buying patterns. Then, each group is observed at various stages to predict the future buys and analyse how the customers interact with the company. It also provides an insight into the different variables that influence an audience. It gives you an idea of the motives, priorities, and decision-making methods being considered during the customer's purchasing journey. This analysis helps you understand how customers feel about your company, as well as if that perception aligns with their regular buying pattern.

Customer Analytics is an idea to study the customer segmentation and predict the purchasing behaviour of the buyers using Machine Learning Algorithm. What makes this particular model different is the fact that it targets the Loyal and best customers of the company. These particular group of customers are the most frequent buyers with maximum purchases and targeting this particular group of customers will be quite beneficial for the company as it will help in the cost reduction due to unwanted advertisement campaign, also by maintaining this range of customers, it can also help us in catching on more customers which might lie in the similar range and buying patterns.

PROBLEM STATEMENT

Customer relationship management is one the most significant managerial tasks in organizations. Many companies may usually adopt a strategy that is known as target marketing. The marketing managers who may consider using target marketing will usually break the market down into groups and to target the most profitable segments. The current challenge is the effective utilisation of the data in CRM processes and selection of appropriate data analytics techniques. Hence, a prerequisite for the development of this customer-centric strategy is the specification of the target markets that the companies will attempt to serve.

Therefore, this project comes forward as solution that begins by the Customer analysis which requires splitting up the customers into various segments that helps us in extracting the most profitable ones i.e. the loyal customers and then predicting their purchasing behaviour based on their past purchases.

RELATED WORK:

Summary of the research papers studied are illustrated below:

Paper1

Title of the paper	Predicting Customer Loyalty Using Various Regression Models
Authors	Guido van der Heijden, Harry Collins, Suhaib Aslam
Year of Publication	August 2019
Publishing Details	Published in University of Twente
Web Link	https://www.researchgate.net/publication/335158533_Predicting_Customer_Loyalty_Using_Various_Regression_Models
Summary	<p>In this paper we describe the process and outcomes of a project targeted towards building various machine learning models for predicting customer loyalty. The project was targeted towards the Kaggle challenge on “Elo Merchant Category Recommendation” to use machine learning to improve the understanding of customer loyalty for the Elo payment brand. We explain the why and how of our data preparation and of the building of our models. We also give pointers on how the models’ performance could be improved further. The results show that polynomial regression and LightGBM algorithms are the best performing algorithms on the given dataset. The results also show that the given dataset is highly sensitive to data outliers’ prediction and to feature selection.</p>

Paper 2

Title of the paper	Machine-Learning Techniques for Customer Retention: A Comparative Study
Authors	Sahar F. Sabbeh
Year of Publication	2018
Publishing Details	Faculty of computing and information sciences, King AbdulAziz University, KSA Faculty of computing and information sciences, Banha University, Egypt
Web Link	https://thesai.org/Downloads/Volume9No2/Paper_38-Machine_Learning_Techniques_for_Customer_Retention.pdf
Summary	<p>This paper tries to compare and analyze the performance of different machine-learning techniques that are used for churn prediction problem. Ten analytical techniques that belong to different categories of learning are chosen for this study. The chosen techniques include Discriminant Analysis, Decision Trees (CART), instance-based learning (k-nearest neighbors), Support Vector Machines, Logistic Regression, ensemble-based learning techniques (Random Forest, Ada Boosting trees and Stochastic Gradient Boosting), Naïve Bayesian, and Multi-layer perceptron. Models were applied on a dataset of telecommunication that contains 3333 records. Results show that both random forest and ADA boost outperform all other techniques with almost the same accuracy 96%. Both Multi-layer perceptron and Support vector machine can be recommended as well with 94% accuracy. Decision tree achieved 90%, naïve Bayesian 88% and finally logistic regression and Linear Discriminant Analysis (LDA) with accuracy 86.7%.</p>

Paper 3

Title of the paper	Analysis of K-Means Clustering Algorithm: A Case Study Using Large Scale E-Commerce Products
Authors	Norsyela Muhammad Noor Mathivanan Nor Azura Md. Ghani Roziah Mohd Janor
Year of Publication	2019
Publishing Details	2019 IEEE Conference on Big Data and Analytics (ICBDA)
Web Link	https://ieeexplore.ieee.org/abstract/document/8987140
Summary	<p>E-commerce has become a crucial platform consists a large database of products with billions number of retailers and consumers. However, these products are placed into different categories according to the structure of different websites. A clustering analysis using K-Means Clustering algorithm helps in providing an insightful pattern on categories of clustered products. This analysis leads to an automatic classification model to classify the products efficiently. This paper presents a step by step cluster analysis using K-Means clustering to group e-commerce products from the online store website in Malaysia. The results show that the e-commerce products were categorized into three clusters. The most frequent words in each cluster provided a useful insight on the category of the clustered products which were hair and face, oral and pets care products. Hence, K-Means clustering analysis able to group a large data set of e-commerce products effectively.</p>

Paper 4

Title of the paper	Developing a model for measuring customer's loyalty and value with RFM technique and clustering algorithms
Authors	Razieh qiasi Malihe baqeri-Dehnavi Behrooz Minaei-Bidgoli Golriz Amooee4
Year of Publication	2012
Publishing Details	The Journal of Mathematics and Computer Science Vol. 4 No.2
Web Link	https://www.isr-publications.com/jmcs/307/download-developing-a-model-for-measuring-customers-loyalty-and-value-with-rfm-technique-and-clustering-algorithms
Summary	<p>In today's competitive world, moving toward customer-oriented markets with increased access to customer's transaction data, identifying loyal customers and estimating their lifetime value makes crucial. Since knowledge of customer value provides targeted data for personalized markets, implementing customer relationship management strategy helps organizations to identify and segment customers and create long-term relationships with them, and as a result, they can maximize customer lifetime value. Data mining techniques are known as a powerful tool for this purpose. The purpose of this paper is customer segmentation using RFM technique and clustering algorithms based on customer's value, to specify loyal and profitable customers. We also used classification algorithms to obtain useful rules for implementing effective customer relationship management. This paper used a combination of behavioral and demographical characteristics of individuals to estimate loyalty. Finally, the proposed model has been implemented on a grocery store's data during 1997 to 1998 in Singapore, to measure customer's loyalty during these two years.</p>

Paper 5

Title of the paper	CRM Analytics Framework
Authors	Joseph P. Bigus, Upendra Chitnis, Prasad M. Deshpande
Year of Publication	January 2009
Publishing Details	Proceedings of the 15th International Conference on Management of Data
Web Link	https://www.researchgate.net/publication/221324954_CRM_Analytics_Framework
Summary	<p>Implementing a CRM Analytics solution for a business involves many steps including data extraction, populating the extracted data into a warehouse, and running an appropriate mining algorithm. We propose a CRM Analytics Framework that provides an end-to-end framework for developing and deploying pre-packaged predictive modeling business solutions, intended to help in reducing the time and effort required for building the application. Standardization and metadata-driven development are used in the solution; this makes the framework accessible to non-experts. We describe our framework that makes use of industry standard software products and present a case study of its application in the financial domain.</p>

Paper 6

Title of the paper	An Analysis of the Potential Target Market through the Application of the STP Principle/Model
Authors	Johnson Kampamba
Year of Publication	August 2015
Publishing Details	Published in Mediterranean Journal of Social Sciences 6(4):324-340
Web Link	https://www.researchgate.net/publication/281521878_An_Analysis_of_the_Potential_Target_Market_through_the_Application_of_the_STP_Principle/Model
Summary	<p>The purpose of this paper was to establish and identify the target market of a proposed residential development in Gaborone using the STP strategy. Research design/methodology/approach – Both non-probability and probability sampling techniques were used by applying qualitative and quantitative research methods (interviews and questionnaire) were used to get research data in order to meet the objective of the study. A survey instrument using a self-administered questionnaire was developed and administered to the respondents who were working earning a monthly salary was used in order to meet the minimum qualifying criteria for a home loan being 30% of one's income to service mortgage payments</p>

Paper 7

Title of the paper	Segmenting and Targeting and Positioning Your Market: Strategies and Limitations
Authors	Michael Lynn
Year of Publication	December 2011
Publishing Details	Published in Hospitality Administration and Management
Web Link	https://scholarship.sha.cornell.edu/cgi/viewcontent.cgi?article=1238&context=articles
Summary	<p>Almost any marketing textbook will tell you that the key to successful marketing can be summed up by the STP strategy—that is, segmentation, targeting, and positioning. This approach suggests that the mass market consists of some number of relatively homogeneous groups, each with distinct needs and desires. STP marketers attempt to identify those market segments, direct marketing activities at the segments which the marketers believe that their company can satisfy better than their competitors, and position their product offering so as to appeal to the targeted segments. Undoubtedly, your hospitality firm uses some form of this approach</p>

Paper 8

Title of the paper	A purchase-based market segmentation methodology
Authors	CY Tsai C.-C. Chiu
Year of Publication	August 2004
Publishing Details	Published in Expert Systems with Applications .
Web Link	https://www.researchgate.net/publication/223343267_A_purchase-based_market_segmentation_methodology
Summary	<p>Market segmentation is critical for a good marketing and customer relationship management program. Traditionally, a marketer segments a market using general variables such as customer demographics and lifestyle. However, several problems have been identified and make the segmentation result unreliable. This paper develops a novel market segmentation methodology based on product specific variables such as purchased items and the associative monetary expenses from the transactional history of customers to resolve these problems. A purchase-based similarity measure, clustering algorithm, and clustering quality function are defined in this paper. A genetic algorithm approach is adopted to ensure that customers in the same cluster have the closest purchase patterns. After completing segmentation, a designated RFM model is used to analyze the relative profitability of each customer cluster. The findings from a practical marketing implementation study will also be discussed.</p>

Paper 9

Title of the paper	Using Segmentation to Improve Sales Forecasts Based on Purchase Intent: Which "Intenders" Actually Buy?
Authors	Vicki G. Morwitz David Schmittlein
Year of Publication	2015
Publishing Details	Published in Sage Publications, Inc.
Web Link	https://www.jstor.org/stable/3172706?seq=1
Summary	<p>The authors investigate whether the use of segmentation can improve the accuracy of sales forecasts based on stated purchase intent. The common current practice is to prepare a sales forecast by using purchase intent and observed historical patterns in purchase rates given level of intent. The authors show that the accuracy of sales forecasts based on purchase intent can be improved by first using certain kinds of segmentation methods to segment the panel members. The main empirical finding is that more accurate sales forecasts appear to be obtained by applying statistical segmentation methods that distinguish between dependent and independent variables (e.g., CART, discriminant analysis) than by applying simpler direct clustering approaches (e.g., a priori segmentation or K-means clustering). The results further reveal that meaningful segments are present and identifiable that vary in their subsequent purchase rates for a given level of intent. This identification has important implications for areas such as target marketing, as it indicates which customer segments will actually fulfill their intentions.</p>

Paper 10

Title of the paper	Study on Customer Loyalty Prediction Based on RF Algorithm
Authors	Jiong Mu, Lijia Xu , Xuliang Duan and Haibo Pu
Year of Publication	8, AUGUST 2013
Publishing Details	Published in Journal Of Computers.
Web Link	https://www.researchgate.net/publication/222519381_Predicting_Customer_Retention_and_Profitability_by_Using_Random_Forests_and_Regression_Forests_Techniques
Summary	<p>In an era of strong customer relationship management (CRM) emphasis, firms strive to build valuable relationships with their existing customer base. In this study, we attempt to better understand three important measures of customer outcome: next buy, partial-defection and customers' profitability evolution. By means of random forests techniques we investigate a broad set of explanatory variables, including past customer behavior, observed customer heterogeneity and some typical variables related to intermediaries. We analyze a real-life sample of 100,000 customers taken from the data warehouse of a large European financial services company. Two types of random forests techniques are employed to analyze the data: random forests are used for binary classification, whereas regression forests are applied for the models with linear dependent variables. Our research findings demonstrate that both random forests techniques provide better fit for the estimation and validation sample compared to ordinary linear regression and logistic regression models.</p>

WORK PLAN

The Customer Loyalty and Purchasing Behaviour Prediction Model is a machine learning model. In order to implement this model using accurate algorithms, we have surveyed various Research Papers in order to come forward with accurate and efficient results and hence formulated a flow of project that summarizes the various steps performed to serve our task:

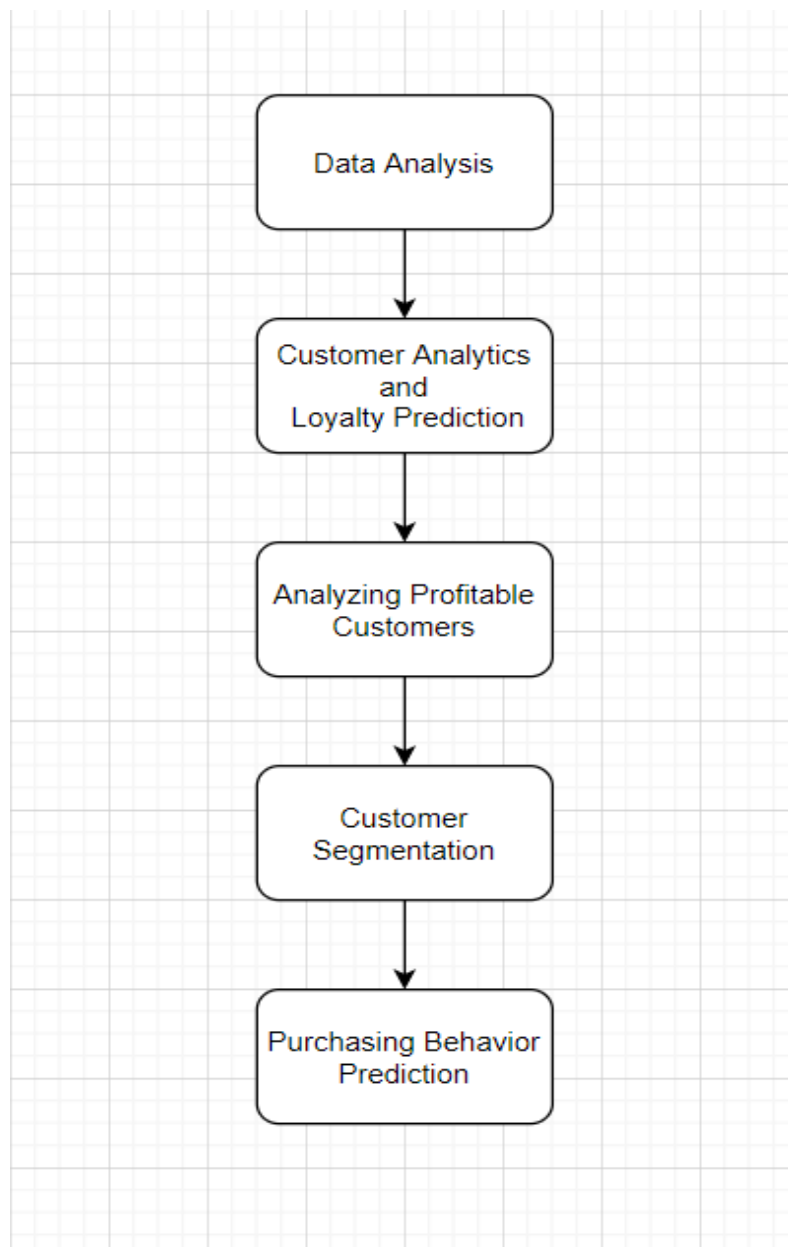


Fig1: The flow of the project

DETAILED ANALYSIS

Step1: Data Analysis

The first task was selecting an appropriate customer retail dataset that would help us in analysing our problem, which was further followed by Data Cleaning which is the first critical step in any machine learning model, it helped in exploring the data for relevant features. It was observed that features like the amount of money spent, number of transactions made by the buyer, the difference between the last purchase and first purchase, types of products bought etc, played an important role in developing a relationship between the customer and the merchants. Formulation of a customer relationship management system is an important part as it creates a relationship with customers that in turn creates loyalty and customer retention which results in increased profit for business.

Step2: Customer Analytics and Loyalty Prediction

Customer analytics can be defined as a process by which the data from customer behaviour is used to help make key business decisions via market segmentation and predictive analytics. The segmentation done will influence marketing and sales decisions, and potentially the survival of a company. The first step involves determining different categories of the customers by using the features extracted from dataset for instance the *customer id, invoice date, unit price, quantity* etc. To serve our purpose, the RFM segmentation model is applied.

RFM Segmentation Model: It allows the marketers to target specific clusters of customers that are much more relevant for their particular behaviour – and thus generate much higher rates of response, plus increased loyalty and customer lifetime value. RFM stands for:

- RECENCY (R): Days since last purchase
- FREQUENCY (F): Total number of purchases
- MONETARY VALUE (M): Total money this customer spent.

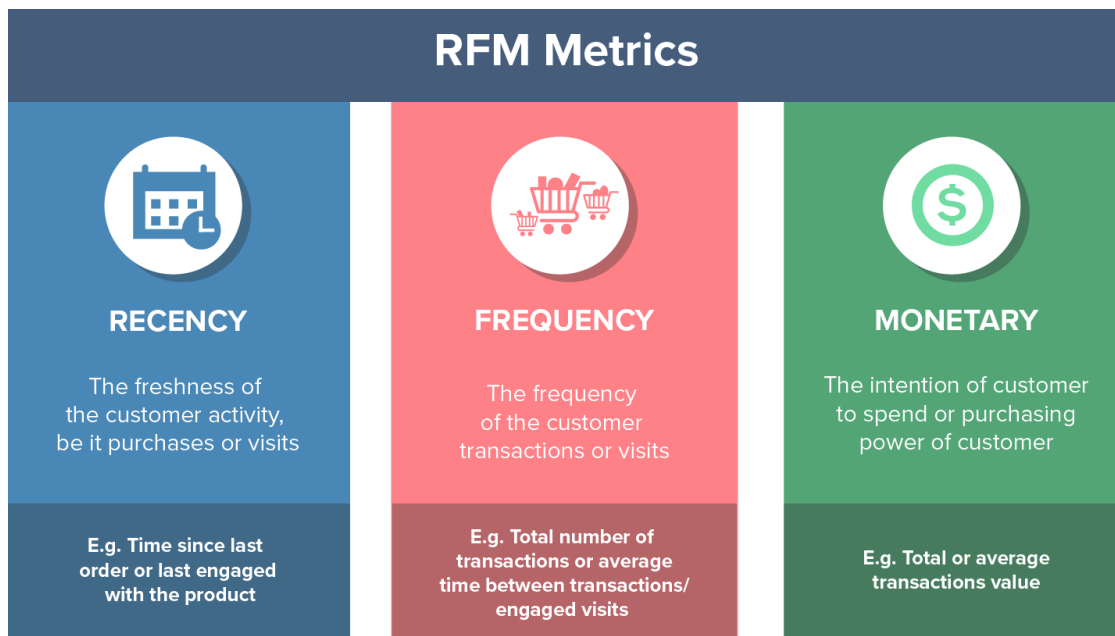


Fig2: The RFM Metrics explaining the R-F-M in concise form

These attributes have been created against each customer. This model is performed in 3 major steps:

1. The first step in building an RFM model is to assign Recency, Frequency and Monetary values to each customer.
2. The second step is to divide the customer list into tiered groups for each of the three dimensions (R, F and M), using Excel or another tool such as k-means cluster analysis – can be performed by software, resulting in groups of customers with more homogeneous characteristics.
3. The third step is to select groups of customers to whom specific types of communications will be sent, based on the RFM segments in which they appear. And further extracting the profitable segments.

The RFM analysis comes out as an efficient way as it utilizes objective, numerical scales that yield a concise and informative high-level depiction of customers.

It is simple – marketers can use it effectively without the need for data scientists or sophisticated software and It is intuitive – the output of this segmentation method is easy to understand and interpret.

Step3: Analyzing Profitable Customers

The RFM Model provided us various categorizes of customers to look upon and hence target our most profitable segments from the following:

Best Customers: '444' (Highest frequency as well as monetary value with least recency)

Loyal Customers: '344' (High frequency as well as monetary value with good recency)

Potential Loyalists: '434' (High recency and monetary value, average frequency)

Big Spenders: '334' (High monetary value but good recency and frequency values)

At Risk Customers: '244' (Customers shopping less often now who used to shop a lot)

Can't Lose Them: '144' (Customers shopped long ago who used to shop a lot)

Recent Customers: '443' (Customers who recently started shopping a lot but with less monetary value)

Lost Cheap Customers: '122' (Customers shopped long ago but with less frequency and monetary value)

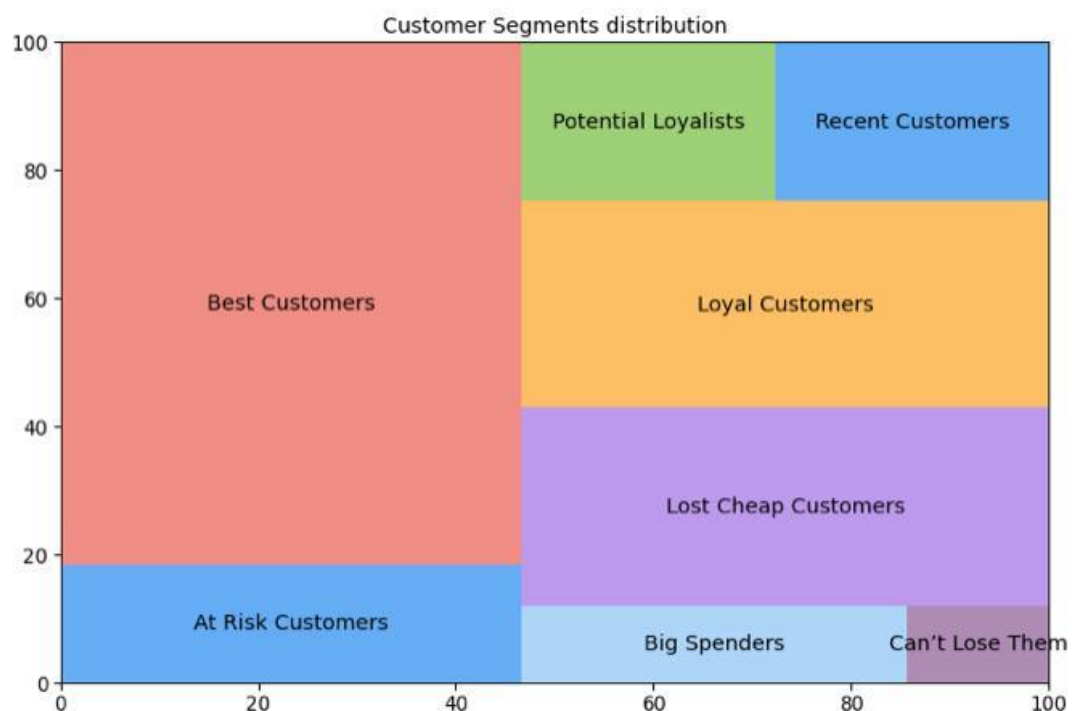


Fig 3: Various customer categories.

The score against each category defines the R-F-M metrics. For instance, if we consider the *Best Customers*, this group consists of those customers who are found in R-Tier-4, F-Tier-4 and M-Tier-4, meaning that they transacted recently, do so often and spend more than other customers and this similarly follows up for the other described categories. Now when viewed from the business perspective, the marketing team of any e-commerce business platform would mainly want to target the categories yielding most profit. It was observed that segments like *Best Customers*, *Loyal Customers* and *Potential Loyalists* turned out to be the most profitable segment among all. And hence the customers belonging to these particular segment blocks were extracted i.e. the data against each *customer ids* with RFM score equal to the values – ‘444’, ‘344’ and ‘434’.

This dataset is now forwarded towards the Customer Segmentation part for predictive analysis of the data and eventually follow up with the results of behaviour prediction.

Step 4: Customer Segmentation

Customer segmentation is a process where we divide the consumer base of the company into subgroups. We need to generate the subgroups by using some specific characteristics so that the company sells more products with less marketing expenditure. The segments formed enables the merchants to understand the patterns that differentiate your customers.

It is the problem of uncovering information about a firm's customer base, based on their interactions with the business. In most cases this interaction is in terms of their purchase behavior and patterns. We explore some of the ways in which this can be used. Here in this model we are using this notebook on an online Retail dataset to explore customer segmentation through a task of unsupervised learning method. Then we go further and apply association rule mining approach to find interesting rules and patterns in this transaction database. These customer segmentation, rules and patterns can be used to make interesting and useful decisions as far as user interest is concerned. This segment-wise marketing will help the company sell more products with lower marketing expenses. Thus, the company will make more profit. This is the main reason why companies use customer segmentation analysis nowadays. Customer segmentation is used among other domain such as the retail domain, finance domain, and in customer relationship management (CRM)-based products.

Companies are using the STP approach to make the marketing strategy firm. STP stands for Segmentation-Targeting-Positioning. There are 3 stages for the following:

1. Segmentation: Here we create segments of our customer base using their profile characteristics as well as consider features provided in the preceding figure. Once the segmentation is firm, we move on to the next stage.
2. Targeting: Now, the marketing teams evaluate segments and try to understand which kind of product is suited to which particular segments. The team performs this exercise for each segment, and finally, the team designs customized products that will attract the customers of one or many segments. They will also select which product should be offered to which segment.
3. Positioning: In this last stage, companies study the market opportunity and what their product is offering to the customer. The marketing team should come up with a unique selling proposition. Here, the team also tries to understand how a particular segment perceives the products, brand, or service. This is a way for companies to determine how to best position their offering. The marketing and product teams of companies create a value proposition that clearly explains how their offering is better than any other competitors. Lastly, the companies start their campaign representing this value proposition in such a way that the consumer base will be happy about what they are getting.

Further we have grouped the products into different classes. The k-means clustering method of sklearn uses a Euclidean distance that can be used for the same. Now, in order to define the number of clusters that best represents the data, we have used the silhouette score.

K-Means/Centroid-based clustering:

It organizes the data into non-hierarchical clusters, in contrast to hierarchical clustering, k-means is the most widely-used centroid-based clustering algorithm. Centroid-based algorithms are efficient but sensitive to initial conditions and outliers.

K-means will randomly initiate centroids at random locations and slowly fit each data point to the nearest centroid. Each data point represents one customer, and the customer closest to

the same centroid will be in the same group. The centroids' locations are adjusted automatically based on the last nearest customer allocated to them. Doing so, it will learn on its own to find other customers with similar characteristics. One method for finding the optimal number of groups is to use Silhouette Score. It takes into consideration both the intra-cluster and inter-clusters distance and returns a score; the lower the score, the more meaningful the clusters formed.

Silhouette refers to a technique or a method of interpretation and validation of consistency within clusters of data. The scores obtained were considered equivalent since, scores of 0.1 ± 0.05 will be obtained for all clusters with $n_clusters \gg 3$.

On the other hand, it was found that beyond 5 clusters, some clusters contained very few elements, we can represent the silhouette scores of each element of the different clusters. Therefore, the dataset was segregated into 5 clusters. In order to ensure a good classification and have insight on the quality of classification, we can represent the silhouette scores of each element of the different clusters.

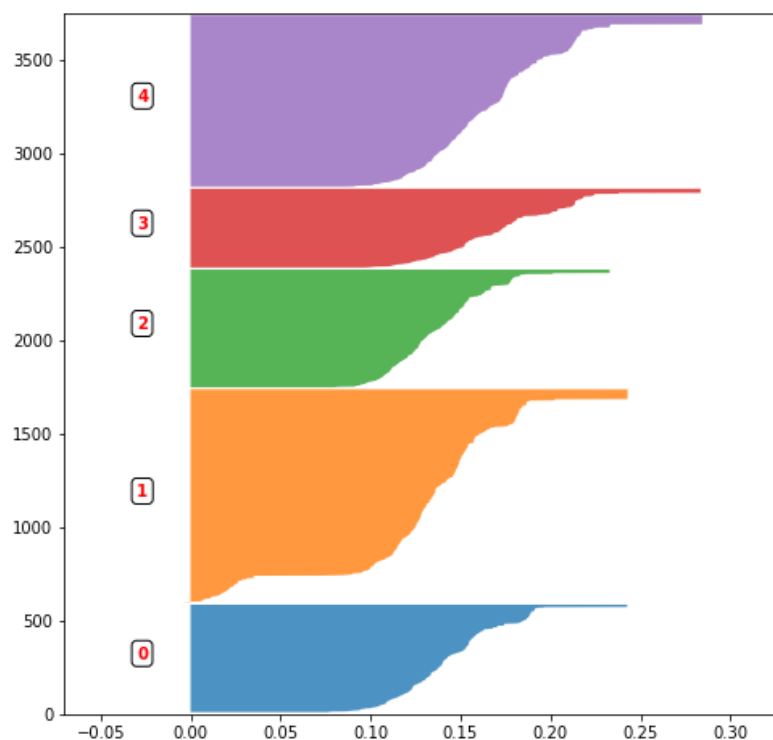


Fig 4: Silhouette scores of each element of the different product clusters

After the creation of product clusters, we move towards the creation of customer clusters or the customer categories which is further followed by classification.

7]:

Jnnamed: 0	CustomerID	count	min	max	mean	sum	categ_0	categ_1	categ_2	categ_3	categ_4	LastPurchase	FirstPurchase	cluster
0	12347	5	382.52	711.79	558.172	2790.86	32.408290	29.105724	11.173617	18.636191	8.676179	59	297	1
1	12359	3	547.50	1803.11	1153.310	3459.93	15.019090	9.916386	3.985052	44.655528	26.423945	119	261	4
2	12362	5	303.76	829.99	510.908	2554.54	17.343631	34.424985	5.787343	34.123560	8.320480	2	225	0
3	12380	2	607.55	626.01	616.780	1233.56	12.569312	22.325627	7.052758	51.437303	6.615000	8	115	4
4	12381	1	1227.39	1227.39	1227.390	1227.39	10.522328	23.602930	8.455340	43.776632	13.642770	49	49	7

Fig 5: Customers classified in the different client categories.

Step5: Purchasing Behavior Prediction

Predictive behavior modeling is applied to historical and transactional data in order to predict the future behavior of customers. It is typically used to select the most effective marketing actions to run on each group of customers, in order to identify which customers will likely change their spending level. The steps for the same model have already been applied. This last step includes the classification of customers in order to define the class to which a client belongs.

We have firstly defined a class that allows to interface several of the functionalities common to different classifiers. A few classifiers were trained in order to categorize the customers. The classifiers used are as follows:

- **Support Vector Machine Classifier (SVC):** The objective of this classifier is to fit to the provided data and returning a hyperplane that divides, or categorizes our data, and when provided with some features returns the predicted class.
- **Logistic Regression:** The logistic model is used to model the probability of a certain class or event occurring having been given some previous data. It works with binary data, where either the event happens or the event does not happen.

- **K-nearest Neighbors:** It is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure.
- **Decision Tree:** A decision tree classifier is a tree in which internal nodes are labeled by features. It categorizes an object x_i by recursively testing for the weights that the features labeling the internal nodes have in vector x_i , until a leaf node is reached.
- **Random Forest:** It is a classification algorithm consisting of many decision trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.
- **Gradient Boosting Classifier:** These are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model.

The results of the different classifiers can be combined to improve the classification model. This can be achieved by using the voting classifier model. It combines multiple different models into a single model, which is stronger than any of the individual models alone.

The whole analysis was based on the data of the first 10 months, now to examine the predictions of the different classifiers that have been trained, we have tested the model on the last 10 months of the dataset to obtain our results and hence check the accuracy of the developed model.

EXPERIMENTAL WORK:

Dataset:

This is a transactional data set which contains all the transactions history for an online retail store. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers i.e. they purchase in huge quantities.

The dataset has major attributes as:

1. 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.
2. StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
3. Description: Product (item) name. Nominal.
4. Quantity: The quantities of each product (item) per transaction. Numeric.
5. InvoiceDate: Invoice Date and time. Numeric, the day and time when each transaction was generated.
6. UnitPrice: Unit price. Numeric, Product price per unit in sterling.
7. CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
8. Country: Country name. Nominal, the name of the country where each customer resides.

	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	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom

Fig 6: The online retail dataset

InvoiceNo	StockCode	Description	Quantity	UnitPrice	CustomerID	Country	Segment	Score	SegmentType
536367	84879	ASSORTED COLOUR BIR	32	1.69	13047	United Kingdom	344	11	Loyal Customers
536367	22745	POPPY'S PLAYHOUSE BE	6	2.1	13047	United Kingdom	344	11	Loyal Customers
536367	22748	POPPY'S PLAYHOUSE KI	6	2.1	13047	United Kingdom	344	11	Loyal Customers
536367	22749	FELTCRAFT PRINCESS C	8	3.75	13047	United Kingdom	344	11	Loyal Customers
536367	22310	IVORY KNITTED MUG C	6	1.65	13047	United Kingdom	344	11	Loyal Customers
536367	84969	BOX OF 6 ASSORTED CC	6	4.25	13047	United Kingdom	344	11	Loyal Customers
536367	22623	BOX OF VINTAGE JIGSA	3	4.95	13047	United Kingdom	344	11	Loyal Customers
536367	22622	BOX OF VINTAGE ALPH	2	9.95	13047	United Kingdom	344	11	Loyal Customers
536367	21754	HOME BUILDING BLOC	3	5.95	13047	United Kingdom	344	11	Loyal Customers
536367	21755	LOVE BUILDING BLOCK	3	5.95	13047	United Kingdom	344	11	Loyal Customers
536367	21777	RECIPE BOX WITH MET	4	7.95	13047	United Kingdom	344	11	Loyal Customers
536367	48187	DOORMAT NEW ENGLA	4	7.95	13047	United Kingdom	344	11	Loyal Customers
536368	22960	JAM MAKING SET WITH	6	4.25	13047	United Kingdom	344	11	Loyal Customers
536368	22913	RED COAT RACK PARIS	3	4.95	13047	United Kingdom	344	11	Loyal Customers
536368	22912	YELLOW COAT RACK PA	3	4.95	13047	United Kingdom	344	11	Loyal Customers
536368	22914	BLUE COAT RACK PARIS	3	4.95	13047	United Kingdom	344	11	Loyal Customers
536369	21756	BATH BUILDING BLOCK	3	5.95	13047	United Kingdom	344	11	Loyal Customers
536370	22728	ALARM CLOCK BAKELIK	24	3.75	12583	France	444	12	Best Customers
536370	22727	ALARM CLOCK BAKELIK	24	3.75	12583	France	444	12	Best Customers
536370	22726	ALARM CLOCK BAKELIK	12	3.75	12583	France	444	12	Best Customers


Fig 7: The extracted dataset of the best, loyal and potential loyal customers along with the RFM score and segment.

RESULT AND ANALYSIS

The dataset consisted of 10 months transactional data. The products were grouped in 5 different categories. Word cloud is a method of visually representing textual data. The highlighted or the bold text displays its importance in the particular cluster.

Fig 8: Word Cloud of type of product in each cluster.

Firstly, the various classification models were applied on the train dataset that consisted of the 8 months history. The precision results for each classifier have been shown below, which is followed by the result of the final Voting classifier that groups all the classifiers to create a combined model in order to improve the accuracy.



```
Support Vector Machine:
Precision: 75.00 %
LogisticRegression:
Precision: 86.49 %
k-Nearest Neighbors:
Precision: 64.86 %
Decision Tree:
Precision: 87.16 %
Random Forest:
Precision: 89.19 %
adaboost:
Precision: 42.57 %
Gradient Boosting:
Precision: 87.84 %
```

Fig 9: Precision of train data using different classifiers.

The accuracy of the obtained model from the same was equivalent to 88.51% which is shown below: -

Precision : 88.51%

Fig 10: Precision of train data using voting classifier.

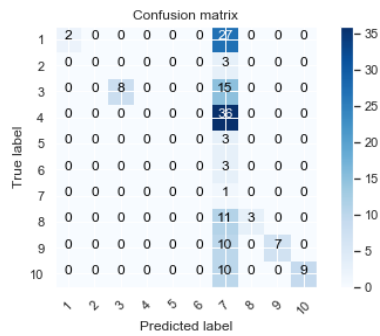


Fig 11(a): SVC

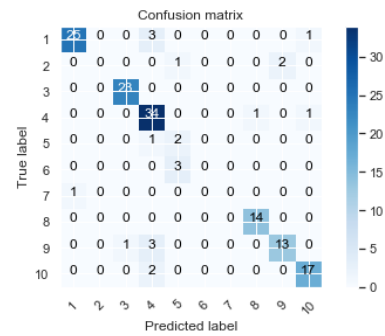


Fig 11(b): Logistic Regression

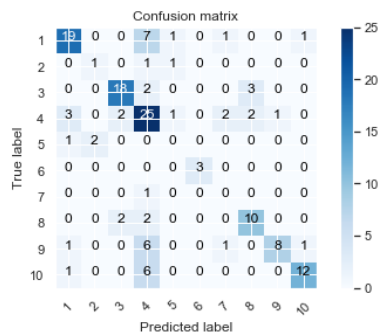


Fig 11(c): KNN

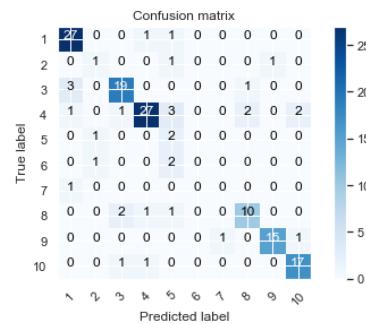


Fig 11(d): Decision Tree

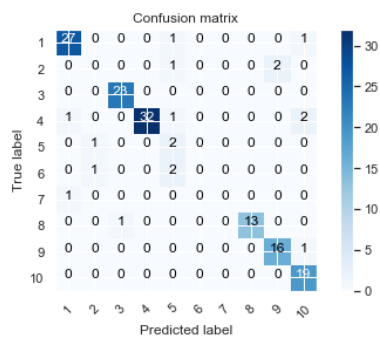


Fig 11(e): Random Forest

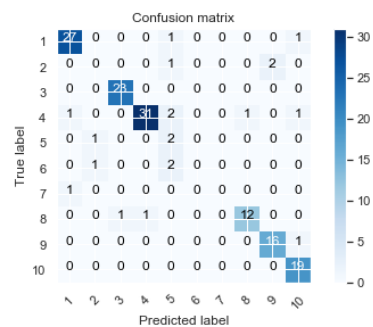


Fig 11(f): Gradient Boosting

Fig 11: Confusion Matrix of different classifiers.

The accuracy of the results seems to be correct. Confusion Matrix have been used to look at how the predictions and real values compare to the breasts of the different classes.

Support Vector Machine
Precision: 80.56 %

Logostic Regression
Precision: 81.48 %

k-Nearest Neighbors
Precision: 62.43 %

Decision Tree
Precision: 73.54 %

Random Forest
Precision: 80.56 %

Gradient Boosting
Precision: 81.88 %

Fig 12: Precision of test data using different classifiers that have been trained.

Precision : 82.41%

Fig 13: Precision of test data using voting classifier.

Finally, the quality of the predictions of the different classifiers was tested over the last two months of the dataset and similar classifiers were applied, the precision results for the same have been displayed above and the final test accuracy from the voting classifier thus obtained was equivalent to 82.41%.

This final result obtained therefore depicts that 82% customers were given the right class and the performance of the classifier therefore seems correct.

FUTURE SCOPE

This project acts as a bridge between two set of Models, first is the Loyal Customer Prediction Model and the other is the Customer Behavior Analysis.

The categorization performed using the RFM Segmentation can be further used to provide the loyal customers with certain rewards and benefits in form of offers or discounted coupons that would rise up the purchase level and therefore retain those customers.

The segments where the RFM metrics were found low are the customers that require more awareness about the e-commerce business involved, a concept of 'Loyalty card' can be introduced which might attract the potential customers who can eventually be turned into the loyal ones.

The purchasing behavior model created can be further combined with a recommendation system and increase its efficiency of prediction.

CONCLUSION

The work described in this project is based on an online retail dataset providing details of purchases made on an E-commerce website over a period of one year. Work done in this project demonstrates how customer-centric business for online retailers can be created by means of data mining techniques and further how can we apply machine learning models to predict the behavior of the customers.

In the first stage of the project it has been found that in the analysis there are two steps in the whole data mining process that are very crucial and the most time-consuming: data preparation and model interpretation and evaluation. The distinct customer groups characterized in this project can help the business better understand its customers in terms of their profitability, and accordingly, adopt appropriate marketing strategies for different consumers. According to this idea we have extracted our loyal and best customers and further applied the formulated machine learning model on them.

Second stage of this project describes different products sold by the website, which was the basis of the first classification. The products were therefore grouped in 5 different categories. Furthermore, classification of the customers was performed by analyzing their historical transactions over a period of 10 months and then further classifying the customers into 11 major categories based on the type of products they usually buy, the number of visits they make and the amount they spent during the 10 months. Once these categories were established, several classifiers were trained whose objective was to be able to classify consumers in one of these 11 categories and this from their first purchase.

Finally, the quality of the predictions of the different classifiers was tested over the last two months of the dataset. The data were then processed in two steps. First, all the data was considered to define the category to which each client belongs, and then, the classifier predictions were compared with this category assignment. It was hence found that 82% of clients are awarded the right classes. The performance of the classifier therefore seems correct.

References:

- [1] Aly, M. (2005). Survey on Multiclass Classi Methods. Retrieved from <https://www.cs.utah.edu/~piyush/teaching/aly05multiclass.pdf>.
- [2] Kaggle | Elo Merchant Category Recommendation. (n.d.). Retrieved January 29, 2019, from <https://www.kaggle.com/c/elo-merchant-category-recommendation>.
- [3] Adams, J. S. (1965). Inequity in social exchange. *Advances in Experimental Social Psychology*.
- [4] Guha, Sudipto, and Nina Mishra. "Clustering data streams: - In Data Stream Management". Springer Berlin Heidelberg, 2016.
- [5] D. Kim , S. Lee, J. Chun, "A semantic classification model for e-catalogs", *Proceedings of the IEEE Conference on E-Commerce*, 2004.
- [6] S. Chong, N. Rampalli, F. Yang, A. Doan Chimera, "Large-scale classification using machine learning rules and crowdsourcing", *Proceedings of the VLDB Endowment*, vol. 7, no. 13, 2014.
- [7] H.M. Chuang and C.C. Shen, A study on the application of data mining techniques to enhance customer lifetime value based on the department store industry, the seventh international conference on machine learning and cybernetics, PP.168-173.
- [8] M. Khajvand, K. Zolfaghar, S. Ashoori, S. Alizadeh . "Estimating customer lifetime value based on RFM analysis of customer purchase behavior: case study,)2011) *Procedia Computer Science*, Vol.3, pp.57–63
- [9] N. Ab e, N. K. Verma, C. Apt´e, and R. Schroko. Cross channel optimized marketing by reinforcement learning. In KDD, 2004.
- [10] C. Apt´e, E. Bib elnieks, R. Natara jan, E. P. D. Pednault, F. Tipu, D. Campbell, and B. Nelson. Segmentation-based modeling for advanced targeted marketing. In KDD, 2001.
- [11] Berkowitz, E. N., Kerin, R. A., & Rudelius, W. (1989). *Marketing*. Homewood Illinois: Irwin.
- [12] Botswana Review. (2014). Botswana Review of Commerce and Industry. Gaborone: B & T Directories (Pty) Ltd.
- [13] 2 G. Hallberg, *All Consumers Are Not Created Equal* (New York: Wiley, 1995) .23

- [14] B. Berman, "Developing an Effective Customer Loyalty Program," California Management Review 49(1) (2006): 123-148.
- [15] Building Data Mining Applications for CRM.
- [16] Clustering analysis for researchers Successful direct marketing method H C Romesburg H. C. (1984). Clustering analysis for researchers. Belmont: Lifetime Learning Publications. Stone, B. (1995). Successful direct marketing method. Lincolnwood, IL: NTC Business Books.
- [17] Jusco. (2002). JUSCO Annual Report. Annual Report, Kuala Lumpur.
- [18] Johnson, K. (1999). Making Loyalty Program more rewarding . Director Marketing, 11(61), 24-27.
- [19] Brito, Pedro Quelhas, Carlos Soares , Sérgio Almeida, Ana Monte, and Michel Byvoet. "Customer segmentation in a large database of an online customized fashion business." Robotics and Computer-Integrated Manufacturing , Elsevier ,2015.
- [20] F.R.Frederick, and P.Schefter, "E-loyalty: Your Secret Weapon on the Web," Harbard Business Review, vol. 78, pp. 105-113, July 2000.