The Feasible Infrastructure of Customer Segregation using Data Mining Approach

Darshana A Naik, Vivek K M, Seema S

Abstract: The customers are the basic assets of the company and their satisfaction plays a greater role in the development of a company. Customers are potentially the future source of profits. And for classifying the customers more accurately we need to have some understanding on the history of their transactions. In our project we made use of the concept of outline data. The outline data contains the useful information related to the customers. By our approach we are using machine learning and data mining algorithms to effectively classify the customers based on their purchase activities. The algorithm classifies the customers by assigning the score based on the calculation. We sorted the classified customers and further determined the loyalty of the customers. The loyalty with respect to each customer is determined based on the factors like frequency of purchase and usage of the offers provided. The classified data can be used by the marketing teams to focus on promotions relating to the loyal customers.

Keywords: Customer Segregation; Data Mining; OutlineData; Loyalty; Offers

I. INTRODUCTION

The marketing industries struggles a lot in discovering potential in customers that they manage huge number of manual procedures to decide the customer's faithfulness. This project intake the customer purchase activity data and then classifies a customer by assigning the score in the range from one to fifty and further the customers are sorted based on their score. This makes the marketing team to precisely determine the loyalty of the customers which further helps in planning the promotions for the better development of the organization. Normally the dataset will not be having the complete data due to missing information from the hesitant customers who don't provide all data, misguided judgement and human mistakes [1]. Therefore, filtering the data from huge dataset and to apply the data mining algorithms to fill in the "Missing "values of client data as well as to discover potential clients has set off the difficult problem that needs to be answered in the business institution by using the concept of big data [2]. The fundamental objective of each diligence is to realize every client independently and utilize this idea for making this understandable for each and every client to do business with industry preferably than with the contender industry [3].

Numerous organizations frequently make use the techniques of data mining for Customer relationship management, whose assistance gives more tailored, unique solution remitting solitary client's needs, rather than focusing on lump retailing ideas. By examining the buying and connection designs on website, organizations can make advertisements and advancements to customer's profiles, so that the clients are rarely inclined to be irritated with the undesirable solicitations like spam mails. These activities will bring in considerable savings in terms of cost to organizations. In spite of the fact that the data mining techniques is extending rapidly, many of the marketers are still depending on their experience rather than trusting on the results from the data mining techniques.

Data mining techniques simply can't replace the noteworthy job of the domain matter experts and their insights on the business knowledge. we can increase valuable outcomes by consolidating the Data mining techniques and business expertise. The data mining techniques can find the hidden patterns where in the most experienced business experts may have failed to notice. Subsequently, the merging of the business expertise's with the ability of the data mining techniques can aid the organizations in gaining an upper hand in their efforts to upgrade the customer management.

A great vertical extent big data might accommodate the huge amount of the repetitive and unrelated information that may influence the fulfillment of the algorithms used. So, the feature selection is considered utter most cardinal for the algorithms in machine learning.

II. DATA MINING TECHNIQUES

Data Mining is viewed as the most significant stage in the knowledge identification process. The procedure of extricating intriguing specimen from an enormous quantity of data is known as the data mining [4]. Data mining solves the problem by investigating the information effectively present in the datasets. The importance of data mining is the superficial procedure of distinguishing legitimate, narrative, helpful, and at last justifiable examples in data.

The data mining job can be divided into two groups namely: predictive and descriptive. The Predictive models perform induction on the present data so as to make prediction, as it were, attempting to foresee unknown values or future patterns or practices relying upon different factors or biblical values introduced in the mined database [4]. Descriptive mining model attempt to separate useful examples that can explain the mined information and investigate its properties and describe the generic properties of the information in the database. The Data mining implementation process can be shown as: -
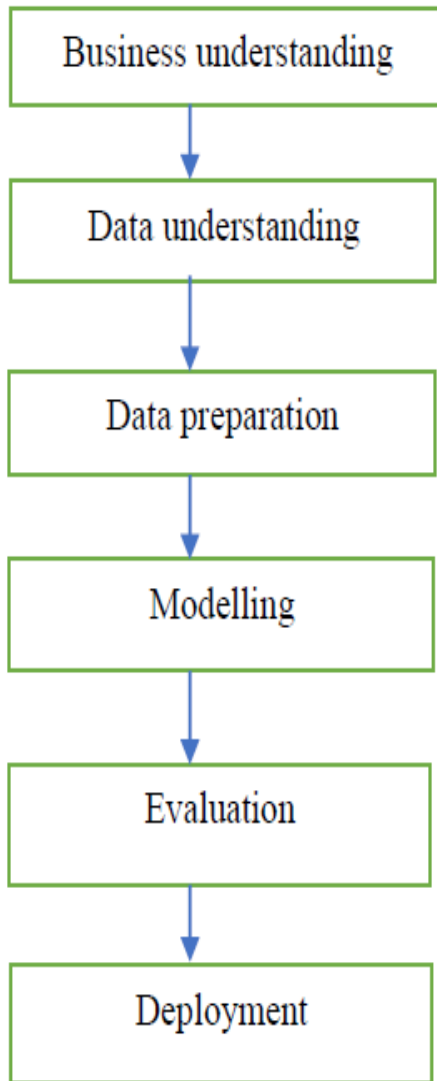
Figure 1. Stages in Implementation of Data Mining Process

In understanding the business phase data mining goals are established. To start with, you need to obtain business and client goals. Then you need to distinguish what exactly does your client needs. Assess the present data mining situation. Factor in assets, assumptions, limitations, and other critical components into your assessment. In the data understanding phase, the data is validated to known if it is applicable for the data mining goals. To start with data is gathered from multiple data sources present in the organization. In the data preparation phase, almost around ninety percent of the time will be spent for the project. The information from the dissimilar origin must be selected, wiped, designed, anonymzed, and developed if required. In the modelling phase, the data patterns are determined using the suitable mathematical models available. In the evaluation phase, patterns distinguished are assessed against the business objectives. In the last phase, the working data mining model can be produced to the end users for their business purpose.

The data mining technique relies upon the application area and the shape of the specimen which are required to be separated. We must realize that there is not solitary technique that is suitable for each and every feasible problem. The discovery of the regular frequent patterns, association, and connection among tremendous measures of information are helpful in numerous applications. Association rule mining algorithm utilize one of the two

basic methodologies: Depth first search and Breadth first search. Numerous productive and versatile algorithms are created for the purpose of rule mining as a result of which the correlations besides the association protocols are inferred. The primary distinction between the patterns that occurs sequentially and the association rules is that the component with respect to time is considered.

A use of data mining techniques in customer relationship management and customer segmentation problem is a prominent course in the organizations and retail companies. This has lured additional notice from scientists as well as specialists over the past few years.

III. LITERATURE SURVEY

The newer feature selection method solved physical world customer relationship management issue that had the turbulent and highly impure dataset. Here the SVM gave more precision and responsiveness and the Naïve bayes provided with elevated Receiver operating characteristic (ROC) and specificity along with the precision and the recall values [1].

The procedure of outlier inspection targeting at giving segments to assist industry arrange the consumers as per the client resources, in order to distinguish customers with great resources, and thereafter develop and enable customers in the well-focused manner, which significantly prevents the wastage of the resources happened by the decentralization [2].

The concepts of the value-based, Behavioral-based, value-at-risk based under the customer segmentation and the targeted marketing campaigns were discussed in depth in this paper and in addition the clustering techniques were used to obtain the optimized results under this model [3].

Data mining itself has provided the foundation for numerous domains together with the database systems, Machine learning, statistics, envision, and innumerable other application domains has achieved a greater advancement over the decade [4].

The machinery or automated improvement has empowered innovative data mining ways to that can be incorporated in the view for locating the most feasible Customer relationship management, where data mining plays a major part in analyzing the client's data [5].

This model introduced the new approach of interval decision information for customer classification problem. This approach had mainly four steps namely: interval information along with the decision matrix is used for normalizing, attributes weights were calculated, customers distance from the ideal point is measured, finally the customers were grouped based on the results obtained in. This model which is innovative could be easily incorporated into the computer information system-based equipment's [6].
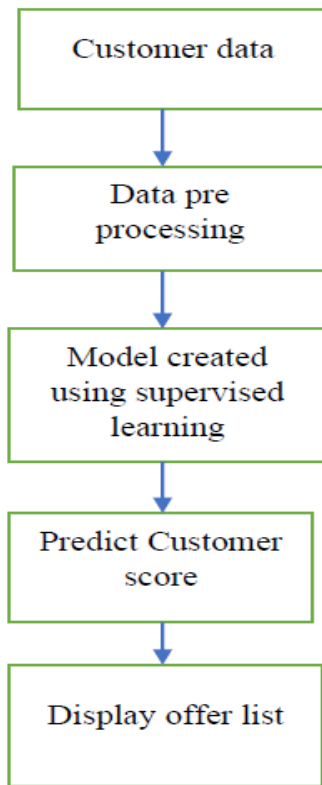
IV.   METHODOLOGY

Figure 2. Data Flow Diagram

The data flow diagram shows the order in which the design of the model proceeds. We have made use of the customer dataset which consists a total of 97606 entries and has eight features or attributes with one target class, which makes it an overall of nine fields. The various fields or attributes present are category, Description, quantity, price of each entity, country, Product, Total price and fscore or customer score which will be our target class. In general, the data preprocessing involves the transforming of the raw into a comprehensible setup. The Physical world statistics is frequently insufficient, inconsistent and lacking in definite etiquette and most credible to include blunder. But preprocessing the data is a proven technique of handling such issues. In our model for determining the customer or the fscore we have preprocessed the data and provided the category, quantity, total price, and product as the input to the data mining and machine learning algorithms.

We have made use of the supervised machine learning algorithms for the calculation of the customer score. We have taken random forest, Support Vector Machine (SVM) and K Nearest Neighbors algorithms from each of the ensemble model, classification model, and probability model respectively.

The customer score is calculated initially by taking into the consideration the attribute values of the category, total price, quantity, and product name. This input is provided to each of the three algorithms used in the project and the customer score is evaluated each time. As the customer does the transactions frequently his fscore also increases each and every time upon his purchase. Based on his transactions the customers fscore also changes accordingly. This means more the number of transactions higher the fscore of the customer will be. And based on the customer score or the fscore the offers are provided to the customer. In our model we have provided the upper bound range of fifty for the maximum offers that can be given to the customers. Finally,

the offers are revealed to the customers depending on the current fscore. And the loyalty of the customers can be determined by how well the customer uses the provided offers to him. We made use of flask which is basically a web framework written in the python programming language. As flask doesn't require particular libraries or the device it is considered as the micro framework. In addition, flask supports extensions that can add application features as if they were implemented in the flask itself.

V.   CUSTOMER SEGREGATION

Customer segregation is basically the procedure of partitioning the clients towards well defined, significant, and similar sections depending on the different qualities and attributes. This empowers institutions to comprehend their clients and develop distinct approach [7]. Conventionally institutions, irrespective of the diligence they work for, lean to make use customer segmentation blueprints which are built on the statistics and merit information. From the last one or two decades institutions are choosing their retailing ventures and building up their new items depending on these easy to understand, business protocol chunks.

The enormous difficulties while following the conventional review based statistical surveying are that it provides a huge amount of data about smaller number of customers. However, to utilize the outcome of statistical surveying it needs realizing the behaviors of each and every client who are present. Here the statistical surveying might discover intriguing segments of customers. After that these are required to be projected onto the current customers database utilizing the accessible information. Social information can be especially helpful for the same purpose; this kind of data could normally be encapsulated from settlement and the billing events in the past. One prerequisite of the statistical surveying is the clients should be recognized hence the behavior of the statistical surveying members is identified.

A sought-after approach of data mining along with the existing clients is the customer segregation. A segregation venture begins with contrast of career goals and terminates with the conveyance of separated retailing ideas for the chunks. A wide range of division types depending on the particular benchmark or properties pre-owned for the bifurcation. In particular, clients could be divided based on the individual value. This kind of bifurcation relies upon the particular retailer goals [4]. The sort of the segregation used relies upon the particular business objective and our objective. Various benchmark and bifurcation techniques are suitable for various circumstances and retailer goals.

Data mining could likewise be utilized in view of improvement of bifurcation blueprints depending on the present and anticipated estimation of the clients. These chunks are fundament so as to organize client managing and marketing involvement as per the significance of each and every client. More frequently, a business would be interested to perform the process of segmentation which puts each client towards some handily depicted chunk. Regularly, these chunks are developed as for promoting objective, for example, membership renewal. Decision tree strategies depicted in are perfect for this kind of segregation.

VI. RESULTS AND DISCUSSION

Table 1. An Overview of Analytics

Dataset Details		Accuracy Score	
No. of Records	97606	Random Forest	99.903%
No. of Columns	11	SVM	89.9%
No. of Classes	10	KNN	73.179%

The Overview of the Data Analytics part shows us that the customer dataset consists a total of 97606 records of the customers and these records are distributed across the 11 columns in the dataset and the dataset contains a total of 10 unique classes.
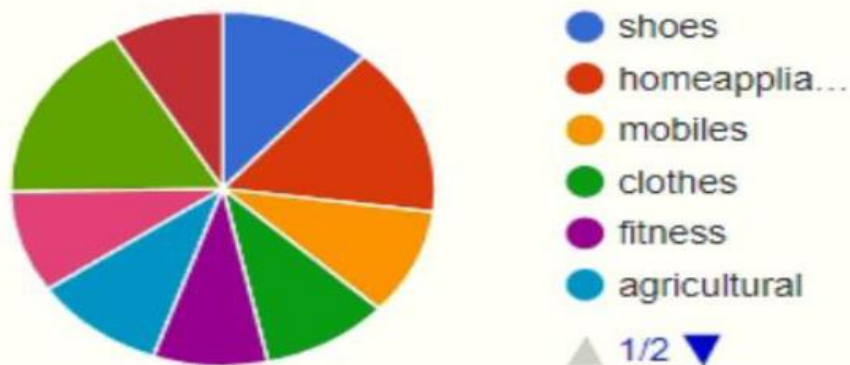

Figure 3. Categories

The Data Visualization part of the diagram shows the different categories that the customer can choose in order to make his transactions. This Means these are the various available options for the customers to choose for the transactions. In the above diagram each of the category is represented by the unique colors. And the area occupied by each of the individual categories represents the percentage of the each of the category which makes up the complete customer dataset. Now from any of these individual categories or from the combination of the categories the customer can select for the transaction. And based on the purchase activities the customer score is calculated and finally the offers list will be displayed to the customer.

The different categories from which the customers can make his transactions includes shoes, home appliances, mobiles, clothes, fitness, agricultural, biscuits, chocolates & laptops. Now the customers can select from any of these categories and under the selected category the customer has to further select the product, total quantity for his transaction. For an example a customer can choose clothes category and under cloth category he can choose the jeans as product name and he can decide to purchase a total of 3 jeans for his transactions. Initially the customer has to register by making use of his email id and password and later the customer can login using the email id to view the offers provided to him. The offers provided will be based on the customer score or the fscore. And for availing a greater amount of discount the customer has to increase the total number of transactions.
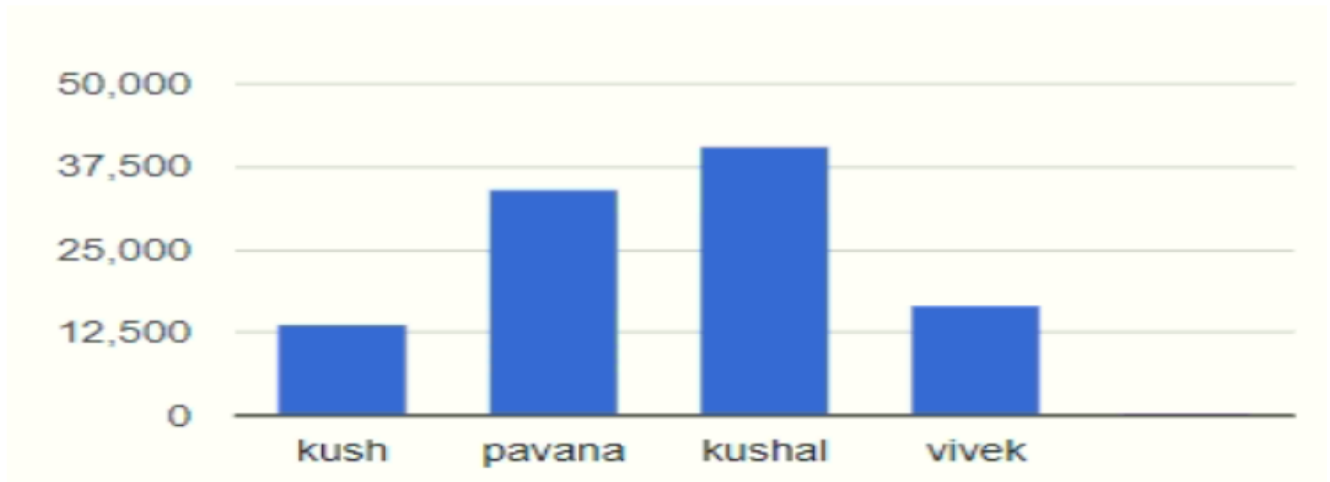
Figure 4. Customer Transaction

The above graph shows the customer transaction of customers versus the customer score. This graph is plotted by taking the updated value of the customer score or fscore of each customer. And the percentage of offer provided is equal to the current value of the customer score of the customer. So, the customer with higher purchase activities will have good customer score and accordingly will be given the better offers. In the above graph the customer kushal is having the highest customer score amongst the other customers as Kushal has done more number of transactions.
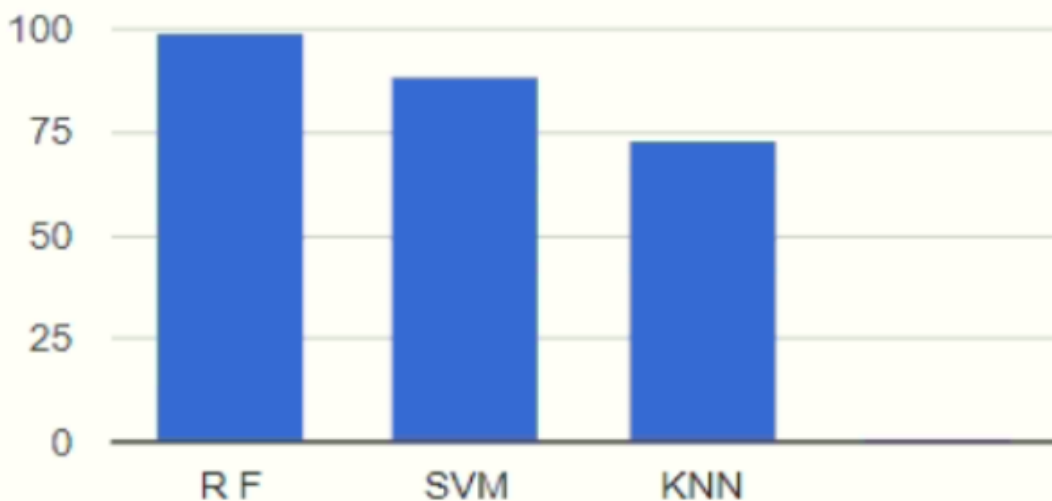

Figure 5. Accuracy Graph

The above graph shows the Algorithms used versus the accuracy percentage of each algorithm. We have made use of 3 algorithms to calculate the customer score. Out of these three algorithms Random forest, Support Vector Machine & K - Nearest Neighbors the Random forest with 99.903% of accuracy has given the better prediction score when compared to other algorithms. The SVM and K-NN has the accuracy percentage of 89.9% and 73.179% respectively. Three algorithms are used to compare the performance of each algorithms for predicting or calculating the customer score or fscore.

VII.   CONCLUSION

In our paper we sketched out how data mining can assist an organization to effectively implement the customer relationship management objectives and accomplish individualized and more productive customer management via customer point of view. Customer segregation is the way towards recognizing groups that have regular attributes. The primary goal of the customer segmentation is to comprehend the customer base and increase client knowledge that will empower the design and development of differentiated marketing strategies. To be precise we have applied the concept of value-based customer segmentation technique i.e. the customers are ranked and segregated according to current and expected or estimated customer value.

REFERENCES

1. "Customer Relationship management classification using data mining techniques". S. UmmugulthumNatchiar, Information Technology, Sethu Institute of Technology, Virudhu Nagar, India. S. Baulkani, Electronics and Communication Engineering, Government College of Engineering, Tirunelveli, India.
2. "Customer classification of discrete data concerning customer assets based on data mining". ZuoLei, GuoJunfeng, Harbin Finance University, Harbin, Heilongjiang,150030, China 523406570@qq.com
3. "Using Mata Mining techniques in customer segmentation". Hasan Ziafat Int. Journal of Engineering Research and Applications, ISSN: 2248-9622, Vol. 4, Issue 9 (Version 3), September 2014, pp.70-79.
4. J. Han and M. Kamber, Data Mining: Concepts and Techniques (2nd ed.). Morgan Kaufmann, 2006.

5. Ngai, E. W. T., Xiu, L., & Chau, D. C. K.," Application of data mining techniques in customer relationship management: A literature review and classification," Expert Systems with Applications, 2009, 2, 2592–2602.
6. "A New Customer Classification method for customer asset value management".
7. K. Tsiptsis and A. Chorianopoulos, "Data Mining Techniques in CRM: Inside Customer Segmentation", John Wiley and Sons, Ltd., Publication, 2009.