Retail Customer Churn Analysis using K-Means Clustering and the RFM Model

Bivas Ranjan Parida
College of Engineering Bhubaneswar, Biju Pattnaik University of Technology, Odisha, India

Abstract— In this current world of business, Customer Churn is one of the major concerns for various business owners or the organizations for maintaining existing and attracting new customers. Analysis of various types of customers can be conducted by researching customer relationship management which in turn provides strong support for business decisions. Customer churn occurs when certain customers are no longer loyal or a part of a particular business. Losing customers will not only result in losses but also develop threat to the organization. Because of multiple competitors in the same business, the re-engagement of customers who are less interested is essential rather than engaging a new one. It is observed that acquiring new buyers is costlier than retaining the present customer. Churn prediction is a new promising method in customer relationship management to analyze customer behavior by identifying customers with a high probability to discontinue the company based on analyzing their past data and also identify strategies for improvement. Once a customer becomes a churn, the loss incurred by the corporate isn't just the lost revenue but also the prices involved in additional marketing in order to attract new customer. Reducing customer churn is a key business goal. In this project, we've taken a dataset from UCI Machine Learning repository. This dataset contains records of transactions that happened between December 1, 2010 and December 1, 2011. This is recorded from a web retail gift store based in the United Kingdom. Here segmentation of customers has been done by using RFM technique and K-means algorithm.

Keywords— Customer Churn, RFM model, K-Means Clustering, customer segmentation

I. INTRODUCTION

In recent years, with the rapid development of electronic commerce, the numbers of e-commerce businesses are booming more and more and on a large scale, and the service has become increasingly homogeneous, making competition more intense among e-commerce businesses. Under an ecommerce environment, companies use the Internet platform to service customers, customers browse the network platform, the buying process produces a large amount of data traffic, and the traffic in the form of data is easy to access any ecommerce businesses. Based on the unique advantages of ecommerce businesses, large amounts of data through data mining, information needs to get customers, provide customers with personalized service, and constantly improve customer satisfaction and loyalty, which has become the main goal of the e-commerce world. In the retail sector, some customers stick around while others stop shopping at a particular store after a certain period of time. Detecting which customers have decided to buy elsewhere and which of them are idle at the instant, may be a difficult task to any organization. A customer likely to break the relationship or lower the purchase rate is known as churn. There are many reasons for customer churn. Depending on which customer churn can be divided into two categories: active and passive.

Active churn, namely voluntary leaving. Means a customer does not do online shopping due to his/her own reasons such as changing jobs, quality of service, business competition, loss of professional etc. Passive churn, also known as involuntary leaving, refers to the type of customer churn that the enterprise should be responsible for. This occurs because the enterprise decides to cancel customers' accounts for some reason, usually due to their credit problems. Customer churn is the tendency of customers to stop purchasing with a company over a time period. Customer churn is also called customer attrition or customer defection. Churning reduces growth. Therefore, companies should have a proper defined method to compute customer churn rate for a given time. By keeping track of churn rate, organizations are often equipped to succeed in terms of customer retention. Retailers need a good strategy to manage customer churn. Measuring the churn rate is kind of crucial for retail businesses because the metric reflects customer response towards the service, quality, price and competition. Churn prediction envision the likelihood of customers to churn. It pares the investment on gaining new customers and helps to retain the existing customer. The marketing efforts and amount spent on attracting a new customer is higher and more difficult than clinging to existing customers. Customers who are unlikely to make a purchase or willing to shift the shopping site because of cautiousness with money, expecting standard and assortment in products can be convinced and clutched. The customers who are ending the relationship due to valuable and unavoidable reasons are free to leave.

Result is firm, though we invest in involuntary churners. Target marketing aids to reach the customers and connect with them. The voluntary churners are often stopped by extending discounts, amending the products to customers choice and by sending out trigger mails. Concentrating only on voluntary churners will scale down the cost of offering benefits to yet and all churned customers.

As the development of the online retail market intensifies competition among the industry. For the e-commerce industry, the customer churn rate is high, business operators need to consider how to minimize the customer churn rate of online shopping. Because the customer's behavior is predictable, through the relevant data collected to hold out the relevant analysis can find the customer's future trading tendencies. For business operators, reduce the number of lost customers, an effective way is to find the customer who has the loose tendency and do the relevant pre-control work. In recent years, online shopping customer churn prediction has become a crucial direction of e-commerce business research.

II. RELATED WORK

In this paper [1] various algorithms are compared and contrasted in predicting customer churn for a retail

business is done and recommendation is given based on the cluster the customer belongs to. Different prediction algorithms were analyzed and studied and the best among them was chosen. Comparative study of each churn predictive algorithm was done and Pareto/NBD was chosen among them. The input parameters for the model were fine tuned to suit the needs of the model. The products are categorized, and the customers are put into a cluster based on their RMF scores. The commendation is given based on the cluster they belong to. In this paper [2], the authors have considered a data set from UCI Machine Learning repository which is recorded from an online retail gift store based in the United Kingdom This data set is pre-processed by removing NAs, validating numerical values, removing erroneous data points. After that aggregation was performed on the data to generate invoice based and customer -based data sets. A variable churn is attached to each data point. Three algorithms are run on this customer aggregated data. After running all the three models, they arrived at a result that accuracy of Random Forest, Support Vector Machines and Extreme Gradient Boosting are increasingly higher. However, the time taken to finish computation also increases in the same order. In the next paper [3], using decision tree algorithms in data mining techniques the authors analyze the basic information ofecommerce business customers, and found out the features of customer churns. In this paper [4], the project proposed churn prediction in e-retail with structured data, imperative features, machine learning techniques. On cross validation better algorithms are picked and elected based on voting. Ensemble algorithm is the closing result and suggests appropriate preventive measures on churners. The authors also studied the existing system. In existing model predictions done with unstructured or semi-structured data. Related works stated results with one particular machine learning algorithm. A comparative study on various algorithms comes out with the best model based on accuracy. Accuracy of the algorithm varies depending on customer data. Xiaojun Wu et al. [5] noted that in the usual scenario, the e-commerce customer churn datasets are imbalanced. It is found that the count of churn customers greatly exceeds non churned consumers. Due to this traditional algorithm favors the majority set and does not give accurate results. Hence, the paper proposes improved SMOTE technique to balance the data. It then uses AdaBoost on this data to classify the customers. In this paper [6], it takes the non-contract scenario of online shopping customers as an example, select transaction data of a domestic e-commerce website for empirical research. On the basis of the single model-BP neural network and support vector machine, apply the integrated learning theory to the online shopping customer classification. The empirical results show that the combined forecasting model has a significant improvement in the hit rate, coverage rate, accuracy rate and lift degree, and so on. In the next paper [7], the authors have implemented an Exploratory Data analysis using Visualization, statistical tests for feature selection and Data mining methods for predicting the likely churners by utilizing a Logistic Regression Model. It is observed that the logistic regression model has predicted better results in the prediction process of churn. By increasing the threshold values and selecting the right features with various combinations, it will deliver a better result. In this paper [8], it reviews the most popular machine learning algorithms used by researchers for churn predicting, not only in the banking sector but also other sectors which highly depend on customer participation. Each churn prediction model studied here has low accuracy and prediction. The next paper [9] applies many techniques of data mining to the research of customer churn, such as clustering analysis, decision tree, neural network, etc, establishes an ecommerce customer churn model and analyses the factors which influence customer retention. There were five major factors about trust in the questionnaire: safety and reliability of website, sustaining and stable business, practical product advertisement, strength of company, cost performance of products. In paper [10], through the study, it has been evaluated that the perception of online shopping is influenced by educational level and gender, age and gender, and income and gender. The initially taken 20 variables are reduced to only 4 factors, while we evaluate the influencing factors for the online purchase. Out of these factors the marketer's integrity towards the service quality is basically considered as the most important factor for influencing online trust of consumers. Our study's analytical results further indicate the relationships between the consumers' perceptions of the factors which are influencing their intention to buy via online; more precisely, the consumers' perceptions of security and privacy, content and design on website, service quality and the factors of customer delight. These analytical results gathered are generally consistent with the findings of the previous researchers. The factors which have received the most consistent support are service quality and web security based on which the consumers' trust for online shopping is formed and influenced. In [11], here various cluster algorithms are taken to analyze their performance by using patient churn data set. The comparison between different clustering was done. It showed that the K-means analysis algorithm is suitable for customer churn analysis, but it requires pre-processing steps for mislaid values. K-medoids take more iterationwhen the number of clusters are increased and Euclidean distance function is used for distance measure. In fuzzy c means cluster we need to define the number of iterations and clusters. Hierarchical clustering does not cluster allobjects in a single step, it takes more time and iterations to cluster objects when the data set is large. The DBSCAN method has distance functions which are cosine and Euclidean distance functions. In paper [12], authors have proposed hybrid Classification techniques that have shown their superiorities over single algorithm techniques. The approach is to compare 2 kinds of hybrid methods by classification with classification and clustering with classification. Here, the prediction of the customer churn in the telecom industry is done. They implemented Hybrid Decision Tree and Logistic Regression Classifiers. This technique provides better results, but it takes a lot of time for execution. To tackle this problem; authors have proposed another hybrid model, hybrid Fuzzy unordered rule induction algorithm (FURIA) with Fuzzy C-Means Clustering for predicting customer churn. FURIA is quite fast in execution without much compromise on theaccuracy. Also, clustering with FURIA helps to move from predictive to prescriptive analysis by providing the reason for churn.

B. Model Description

As shown in the Figure (1), the model consists of the following steps to acquire the cluster which the is nothing but customer segmentation. Here at first the whole dataset has been acquired and loaded then by applying removing duplicate values, handling missing values, scaling data the whole preprocessing or cleaning of the data has been done. Applying RFM model to get the Recency, Frequency, Monetary score for each customer and lastly applying K- means algorithm to assign the customer to the cluster to which they belong.

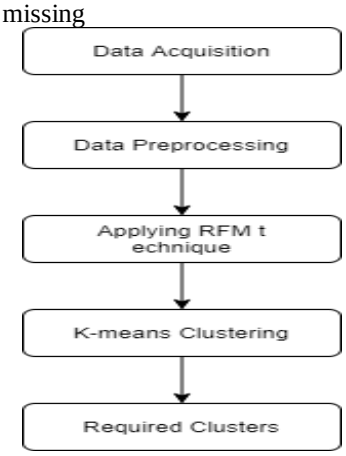

Fig. 1. Churn Analysis Model

III. METHODOLOGY

A. *Data Description*

The dataset used is a transnational data set [13] which contains all the transactions for a UK-based online retail is used for clustering purposes. It has 541909 rows and 8 columns. The data set has following 8 attributes –

1. Invoice No: Invoice number. Unique integral number assigned to each transaction. Cancelled transaction is indicated by 'c'.
2. Stock Code: Product code. Unique nominalnumber assigned to each distinct product.
3. Description: Product name. Nominal.
4. Quantity: The quantity of each product in every transaction. Numeric.
5. Invoice Date: Invoice Date and time. Transaction date and time. Numeric.
6. Unit Price: Unit price. Numeric, Product price per unit.
7. Customer ID: Customer number. Nominal, unique integral number assigned to each customer.
8. Country: Country name of the customer.

C. *Data Preprocessing*

The dataset mentioned in the previous section is in the formof a CSV file. The CSV file is imported, cleaned, aggregated as per the requirements, and made into a new data frame. The retail sales dataset has a total size of 349096 tuples when compared to the original raw dataset which had around 541909 tuples. The pre-processing step involved cleaning of data by removing the NA values and also defining one more attribute called Revenue which is the product of Quantity and Unit price. This process also involved creation of two more datasets using aggregation which are customer aggregate data and invoice aggregate data. The missing entries in the customerID field, duplicate entries, cancelled transactions that completely got cancelled and the invoices which do not relate to customers are removed.

D. *RFM Technique*

RFM (Recency, Frequency, Monetary) analysis is a marketing model for customer segmentation. It is based on customer behavior. It groups customers based on their transactional history that is how recently, how often and how much did they buy. RFM helps divide customers

into various clusters to identify customers who are more likely to discontinue the business relationship.

1. Recency: The freshness of customer activities: Time since last transaction.
2. Frequency: The Frequency of customer transaction. E.g., the total number of recorded transactions.
3. Monetary: The total amount paid. E.g., the total transaction value.

RFM factors illustrate these facts:

1. The newer purchase, the more responsive the customer is to promotions.
2. The more frequently the customer buys, the more engaged and satisfied they are.
3. Monetary value differentiates heavy spenders from low value purchasers.

E. *RFM Implementation*

Following are the steps for RFM Calculation -

2. Recency calculation- Using the most recent date in the complete dataset, the recency for all customers is calculated by subtracting the most

recent date from the customers recent date of transaction. The most recent date from the complete dataset is found to 2011-12-09. This is used as a benchmark date for the calculation of recency value for each customer.

3. Frequency calculation- Adding the number of times a particular customer has purchased, gavethe frequency.

4. Monetary calculation- Using the quantity bought in order and unit price of product, the total money spent by customer on each transaction is calculated as total. Sum of this total of each transaction is taken for each customer to get the total money spent by the customer.

	CustomerID	Recency	Frequency	Monetary
0	12346.0	325	1	77183.60
1	12747.0	2	103	4196.01
2	12748.0	0	4596	33719.73
3	12749.0	3	199	4090.88
4	12820.0	3	59	942.34

Fig. 2. RFM values for each customer

	Recency	Frequency	Monetary
count	3921.000000	3921.000000	3921.000000
mean	91.722265	90.371079	1863.910113
std	99.528532	217.796155	7481.922217
min	0.000000	1.000000	0.000000
25%	17.000000	17.000000	300.040000
50%	50.000000	41.000000	651.820000
75%	142.000000	99.000000	1575.890000
max	373.000000	7847.000000	259657.300000

Fig. 3. Description of RFM values

From Figure (3), we can observe that the average recency of the customers are 92 days (approx..), on an average the customers are purchasing the product 90 times andspending an average 1863.91 price.
It can be observed from Figure (2), the values of monetary are quite large in comparison to recency and frequencyvalues. So, using log transformation, they are scaled properly to be given to the K-Means algorithm as input.

Based on the Recency, Frequency and Monetary values R_score, F_score and M_score is calculated and grouped together to get the RFM_score which is sum of R_score, F_score and M_score as shown in below Figure (4), RFM Scores have been calculated now we will use this score to make segments of the customers and define level of loyalty.

CustomerID	Recency	Frequency	Monetary	R_score	F_score	M_score	RFM_Group	RFM_Score
12346.0	325	1	77183.60	4	4	1	441	9
12747.0	2	103	4196.01	1	1	1	111	3
12748.0	1	4596	33719.73	1	1	1	111	3
12749.0	3	199	4090.88	1	1	1	111	3
12820.0	3	59	942.34	1	2	2	122	5

Fig. 4. RFM scores for each customer

F. K Means Clustering Algorithm

Clustering is defined as the process which divides the whole data into groups or clusters supporting the patterns within the data. To process the training data, the K-meansalgorithm in data processing starts with a primary group of randomly selected centroids, which are used because the beginning points for each cluster, then performs iterative calculations to optimize the positions of the centroids.

K-Means Algorithm-

1. Initialize k points, called means, randomly.
2. Categorize each item to its closest mean and update the mean's coordinates, which are the averages of the things categorized therein mean thus far.
3. Repeat the method for a given number of iterations or till the clusters forming are the

same as previous and at the top, have the ultimate clusters.

G. Implementation of K-means Algorithm

In this project, for creation of customer segmentation using K-Means algorithm based on the R, F, and M Scores, it is essential to decide the number of clusters to form i.e. the value of K. For deciding the value of k Elbow Technique isused.

Elbow technique: The elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters. By default, the distortion score is computed, the sum of square distances from each point to its assigned center and picking the elbow of the curve as the number of clusters to use.
We can observe from Figure (5), as the number of clusters increases the sum of square distances are becoming lesser. And will take the count of clusters where this elbow is bending. In our case, the sum of square distance is dramatically decreasing at K = 3, so this is the optimal value to choose for no of clusters.

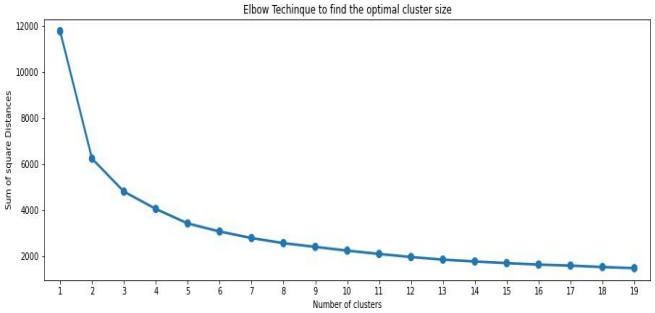

Fig. 5. Elbow Technique to find number of cluster

Visualization for Recency, Frequency and Monetary basedon Cluster groups is shown below in Figure(6),(7),(8) .

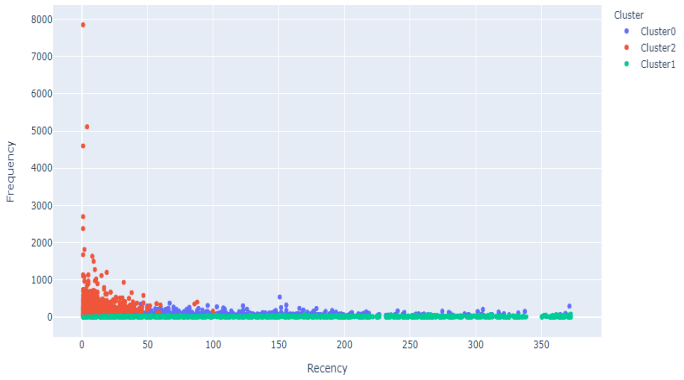

Fig. 6. Recency vs Frequency

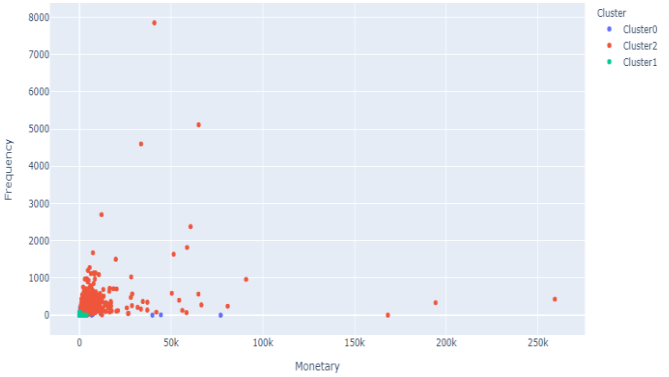

Fig. 7. Monetary vs Frequency

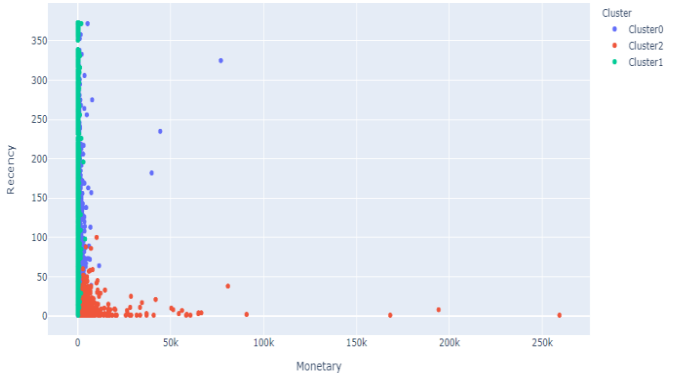

Fig. 8. Monetary vs Recency

At last, from the above graphs, number of customers for each cluster is calculated. In figure, three clusters can be seen. Out of these three, the red colored cluster i.e., Cluster 2 represents Best customers or we can say Champion Customers having total number of customers as 838. Another cluster, the blue colored i.e., Cluster 1 represents the loyal customer having a total number of customers as 1662. And the last cluster i.e., Cluster 0 which is colored cyan gives the customers at risk. Total customers at risk are 1421.At last from each cluster, top 5customers with detailed information are also retrieved.

IV. CONCLUSION

In the current increasingly e-commerce environment, analyzing the customers' behavior, attracting new potential customers and maintaining existing customers by increasing their satisfaction are the best ways to sustain and win the market competition. Customer churn analysis allows the business to prepare a solid base for future marketing analysis and campaigns. Customer churn analysis also supports customer-driven product development with keeping customers engaged and loyal over time. In this project Clusters have been created based on the values of recency, frequency and monetary which was calculated by using RFM Model with the help of K-Means Clustering. The number of customers present in each cluster is calculated. At the end the customers which are at risk and need more attention of the business are identified. Based on this identification, proper techniques can be applied to retain such customers.

REFERENCES

[1] Punya P Shetty, Varsha C M, Varsha D Vadone, Shalini Sarode, Pradeep Kumar D," Customers Churn Prediction with Rfm Model and Building a Recommendation System using Semi-Supervised Learning in Retail Sector", International Journal of Recent Technology and Engineering (IJRTE),2019, ISSN: 2277-3878, Volume-8, Issue-1

[2] Shoaib, Tahmina,2018, "Customers churn prediction in retail store", 10.13140/RG.2.2.30545.38242

[3] F. Guo and H. Qin, "The Analysis of Customer Churns in e-Commerce Based on Decision Tree," 2015 International Conference on Computer Science and Applications (CSA), Wuhan,2015, pp. 199-203, doi: 10.1109/CSA.2015.74.

[4] M Jaeyalakshmi, S Gnanavel, K S Guhapriya, S Harshini Phriyaa, K Kavya Sree," Prediction of Customer Churn on e-Retailing", 2020, International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8 Issue-6.

[5] Xiaojun Wu and Sufang Meng, "E-commerce customer churn prediction based on improved SMOTE and AdaBoost," 2016 13th International Conference on Service Systems and Service Management (ICSSSM), Kunming, 2016, pp. 1-5, doi: 10.1109/ICSSSM.2016.7538581

[6] Xia, Guoen & He, Qingzhe. (2018). The Research of Online Shopping Customer Churn Prediction Based on Integrated Learning. 10.2991/mecae-18.2018.133

[7] Ahmad, A.K., Jafar, A. & Aljoumaa, K. "Customer churn prediction in telecom using machine learning in big data platform", J Big Data 6, 28 (2019). https://doi.org/10.1186/s40537-019-0191-6

[8] Kumar, A. S. and D. Chandrakala. "A Survey on Customer Churn Prediction using Machine Learning Techniques." International Journal of Computer Applications 154 (2016): 13-16.

[9] H. -l. Wu, W. -w. Zhang and Y. -y. Zhang, "An Empirical Study of Customer Churn in E-Commerce Based on Data Mining," 2010 International Conference on Management and Service Science, Wuhan, China, 2010, pp. 1-4, doi: 10.1109/ICMSS.2010.5576627

[10] Singh, Amit Kumar & Ajmani, Aayush," Future of B2C E-commerce (Buyers Perspective) in India: An Empirical Analysis," Asia-Pacific Journal of Management Research and Innovation,2017,12. 2319510X1668898.10.1177/2319510X16688986.

[11] I. Franciska and B. Swaminathan, "Churn prediction analysis using various clustering algorithms in KNIME analytics platform," 2017 Third International Conference on Sensing, Signal Processing and Security (ICSSS), Chennai, 2017, pp. 166-170, doi: 10.1109/SSPS.2017.8071585.

[12] A. S. Choudhari and M. Potey, "Predictive to Prescriptive Analysis for Customer Churn in Telecom Industry Using Hybrid Data Mining Techniques," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-6, doi: 10.1109/ICCUBEA.2018.8697532.

[13] Online Retail Dataset from UCI ML Repo, https://www.kaggle.com/jihyeseo/online-retail-data-set-from-uci- ml-repo.