2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT) | 979-8-3503-9763-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICEEICT56924.2023.10156962

Evaluative study of cluster based customer churn prediction against conventional RFM based churn model

Harish A S[1*] and Malathy C[2]
Department of Networking and Communications, School of Computing, SRM Institute of Science and Technology, Kattankuluthur, Chennai, Tamil Nadu, 603203, India
[1]hs3571@srmist.edu.in
[2]malathyc@srmist.edu.in

***Abstract*— Churn prediction in retail industry is a very critical measure as retention strategies play a crucial role in any successful business model. In the era of e-commerce and customized customer solutions, it is essential for businesses to be on their toes to predict the changes in customer behavior as early as possible in their life cycle. RFM analysis is one of the age old techniques that have proven to be an accurate measure of customer behavior and although it aids in customized targeting, it also requires time to amass enough data. It sometimes means that by the time the business collects substantial data about a customer to make strategic decisions, they are already half way through their change in interests or behavior. Therefore, it is essential to study the possibilities of predicting future states of a customer well in advance. This study focuses on evaluating the accuracy of Machine Learning Models in predicting the churn of retail customers based on their k-means clusters against their actual RFM features. This evaluative comparison would help to establish how reliable it is to utilize a model's results to feed another predictive model and its further applications.**

Keywords*— *Churn, Retail Analytics, Predictive Modeling, RFM Model, Churn Prediction, Clustering, Machine Learning, Logistic Regression, Support Vector Machine, Random Forest, CART, GBM, Gradient Boosting

I. Introduction

Retail industry is fast paced and dynamic where the status and interests of a customer evolves rapidly. This means that before the business is able to design and implement a strategic solution in the market, the customers have already shifted their behavior.

Conventional business models dealt with this situation by approximating the rate of key metrics such as acquisition, attrition, revenue spike etc. But in today's era of data and information management, predictive analytics can be a boon to retail businesses to accurately predict the near future states of a customer to implement tailor made solutions for their various customers.

The businesses are now investing more and more in Customer Relationship Management (CRM), which is a combination of business, technological and analytical tactics that helps to build long lasting relationship with the customers. It consists of four major dimensions – to identify potential customer, to attract them with customized offers, to retain their relationship and to develop their relationship with the business. While the identification and attraction aspects go hand in hand, the retention and development strategies usually co-exist and support one another. In this study, the aim is to explore more around the latter and focus on customer churn prediction which is critical for a business to better retain and improve the customer net worth.

RFM metrics are essentially one of the most effective parameters [1] in the retail industry widely used for various predictive applications and this has been utilized for churn prediction in various studies. But calculation of absolute Recency, Frequency and Monetary values of a customer needs substantial amount of historic data [2] before the business can reliably utilize it in a model. Therefore, this study aims in utilizing the clusters obtained by k-means clustering, to predict the churn prediction and compare it with the churn prediction results obtained by the same set of algorithms. This would help to ease the process of churn prediction in cases where absolute data is unavailable for all customers.

The churn prediction will be done using logistic regression, CART, Random Forest, Support Vector Machine and Gradient Boosting algorithms [3, 4]. This will be one using absolute RFM metrics as well as the k-means clusters [1] obtained for the same populations and the comparative results will help to establish how reliable the clusters are in terms of churn prediction.

II. Customer Churn

Churn rate, otherwise known as customer attrition rate, is the rate at which customer off board a business and this may be voluntary or involuntary [5]. Voluntary churn is when the customer makes a conscious decision of leaving the business due to losing interest, dissatisfaction, acquired by competitor etc. Involuntary churn is when the customer leaves due to external factors such as geography etc.

The customers of a retail business can be classified into three broad categories as occasional, repeat and loyal customers. Occasional buyers are the casual ones who hover

around the businesses or may be visiting on rare occasions such as purchasing gifts or visiting a business in a different geography etc. They usually do not have high potential of revenue for the business. The loyal category of customers are the ones who visit the business time and again and generally show a keen interest and affliction towards the brand in general, bringing considerable revenue in flow. The largest group of customers would essentially fall in repeat category and are the ones that move across competitors and usually churn if not catered to their interests and requirements. They are also the ones with high revenue potential and therefore identifying them accurately would help with taking strategic actions to keep them loyal [6].

$$\text{Churn Rate } \% = \frac{\text{Number of Customers Lost in T Period}}{\text{Total Number of Customers in T Period}} \text{ X } 100 \quad (1)$$

Customer Churn can be defined in various ways specific to businesses and it is usually marked by a termination of relationship between the customer and business in contractual or subscription based industries [4]. In case of retail industry, purchase pattern [7] is usually a reliable method of establishing customer churn as there would be no solid sign of termination of the customer relationship. Deterioration in number of visits or volume of purchase usually indicates the customer's loss of interest or a positive affliction towards other competitors [8].

2.1 Churn Prediction

It is established by various researches that a business incurs more charge in acquiring a new customer than retaining an existing customer with customized incentives [9, 10]. Therefore, it is a common practice throughout time to provide attractive offers and benefits to customers who attrite or terminate their relationship [6]. In retail industry, this is a major challenge because aside from the obvious difficulty in pin pointing churning customers there is also a factor of time where any delays in strategic remedies will largely impact the retention prospect. This is why predicting churning customers as well in advance as possible is of much value to retail businesses.

Although conventional mathematical models have been used to predict or calculate churn prospect of a customer, the machine learning algorithms [11] have proven to be one of the robust and scalable solutions for churn prediction that is being widely implemented in commercial sector today.

III. Machine Learning Models

Machine Learning is an artificial intelligence based methodology that is used to build models and statistical algorithms. It gives computers an ability to learn without a manual programming or guidance [12]. In simple terms, machine learning algorithms are when computers are fed observations which are processed to train themselves to detect various patterns and associations that can be used to make better decisions and predictions [13]. The algorithms also improve the accuracy of prediction by learning from the past predictions cycle. Supervised Learning Algorithms are a sub category of machine learning techniques where training is done using examples based on absolute data from a sample population. The training dataset is labeled containing both the required input and output. An extension of this is an Ensemble Learning Algorithm where several base learning models are combined to make one optimal predictive model. In this study, five machine learning algorithms for churn prediction are used where the absolute values of churn (Yes/No) is pre calculated and is used to train and test for accuracy.

3.1 Logistic Regression

Logistic Regression is an effective algorithm primarily used for binary classification applications. It bases the outcome on the probability of an event taking place by training itself on a given labeled dataset and bounds the dependent variable from 0 to 1 based on the probability. This is indicated in equation 2, where x is the independent variable and logistic function denotes the dependent variable.

$$logistic\ function(x) = \frac{1}{1+e^{-x}} \quad (2)$$

In case of classification problems, the dependent variable is declared based on the positive or negative orientation of the probability. For example, the logistical probability of a customer churning is used to predict the absolute cases of churn.

3.2 Classification and Regression Tree (CART)

CART model, also known as Decision Tree, is another popular machine learning algorithm also used for both regression and classification. It has a hierarchical tree structure and consists of a root node, branches, internal node and leaf nodes. It works by recursively partitioning the data into sub spaces and continues until the outcomes are as homogenous as possible. The individual features act as the parameter for internal nodes and therefore it is essential to be of categorical in nature. In case of continuous variables, the values are discretized before being utilized in the CART model. The selection of root node at each level is the challenging aspect of a decision tree and the smaller the tree, the better balanced and accurate the tree is.

3.3 Random Forest

Random Forest is an ensemble learning model that is highly robust and is able to deal with highly complex and large scale data with better accuracy. It creates a multitude of decision trees on subsets of datasets as base learners and combines the collective solution to finally select the best solution in terms of features or nodes. It eliminates the shortcomings of a CART model by avoiding over-fitting and they work particularly well when the individual trees are as diverse from each other. While decision trees consider all the possible feature splits in the design and solution, random forests deals with various combinations of features to identify the right subset of features that supports the prediction as illustrated in Fig 1.

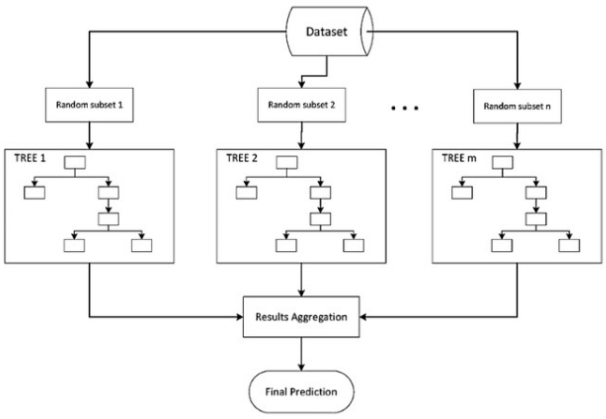


Figure 1: Random Forest combining multiple decision trees

3.4 Support Vector Machine (SVM)

SVM algorithm can be used for a variety of predictive applications, but is predominantly used for classification applications. It works by plotting each observation as a point in n-dimensional space with the number of features being *n* and the values of each feature being the point on the respective co-ordinates. The co-ordinates are referred as Support Vectors and hence the name. The classification is then performed by finding the hyper plane that splits the two classes effectively. The dimension of the hyper plane may vary depending on the number of features with the simplest being just a line in cases with to features. This plane is then taken as reference to classify the test dataset.

3.5 Gradient Boosting Machine (GBM)

Gradient Boosting Machine is a powerful algorithm that employs a bunch of machine learning techniques by using the loss function to identify the weak learners and potentially identifying the reliable learner in the process. The residual errors of each predictor is then fed back to train the subsequent predictors thereby achieving a stronger learner. In applications with large scale complex data, this prediction technique has acquired an important standpoint owing to its prediction speed and accuracy.

IV. RFM AND K-MEANS CLUSTERING

RFM stands for recency, frequency, and monetary value. It assigns a clear measure to their last purchase, how often they have been engaging with the business and how much they have spent. This has proven to be an effective metric for various strategic models in retail, especially for customer segmentation and customized marketing [2]. It is a reliable and simple model, yet there are some limitations where the customer's behavior can only be reliably implied over the given period and does not consider other customer features. Therefore, it may not be well adjusted to predict future customer activity. However, combining RFM metrics with other predictive algorithms may be an effective strategy to combat the shortcoming and to effectively combine the benefits of both aspects.

The k-means clustering on RFM is a tried and tested model that is proven to be of utmost benefit to segment customers and to make specific decisions that are tailor made to meet customer expectations [1]. This assigns an approximation factor to each customer effectively grouping them into similar buckets and helps business to deal with each group in a strategic manner. Profiling directly on RFM scores is also a commonly used method, but clustering aids with effective grouping by capturing the hidden patterns and aspects of the population.

The clustering can also be implemented by taking other available features into consideration while also including all the positive prospects of RFM metrics. There is also an exhaustive study and research being done in the field of clustering algorithms that have established effective algorithms and methods to predict the movement of clusters and future behavior of clusters. This implies that clustering a base would further open various dimensions and prospects of future predictions and analysis.

In this study, the aim to evaluate the impact in accuracy of churn prediction models when the customers are clustered based on RFM and the cluster results are being passed on as a feature to the model. This is compared against the model prediction on absolute RFM values of the customer. This would help to establish the extent of approximation and whether or not it is reliable to utilize clusters as a model input for various applications.

V. CHURN MODEL IMPLEMENTATION

The study has been conducted on a transactional dataset from a retail departmental store that has been collected over a span of 12 months. The dataset consists of customer ID and invoice details with date and amount being primarily utilized to derive the features.

The dataset, after being passed through the default preprocessing phase of missing variable and outlier treatment, is transformed to add the required features and the RFM metrics. The Invoice Date, Unit Price and Quantity fields are the key fields which are in turn used to arrive at the Recency, Frequency and Monetary calculated fields.

The dataset is then processed to add the average number of visits per month and the gap in number of months between the last and the previous to last visit of the customer. Although this is similar in nature to the RFM metrics, this is used as a simple feature to calculate the churn mathematically for the purpose of this study. RFM being a long term calculated value based on the entire 12 month span of our dataset, the visits per month and gap in visit would only require the last few visits of the customer to establish.

Customer churn, in this case, is defined as a phenomenon when a customer has not been visiting the business for more than 3 months and has shown an incremental gap in months from their usual pattern. For example, if a customer has not been visiting the business for 4 months then they are considered churn if they usually show a purchase pattern of once every 2 months and are not considered churn if they had visited once every 4 months in the past. This is mathematically established and calculated across the dataset of 4,259 records and is used for churn model implementation.

The initial dataset is meanwhile normalized to add recency, frequency and monetary log values to calculate the clusters using k-means algorithm as shown in figure 2. The elbow method is used to establish the k value to be 6.

	count	mean	std	min	25%	50%	75%	max
recency_log	4259.0	-2.86e-17	1.0	-2.63	-0.69	0.11	0.86	1.50
frequency_log	4259.0	-1.72e-16	1.0	-1.02	-1.02	-0.24	0.54	4.93
amount_log	4259.0	6.36e-16	1.0	-4.39	-0.67	-0.04	0.67	4.53

Figure 2: Log normalized RFM metrics

The final dataset is feature selected to have the average visits per month, gap in visits, RFM metrics, cluster value and the absolute churn as a factor across each customers. The dataset is also validated to ensure there is no outlier or skew in distribution prior to proceeding with the machine learning model implementation. It is noted that the clusters are reliably evenly distributed as shown in figure 3. The churn distribution is also balanced with substantial positive and negative population i.e. 1,250 churners out of 4,259 customers and is appropriate for the model to perform.

The dataset is then split into train and test data in the ratio of 70:30. This will be utilized throughout the study.

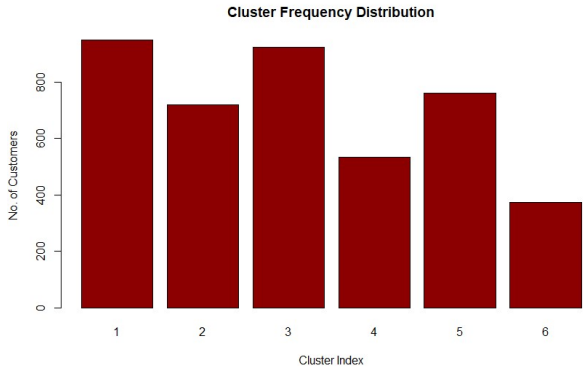


Figure 3: Customer distribution across clusters

The train and test datasets are both treated to keep and drop the required sets of features. In case of RFM based model, Visits, Gap in Months and the RFM metrics along with churn which is the dependent variable were retained The Corr plot of the resultant data is shown in figure 4.

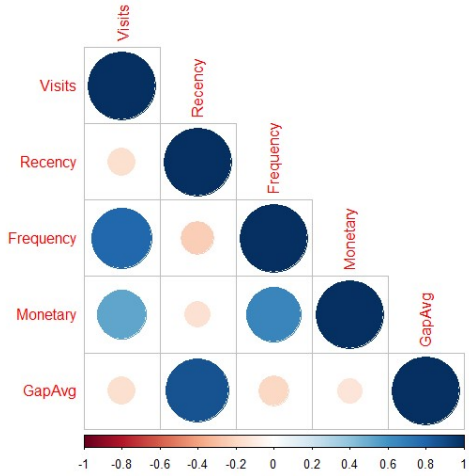


Figure 4: Corr Plot of RFM based input

In case of the cluster based model, only the assigned cluster of the customer is retained along with the visit and gap data fields. It is to be noted that the six cluster categories are binary encoded to arrive at 6 separate features that denote the respective groups. The resultant Corr plot of the data is shown in figure 5. It is notable that the correlation reduces with cluster based data input. The next steps would be to utilize the two sets of datasets and predict churn using various machine learning algorithms for evaluation of cluster based model accuracy.

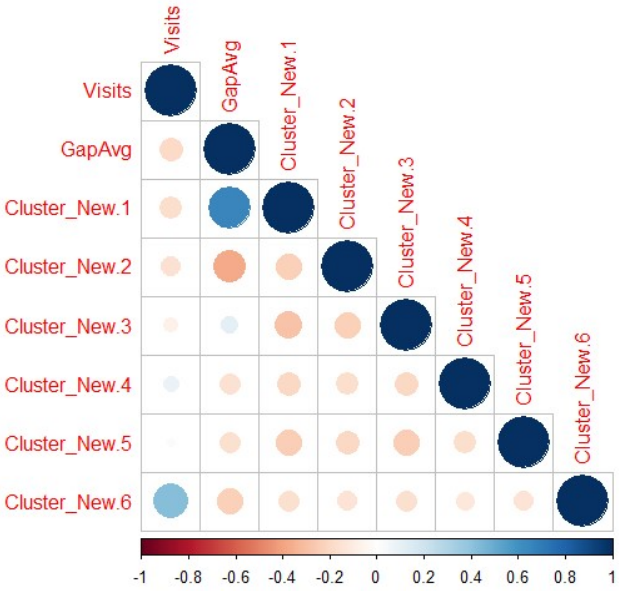


Figure 5: Corr Plot of Cluster based input

The train and test datasets from both sets of data are normalized and passed through the models and predictors of logistic regression, CART, Random Forest, Support Vector Machine and Gradient Boosting Algorithms. The confusion matrices, accuracy, specificity, sensitivity and the balanced accuracy are then compared to study the results.

VI. Evaluation Of Cluster Based Churn Model

The model prediction results are tabulated as shown in tables 1 and 2. A model's performance may not be simply denoted by the confusion matrix or the accuracy, rather the rate of accurate positive predictions also known as sensitivity and the proportion of the negatives rightly predicted also known as specificity also play a crucial factor in a model's performance. In cases where a dataset is imbalanced, with classification tending towards one category, the balanced accuracy better indicated the model's validity.

Table 1: RFM based model performance comparison

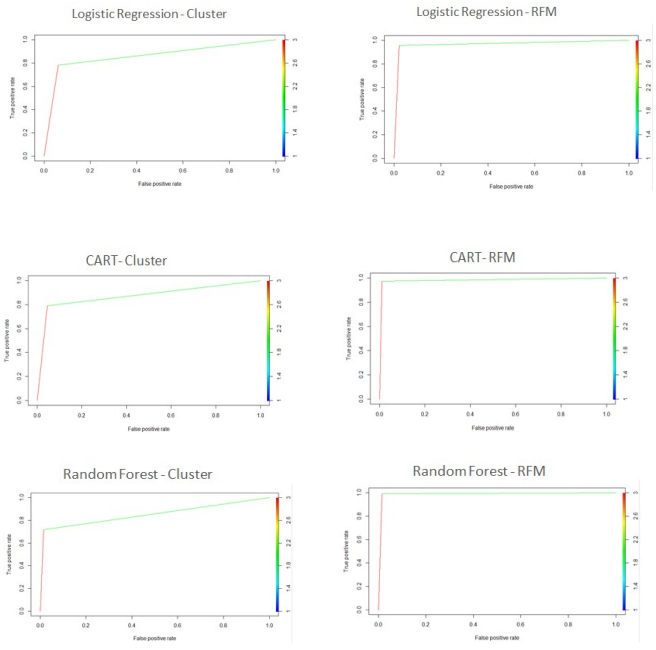
	Logistic Regression	CART	Random Forest	Support Vector Machine	Gradient Boosting Machine
Accuracy	0.9718	0.9851	0.9875	0.9757	0.9703
Sensitivity	0.9472	0.9759	0.9662	0.9526	0.9332
Specificity	0.9822	0.9889	0.9966	0.9855	0.9865
Balanced Accuracy	0.9647	0.9824	0.9814	0.9691	0.9598

Table 2: Cluster based model performance comparison

	Logistic Regression	CART	Random Forest	Support Vector Machine	Gradient Boosting Machine
Accuracy	0.8929	0.9061	0.8991	0.9077	0.9030
Sensitivity	0.8420	0.8783	0.8750	0.9509	0.8638
Specificity	0.9118	0.9160	0.9074	0.8953	0.9175
Balanced Accuracy	0.8769	0.8972	0.8912	0.9231	0.8906

The RFM based churn prediction models all perform with high accuracy metrics, which is as expected as the Churn calculation was also based on the customer's purchase history and the high correlation between the metrics was rightly observed. However, upon evaluating the cluster based model performance it is observed that the accuracy is slightly lesser, which is understandable owing to the fact that clustering is a form of approximation done by grouping a set of similar customers together. But despite the approximation, it proves to be a valid feature that can be reliably used for churn prediction. The balanced accuracy is still maintained at an excellent level and it is evident that Support Vector Machine algorithm is able to maintain almost the same sensitivity and considerably good balanced accuracy in case of cluster based model.

The ROC curve of all five algorithms for both RFM and cluster based datasets is depicted in figure 6. It can be observed that although there is a loss in accuracy and performance, the cluster based models are still doing acceptably well and are therefore considered a reliable method to be utilized for churn prediction in various applications as necessary.



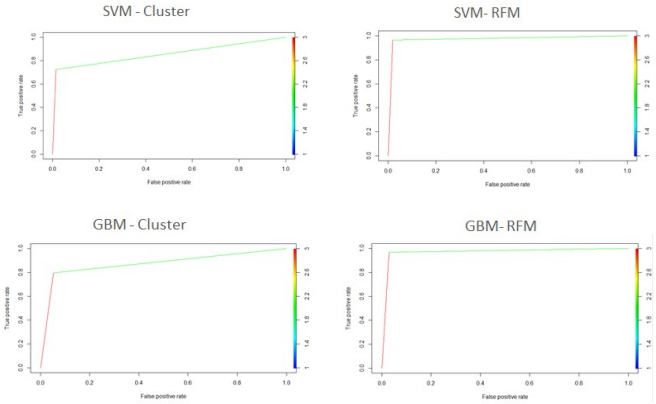


Figure 6: ROC curve comparison across five algorithms for RFM vs. Cluster based data

VII. Conclusion

The study successfully establishes that a dataset clustered can act as a reliable input for further predictive modeling analysis despite going through some level of approximation. This would essentially open up various applications where machine learning algorithms can be leveraged to predict the customer behavior well in advance effectively for the industries that are as dynamic as the retail domain. The scope of this study can be extended to design a prediction on prediction model, where the future churn behavior of a customer is established by taking their future clusters into consideration. The RFM metrics require absolute historic data of the customer for a given period, but cluster prediction can be implemented at a much earlier state using models such as Markov Chain etc.

References

[1] C. Jie, Y. Xiaobing, and Z. Zhifei, "Integrating OWA and data mining for analyzing customers churn in e-commerce," The Editorial Office of JSSC and Springer- Verlag Berlin Heidelberg, vol. 28, pp. 381-391 2015.

[2] Chen, Y.L., Kuo, M.H., Wu, S.Y., Tang, K.: Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data. Electron. Commer. Res. Appl. 8(5), pp. 241–251 (2009)

[3] Weiyun Ying, Xiu Li, Yaya Xie and E. Johnson, "Preventing customer churn by using random forests modeling," 2008 IEEE International Conference on Information Reuse and Integration, 2008, pp. 429-434.

[4] Mirkovic, Milan, Teodora Lolic, Darko Stefanovic, Andras Anderla, and Danijela Gracanin. "Customer Churn Prediction in B2B Non-Contractual Business Settings Using Invoice Data." Applied Sciences 12, no. 10 (2022): 5001.

[5] J. Hadden, A. Tiwari, R. Roy and D. Ruta, "Computer assisted customer churn management: State-of-the-art and future trends," Computers & Operations Research, vol. 34, no. 10, pp. 2902-2917, October 2007

[6] A Lemmens, & S. Gupta, "Managing Churn to Maximize Profit", Harvard Business Schol Working Paper, (14- 020), (2013)

[7] Buckinx, W., Van den Poel, D.: Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. Eur. J. Oper. Res. 164(1), pp. 252–268 (2005)

[8] Ahn, Jaehyun, Junsik Hwang, Doyoung Kim, Hyukgeun Choi and Shin-sung Kang. "A Survey on Churn Analysis in Various Business Domains." IEEE Access 8 (2020): 220816-220839.

[9] Dingli, Alexiei, Vincent Marmara, and Nicole Sant Fournier. "Comparison of deep learning algorithms to predict customer churn

within a local retail industry." International journal of machine learning and computing 7.5 (2017): pp. 128-132.

[10] García, David & Nebot, Àngela & Vellido, Alfredo. (2017). Intelligent data analysis approaches to churn as a business problem: a survey. Knowledge and Information Systems. 51. pp. 1-56.

[11] Patil, Annapurna P., M. P. Deepshika, Shantam Mittal, Savita Shetty, Samarth S. Hiremath, and Yogesh E. Patil. "Customer churn prediction for retail business." In 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), pp. 845-851. IEEE, 2017.

[12] Miguéis, Vera L., Dirk Van den Poel, Ana S. Camanho, and João Falcão e Cunha. "Modeling partial customer churn: On the value of first product-category purchase sequences." Expert systems with applications 39, no. 12 (2012): pp. 11250-11256.

[13] Niccolò Gordini, Valerio Veglio, “Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry.” Industrial Marketing Management, Volume 62,2017 pp. 100-107