Mining Customer Churns for Banking Industry using K-means and Multi-layer Perceptron

Saw Thazin Khine
Faculty of Computer Science
University of Information Technology
Yangon, Myanmar
sawthazinkhine@uit.edu.mm

Win Win Myo
Faculty of Computing
University of Information Technology
Yangon, Myanmar
winwinmyo@uit.edu.mm

Abstract*— Customer churns prediction is a major concern for banking industries all over the world, and market development is happening a more considerable task, and a more important challenge is taking place in business growth. The revenue of the banking sector totally depends on its important customers. By reducing the churn customers, commercial banks have many benefits on both gaining more profits and improving core competitiveness among the competitors. In this study, a mining customer churns using K-means and Multi-Layer Perceptron (MLP) is proposed for forecasting of customer churns in the banking industry. Churn-Modelling dataset from Kaggle site is used in this study. To develop the customer churns prediction model, the dataset is first cleaned, preprocessed and then K-means is used to find similar customer groups. To improve cluster quality, Silhouette method is applied before K-means. After that, Multi-Layer Perceptron classifier is applied to predict whether each customer from each group can leave or not from a bank. The proposed model gets good accuracy with low training time in all customers groups. The results show that the proposed mining customer churns prediction is able to find a certain number of customer churns identifying customer behavior and churn happens. To compare with the proposed model, the methods of K-means + SVM and the previous study on the same dataset are used.

Keywords—Customer Churn, Data Mining, K-means, Silhouette Method, MLP, Banking, Multi-layer Perceptron

I. INTRODUCTION

As people like to choose their money to invest at bank, and many companies prefer to pay salary via bank, banking sectors are most popular in every social standards. There are many competitors in the today banking environment, and these competitors are always standing by a specific good and service to offer better quality and lower prices. Therefore, customers can choose loyalties as they like among bank competitors. Customer churns occur if a user immediately drops out of a bank, or cuts of the use of banking services. There are two kinds of churning, such as voluntary or involuntary churning. Users removed from a bank are becoming involuntary churners. Voluntary churns occur when customers have unexpectedly stopped using the products. This study strongly focuses on voluntary churn happening.

In a banking zone, customers can become churners by the interest rates, disliking services, not having customer-friendly bank staffs, not getting the latest technology, and so on. The banking industry obtains many advantages on not only getting many profits but also improving core competence among challengers by reducing customer churn. On the other hand, much effort and money is required to attract new customers. Thus, the customer churns prediction is an important research area for predicting churn customer among the banking industries.

In every economic environment, there exist a large number of customers, and each customer has a huge amount of data. Thus, forecasting customer churns from a wide range of data is not practicable. Thus, effective data mining techniques need to be applied to the customer churns from a huge database. In this modem age, data mining is so well-known in many research works according to its versatile accessibility for too much data and the skillful transformation of those data into useful information [1]. Correctly predicting churn happening with customer's transaction, and then making marketing development to those customers will increase the profit of the bank with comfortability.

Nonetheless, many professionals have addressed many specific data mining models of churn prediction, but reliability was still poor, and some methods took too much computation time. In this study, prediction model is developed using combined data mining techniques (Silhouette + K-means + MLP) to forecast possible churns with satisfactory performance. Silhouette method is used to obtain reliable number of cluster. For data clustering, K-means is very popular because it can easily group similar characteristics of customers within the same clusters. We used MLP classifier which gives the best accuracy of performance because of its forward and back-propagation algorithms, but training time was increased. So, K-means is used before MLP classifier to get better performance with low training time. By using well-known K-means, bank can discover behavior characteristics of customer groups and can make marketing actions on each particular those customer groups.

Customer churn mining is developed using Churn_Modelling dataset from (www.kaggle.com). It consists of 14 features with 10,000 records. We mainly focus on three features in this dataset (HasCrCard, IsActiveMember, NumberOfProducts) for grouping customer. *HasCrCard* is whether a customer uses bank's credit card or not. *IsActiveMember* is an active person who used bank's products in a specific period. And *NumOfProducts* means the count of using a number of products that the customers used. After similar customer groups are obtained, Multi-Layer Perceptron (MLP) classifier is applied to forecast churner on each particular customer group. Another classifier, developed K-means + SVM, Decision Tree in both Apache Spark ML and Apache MLlib packages from previous work were illustrated to make comparison with the proposed model. The comparison results indicate that the model being proposed is better than other works.

The remaining sections are organized with five parts. Part II discusses literature review. Part III presents the system design in more detail. Part IV discusses experimental setup. Part V discusses the results of proposed model and compares with other works, and then, Part VI concludes and points out.

II. RELATED WORKS

Many research professionals focused on customer churn prediction in different domains such as bank and insurance environment, e-commerce system, IoT companies, Wholesale and retail services, telecom industry, and so on. Churn forecasting were mainly considered on behavioral attributes, socio-demographic features, account level, and balance of customers. And they suggested good data mining algorithms, including Neural Network (NN), Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Logistic Regression (LR), Random Forest (RF), Genetic Algorithm (GA), Markov model, and etc. for predicting customer churn.

In 2015, A. Hudaib et al. [2] proposed K-means + MLP, Hierarchical clustering + MLP, and Self-organizing Map (SOM) + MLP hybrid models on real data of Jordanian Telecommunication Company. Firstly, they clustered on the whole dataset using different kinds of clustering models such as K-means, hierarchical, and SOM. Afterwards, they chose the largest two clusters to train MLP classifier neglecting small clusters. And they compared those proposed models to MLP and C4.5 single models. The results showed that the proposed K-means + MLP model outperformed among those three hybrid models.

In 2017, A. P. Patil et al. [3] used online retail dataset to compare different algorithms for churn analysis. They develop three predictive models using SVM, RF, Extreme Gradient Boosting (EGB) to run on that dataset. Although they were getting high accuracies with the three models, the computation time in the same model was still increased.

In 2017, Fa-Gui Liu et al. [4] developed telecom customer churns prediction using hybrid method of Fuzzy K prototypes (FKP) and SVM for the purpose of reducing time to operate and improving performance to predict. They used fuzzy K-prototypes because dataset consisted of both categorical and numerical attributes. They also made the comparison with other models such as (Fuzzy C-Means) FCM-SVM and K-means-SVM. They found that the proposed FKP-SVM model outperformed among other models.

In 2018, F. Abdi et al. [5] discussed two-stage framework data mining techniques to develop prediction of telecom user churn happening. K-means algorithm with classification methods: NN and Decision Tree (DT) were proposed for different purposes. Firstly, K-means was used to cluster six groups of customers using their socio-demographic features on the whole dataset. After that, level of each customer was identified based on their behavioral attractiveness. And then, they used those clusters to predict customer level attractiveness using DT and NN. As a second step, they used both NN and DT to classify churner on both demographics and behavioral features. Prediction accuracy of NN achieved 68% on socio-demographic features and 76% on behavioral features. DT achieved 69% on socio-demographic features and 74% on behavioral features. Although they proved that their method can help improving managers' ability to manage customer relationships, the accuracy of churn prediction on both DT and NN are still needed to improve.

In 2018, H. Sayed et al. [6] proposed banking customer churns analysis using DT on both Apache Spark ML and MLlib packages. They used the same dataset with this study. They pointed out that model's accuracy in ML package is better than MLlib package. Although ML package took longer time to train, it took fewer time to evaluate. Inversely, MLlib package took fewer time to train but it took longer time to evaluate.

In other previous works, they use (clustering + classification) methods for aiming of outlier filtering (data reduction). Unlike the previous works, Hybrid model of (Silhouette + K-means + MLP) for bank customer churns prediction is developed for the purpose getting high prediction performance with low training time by running individual group of customers without running the whole dataset at a time. And also, bank marketing professionals can make effective marketing actions on each customer groups.

III. PROPOSED SYSTEM DESIGN

The proposed model for customer churn mining is designed in Fig. 1.

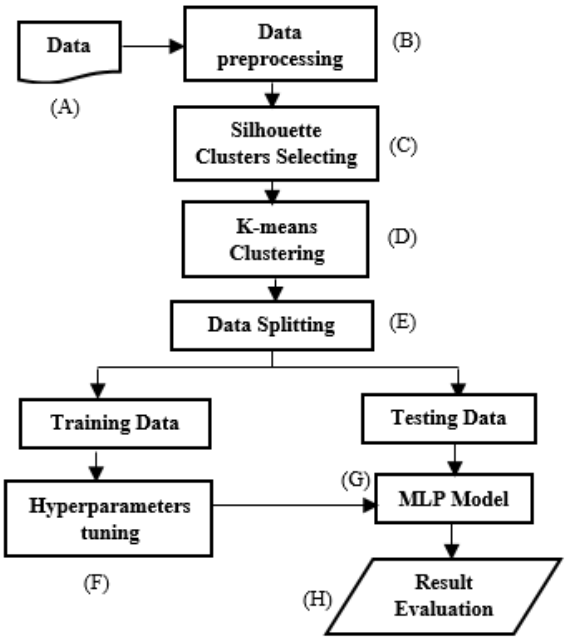

Fig. 1. Proposed System Design of a novel hybrid customer churn prediction

A. Data

The proposed Mining Customer Churn Prediction is developed using Churn-Modelling dataset on bank domain created by S. lyyer [7]. It includes 14 features with 10k customer data records. It consists of two types of attributes: continuous and categorical attributes as shown in Table I.

TABLE I. TOTAL 14 ATTRIBUTES WITH TWO TYPES

Continuous Attributes	Categorical Attributes
RowNumber	Gender
CustomerId	Geography
Surname	IsActiveMember
Age	HasCrCard
CreditScore	Exit
Tenure	
NumOfProducts	
Balance	
EstimatedSalary	

RowNumber is as a customer number. *CustomerId* is as Unique Id of customer. *Surname* is surname of each customer. *Age* is the age of customer. And credit card usage score is described as *CreditScore*. *Tenure* is duration of usage in months. *NumOfProducts* means the count of using a number of products that the customers used. *Balance* is as the amount of money having in an account. *EstimatedSalary* is as each customer's estimated salary. *Gender* is as gender of each customer, and *Geography* is the location of bank exist. And *IsActiveMember* is an active person who has used the products of the bank for a specific period. *HasCrCard* is a customer used a credit card from a bank or not. And *Exit* is defined as the churn customers.

B. Data Preprocessing

There are three data preprocessing processes such as drop irrelevant columns, do one-hot Encoding, and standardize data as mentioned in Fig. 2.

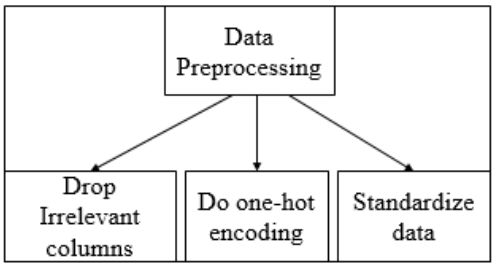

Fig. 2. Data preprocessing

Firstly, irrelevant attributes are dropped: *RowNumber* and *Surname*. Because, *RowNumber* is just a serial number and Surname is the name of each customer. So, that columns are not greater impact on customer churn forecasting. Secondly, for converting categorical attributes (such as Geography and Gender) to numerical attributes, the one-hot encoding method is used. It is a very useful data encoding method because only numerical data can be calculated by the proposed model.

Gender has two values 'Female' and 'Male', and also Geography has three values 'France', 'Spain', and 'Germany'. One-hot encoding separates each attribute to the multiple columns that the value exists. And it encodes each attribute as -1 or 1. For example, let Customer's Gender be Female and who are from France. So, it encoded Female be "1" and male be "–1" and also France be "1" and others be "-1". One-hot encoding result for Geography and Gender is as shown in Table II.

TABLE II. ONE-HOT ENCODING RESULT FOR GEOGRAPHY AND GENDER

Gender_ Male	Gender_ Female	Geography_ France	Geography_ Spain	Geography_ Germany
-1	1	1	-1	-1

Data standardization is needed after one-hot encoding as the third process because the dataset consisted of various ranges of attributes such as Age and NumOfProducts. The range of age is between 18 and 92, and the range of NumOfProducts is between 1 and 4. To scale such kinds of attributes similarly, Min-max scaler is implemented in this work. The efficiency of MLP classifier can be improved by using the min-max scaler.

For example, the standardization of age value (42) is calculated as follow.

Let min value of age among dataset = 18

max value of age among dataset = 92

min-max scaler value for Age (42) = (42-18)/ (92-18)

= 0.3243

After preprocessing, 15 attributes with 10k data records exists in the dataset.

C. Silhouette Clusters Selecting

The Silhouette method is a key to finding the right optimal cluster groups to perform efficient grouping of customer similarities in K-means. It is used within own cluster and its neighboring clusters for the purpose of measuring similarities. Its range is defined as [-1,1]. The score of Silhouette Method is considered as in (1). And also, the main distance measurements a(i) and b(i) are formulated as in (2) and (3). Highest silhouette score is selected as the reliable optimum number of clusters.

Silhouette Score S(i):

$$S(i) = \begin{cases} \dfrac{b(i)-a(i)}{\max\{a(i),\ b(i)\}}, & if |C_i| > 1 \\ 0, & if C_j = 1 \end{cases} \quad (1)$$

$$a(i) = \frac{1}{|C_i|-1}\sum_{j\in|C_i|,\ i\neq j} d(i,j) \quad (2)$$

$$b(i) = min_{i\neq j}\frac{1}{|C_j|}\sum_{j\in|C_j|} d(i,j) \quad (3)$$

Where,

a(i) = the distance measurement of point i between cluster itself

b(i) = the dissimilarity measure of the point i into neighboring cluster

C_i = total number of data points in own cluster itself

C_j = total number of data points in adjacent cluster itself

$d\ (i, j)$ = the average distance between points *i* and j

Silhouette Method's result is as shown in Fig. 3. According to the silhouette curve, K=6 is the best optimal number of cluster because (K=6) has the highest silhouette score among K values 2 to 9.

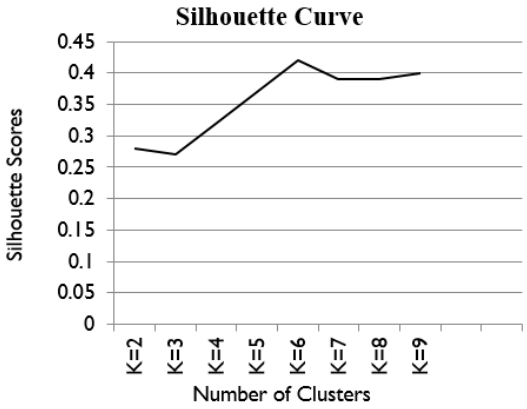

Fig. 3. Silhouette Curve

D. K-means Clustering

Clustering, which is one of the most popular unsupervised methods, can find similarity groups of customers from a large amount of data. Clustering of K-means is used to cluster the data with similar structures.

Macqueen's K-means steps are:

(1) Prepare K center locations ($c_1, \ldots, c_k$).

(2) Each data point (x_i) is assigned to the nearest centroid cluster c_k.

(3) Centroid recalculation of newly formed clusters c_k as the mean of all x_i.

(4) Determine Distance D by Euclidean method

(5) If D has converged, return the value ($c_1, \ldots, c_k$).

By showing the Silhouette curve in Fig.3: section C, the optimal effective cluster group is K=6.

Therefore, 6 customer groups are clustered based on three features (HasCrCard, IsActiveMember, and NumOfProducts) by using the K-means algorithm due to the silhouette result. These three features are mainly considered because HasCrCard and IsActiveMember features are highly related with churn label. Due to the data correlation with class label, the customers who have no credit card might become more churners, and the inactive customers may happen more churner. So, we paired these two important features with other features. And we found that HasCrCard, IsActiveMember, and NumOfProducts features pair can give high accuracy of classification.

Six cluster groups include:

- Cluster 0: Inactive customers using Credit Card and only one product
- Cluster 1: Active customers both using Credit Card and 2, 3, or 4 products
- Cluster 2: Active customers with no Credit Card and using 1, 2, 3, or 4 Products
- Cluster 3: Inactive customers without using Credit Card, and 1, 2, 3, or 4 Products
- Cluster 4: Inactive customers using Credit Card, and 2, 3, or 4 products
- Cluster 5: Active customers using Credit Card and only one product

E. Data Splitting

It is appropriate to transform the samples in all six clusters resulting from K-means into 60% train samples and 40% test samples in each group. The total number of samples, training data, and testing data in each cluster are shown in Table III.

TABLE III. TOTAL NUMBER OF SAMPLES, TRAINING SAMPLES, AND TESTING SAMPLES FOR EACH CLUSTER

Clusters	Total samples	Training samples	Testing samples
Cluster 0	1803	1082	721
Cluster 1	1832	1099	733
Cluster 2	1544	926	618
Cluster 3	1401	841	560
Cluster 4	1645	987	658
Cluster 5	1775	1065	710

F. Hyperparameters Tuning

We need to use the parameters tuning phase for selecting the best parameters which give the best accuracy of the model. GridSearch is used for finding optimal hyper-parameters in MLP. It can exhaustively search the best parameters from many parameters' values of search space of a learning algorithm. It is used because it is more understandable and easier to implement.

MLP hyperparameters are activation (for hidden layer), solver (for weight optimization), alpha (for regularization term), learning_rate (for scheduling weight updates), learning_rate_init (for updating weights), momemtum (for gradient decent update). The search space and best search parameters in MLP are illustrated in below Table IV.

TABLE IV. BEST SEARCH SPACE OF HYPERPARAMETER TUNING IN MLP

Hyperparameters	Search Space	Best Search
Activation	relu, softmax, tanh	relu
Solver	lbfgs, sgd, adam	adam
Alpha	0.001-0.05	0.05
Learning_rate	Constant, Adaptive	Adaptive
Learning_rate_init	0.001, 0.005,0.05	0.005
Momemtum	0.5-0.9	0.8

For K-means + MLP, the same best search space of each parameter above table IV is used in all 6 clusters to improve accuracy.

G. MLP Model

Artificial Neural Network (ANN), one of the most widely used data mining techniques, is used for solving different kinds of classification problems [8]. In this study, Multi-layer Perceptron (MLP) is implemented in each customer group to forecast bank customer churns. It works in both forward and backward algorithms. It can precisely predict churn customers because of its back-propagation algorithm. We can get a strong prediction model by reducing the error as much as back-propagation can. Four layers were composed in MLP model: one input layer, two hidden layers and one output layer as shown in Fig. 4. Input layer's nodes are considered as the number of all features in the whole data except label data. And, output layer's node is label class in the data. Consideration of neural nodes within two hidden layers is formulated with by dividing 2/3 in its total number of all attributes [9]. There exists n=14 input features for $a_1, a_2, \ldots, a_n$ and total number of hidden neurons k=10, for two hidden layers ($h_1, h_2, \ldots, h_k$). The output layer y=1, it is a class label for predicting exit or not as described in Fig. 4.

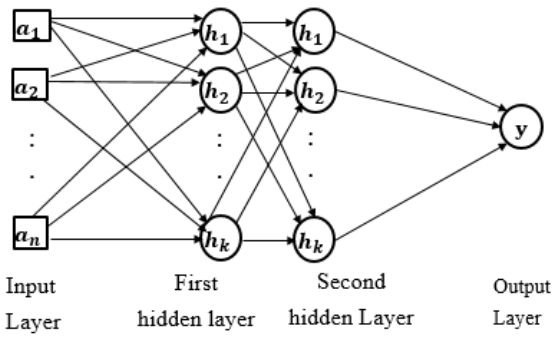

Fig. 4. Proposed MLP Model

H. Result Evaluation

After model was been developed, Confusion Matrix evaluation method is utilized to measure the performance of model on the test data. In Table V, Confusion matrix is described.

TABLE V. CONFUSION MATRIX

	Class 1 Predicted	**Class 2 Predicted**
Class 1 Actual	TP	FN
Class 2 Actual	FP	TN

Accuracy is calculated as the sum of TP and TN is divided by all the sum of TP+TN+FP+FN as shown in (4).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

Where,

TP = true positive rate, which means actual class one is predicted as class one.
TN = true negative rate, which means actual class two is predicted as class two.
FP = false positive rate, which means actual class two is predicted as class one.
FN = false negative rate, which means actual class one is predicted as class two.

IV. EXPERIMENTAL SETUP

The customer churns prediction model aims to improve the relationship between a bank and its potential customers. By knowing each customer's condition, bank marketing professionals make desire plan what customers want.

Python version of 3.7 on Spyder 4.1 IDE is used to construct that prediction model. The scikit-learn software machine learning library, that consisted of different individual packages for data mining methods is used.

Firstly, data file (churn_modelling.csv) is imported using pandas library. After that, data is preprocessed by removing RowNumber and Surname using drop function. And Geography and Gender are converted into numerical value using one-hot encoding. And then, data standardization is performed using min-max scaler. After data preprocessing is performed, K-means is applied for data clustering. Before K-means clustering, Silhouette method is used to find optimal number of clusters by importing silhouette_score function.

According to silhouette result in section C from Chapter III, six cluster groups are divided by calling K-means from sklearn library (from sklearn.cluster import KMeans). To extract each customer group, data is sorted by Cluster (sort_values(by="Cluster"). Each cluster group form section D, Chapter III is obtained. The dataset is divided into 60% training and 40% testing as described in Section E, Chapter III. And 60% of data is assigned to df_train and 40% of data is assigned to df_test. As this model is classification model, the dataset is separated into predictor values and predictand values using (.loc) function. For training model, 14 predictors values (df_train.loc[:, df_train.columns != 'Exited']) and one predictand value (df_train.Exited) is defined. For testing model, 14 predictors values (df_test.loc[:df_test.columns ! = 'Exited']) and one predictand value (df_test.Exited) is defined. And MLP model is developed using training predictor and predictand values by importing (from sklearn.neural_network import MLPClassifier). And then, prediction is made on the test predictors and predictant using developed model. Lastly, model evaluation is measured using confusion matrix and accuracy score.

V. RESULTS

This study is dealing with the hybrid models of K-means and MLP classifier to improve the performance of customer churn prediction and support useful information to make marketing reactions on each particular group of customers.

In Table VI, the accuracy of only SVM and MLP are shown.

TABLE VI. ACCURACY OF SVM AND MLP MODELS

Models	**Accurcay %**	**Training Time (s)**
SVM	82%	10 seconds
MLP	85%	9 seconds

In Table VII, the accuracy of classifying churn in each cluster using K-means+SVM is described.

TABLE VII. RESULTS ON ACCURACY OF (K-MEANS+SVM) MODEL ON EACH CLUSTER

	K-means + SVM
	Accuracy %
Cluster 0	76%
Cluster 1	92%
Cluster 2	83%
Cluster 3	74%
Cluster 4	90%
Cluster 5	83%

In Table VIII, the accuracy of classifying churn in each cluster using each test set is described. From the results, the accuracy of the proposed model (K-means + MLP) in cluster 1 gives best accuracy among other clusters.

TABLE VIII. RESULTS ON ACCURACY OF PROPOSED MODEL ON EACH CLUSTER

	K-means + MLP
	Accuracy %
Cluster 0	76%
Cluster 1	94%
Cluster 2	85%
Cluster 3	84%
Cluster 4	93%
Cluster 5	83%

In Table IX, DT in ML package gave good accuracy but it took high training time. On the other hand, DT in MLlib package took small amount of time to train, but its accuracy is lower than ML package. As a result, the proposed model gives best accuracy 86% with low training time among K-means + SVM, DT models in both using Apache Spark ML and MLlib packages. And even the hybrid method of clustering and classification models achieve good accuracy with less elapsed time for training by comparing Table VI and Table IX.

TABLE IX. COMPARISON RESULTS WITH K-MEANS + SVM AND PREVIOUS WORK[6]

	Accuracy %	**Training Time (seconds)**
Proposed model	86	6 seconds
K-means + SVM	83	7 seconds
Decision Tree (ML package)	79	25 seconds
Decision Tree (MLlib package)	73	6 seconds

VI. CONCLUSION

In this work, mining customer churns of banking industry is proposed using hybrid model (Silhouette + K-means + MLP). In which, the silhouette technique can choose the optimal right clusters and the K-means can cluster the similarities nature of customer groups in the banking industry. By using the proposed combining model, the overall accuracy succeeds 86% with only 6 seconds. Its accuracy is 3% better than other works as shown in Table IX. As a result, the performance of the hybrid method claimed good classification accuracy with low training time for predicting customer churn in the banking system.

It highlighted that the hybrid model could be applied in other business domain to identify the occurrence of customer churn and make marketing retention plans to protect their precious users from churn happening.

REFERENCES

[1] N. Ahmad Naz, U. Shoaib, and M. Shahzad Sarfraz, A Review on Customer Churn Prediction Data Mining Techniques, "Indian Journal of Science and Technology, Vol 11(27), July 2018".

[2] A. Hudaib, R. Dannoun, O. Harfoushi, R. Obiedat, and H. Faris, Hybrid Data Mining Models for predicting Customer Churn, "International Journal of Communications, Network and System Sciences, (IJCNSNSS), (2015)".

[3] A. P. Patil, S. Shetty, D. M P, S. S Hiremath, S. Mittal, and Y. E Patil "Customer Churn Prediction for Retail Business", "(ICECDS-2017)".

[4] Fa-Gui Liu, Z. Zhang, and X. Yang, "Using Combined Model Approach for Churn Prediction in Telecommunication (FKP-SVM)", "3rd Annual International Conference on Electronics, Electrical Engineering and Information Science (EEEIS 2017)".

[5] F. Abdi and S. Abolmakarem, "Customer Behavior Mining Framework (CBMF) using clustering and classification techniques", "Journal of Industrial Engineering International (2018)".

[6] H. Sayed, Manal A. Abdel-Fattah, S. Kholief, Predicting Potential Banking Customer Churn using Apache Spark ML and MLlib Packages: a Comparative Study, " International Journal of Advanced Computer Science and Applications, (IJACSA), Vol. 9, No. 11, (2018)".

[7] S. lyyer, "Churn_Modelling dataset", April 4, 2019, www.kaggle.com

[8] W. W. Myo, P. Aiyarak, and W. Wettayaprasit, "A noble feature selection method for human activity recognition using linearly dependent concept (LDC)," Procedia International Conference on 7th International Conference on Software and Computer Applications, 2018, no. 1, pp. 173–177.

[9] W. W. Myo, W. Wettayaprasit and P. Aiyarak, "Designing Classifier for Human Activity Recognition Using Artificial Neural Network," 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), Singapore, 2019, pp. 81-85.