Application of Classification Methods in Forecasting Broadband Internet Subscribers Leaving the Network

Dong-Ho Le and Van-Dung Hoang*

***Abstract*—The cancellation of subscribers is always a matter of special concern for service providers in general and VNPT An Giang in particular because customers are the ones who bring in revenue and bring value to service providers. To achieve growth and maintain profitability, service providers must find ways to develop new subscribers while also maintaining a stable number of existing subscribers. Therefore, it is very important to research solutions to identify and forecast subscribers who are likely to withdraw from the data network in order to have a customer care strategy to reduce leaving the network. In this paper, we present an approach to exploiting broadband internet subscriber data from the available data warehouse at VNPT An Giang, building a forecasting model for broadband internet subscribers leaving the network before 1 month and 3 months. First, collect data about the 12-month usage history of broadband internet subscribers including 102,920 active subscribers and 24,376 disconnected subscribers with 12 related attributes per subscriber; then the data is preprocessed to remove null data, negative numeric data, and duplicated data; In order to reduce the number of input attributes, select the attributes that are considered to be the most useful for the model, we use the SelectKBest method of the Sklearn library to evaluate and select 8 attributes with high scores. Based on historical data of 6 months/12 months, divide the data into 12 different data sets (in which 6 data sets are for building and evaluating the model to predict that subscribers leave the network before 1 month; 6 datasets for building and evaluating predictive models of subscribers leaving the network before 3 months) (see Table 6 and Table 7). To select the set of attributes and the most suitable model for the forecasting problem of broadband internet subscribers leaving the network, we propose to use 4 machine learning methods including Decision Tree (DT), Support Vector Machine (SVM), Multilayer Perceptron (MLP), Long short-term memory (LSTM). To evaluate the performance of machine learning methods on predictability, we use the Area Under the Curve measure (AUC).**

***Index Terms*—Forecasting broadband internet, decision tree, support vector machine, multilayer perceptron, long short-term memory.**

I. Introduction

With the development of science and technology, especially in the context of the 4th industrial revolution and the development of digital technology, which is progressing very quickly, with breakthroughs, far-reaching, and multidimensional impacts on a global scale. With the mission of becoming a provider of digital infrastructure including telecommunications networks and cloud computing platforms, the telecommunications industry continues to have impressive growth. According to [1], In 2019, total telecommunications revenue reached more than $5.6 billion, and international connection bandwidth capacity reached more than 10 Tbps, an increase of more than 8 times compared to 2015. A fiber optic cable network has been deployed to commune with nearly 01 million km of fiber optic cable, an increase of 1.9 times compared to 2017. The total number of Internet subscribers reached more than 75 million (including nearly 61 million mobile broadband and nearly 15 million fixed broadband), an increase of 1.5 times compared to 2015. The number of ".vn" domain names reached over 503,000 domain names, leading the ASEAN region in terms of registrations using national domain names. The rate of subscribers to IPv6 applications reaches nearly 40%, ranking 2nd in ASEAN and 8th in the world.

By the end of 2019, the total number of broadband internet subscribers was 14,802,372 and the rate of subscribers per 100 people was 15.34% [1]. Of which VNPT accounted for 39.33%, Viettel accounted for 38.61%, FPT accounted for 15.56%, SCTV accounted for 5.54%, CMC accounted for 0.48% and other enterprises (SPT, QTSC, VTC, HTC, ...) accounted for 0.48% (see Fig. 1). Particularly in An Giang province, there are 4 main broadband internet service providers: VNPT, Viettel, FPT, and SCTV. According to VNPT An Giang's statistics, in 2019 the total number of broadband internet subscribers in An Giang province was nearly 34,000, in which VNPT accounted for 41%, Viettel accounted for 39%, FPT accounted for 14% and SCTV accounted for 6% (see Fig. 2).

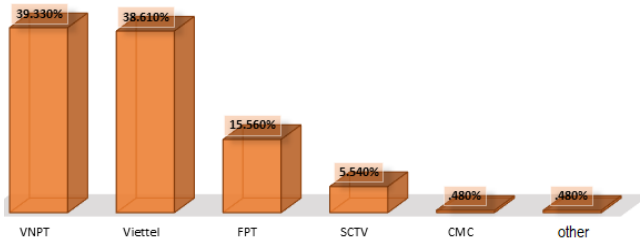


Fig. 1. Broadband internet subscription market sharing [1].

The broadband internet market is gradually approaching saturation, plus the fierce competition of service providers in terms of price, utility, service quality, etc. has led to a large number of customers switching from one service provider to another and vice versa (this problem is called network subscription cancellation). Statistics at VNPT An Giang in 3 years 2018, 2019, and 2020 the rate of broadband internet subscribers leaving the network on

Manuscript received July 25, 2022; revised August 22, 2022; accepted September 21, 2022.

Dong-Ho Le is with VNPT An Giang, Vietnam.

Van-Dung Hoang is with HCMC University of Technology and Education, Vietnam.

*Correspondence: dunghv@hcmute.edu.vn

doi: 10.18178/ijml.2023.13.1.1125

average per year accounted for 38% compared to the number of new developments.

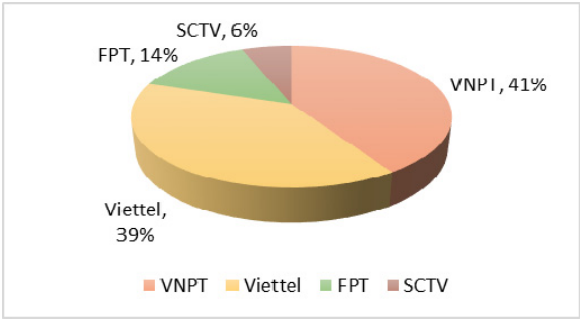


Fig. 2. Broadband internet subscription market sharing in province[1].

According to the classification at VNPT An Giang, there are 2 types of subscribers cancelling the network: Leaving the network passively and actively, both are defined as follows

Passive cancellation: means the service provider cancels the customer's service, with the most common reason being a debt of non-payment of charges.

Active cancellation: means that customers actively stop using the service and switch to using another provider's services, with possible reasons such as having poor service quality, high rates, poor customer service and support...; In addition, for active leavers, there are subscribers who leave the network and do not switch to any service provider (accounting for a very small number), because there is no need to use or change their residence. Over the time of service provision, the customer data warehouse stored at VNPT An Giang has grown larger and larger, which is a valuable asset of the business. From this data storage, it can be analyzed, evaluated, and used for many different problems for the purpose of developing, taking care of customers, and building the development strategy of the unit. Including the problem of predicting subscribers leaving the network, specifically in this article, we will focus on forecasting subscribers cancelling the broadband Internet, and in order for service providers to have enough time to develop customer care and retention policies, the model needs to predict that subscribers are likely to leave the network 1 to 3 months in advance.

II. Related Works

There are many proposed approaches to solve forecasting problems and are applied in many different fields such as: The telecommunications sector has a forecast of subscribers leaving the network, the financial sector has a forecast of customers using credit cards, and the retail sector has a forecast that customers will leave in the future and do not return to purchase. In the previous researches on predicting subscribers leaving the network in the telecommunications sector, we found many researches related to forecasting mobile subscribers leaving the network in telecommunications companies around the world, but No research has been found that is interested in forecasting broadband internet subscribers leaving the network in telecommunications companies in Vietnam. Most of them use traditional machine learning methods such as SVM [2] [5], Decision Tree [7], Random Forest [4], Neural Networks [5], [6], Logistic regression [7], [8], XGBOOST [12], ProfTree [13]; In addition, there are a few research using deep learning methods such as Convolutional Neural Network [10] and ANN [18]. These studies mostly used data sets of limited quantity, including available attributes provided by telecommunications companies, or published on the internet. Another problem of previous studies is not to mention predicting customers leaving the network before a period of time so that service providers have time to develop customer care and retention policies. Some of the relevant methods are as follows:

In 2013, Brandusoiu and Toderean. [2] presents a method to build predictive models of subscribers leaving the network in mobile telecommunications companies. The author uses SVM to train the model with 4 kernels (RBF, LIN, POL, SIG) combining parameters (C, Gamma, Bias, Degree) to compare the results. The model gives an overall accuracy of 88.56%.

In 2014, Farquad and Ravi *et al.* [3] built a system to predict the volatility of customers using bank credit cards. The system is like an early warning expert for bank management. Use a hybrid approach to extract rules from an SVM for Customer Relationship Management purposes. The proposed hybrid approach consists of three stages. (1) In the early stages; SVM-RFE (SVM- Recursive Feature Elimination) is used to reduce the feature set. (2) The dataset with reduced features is then used in the second stage to obtain the SVM model and extracted support vectors. (3) The rules are then created using the Naive Bayes Tree (NBTree) in the final stage. The data set analyzed in this study is on Bank Credit Card Customer Volatility Prediction (Business Intelligence Cup 2004) and it is very unbalanced with 93.24% loyal customers and 6.76 % of customers who stopped using the service. As observed from the experimental results, the proposed hybridization outperformed all other techniques tested and achieved an accuracy of 91.85%.

In 2015, Huang and Zhu *et al*. [4] developed a model to predict prepaid mobile subscribers switching to other service providers. Data was provided by carriers in China and collected for 9 consecutive months from 2013 to 2014. The author used a number of unsupervised and supervised machine learning methods to extract the most important features, then used a random forest classifier (RF) to predict. The overall prediction efficiency of the model is 93% with the AUC measure.

In 2016, Brânduoiu and Toderean *et al*. [5] presented an advanced data mining method for predicting customer churn in the prepaid mobile telecommunications industry. The author proposes to apply three machine learning algorithms: Neural Networks (NN), SVM, and Bayesian networks (BN). The dataset used in this study includes call detail logs and was obtained from the University of California, Department of Information and Computer Science, Irvine, California. It contains information about the usage of the mobile telecommunication system and has a total of 3,333 customers with 15 continuous attributes and

[1] Data source from VNPT An Giang

5 discrete attributes, and a classification attribute with two Yes/No classes. The models have an overall accuracy of 99.10% for BN, 99.55% for NN, and 99.70% for SVM.

In 2016, Brandusoiu and Toderean. [6] Presenting a method to use Neural Networks architecture to predict prepaid mobile subscribers in the telecommunications industry who are likely to leave the network. Using data set "Churn data set" available on the internet (UCI repository of machine learning Databases), including call details of 3,333 subscribers. The overall performance of this model for predicting leaving and not leaving the network is 99.55%.

In 2016, Dalvi and Khangdge *et al.* [7] built predictive models of mobile subscribers leaving the network for telecommunications companies, using data mining and machine learning techniques, specifically logistic regression and decision trees, based on available data sets. The author has provided an analytical tool to predict customer churn based on a comparison between decision trees and logistic regression. Choosing the right combination of attributes and appropriate threshold values can produce more accurate customer churn prediction results. The proposed model shows that data mining techniques can be a promising solution for customer churn management.

In 2017, Can and Albey [8] studied the prediction of prepaid mobile subscribers leaving the network using the Pareto/NBD model and with two benchmarks: a logistic regression model based on RFM data and a logistic regression model based on additional features. Data provided by one of the mobile operators in Turkey, the RFM data set consists of prepaid mobile subscribers who made their first activation with a prepaid charging method and did not change their charging method during the selected time period. The selected period is 2 years from February 1, 2015, to January 31, 2017. Results obtained from the Logistic Regression model run only with RFM data give the highest accuracy of 95% with the Accuracy measure.

In 2017, Umayaparvathi, and Iyakutti [9] developed three deep neural network architectures and built a corresponding customer departure prediction model using two telecommunications datasets Cell2Cell and CrowdAnalytix. Test results show that deep learning models perform as well as traditional classifiers like SVM and random forest.

In 2018, Mishra and Abinash *et al.* [10] proposed deep learning approaches to predict customer churn in telecommunications companies. The author uses the Convolutional Neural Network (CNN) technique and the data set from the website http://www.ics.uci.edu/~mlearn/MLRepository.html. It has shown that the predictive model reached accuracy of 86.85%, error rate of 13.15%, precision of 91.08, recall of 93.08%, and F-score of 92.06%.

In 2019, Mena and Caigny *et al.* [11] presented a method to predict customer churn in the financial industry with sequential data and deep neural networks. The author uses an LSTM model for sequential data, using hit, frequency, and monetary value data from a financial service provider in Europe. The performance gain of the LSTM model is better than that of the logistic model.

In 2019, Ahmad and Jafar *et al.* [12] presented a method to predict customer churn in the telecommunications sector using machine learning in a big data platform. The author uses machine learning techniques based on big data and builds a new way of engineering and attribute selection. To measure the performance of the model, the standard measure Area Under the Curve (AUC) was used and the obtained AUC value was 93.3%. Another major contribution is the use of customers' social networks in predictive modeling by extracting Social Network Analysis (SNA) attributes. The use of SNA enhanced the model's performance from 84 to 93.3% compared to AUC. The model was prepared and tested through the Spark environment by working on a large data set generated by large raw data transformation of mobile subscribers provided by telecommunications company SyriaTel. The dataset collects all customer information for more than 9 months and is used to train, test, and evaluate the system at SyriaTel. The model tested four algorithms: Decision Tree, Random Forest, Gradient Enhancement Machine Tree "GBM" and Extreme Gradient Enhancement "XGBOOST". However, the best results were obtained by applying the XGBOOST.

In 2020, Höppner and Stripling *et al.* [13] introduced a new classifier that integrates the Maximum Profit measure for Customer churn (EMPC) directly into the model building. This technique, called ProfTree, uses an evolutionary algorithm to learn profit-driven decision trees. In a benchmark study with real data sets from different telecommunications service providers, it was found that ProfTree achieved significant profitability improvements over mainstream tree-based methods.

In 2021, Seymen and Dogan *et al.* [14] propose a deep learning model to predict whether customers in the retail industry will leave in the future or not, the author compares the proposed model with a logistic regression model and artificial neural network model. The dataset used includes 10,000 transactions in 27 months by customers from two major Turkish cities, Istanbul and Ankara. As a result, the deep learning model achieved better classification and prediction success than the compared models.

In 2021, Pondel and Wuczynski *et al.* [15] developed a deep learning model to predict customer churn in the retail industry, the goal was to create a model that calculates the probability that customers will return to the same supplier and how many days they will return. The usage dataset consists of 626,275 rows and 131 columns, each row related to a single purchase and an aggregate history of all previous customer purchases. The author uses two basic artificial neural network topologies, a multilayer perceptron (MLP), with one or two fully connected dense layers used. In addition, a repeating layer as a first hidden layer (RNN), optionally supported by an additional dense layer was used. The model achieved prediction performance with 74% accuracy, 78% precision, and 68% recall.

Domingos and Ojeme *et al.* [16] and Dalli [17] present empirical analysis on the impact of different hyperparameters when using deep neural networks (DNN)

to predict customer churn in the banking sector. In this paper, the data set used is loaded from Kaggle. Experiments have been performed on both DNN and MLP models by changing the activation functions used in the hidden and output layers; changing the batch sizes, and changing the training algorithm. Results DNN gives better performance than MLP when using a rectifier in the hidden layer, and sigmoid in the output layer; best batch size when smaller than test dataset size; RMSProp algorithm has better accuracy than other algorithms.

In 2022, Samah and Suresh *et al.* [18] deployed a deep learning model using Deep-BP-ANN to predict customer churn in the telecommunications industry, using two datasets IBM Telco (7043 customers) and Cell2Cell (51047 customers). The results show that Deep-BP-ANN has better predictive performance than machine learning techniques XG Boost, Logistic_Regression, Naïve_Bayes, and KNN.

In 2022, Seymen and Ölmez *et al.* [19] presents a model that uses Ordinary Artificial Neural Network (ANN) and Convolution Neural Network (CNN), to predict whether customers in the retail industry will leave in the future or not. The data was used from the supermarket chain, including 27-month retail scanner data for 5747 customers. The performance result of CNN is better than ANN, but the difference between them is not too much (CNN: AUC 0.976; ANN: AUC 0.963).

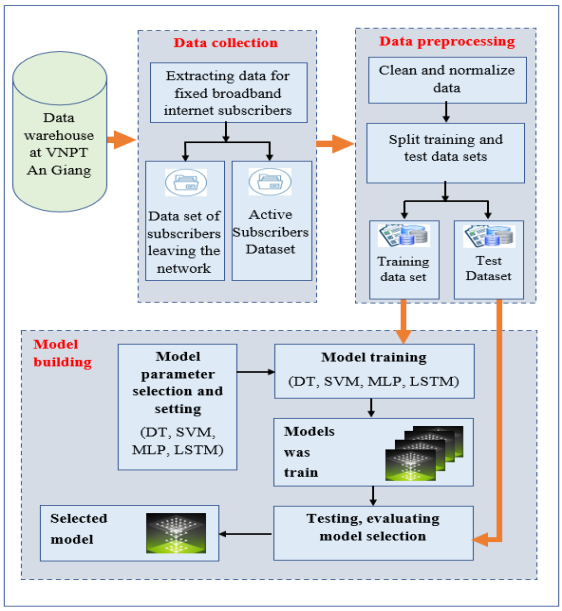


Fig. 3. Proposed flowchart of the predicting system.

III. Proposed Solutions

A. Forecasting Model

We propose a general model to solve the problem of forecasting broadband internet subscribers leaving the network, including 3 main functional blocks: "Data collection" function block, "Data preprocessing" function block, "Building model" function block, presented schematically in Fig. 3.

Function block "Data collection" performs data extraction from the data warehouse at VNPT An Giang, including the attributes that are most likely to affect the cancellation of broadband internet subscribers. This function block has the input data as the data warehouse at VNPT An Giang and the output results are two datasets: the data set of broadband internet subscribers cancelling the network and the data set of active broadband internet subscribers.

Function block "Data preprocessing" performs cleaning, normalizing data, building training, and testing data sets.

Data cleaning: data after being collected will include empty data and negative numeric data. Duplicated and incomplete data will be processed and eliminated.

Standardized data: normalize and scale the data feature range on the range [0,1] to fit the input of the classification methods.

Build training and test datasets: After going through the preprocessing steps, the data is built into data sets for training (accounting for 70%) and testing (accounting for 30%) the predictive model (detailed in Section II.B).

This function block has as input two data sets at the output of the "Data acquisition" function block, and the outputs are normalized data sets used for training and testing the model.

Function block "Building models" perform selection, parameter setting of classification methods, model training and testing, evaluation, and selection of the most suitable model.

Model parameter selection and setting: because the input data has a time factor, related to the usage history of fixed broadband internet subscribers, it will be suitable for the LSTM classification method, in addition to having a comparison, evaluation, and selection model using the most appropriate classification method. In this paper, we choose four classification methods DT, SVM, MLP, and LSTM, with the values of the main hyperparameters of the model selected on an experimental basis.

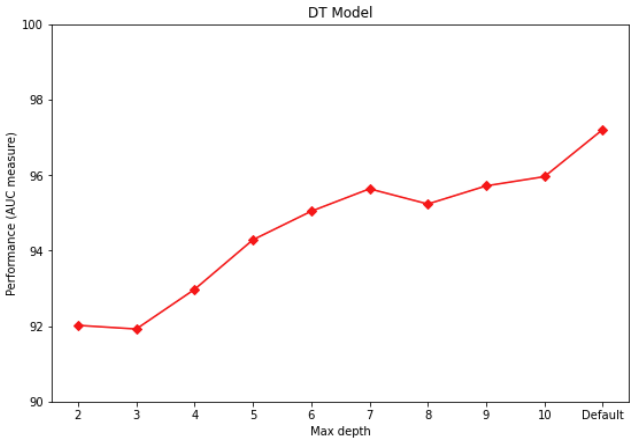


Fig. 4. Selecting the max_depth parameter value in DT.

For the DT classification method, we perform experiments with parameter values of "*max_depth*" from 2 to 10 respectively and the default value of the library (None), the results show that the default value of the library will give the best performance (see Fig. 4), so we use the default value of the library for all parameters.

For the SVM classification method, we perform experiments with "*kernel*" parameter values of rbf (default value of the library) and linear, the results show that using

kernel rbf will give performance higher than kernel linear (see Fig. 5), so we use the default value of the library for all parameters.

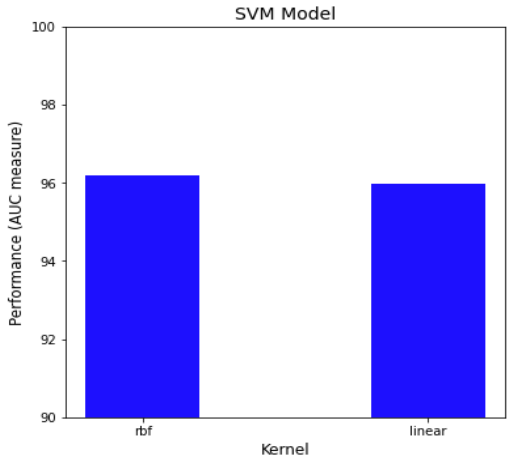


Fig. 5. Selecting the "kernel" parameter value in SVM.

For the MLP classification method, we perform experiments with the "*max_iter*" parameter values from 100 to 1000, the results show that the model achieves the degree of convergence from the max_iter value of 500 (see Fig. 6), therefore we choose to use max_iter value of 500, also for the rest of the parameters we use the default value of the library.

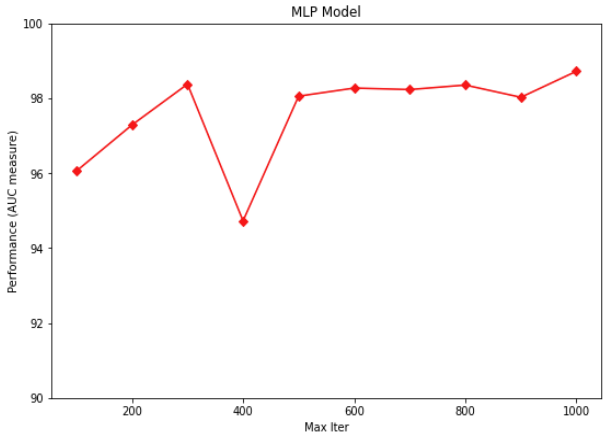


Fig. 6. Selecting the "max_iter" parameter value in MLP.

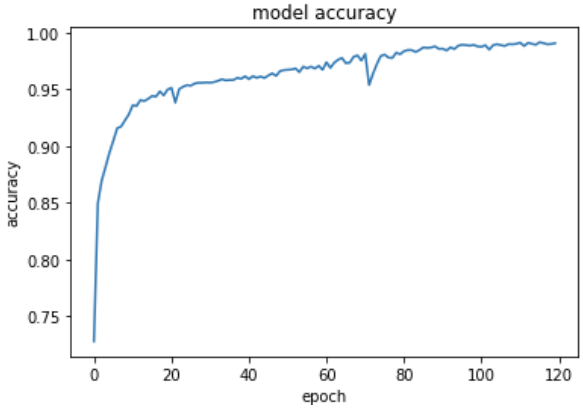


Fig. 7. Selecting the "epoch" parameter value in LSTM.

For the LSTM classification method, we perform experiments with epoch values up to 120, the results show that the model achieves the degree of convergence from the 100th epoch (see Fig. 7), so we use an epoch value of 100.

As a result, we use the model parameter values as shown in Table I.

TABLE I: MODEL PARAMETERS

	Parameter name	Setting value
DT	criterion	gini
	splitter	best
	max_depth	None
	min_samples_split	2
	min_samples_leaf	1
	min_weight_fraction_leaf	0.0
	max_features	None
	random_state	None
	max_leaf_nodes	None
	min_impurity_decrease	0.0
	class_weight	None
	ccp_alpha	0.0
SVM	C	1.0
	kernel	rbf
	degree	3
	gamma	scale
MLP	hidden_layer_sizes	(100,)
	activation	relu
	solver	adam
	batch_size	auto
	learning_rate_init	0.001
	max_iter	500
LSTM	LSTM	128
	activation	sigmoid
	dropout	0.5
	loss	categorical_crossentropy
	optimizer	adam
	metrics	accuracy
	epochs	100
	batch_size	512

Model training: perform model training in turn by 4 classification methods (DT, SVM, MLP, and LSTM) on training data sets, resulting in trained models.

Testing, evaluating model selection: using the test data sets put into the trained model to perform testing, analysis, and evaluation of predictive results of the models, and to propose models. best-fit model and attribute set.

This function block has as input the training and test data sets (from the output of the "Data preprocessing" function block), and the output is the model that best fits the prediction problem. informing broadband internet subscribers leaving the network.

B. Data Collection and Processing

1) Data collection

The data source available at VNPT An Giang's data warehouse is very large and includes information and usage history of many different types of subscribers (broadband Internet, MyTV, mobile phones...), which have been stored and exploited by business management software systems such as Customer management system; Subscriber information management system; Detailed management system for service use; Charge management system; Financial and accounting management system; Complaint information management system, customer satisfaction. We have collected data regarding fixed broadband internet subscribers such as:

Information about customers: location, type of customer, age, and gender;

Information related to the service used by the customer: the time the customer has used the service, the speed of the internet connection, the amount incurred monthly, the

amount owed by the customer, the customer paying in advance;

Information related to customer usage behavior: monthly internet traffic, number of days using the service in a month;

Customer support information: number of service failure reports.

In addition, there is information related to customer identification (full name, address, identity card number …) and information related to device parameters (device name, port on the device, number cable cabinet, cable number…) and id attributes, we will not collect.

In this article, we consulted an expert on customer relationship management at VNPT An Giang to advise and extract the 12 most important attributes related to broadband internet subscribers, and 1 classification attribute (see Table II).

TABLE II: ATTRIBUTES EXTRACTED ACCORDING TO EXPERT ADVICE

ATTRIBUTES	DESCRIPTIONS
TUOI	Customer age
GIOI_TINH	Customer's gender (Female: 0; Male: 1)
THANG_SU_DUNG	Total number of months using the service
DOI_TUONG_KH	Subject customers: (Personal: 0; Business: 1)
KHU_VUC	Customer management area, divided by route and administrative boundaries
TOCDOTHUC	Line speed (Mbps)
SO_THANG_DCT	Remaining months of prepayment
CUOC_PHAT_SINH	Charges incurred
NO_CUOC	Amount owed by customer
LUU_LUONG	Traffic usage in the month
NGAY_LUU_LUONG	Number of days with traffic in the month
BAO_HONG	Number of failure reports in a month
ROI_MANG	Class Attribute (Active: 0; Leaving network: 1)

TABLE III: NUMBER OF SUBSCRIBERS COLLECTION

Data set	Number (Subscribers)	
	Raw data	Processed data
Subscribers using the network	109,784	102,920
Subscribers cancelling the network	39,406	24,376
	149,190	**127,296**

The usage history of the current month is called n; *n*-1, *n*-2, ... is the usage history of the months before the current month 1 month, 2 months…

The collected broadband internet subscriber data includes 12 attributes (according to expert advice in Table II) and 14-month usage history data of each subscriber, that is, each attribute will be collected historical data for 14 months respective to 14 columns of data (from month *n*-1 to month *n*-14), so the data set will have 14×12 + 1 = 169 data columns, the number of subscribers collected, there are 109,784 active subscribers and 39,406 subscribers stopped using the network. More specifically, the set of active subscribers is obtained at the time of December 2021 and the set of subscribers who have left the network is obtained according to the time of leaving the network from December 2021 and earlier. After removing null data, negative numeric data, and duplicated data, there are 102,920 subscribers remain active and 24,376 subscribers left the network (see Table III).

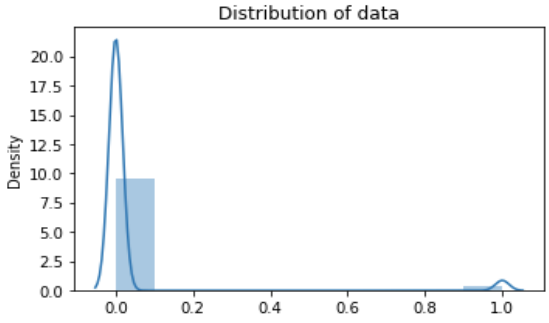
2) Attribute Selection

Attribute selection involves the process of selecting a subset of related attributes from an initial set of attributes, reducing the number of attributes for input to the model in order to reduce the cost of data collection. data and calculation costs. Moreover, it not only gives more accurate results but is also more compact and easy to understand. In the classification problem, attribute selection aims to select a set of highly discriminant attributes, in other words, to choose an attribute capable of distinguishing samples belonging to different classes.

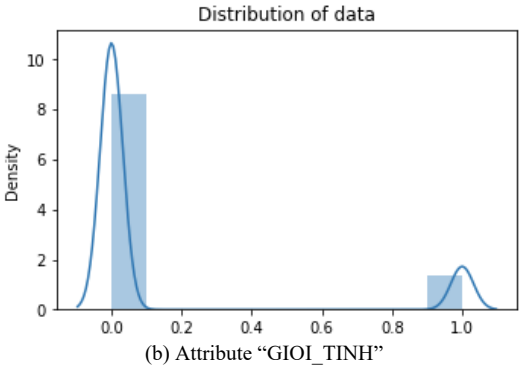
In order to select the attributes that are considered most useful for the model, we use the SelectKBest method of the Sklearn library to evaluate the scores of the attributes, resulting in 1 attribute with a score of 182879; 3 attributes with scores in the range [23731, 28807]; 4 attributes have scores between [1253, 3025] and 4 attributes have scores between [49, 591] (details in Table IV).

TABLE IV: SCORE OF ATTRIBUTES

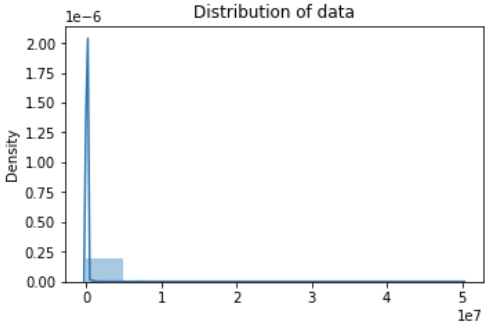
ATTRIBUTES	SCORES
NGAY_LUU_LUONG	182,879
NO_CUOC	28,807
LUU_LUONG	28,188
THANG_SU_DUNG	23,731
KHU_VUC	3,025
TUOI	2,491
SO_THANG_DCT	1,674
TOCDOTHUC	1,253
BAO_HONG	591
DOI_TUONG_KH	297
GIOI_TINH	217
CUOC_PHAT_SINH	49



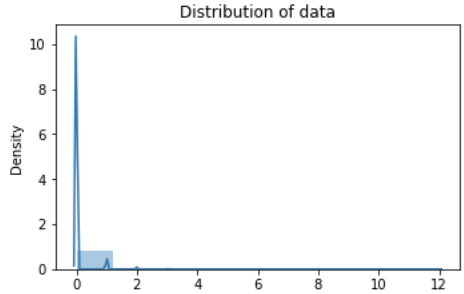
(a) Attribute "DOI_TUONG_KH"



(b) Attribute "GIOI_TINH"



(c) Attribute "CUOC_PHAT_SINH"



(d) Attribute "BAO_HONG"

Fig. 8. Data distribution density of the 4 attributes with the lowest scores.

Based on the scores in Table IV, found that 4 attributes "BAO_HONG", "DOI_TUONG_KH", "GIOI_TINH" and "CUOC_PHAT_SINH" have very low scores (under 600 points) compared to the remaining attributes, we remove them. , do not use. Where the attributes "DOI_TUONG_KH" and "GIOI_TINH" take the value 0 or 1 (see the data distribution graph in Fig. 8a, Fig. 8b), which can cause the model to be skewed. In addition, the attribute "CUOC_PHAT_SINH" is the amount the customer has to pay monthly while using the service (see the data distribution chart in Fig. 8c), which shows that it has little effect on the problem predict customers leaving the network. And the attribute "BAO_HONG" is the number of times a customer has a service problem in a month and needs to contact the supplier for reparation, whose distribution mainly falls in the value of 0 (mostly) and 1, (see data distribution chart in Fig. 8d), it has little effect on customer churn factor.

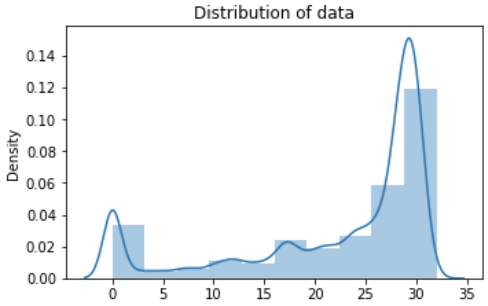


Fig. 9. Data distribution density of the "NGAY_LUU_LUONG" attribute.

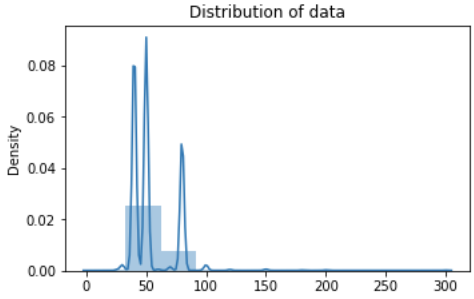


Fig. 10. Data distribution density of the "TOC_DO_THUC" attribute.

For the group of 8 attributes with higher scores, in which the attribute with the highest score "NGAY_LUU_LUONG", is the number of days in the month where the customer's traffic is generated. The low usage of the service by customers (low number of days of traffic generation in a month) will lead to the risk of leaving the network and vice versa (See the data distribution chart in Fig. 9); attribute "TOC_DO_THUC" has the lowest score in this group of 8 attributes, representing the internet access speed that customers register to use, it helps to group customers with similar access speed (See chart data distribution diagram in Fig. 10), this is very useful in classifying and predicting customers leaving the network.

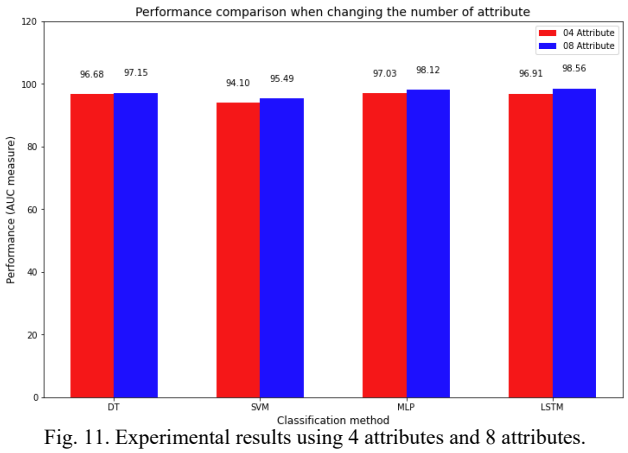


Fig. 11. Experimental results using 4 attributes and 8 attributes.

TABLE V: 8 ATTRIBUTES WITH THE HIGHEST SCORES ARE SELECTED

ATTRIBUTES	SCORES
NGAY_LUU_LUONG	182,879
NO_CUOC	28,807
LUU_LUONG	28,188
THANG_SU_DUNG	23,731
KHU_VUC	3,025
TUOI	2,491
SO_THANG_DCT	1,674
TOCDOTHUC	1,253

TABLE VI: DATA SETS FOR TRAINING AND EVALUATION OF PREDICTIVE MODELS 1 MONTH IN ADVANCE

Data set	Number of columns of data	Number of subscribers leaving the network	Number of active subscribers
1. Twelfth months usage history (months n-1 to n-12)			
The dataset has a scale of 1:4	**97**	**24.376**	**97.504**
Training set (70%)	97	17.063	68.252
Test set (30%)	97	7.313	29.252
The dataset has a scale of 1:2	**97**	**24.376**	**48.752**
Training set (70%)	97	17.063	34.126
Test set (30%)	97	7.313	14.626
The dataset has a scale of 1:1	**97**	**24.376**	**24.376**
Training set (70%)	97	17.063	17.063
Test set (30%)	97	7.313	7.313
2. Six months usage history (months n-1 to n-6)			
The dataset has a scale of 1:4	**49**	**24.376**	**97.504**
Training set (70%)	49	17.063	68.252
Test set (30%)	49	7.313	29.252
The dataset has a scale of 1:2	**49**	**24.376**	**48.752**
Training set (70%)	49	17.063	34.126
Test set (30%)	49	7.313	14.626
The dataset has a scale of 1:1	**49**	**24.376**	**24.376**
Training set (70%)	49	17.063	17.063
Test set (30%)	49	7.313	7.313

To be more objective in choosing attributes to use, we performed two experiments. The first experiment uses the four attributes with the highest scores: "NGAY_LUU_LUONG", "NO_CUOC", "LUU_LUONG" and "THANG_SU_DUNG"; The second experiment uses all 8 attributes. As a result, using 8 attributes will give

higher predictive performance than using 4 attributes (see Fig. 11). From the above experimental results, we choose to use the eight attributes with the highest scores (see Table V), so the data set will have 14 × 8 + 1 = 113 columns of data (details of data columns in Appendix 1).

The data in Table III shows that the number of subscribers leaving the network accounts for much less than the number of active subscribers. In order to evaluate and select the most suitable data set and model, we built 12 datasets based on historical data of 12 months/6 months, and the ratio of data of subscribers leaving the network/active subscribers equal to 1:4, 1:2, and 1:1, 6 datasets of which are used to train and evaluate the model to predict the subscribers leaving the network before 1 month (see Table VI); 6 datasets to train and evaluate the model to predict subscribers leaving the network before 3 months (see Table VII).

TABLE VII: DATA SETS FOR TRAINING AND EVALUATION OF PREDICTIVE MODELS 3 MONTHS IN ADVANCE

Data set	Number of columns of data	Number of subscribers leaving the network	Number of active subscribers
1. Twelfth months usage history (months n-3 to n-14)			
The dataset has a scale of 1:4	**97**	**24.376**	**97.504**
Training set (70%)	97	17.063	68.252
Test set (30%)	97	7.313	29.252
The dataset has a scale of 1:2	**97**	**24.376**	**48.752**
Training set (70%)	97	17.063	34.126
Test set (30%)	97	7.313	14.626
The dataset has a scale of 1:1	**97**	**24.376**	**24.376**
Training set (70%)	97	17.063	17.063
Test set (30%)	97	7.313	7.313
2. Six months usage history (months n-3 to n-8)			
The dataset has a scale of 1:4	**49**	**24.376**	**97.504**
Training set (70%)	49	17.063	68.252
Test set (30%)	49	7.313	29.252
The dataset has a scale of 1:2	**49**	**24.376**	**48.752**
Training set (70%)	49	17.063	34.126
Test set (30%)	49	7.313	14.626
The dataset has a scale of 1:1	**49**	**24.376**	**24.376**
Training set (70%)	49	17.063	17.063
Test set (30%)	49	7.313	7.313

IV. EXPERIMENTAL RESULTS AND EVALUATION

A. Evaluation Criteria

There are many methods to evaluate the effectiveness of a classification model, depending on different data and problems, different methods will be used, some commonly used methods such as Accuracy score, ROC curve, Area Under the Curve (AUC), Precision, Recall, F1 score. In this paper, we use the AUC measure to evaluate the effectiveness of the model and select the model with the highest AUC.

Call:

P (Positive): is a subscriber leaving the network;

N (Negative): is a subscriber who does not leave the network;

TP (True Positive): is the subscriber that actually leaves the network, is predicted to leave the network;

FP (False Positive): is a subscriber that is not actually leaving the network, but is predicted to leave the network;

TN (True Negative): is a subscriber that actually does not leave the network, is predicted to not leave the network;

FN (False Negative): is the subscriber that actually leaves the network, but is predicted to not leave the network;

TPR (True Positive Rate): is the ratio of predicted true leaving subscribers to the total number of leaving subscribers that actually exist, this index will evaluate the accuracy of the model's prediction on class P. The higher its value, the better the predictive model on the P class. TPR value is calculated according to the formula:

$$TPR = \frac{TP}{FN + TP} \quad (1)$$

FPR (False Positive Rate): is the proportion of subscribers who do not leave the network but are predicted to leave the network, out of the total number of subscribers who do not leave the network that actually exists. The lower the FPR of a model, the more accurate the model is because its error on class N is lower. The FPR value is calculated according to formula number 2.

$$FPR = \frac{FP}{TN + FP} \quad (2)$$

Accuracy is computed by the following formula

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

Precision is calculated by following formula

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4)$$

Recall is calculated by following formula

$$\text{Recall} = \frac{TP}{TP + FN} \quad (5)$$

The measure of F1 score is calculated

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

ROC curve is a curve that represents the classification ability of a classification model at thresholds. This curve is based on two metrics True Positive Rate (TPR) and False Positive Rate (FPR). The AUC measure is the area under the ROC curve, see also Fig. 12.

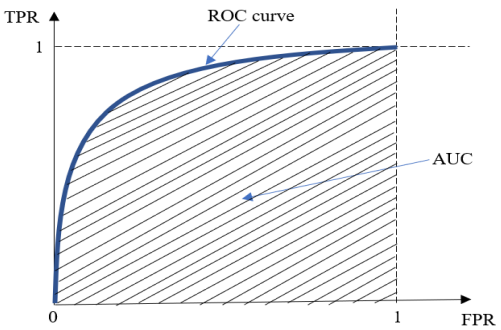


Fig. 12. Measurement of ROC curve and AUC.

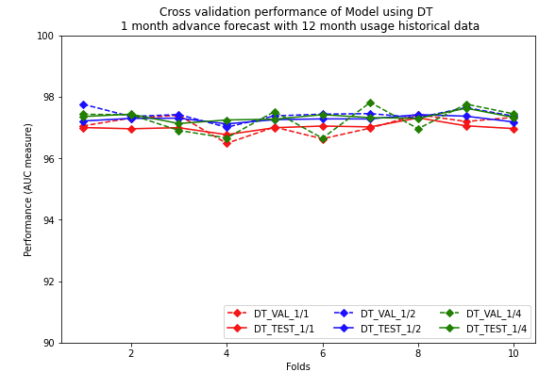
To help evaluate the model more fully and accurately, we use the 10-fold cross-validation technique (Cross validation Kfold = 10) to measure the model performance and then

take the average value.

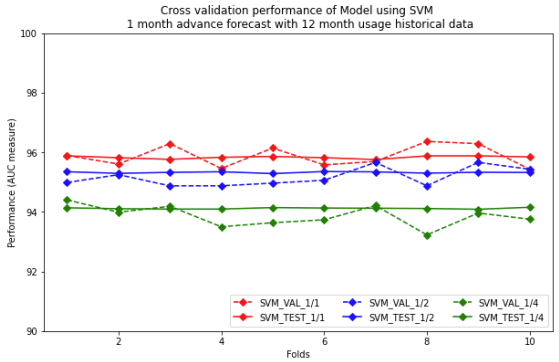
C. Forecast of Subscribers Cancelling the Network 1 Month in Advance

1) Performance of forecast model 1 month in advance with 12 months of usage history data

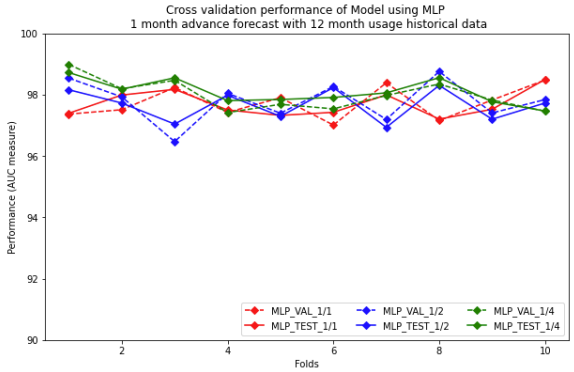
Fig. 13 presents the cross-validation results (KFold is 10) for the model using DT, SVM, MLP, and LSTM classification methods.



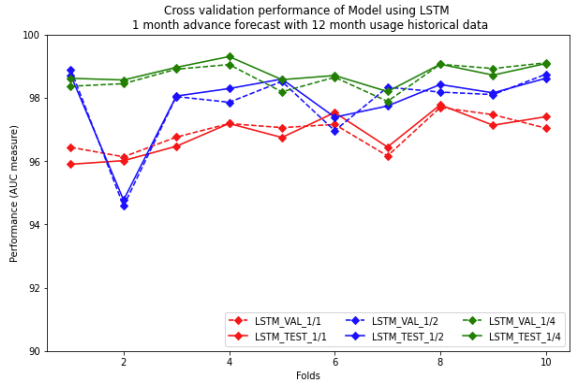
(a) Using DT classification method



(b) Using SVM classification method



(c) Using MLP classification method



(d) Using LSTM classification method

Fig. 13. Performance of 1-month advance forecasting model with 12 months of usage historical data.

Experimental results show that the DT and MLP methods using data sets with the ratio 1:1, 1:2, and 1:4 give equivalent results, which are not affected when the data is out of balance. While the SVM method gives the best results when using a balanced data set of a 1:1 ratio and the lowest when testing on an unbalanced data set of 1:4, the LSTM method gives good results, the highest when using the most data set of 1:4 scale and the lowest when testing on the least data set of 1:1.

In Table VIII, Table IX, Fig. 14 presents the average results performing cross-validation using DT, SVM, MLP, and LSTM methods on data sets with scale 1;1, 1:2, and 1:4 (with historical data for 12 months). Looking at these figures and tables, it can be seen that:

The DT method has the highest average AUC performance of 97.41%, standard deviation of 0.19, the lowest distribution of 97.01%, the highest of 97.76%, and the median of 97.28% when tested on a validation dataset with a ratio of 1:2; and AUC performance of 97.34%, standard deviation of 0.13, the lowest distribution of 97.13%, the highest of 97.63%, and the median of 97.32% when tested on a test dataset of 1:4 ratio.

The SVM method has the highest average AUC performance of 95.87%, standard deviation of 0.35, the lowest distribution of 95.43%, the highest of 96.37%, and the median of 95.79% when tested on a validation dataset of 1:1 ratio; and AUC performance of 95.83%, standard deviation of 0.04, the lowest distribution of 95.76%, the highest of 96.88%, and the median of 95.84% when tested on test dataset of 1:1 ratio.

TABLE VIII: AVERAGE PERFORMANCE PERFORM CROSS-VALIDATION ON VALIDATION DATA (KFOLD = 10) OF 1-MONTH ADVANCE PREDICTION MODEL (DATA SET WITH 12 MONTHS OF HISTORICAL DATA)

Method	Testing on data VALIDATION with 12 months of usage history data					
	The rate of subscribers leaving network / active is 1:4		The rate of subscribers leaving network / active is 1:2		The rate of subscribers leaving network / active is 1:1	
	AUC (%)	σ AUC	AUC (%)	σ AUC	AUC (%)	σ AUC
DT	97.26	0.41	**97.41**	0.19	97.08	0.30
SVM	93.86	0.34	95.16	0.30	**95.87**	0.35
MLP	**97.99**	0.48	97.78	0.65	97.73	0.49
LSTM	**98.66**	0.40	97.82	1.19	96.86	0.61

σ AUC denotes for standard deviation of AUC.

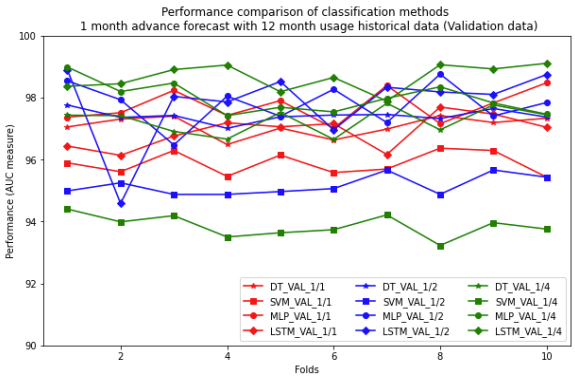
TABLE IX: AVERAGE PERFORMANCE PERFORM TESTING ON TEST DATA FOR PREDICTIVE MODEL 1 MONTH IN ADVANCE (DATA SET WITH HISTORICAL DATA OF 12 MONTHS)

Method	Testing on data TEST with 12 months of usage history data					
	The rate of subscribers leaving network / active is 1:4		The rate of subscribers leaving network /active is 1:2		The rate of subscribers leaving network / active is 1:1	
	AUC (%)	σ AUC	AUC (%)	σ AUC	AUC (%)	σ AUC
DT	**97.34**	0.13	97.27	0.08	97.01	0.13
SVM	94.12	0.02	95.32	0.02	**95.83**	0.04
MLP	**98.09**	0.39	97.66	0.49	97.70	0.41
LSTM	**98.78**	0.31	97.87	1.11	96.86	0.61

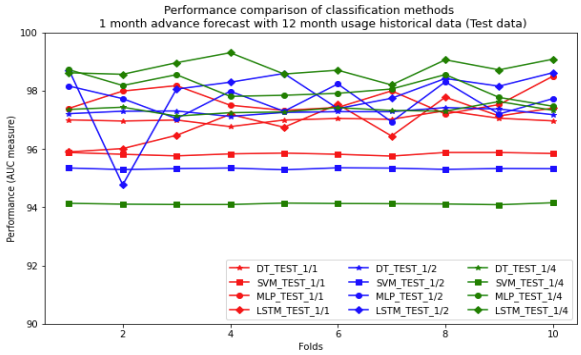
The MLP method has the highest average AUC performance of 97.99%, standard deviation of 0.48, the lowest distribution of 97.42%, the highest of 98.99%, and the median of 97.90% when tested on a validation dataset

with of 1:4 ratio; and AUC performance of 98.09%, standard deviation of 0.39, the lowest distribution of 97.47%, the highest of 98.72%, and median of 97.98% when tested on a test dataset of 1:4 ratio.

The LSTM method has the highest average AUC performance of 98.66%, standard deviation of 0.40, the lowest distribution of 97.89%, the highest of 99.10%, and the median of 98.77% when tested on a validation dataset of 1:4 ratio; and AUC performance of 98.78%, standard deviation of 0.31, the lowest distribution of 98.20%, highest of 99.30%, and median of 98.71% when tested on a test dataset of 1:4 ratio.



(a) Testing on validation dataset



(b) Testing on test dataset

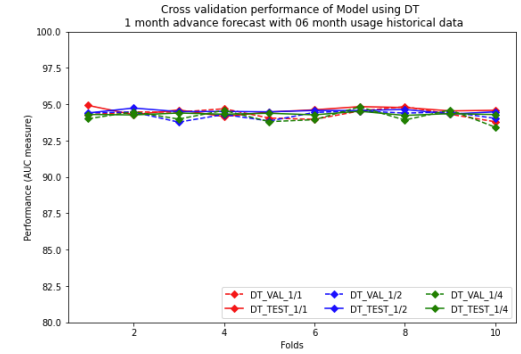
Fig. 14. Comparison of the forecasting performance of the classification methods (12-month usage historical data and 1-month advance forecast).

Thus, with the 1-month prediction model and 12 months of usage history data, the LSTM method gives the highest performance, the AUC reaches 98.66% on the validation dataset with a ratio of 1:4, AUC 98.78% on the test data set with a ratio of 1:4; The SVM method gives the lowest performance, the AUC is 95.87% on the validation data set with a ratio of 1:1 and 95.83% on the test dataset with a ratio of 1:1.

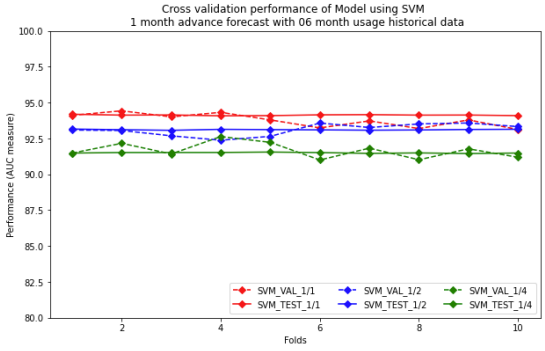
2) Performance of forecast model 1 month in advance with 6 months of usage history data

Fig. 15 presents the cross-validation results (KFold is 10) for the model using the DT, SVM, MLP, and LSTM classification method.

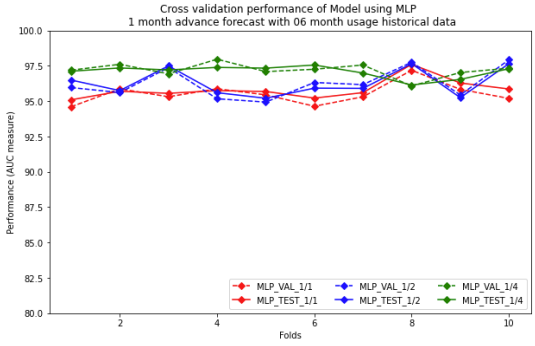
Experimental results show that the DT method gives similar results when testing on data sets with the ratio 1:1, 1:2, and 1:4, not affected when the data is unbalanced, while the SVM method gives the best results when testing on a balanced dataset with a ratio of 1:1 and the lowest when testing on an unbalanced dataset of 1:4, while the MLP and LSTM methods give the best results when using the most data set of 1:4 scale and the lowest when testing on the least data set of 1:1.



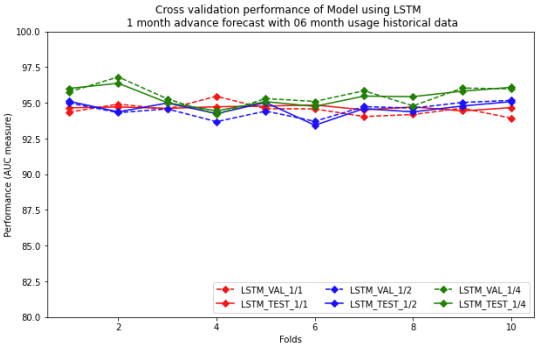
(a) Using DT classification method



(b) Using SVM classification method



(c) Using MLP classification method



(d) Using LSTM classification method

Fig. 15. Performance of 1-month advance forecasting model with 6 months usage historical data.

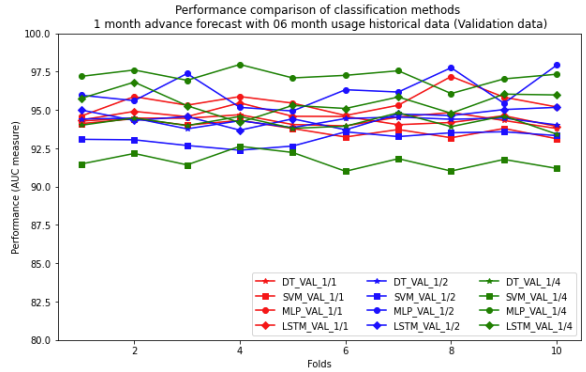
TABLE X: AVERAGE PERFORMANCE PERFORM CROSS-VALIDATION ON VALIDATION DATA (KFOLD = 10) OF 1-MONTH ADVANCE PREDICTION MODEL (DATA SET WITH 6 MONTHS OF HISTORICAL DATA)

Method	Testing on data VALIDATION with 6 months of usage history data					
	The rate of subscribers leaving network / active is 1:4		The rate of subscribers leaving network / active is 1:2		The rate of subscribers leaving network / active is 1:1	
	AUC (%)	σ AUC	AUC (%)	σ AUC	AUC (%)	σ AUC
DT	94.16	0.42	94.27	0.27	**94.34**	0.31
SVM	91.67	0.52	93.11	0.40	**93.77**	0.44
MLP	**97.21**	0.48	96.27	1.02	95.53	0.70
LSTM	**95.51**	0.70	94.53	0.49	94.52	0.42

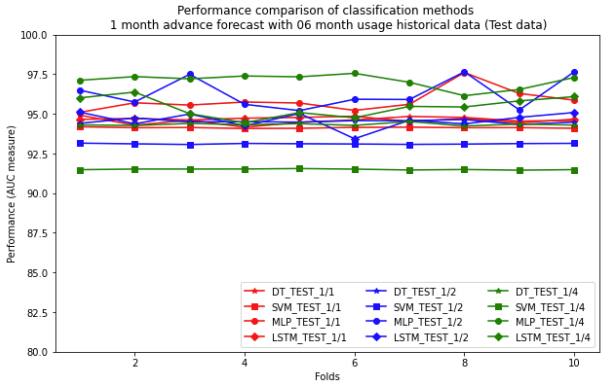
In Table X, Table XI, Fig. 16 presents the average results performing cross-validation using DT, SVM, MLP, and LSTM methods on data sets with a scale of 1:1, 1:2, and 1:4 (with historical data of 6 months of usage). Looking at these figures and tables, we see that:

TABLE XI: AVERAGE PERFORMANCE PERFORM TESTING ON TEST DATA FOR PREDICTIVE MODEL 1 MONTH IN ADVANCE (DATA SET WITH HISTORICAL DATA OF 6 MONTHS)

Method	Testing on data TEST with 6 months of usage history data					
	The rate of subscribers leaving network / active is 1:4		The rate of subscribers leaving network / active is 1:2		The rate of subscribers leaving network / active is 1:1	
	AUC (%)	σ AUC	AUC (%)	σ AUC	AUC (%)	σ AUC
DT	94.33	0.08	94.52	0.11	**94.58**	0.22
SVM	91.50	0.03	93.11	0.03	**94.13**	0.03
MLP	**97.09**	0.41	96.29	0.92	95.84	0.67
LSTM	**95.45**	0.60	94.60	0.50	94.67	0.12



(a) Testing on validation dataset



(b) Testing on test dataset

Fig. 16. Comparison of the forecasting performance of the classification methods (6-month usage historical data and 1-month advance forecast).

The DT method has the highest average AUC performance of 94.34%, standard deviation of 0.31, the lowest distribution of 93.80%, the highest of 94.84%, and the median of 94.39% when tested on a validation dataset with a ratio of 1:1; and AUC performance of 94.48%, standard deviation of 0.22, the lowest distribution of 94.16%, highest of 94.92%, and median of 94.60% when tested on test dataset of 1:1 ratio.

The SVM method has the highest average AUC performance of 93.77%, standard deviation of 0.44, the lowest distribution of 93.12%, the highest of 94.43%, the median of 93.79% when tested on a validation dataset with a ratio of 1:1; and AUC performance of 94.13%, standard deviation of 0.03, the lowest distribution of 94.08%, the highest of 94.19%, and the median of 94.13% when tested on test data set of 1:1 ratio.

The MLP method has the highest average AUC performance of 97.21%, standard deviation of 0.48, the lowest distribution of 96.07%, the highest of 97.97%, and the median of 97.23% when tested on a validation dataset with a ratio of 1:4; and AUC performance of 97.09%, standard deviation of 0.41, the lowest distribution of 96.14%, the highest of 97.56%, and the median of 97.25% when tested on the test dataset with a ratio of 1:4.

The LSTM method has the highest average AUC performance of 95.51%, standard deviation of 0.70, the lowest distribution of 94.19%, the highest of 96.82%, and the median of 95.53% when tested on a validation dataset with a ratio of 1:4; and AUC performance of 95.45%, standard deviation of 0.60, the lowest distribution of 94.44%, the highest of 96.37%, and the median of 95.45% when tested on a test dataset of 1:4 scale.

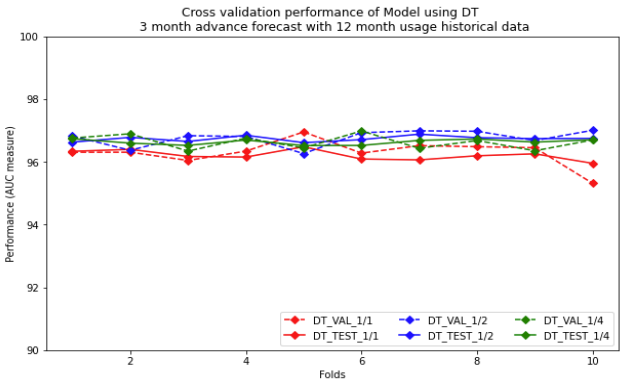
Thus, with the 1-month prediction model and 6 months of usage history data, the MLP method gives the highest performance, the AUC reaches 97.21% on the validation dataset with a rate of 1:4, AUC 97.09% on the test data set with the scale of 1:4; SVM method gives the lowest performance, the AUC is 93.77% on the validation data set of 1:1 and 94.13% on the test dataset with the ratio of 1:1.

In conclusion, with the model that predicts subscribers leaving the network before 1 month, when using data with a history of use in 12 months, 6 months, the SVM method gives the lowest performance, the AUC reaches 93.77% on the validation dataset, with usage historical data of 6 months, balance scale of 1:1; and 94.13% on test datasets with usage historical data of 6 months, balance scale of 1:1; The LSTM method gives the highest performance, the AUC reaches 98.66% on the validation dataset with 12-month usage history data with a rate of 1:4 and 98.78% on the test dataset with 12-month usage history data. with a ratio of 1:4. Realized, the LSTM method gives the highest performance and outperforms the DT, SVM, and MLP methods when the data set has a large enough number of subscribers and usage history.

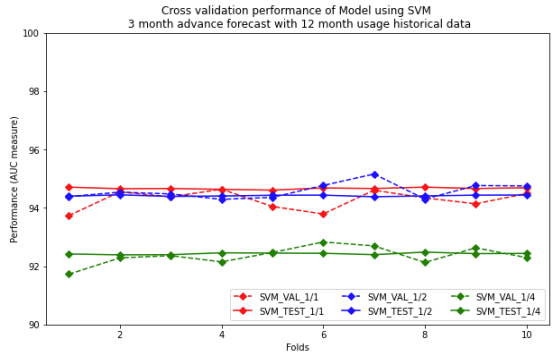
D. Forecast of Subscribers Leaving the Network 3 Months in Advance

1) Performance of forecast model 3 months in advance with 12 months of usage history data

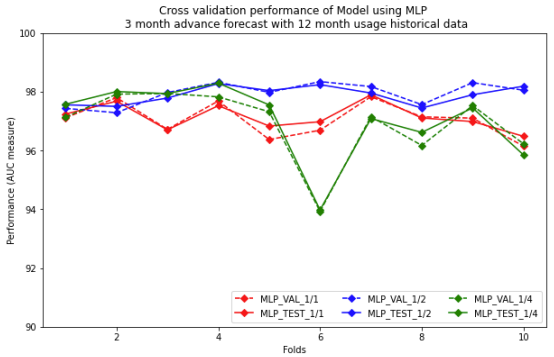
Fig. 17 presents the cross-validation results (KFold is 10) for the model using the DT, SVM, MLP, and LSTM classification method.



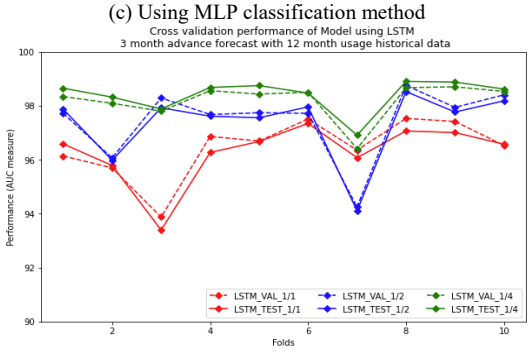
(a) Using DT classification method



(b) Using SVM classification method



(c) Using MLP classification method



(d) Using LSTM classification method

Fig. 17. Performance of 3-month advance forecasting model with 12 months of usage historical data.

TABLE XII: AVERAGE PERFORMANCE PERFORM CROSS-VALIDATION ON VALIDATION DATA (KFOLD = 10) OF 3-MONTH ADVANCE PREDICTION MODEL (DATA SET WITH 12 MONTHS OF HISTORICAL DATA)

Method	Testing on data VALIDATION with 12 months of usage history data					
	The rate of subscribers leaving network / active is 1:4		The rate of subscribers leaving network / active is 1:2		The rate of subscribers leaving network / active is 1:1	
	AUC (%)	σ AUC	AUC (%)	σ AUC	AUC (%)	σ AUC
DT	96.63	0.22	**96.77**	0.25	96.30	0.40
SVM	92.36	0.30	**94.58**	0.26	94.27	0.31
MLP	96.91	1.16	**97.94**	0.37	97.06	0.56
LSTM	**98.20**	0.65	97.46	1.27	96.46	1.04

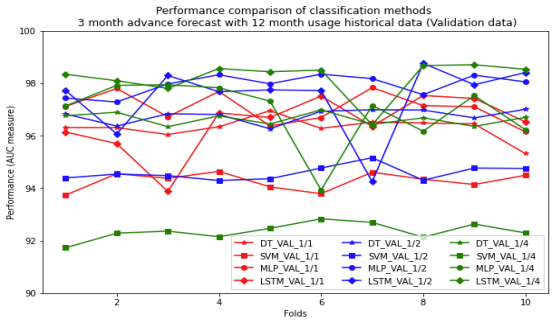
Experimental results show that the DT and MLP methods using data sets with the ratio 1:1, 1:2, and 1:4 give equivalent results, which are not affected when the data is out of balance. While the SVM method gives the best results when using a balanced data set of 1:1 ratio and the lowest when testing on an unbalanced data set of 1:4, the LSTM method gives good results, the highest when using the most data set of 1:4 scale and the lowest when testing on the least data set of 1:1.

In Table XII, Table XIII, Fig. 18 presents the average results performing cross-validation using DT, SVM, MLP, and LSTM methods on data sets with a scale of 1:1, 1:2, and 1:4 scale. (with 12 months of usage history data). Looking at these figures and tables, we see that:

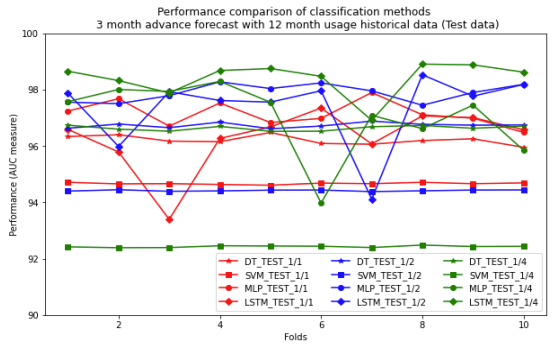
The DT method has the highest average AUC performance of 96.77%, standard deviation of 0.25, the lowest distribution of 96.26%, the highest of 97.00%, and the median of 96.83% when tested on the validation dataset with the ratio 1:2; and AUC of 96.74%, standard deviation of 0.08, the lowest distribution of 96.61%, the highest of 96.88%, and the median of 96.74% when tested on test data set of 1:2 scale.

TABLE XIII: AVERAGE PERFORMANCE PERFORM TESTING ON TEST DATA FOR PREDICTIVE MODEL 3 MONTHS IN ADVANCE (DATA SET WITH HISTORICAL DATA OF 12 MONTHS)

Method	Testing on data TEST with 12 months of usage history data					
	The rate of subscribers leaving network / active is 1:4		The rate of subscribers leaving network / active is 1:2		The rate of subscribers leaving network / active is 1:1	
	AUC (%)	σ AUC	AUC (%)	σ AUC	AUC (%)	σ AUC
DT	96.63	0.08	**96.74**	0.08	96.21	0.15
SVM	92.43	0.03	94.42	0.02	**94.67**	0.03
MLP	97.03	1.23	**97.89**	0.29	97.14	0.42
LSTM	**98.41**	0.57	97.35	1.25	96.28	1.06



(a) Testing on validation dataset



(b) Testing on test dataset

Fig. 18. Comparison of the forecasting performance of the classification methods (12-month usage historical data and 3-month advance forecast).

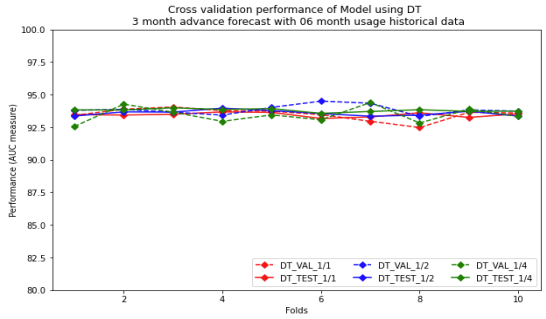
The SVM method has the highest average AUC performance of 94.58%, standard deviation of 0.26, the lowest distribution of 94.29%, the highest of 95.16%, and the median of 94.51% when tested on a validation dataset with a ratio of 1;2; and AUC of 94.67% with a standard deviation of 0.03, the lowest distribution of 94.61%, the highest of 94.71%, and the median of 94.66% when tested on a test data set of 1:1 ratio.

The MLP method has the highest average AUC performance of 97.94%, standard deviation of 0.37, the lowest distribution of 97.28%, the highest of 98.34%, and the median of 98.01% when tested on a validation dataset with a ratio of 1:2; and AUC of 97.89%, standard deviation of 0.29, the lowest distribution of 97.44%, the highest of

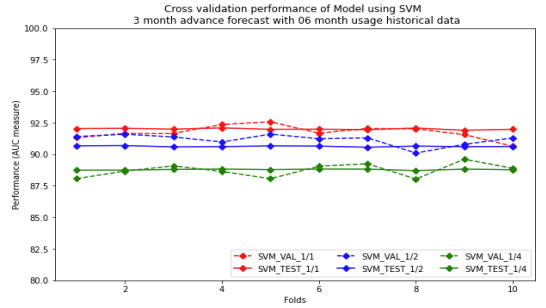
98.28%, and the median of 97.93% when tested on test data set of 1:2 ratio.

The LSTM method has the highest average AUC performance of 98.20%, standard deviation of 0.65, the lowest distribution of 96.41%, the highest of 98.70%, and the median of 98.47% when tested on a validation dataset with a ratio of 1:4; and AUC of 98.41%, standard deviation of 0.57, the lowest distribution of 96.91%, the highest of 98.90%, and the median of 98.64% when tested on a test dataset of 1:4 ratio.

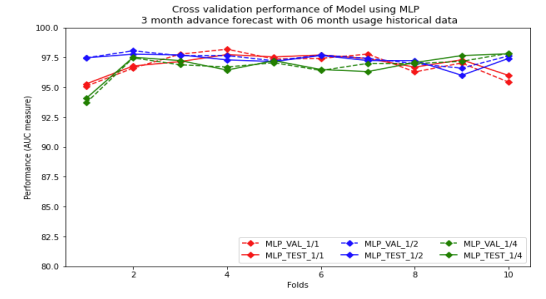
Thus, with the 3-month forecast model and 12-month historical data, the LSTM method gives the highest performance, the AUC reaches 98.20% on the validation dataset with a ratio of 1:4, AUC 98.41% on the test dataset with a ratio of 1:4; The SVM method gives the lowest performance, the AUC is 94.58% on the validation dataset with 1:2 scale and 94.67% on the test dataset with the ratio 1:1.



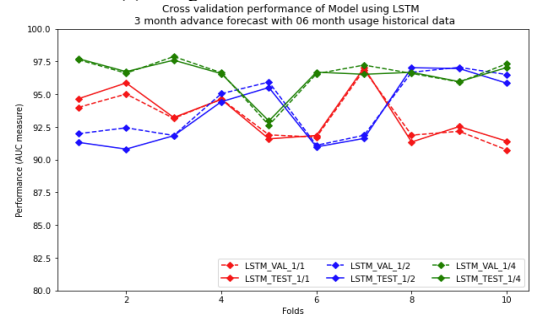
(a) Using DT classification method



(b) Using SVM classification method



(c) Using MLP classification method



(d) Using LSTM classification method

Fig. 19. Performance of 3-month advance forecasting model with 6 months of usage historical data.

2) Performance of forecast model 3 months in advance with 6 months of usage history data

Fig. 19 presents the cross-validation results (KFold is 10) for the model using the DT, SVM, MLP, and LSTM classification method.

Experimental results show that the DT and MLP methods using data sets with the ratio 1:1, 1:2, and 1:4 give equivalent results, which are not affected when the data is out of balance. While the SVM method gives the best results when using a balanced data set of 1:1 ratio and the lowest when testing on an unbalanced data set of 1:4, the LSTM method gives good results, the highest when using the most data set of 1:4 scale and the lowest when testing on the least data set of 1:1.

TABLE XIV: AVERAGE PERFORMANCE PERFORM CROSS-VALIDATION ON VALIDATION DATA (KFOLD = 10) OF 3-MONTH ADVANCE PREDICTION MODEL (DATA SET WITH 6 MONTHS OF HISTORICAL DATA)

Method	Testing on data VALIDATION with 06 months of usage history data					
	The rate of subscribers leaving network / active is 1:4		The rate of subscribers leaving network / active is 1:2		The rate of subscribers leaving network / active is 1:1	
	AUC (%)	σ AUC	AUC (%)	σ AUC	AUC (%)	σ AUC
DT	93.44	0.58	**93.85**	0.35	93.50	0.44
SVM	88.73	0.52	91.16	0.43	**91.74**	0.52
MLP	96.72	1.07	**97.44**	0.39	96.90	0.98
LSTM	**96.51**	1.40	94.04	2.27	93.21	1.79

TABLE XV: AVERAGE PERFORMANCE PERFORM TESTING ON TEST DATA FOR PREDICTIVE MODEL 3 MONTHS IN ADVANCE (DATA SET WITH HISTORICAL DATA OF 6 MONTHS)

Method	Testing on data TEST with 6 months of usage history data					
	The rate of subscribers leaving network / active is 1:4		The rate of subscribers leaving network / active is 1:2		The rate of subscribers leaving network / active is 1:1	
	AUC (%)	σ AUC	AUC (%)	σ AUC	AUC (%)	σ AUC
DT	**93.80**	0.12	93.58	0.20	93.45	0.16
SVM	88.78	0.04	90.63	0.04	**92.00**	0.06
MLP	96.78	1.03	**97.29**	0.48	96.96	0.76
LSTM	**96.45**	1.26	93.63	2.44	93.41	1.92

In Table XIV, Table XV, Fig. 20 presents the average results performing cross-validation using DT, SVM, MLP, and LSTM methods on data sets of 1:1, 1:2, and 1:4 scale. (with 6 months of usage history data). Looking at these figures and tables, we see that:

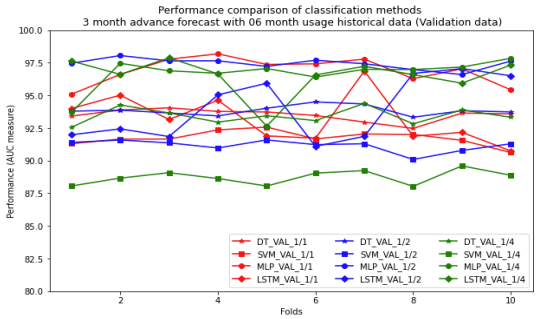
The DT method has the highest average AUC performance of 93.85%, standard deviation of 0.35, the lowest distribution of 93.33%, the highest of 94.50%, and the median of 93.81% when tested on a validation dataset with a ratio of 1:2; and AUC of 93.80%, standard deviation of 0.12, the lowest distribution of 93.56%, the highest of 93.99%, and the median of 93.82% when tested on a test data set of 1:4 ratio;

The SVM method has the highest average AUC performance of 91.74%, standard deviation of 0.52, the lowest distribution of 90.63%, the highest of 92.57%, and the median of 91.66% when tested on a validation dataset with a ratio of 1:1; and AUC of 92.00% with 0.06 standard deviation, the lowest distribution of 91.90%, the highest of 92.09%, and the median of 91.99% when tested on test data set of 1:1 ratio;

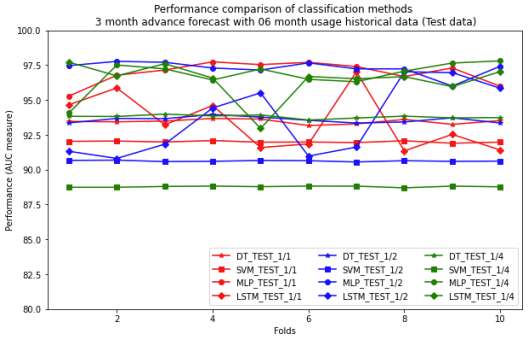
The MLP method has the highest average AUC

performance of 97.44%, standard deviation of 0.39, the lowest distribution of 96.60%, the highest of 98.05%, and the median of 97.55% when tested on a validation dataset with a ratio of 1:2; and AUC of 97.29%, standard deviation of 0.48, the lowest distribution of 96.00%, the highest of 97.77%, and the median of 97.36% when tested on the test data set of 1:2 ratio;

The LSTM method has the highest average AUC performance of 96.51%, standard deviation of 1.40, the lowest distribution of 92.64%, the highest of 97.88%, and the medium of 96.63% when tested on a validation dataset with a ratio of 1:4; and AUC of 96.45%, standard deviation of 1.26, the lowest distribution of 92.96%, the highest of 97.71%, and the median of 96.68% when tested on a test dataset of 1:4 ratio.



(a) Testing on validation dataset



(b) Testing on test dataset

Fig. 20. Comparison of the forecasting performance of the classification methods (6-month usage historical data and 3-month advance forecast).

TABLE XVI: RESULTS OF CROSS-VALIDATION (AUC) OF FORECASTING MODEL 1 MONTH IN ADVANCE, TESTING ON VALIDATION DATA

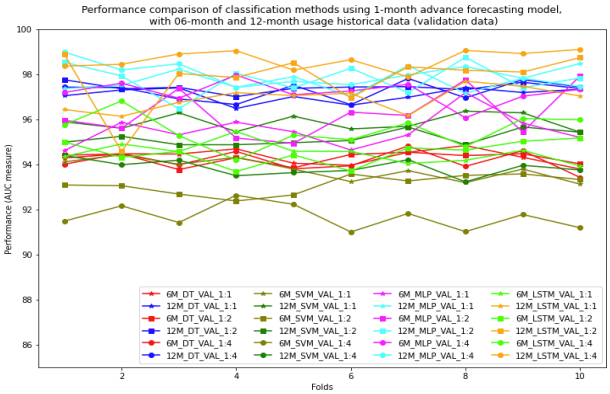
Data set	Forecast model 1 month in advance Testing on validation data							
	6 months of usage history data				12 months of usage history data			
	DT	SVM	MLP	LSTM	DT	SVM	MLP	LSTM
Scale of 1:1	94,34	93,77	95,53	94,52	97,08	95,87	97,73	96,91
Scale of 1:2	94,27	93,11	96,27	94,53	97,41	95,16	97,78	97,82
Scale of 1:4	94,16	91,67	97,21	95,51	97,26	93,86	97,99	98,66

TABLE XVII: RESULTS OF CROSS-VALIDATION (AUC) OF FORECASTING MODEL 1 MONTH IN ADVANCE, TESTING ON TEST DATA

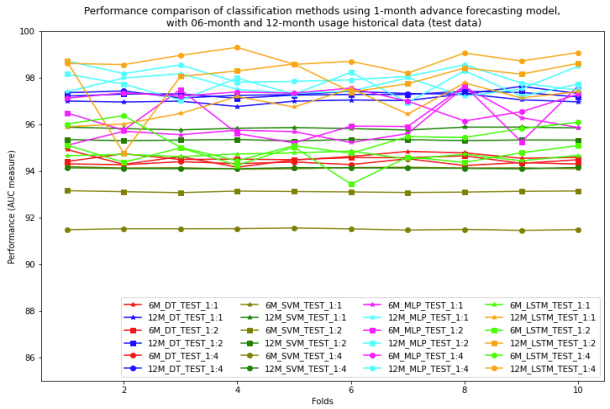
Data set	Forecast model 1 month in advance Testing on test data							
	6 months of usage history data				12 months of usage history data			
	DT	SVM	MLP	LSTM	DT	SVM	MLP	LSTM
Scale of 1:1	94,58	94,13	95,84	94,67	97,01	95,83	97,70	96,86
Scale of 1:2	94,52	93,11	96,29	94,60	97,27	95,32	97,66	97,87
Scale of 1:4	94,33	91,50	97,09	95,45	97,34	94,12	98,09	98,78

Thus, with the 3-month forecast model and 6-month historical data, the MLP method gives the highest performance, the AUC reaches 97.44%% on the validation dataset with the rate of 1:2, AUC 97.29% on the test dataset with a ratio of 1:2; SVM method gives the lowest performance, the AUC reaches 91.74% on the validation dataset of 1:1 scale and 92.00% on the test dataset with a ratio of 1:1.

In conclusion, with the model that predicts subscribers leaving the network before 3 months when using data with a history of use in 12 months and 6 months, the SVM method gives the lowest performance, the AUC reaches 91.74% on the validation dataset with usage historical data of 6 months, balance scale of 1:1; and 92.00% on test datasets with usage historical data of 6 months, balance scale of 1:1; The LSTM method gives the highest performance, the AUC reaches 98.20% on the validation dataset with 12-month usage history data with a rate of 1:4 and 98.41% on the test dataset with 12-month usage history data, a ratio of 1:4. Realized, the LSTM method gives the highest performance and outperforms the DT, SVM, and MLP methods when the data set has a large enough number of subscribers and usage history.



(a) Testing on validation data



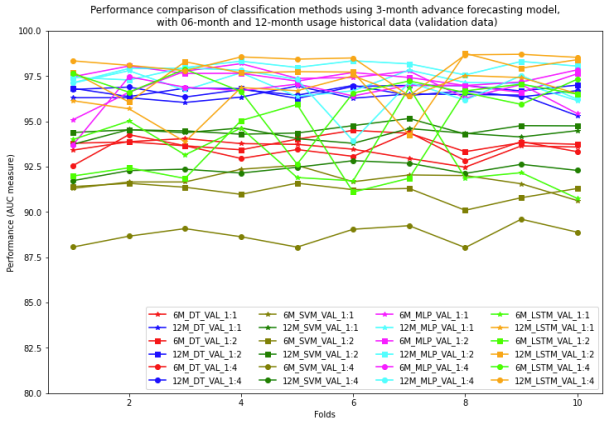
(b) Testing on test data

Fig. 21. Comparison of forecasting performance of the 1-month advance model.

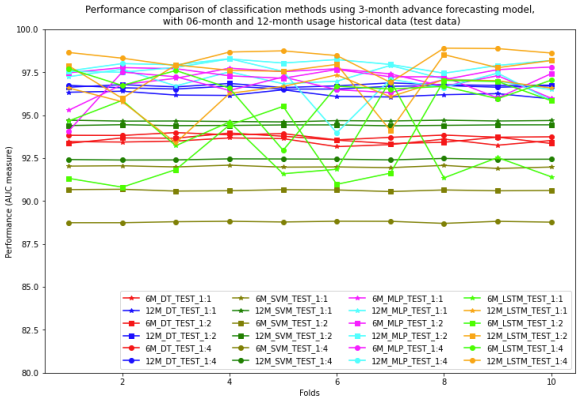
E. General Evaluation

The general evaluation for the 1-month and 3-month advance prediction model, from Table XVI-Table XIX and Fig. 21, Fig. 22 shows that, when using a dataset with a 12-month usage history to train the model, it will achieve stable forecasting performance, and better than using data set with 6 months usage history. This means that, for data with a 12-month usage history, it will provide more complete information about a customer's usage behavior, allowing a more accurate assessment of the likelihood of a customer leaving the network or not; In addition, when changing the ratio of training data sets 1:1, 1:2 and 1:4, there is not much influence on the predictive performance

of DT and MLP models, only the SVM model gives better results at 1:1 scale and the LSTM model gives better results at 1:4 scale. This proves that the DT and MLP classification methods handle the imbalanced data problem very well, while the SVM model works better with balanced data and the LSTM model performs better when there is a lot of data providing enough information to train the model. Regarding the predictive performance of the model, the model using LSTM classification method gives the best results, followed by MLP, DT, and finally, SVM gives the worst results.



(a) Testing on validation data



(b) Testing on test data

Fig. 22. Comparison of forecasting performance of the 3-month advance model.

TABLE XVIII: RESULTS OF CROSS-VALIDATION (AUC) OF PREDICTIVE MODEL 3 MONTHS IN ADVANCE, TESTED ON VALIDATION DATA

Data set	Forecast model 3 months in advance, testing on validation data							
	6 months of usage history data				12 months of usage history data			
	DT	SVM	MLP	LSTM	DT	SVM	MLP	LSTM
Scale of 1:1	93,50	91,74	96,90	93,21	96,30	94,27	97,06	96,46
Scale of 1:2	93,85	91,16	97,44	94,04	96,77	94,58	97,94	97,46
Scale of 1:4	93,44	88,73	96,72	96,51	96,63	92,36	96,91	98,20

TABLE XIX: RESULTS OF CROSS-VALIDATION (AUC) OF PREDICTIVE MODEL 3 MONTHS IN ADVANCE, TESTED ON TEST DATA

Data set	Forecast model 3 months in advance, testing on test data							
	6 months of usage history data				12 months of usage history data			
	DT	SVM	MLP	LSTM	DT	SVM	MLP	LSTM
Scale of 1:1	93,45	92,00	96,96	93,41	96,21	94,67	97,14	96,28
Scale of 1:2	93,58	90,63	97,29	93,63	96,74	94,42	97,89	97,35
Scale of 1:4	93,80	88,78	96,78	96,45	96,63	92,43	97,03	98,41

V. CONCLUSIONS

In this study, we exploit broadband internet subscription data at VNPT An Giang's database and propose an attribute set along with 4 methods of classifying DT, SVM, MLP, and LSTM to build, test, and choose the most suitable model, applied to the problem of forecasting broadband internet subscribers leaving the network 1 month and 3 months before. As shown in the experimental results, using a dataset with a 12-month usage history will give better results than using a 6-month usage data set; and the model using the LSTM classification method gives the highest performance and outperforms the models using the DT, SVM, and MLP methods.

The results of our study have two main contributions: (1) proposed set of attributes including 8 attributes related to forecasting broadband internet subscribers leaving the network (Table V), with historical uses data of 12 months; (2) propose a model using LSTM classification method suitable for the problem of forecasting broadband internet subscribers cancelling the network 1 month and 3 months before.

In this study, although the results are quite good, it is still limited that the broadband subscriber data set is only collected from one service provider, VNPT An Giang, so it is not general. Therefore, the research results are only suitable for VNPT An Giang, so it is necessary to be very careful when applying these results to other service providers. This shows that, in the future, in order to increase the generality of the model, it is necessary to continue researching and collecting data from many different service providers; At the same time, in order to increase the stability and predictive performance of the model, it is also possible to study in the direction of impact hyperparameters affecting the LSTM model. In addition, the topic can also research and develop more forecasts for many other types of services in the telecommunications industry, such as mobile phone services and IPTV television services.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Author Hoang contributed ideas of the article. Author Le contributed to the preparation of data. Author Le participated in implementation and running the experiments. Authors Le and Hoang contributed to the analysis of the results wrote the paper. All authors had approved the final version.

REFERENCES

[1] Ministry of Information and Communications, "Vietnam Information and Communication Technology White Paper 2020," *Information and Communication Publishing House*, 2020.

[2] I. Brandusoiu and G. Toderean, "Churn Prediction in the Telecommunications Sector Using Support Vector Machines," *Ann. ORADEA Univ. Fascicle Manag. Technol. Eng.*, vol. XXII (XII), no. 1, pp. 19–22, 2013.

[3] M. A. H. Farquad, V. Ravi, and S. B. Raju, "Churn prediction using comprehensible support vector machine: An analytical CRM application," *Appl. Soft Comput. J.*, vol. 19, pp. 31–40, 2014.

[4] Y. Huang *et al.*, "Telco churn prediction with big data," in *Proc. ACM SIGMOD Int. Conf. Manag. Data*, , 2015, pp. 607–618, doi: 10.1145/2723372.2742794.

[5] I. Brându oiu, G. Toderean, and H. Beleiu, "Methods for churn prediction in the prepaid Mobile Telecommunications Industry," pp. 97–100, 2016.
[6] I. B. Brandusoiu and G. Toderean, "Churn prediction in the telecommunications sector using neural networks," *Acta Tech. Napocensis*, vol. 57, no. 1, p. 27, 2016.
[7] P. K. Dalvi, S. K. Khandge, A. Deomore, A. Bankar, and V. A. Kanade, "Analysis of customer churn prediction in telecom industry using decision trees and logistic regression," in *Proc. 2016 Symp. Colossal Data Anal. Networking, CDAN 2016*, 2016.
[8] Z. Can and E. Albey, "Churn prediction for mobile prepaid subscribers," in *Proc. 6th Int. Conf. Data Sci. Technol. Appl.*, no. Data, 2017, pp. 67–74, doi: 10.5220/0006425300670074.
[9] V. Umayaparvathi and K. Iyakutti, "Automated Feature Selection and Churn Prediction using Deep Learning Models," *Int. Res. J. Eng. Technol.*, vol. 4, no. 3, pp. 1846–1854, 2017, [Online]. Available: https://irjet.net/archives/V4/i3/IRJET-V4I3422.pdf.
[10] A. Mishra and U. S. Reddy, "A novel approach for churn prediction using deep learning," in *Proc. 2017 IEEE Int. Conf. Comput. Intell. Comput. Res. ICCIC 2017*, December 2018, pp. 1–4.
[11] C. G. Mena, A. De Caigny, K. Coussement, K. W. De Bock, and S. Lessmann, "Churn prediction with sequential data and deep neural networks," *A Comparative Analysis*, pp. 1–12, 2019, [Online]. Available: http://arxiv.org/abs/1909.11114.
[12] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," *J. Big Data*, vol. 6, no. 1, 2019.
[13] S. Höppner, E. Stripling, B. Baesens, S. vanden Broucke, and T. Verdonck, "Profit driven decision trees for churn prediction," *Eur. J. Oper. Res.*, vol. 284, no. 3, pp. 920–933, 2020.
[14] O. F. Seymen, O. Dogan, and A. Hiziroglu, "Customer churn prediction using deep learning," *Adv. Intell. Syst. Comput.*, , pp. 520–529, 2021.
[15] M. Pondel *et al.*, "Deep learning for customer churn prediction in e-commerce decision support," *Bus. Inf. Syst.*, vol. 1, pp. 3–12, 2021.
[16] E. Domingos, B. Ojeme, and O. Daramola, "Experimental analysis of hyperparameters for deep learning-based churn prediction in the banking sector," *Computation*, vol. 9, no. 3, 2021.
[17] A. Dalli, "Impact of hyperparameters on deep learning model for customer churn prediction in telecommunication sector," *Math. Probl. Eng.*, 2022.
[18] W. F. Samah, S. Suresh, and A. K. Moaiad, "Customer churn prediction in telecommunication industry using deep learning," *Inf. Sci. Lett.*, vol. 11, no. 1, pp. 1–15, 2022.
[19] O. F. Seymen, E. Ölmez, O. Doğan, O. Er, and K. Hiziroğlu, "Customer churn prediction using ordinary artificial neural network and convolutional neural network algorithms: a comparative performance assessment," *Gazi Univ. J. Sci.*, vol. 36, no. 2, 2022.

Copyright © 2023 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).

Dong-Ho Le received a degree in electrical and electronic engineering from Ho Chi Minh City University of Technology and Education, Vietnam, in 1996. He is currently the deputy director of Information Technology Center, An Giang Telecommunications. His research interests include forecasting systems, AI, and IoT.

Van-Dung Hoang received the Ph.D. degree from the University of Ulsan, South Korea, in 2014. He was associated and joined as a visiting researcher with the Intelligence Systems Laboratory, University of Ulsan, 2015. He joined the Robotics Laboratory on Artificial Intelligence, Telecom SudParis as a postdoctoral fellow, 2016. He has been serving as an associate professor in computer science, Faculty of Information Technology, Ho Chi Minh City University of Technology and Education, Vietnam. He has published numerous research articles in ISI, Scopus indexed, and high-impact factor journals. His research interests include a wide area, which focuses on pattern recognition, machine learning, medical image processing, computer vision application, visionbased robotics and ambient intelligence.

APPENDIX: DETAILS OF COLUMNS OF DATA TO BE SELECTED

COLUMNS NAME	DESCRIPTIONS
TOCDOTHUC_N14	Line speed (Mbps), previous month current month 14 months
TOCDOTHUC_N13	Line speed (Mbps), previous month current month 13 months
TOCDOTHUC_N12	Line speed (Mbps), previous month current month 12 months
TOCDOTHUC_N11	Line speed (Mbps), previous month current month 11 months
TOCDOTHUC_N10	Line speed (Mbps), previous month current month 10 months
TOCDOTHUC_N9	Line speed (Mbps), previous month current month 9 months
TOCDOTHUC_N8	Line speed (Mbps), previous month current month 8 months
TOCDOTHUC_N7	Line speed (Mbps), previous month current month 7 months
TOCDOTHUC_N6	Line speed (Mbps), previous month current month 6 months
TOCDOTHUC_N5	Line speed (Mbps), previous month current month 5 months
TOCDOTHUC_N4	Line speed (Mbps), previous month current month 4 months
TOCDOTHUC_N3	Line speed (Mbps), previous month current month 4 months
TOCDOTHUC_N2	Line speed (Mbps), previous month current month 2 months
TOCDOTHUC_N1	Line speed (Mbps), previous month current month 1 months
SO_THANG_DCT_N14	Remaining number of prepaid months of the previous month of the current month 14 months
SO_THANG_DCT_N13	Remaining number of prepaid months of the previous month of the current month 13 months
SO_THANG_DCT_N12	Remaining number of prepaid months of the previous month of the current month 12 months
SO_THANG_DCT_N11	Remaining number of prepaid months of the previous month of the current month 11 months
SO_THANG_DCT_N10	Remaining number of prepaid months of the previous month of the current month 10 months
SO_THANG_DCT_N9	Remaining number of prepaid months of the previous month of the current month 9 months
SO_THANG_DCT_N8	Remaining number of prepaid months of the previous month of the current month 8 months
SO_THANG_DCT_N7	Remaining number of prepaid months of the previous month of the current month 7 months
SO_THANG_DCT_N6	Remaining number of prepaid months of the previous month of the current month 6 months
SO_THANG_DCT_N5	Remaining number of prepaid months of the previous month of the current month 5 months
SO_THANG_DCT_N4	Remaining number of prepaid months of the previous month of the current month 4 months
SO_THANG_DCT_N3	Remaining number of prepaid months of the previous month of the current month 3 months
SO_THANG_DCT_N2	Remaining number of prepaid months of the previous month of the current month 2 months
SO_THANG_DCT_N1	Remaining number of prepaid months of the previous month of the current month 1 months
TUOI_N14	Customer age at the previous month of the current month 14 months
TUOI_N13	Customer age at the previous month of the current month 13 months
TUOI_N12	Customer age at the previous month of the current month 12 months
TUOI_N11	Customer age at the previous month of the current month 11 months
TUOI_N10	Customer age at the previous month of the current month 10 months
TUOI_N9	Customer age at the previous month of the current month 9 months
TUOI_N8	Customer age at the previous month of the current month 8 months

COLUMNS NAME	DESCRIPTIONS
TUOI_N7	Customer age at the previous month of the current month 7 months
TUOI_N6	Customer age at the previous month of the current month 6 months
TUOI_N5	Customer age at the previous month of the current month 5 months
TUOI_N4	Customer age at the previous month of the current month 4 months
TUOI_N3	Customer age at the previous month of the current month 3 months
TUOI_N2	Customer age at the previous month of the current month 2 months
TUOI_N1	Customer age at the previous month of the current month 1 months
KHU_VUC_N14	Customer management area in previous month current month 14 months
KHU_VUC_N13	Customer management area in previous month current month 13 months
KHU_VUC_N12	Customer management area in previous month current month 12 months
KHU_VUC_N11	Customer management area in previous month current month 11 months
KHU_VUC_N10	Customer management area in previous month current month 10 months
KHU_VUC_N9	Customer management area in previous month current month 9 months
KHU_VUC_N8	Customer management area in previous month current month 8 months
KHU_VUC_N7	Customer management area in previous month current month 7 months
KHU_VUC_N6	Customer management area in previous month current month 6 months
KHU_VUC_N5	Customer management area in previous month current month 5 months
KHU_VUC_N4	Customer management area in previous month current month 4 months
KHU_VUC_N3	Customer management area in previous month current month 3 months
KHU_VUC_N2	Customer management area in previous month current month 2 months
KHU_VUC_N1	Customer management area in previous month current month 1 months
THANG_SU_DUNG_N14	Total number of months using the service in the month before the current month 14 months
THANG_SU_DUNG_N13	Total number of months using the service in the month before the current month 13 months
THANG_SU_DUNG_N12	Total number of months using the service in the month before the current month 12 months
THANG_SU_DUNG_N11	Total number of months using the service in the month before the current month 11 months
THANG_SU_DUNG_N10	Total number of months using the service in the month before the current month 10 months
THANG_SU_DUNG_N9	Total number of months using the service in the month before the current month 9 months
THANG_SU_DUNG_N8	Total number of months using the service in the month before the current month 8 months
THANG_SU_DUNG_N7	Total number of months using the service in the month before the current month 7 months
THANG_SU_DUNG_N6	Total number of months using the service in the month before the current month 6 months
THANG_SU_DUNG_N5	Total number of months using the service in the month before the current month 5 months
THANG_SU_DUNG_N4	Total number of months using the service in the month before the current month 4 months
THANG_SU_DUNG_N3	Total number of months using the service in the month before the current month 3 months
THANG_SU_DUNG_N2	Total number of months using the service in the month before the current month 2 months
THANG_SU_DUNG_N1	Total number of months using the service in the month before the current month 1 month
LUU_LUONG_N14	Traffic usage in the month, at the previous month of the current month 14 months
LUU_LUONG_N13	Traffic usage in the month, at the previous month of the current month 13 months
LUU_LUONG_N12	Traffic usage in the month, at the previous month of the current month 12 months
LUU_LUONG_N11	Traffic usage in the month, at the previous month of the current month 11 months
LUU_LUONG_N10	Traffic usage in the month, at the previous month of the current month 10 months
LUU_LUONG_N9	Traffic usage in the month, at the previous month of the current month 9 months
LUU_LUONG_N8	Traffic usage in the month, at the previous month of the current month 8 months
LUU_LUONG_N7	Traffic usage in the month, at the previous month of the current month 7 months
LUU_LUONG_N6	Traffic usage in the month, at the previous month of the current month 6 months
LUU_LUONG_N5	Traffic usage in the month, at the previous month of the current month 5 months
LUU_LUONG_N4	Traffic usage in the month, at the previous month of the current month 4 months
LUU_LUONG_N3	Traffic usage in the month, at the previous month of the current month 3 months
LUU_LUONG_N2	Traffic usage in the month, at the previous month of the current month 2 months
LUU_LUONG_N1	Traffic usage in the month, at the previous month of the current month 1 month
NO_CUOC_N14	Amount owed by customers in the month before the current month 14 months
NO_CUOC_N13	Amount owed by customers in the month before the current month 13 months
NO_CUOC_N12	Amount owed by customers in the month before the current month 12 months
NO_CUOC_N11	Amount owed by customers in the month before the current month 11 months
NO_CUOC_N10	Amount owed by customers in the month before the current month 10 months
NO_CUOC_N9	Amount owed by customers in the month before the current month 9 months
NO_CUOC_N8	Amount owed by customers in the month before the current month 8 months
NO_CUOC_N7	Amount owed by customers in the month before the current month 7 months
NO_CUOC_N6	Amount owed by customers in the month before the current month 6 months
NO_CUOC_N5	Amount owed by customers in the month before the current month 5 months
NO_CUOC_N4	Amount owed by customers in the month before the current month 4 months
NO_CUOC_N3	Amount owed by customers in the month before the current month 3 months
NO_CUOC_N2	Amount owed by customers in the month before the current month 2 months
NO_CUOC_N1	Amount owed by customers in the month before the current month 1 month
NGAY_LUU_LUONG_N14	Total number of days with usage traffic, in the month before the current month 14 months
NGAY_LUU_LUONG_N13	Total number of days with usage traffic, in the month before the current month 13 months
NGAY_LUU_LUONG_N12	Total number of days with usage traffic, in the month before the current month 12 months
NGAY_LUU_LUONG_N11	Total number of days with usage traffic, in the month before the current month 11 months
NGAY_LUU_LUONG_N10	Total number of days with usage traffic, in the month before the current month 10 months
NGAY_LUU_LUONG_N9	Total number of days with usage traffic, in the month before the current month 9 months
NGAY_LUU_LUONG_N8	Total number of days with usage traffic, in the month before the current month 8 months
NGAY_LUU_LUONG_N7	Total number of days with usage traffic, in the month before the current month 7 months
NGAY_LUU_LUONG_N6	Total number of days with usage traffic, in the month before the current month 6 months
NGAY_LUU_LUONG_N5	Total number of days with usage traffic, in the month before the current month 5 months
NGAY_LUU_LUONG_N4	Total number of days with usage traffic, in the month before the current month 4 months
NGAY_LUU_LUONG_N3	Total number of days with usage traffic, in the month before the current month 3 months

COLUMNS NAME	DESCRIPTIONS
NGAY_LUU_LUONG_N2	Total number of days with usage traffic, in the month before the current month 2 months
NGAY_LUU_LUONG_N1	Total number of days with usage traffic, in the month before the current month 1 month
ROI_MANG	Class Attribute (Active: 0; Leaving Network: 1)