Customer Churn Prediction in an Internet Service Provider

Duyen DO
Data Scientist Big-Data
FPT-Telecom
HCM, Vietnam
duyen.do.vn@gmail.com

Phuc HUYNH
Data Scientist Big-Data
FPT-Telecom
HCM, Vietnam
phucHM5@fpt.com.vn

Phuong VO
Data Scientist Big-Data
FPT-Telecom
HCM, Vietnam
phuongVTH@fpt.com.vn

Tu VU
EVP FPT-Telecom
FPT-Telecom
HCM, Vietnam
tuVA@fpt.com.vn

***Abstract*—Customer retention is regarded as one of the most concerns in any company, since they provide the fundamental source of revenue for business. Losing customers not only loses the profit, but also may put a whole business in danger. In order to increase customer base, businesses need to improve both acquisition and retention of its customers. Therefore, customer churn prediction is becoming the top concerns that many companies devoted their time and resources to address it. This paper presents the customer churn prediction on an extremely imbalanced data in an Internet Service Provider company to identify the users at the risk of leaving the services. It consists of feature engineering and predictive modeling. In the feature engineering, the most essential features are selected from a large number of created candidates. In the predictive modeling, the imbalance between the number of churners and non-churners was reduced using SMOTE oversampling technique before implementing several models such as AdaBoost, Extra Trees, KNN, Neural Network and XGBoost. Comparing between these models in term of precision and recall, the XGBoost model gives the highest performance. Using the dataset with 98% non-churners and 2% churners, precision and recall of the model are 45.71% and 42.06%, respectively**

***Keywords*-Customer Churn Prediction; imbalanced data; feature importance; Internet Service Provider; SMOTE; XGBoost; AdaBoost; kNN; Neural Network; Extra Trees;**

I. Introduction

Nowadays, the businesses are facing with both internal and external challenges which lead consumers to leave the services. Internal challenges come from poor quality of services such as bad products, the bad customer services, high prices, and so on. On the other hand, external challenges come from direct and indirect competitors. Besides acquiring new customers, the business also need to retain existing ones since their churn customers are new customers of their competitors. Therefore, customer churn prediction gains many attentions from major firms, especially in telecommunication industry. Since the growing needs, this industry continues to expand operations on a global level. In Vietnam, it has also developed rapidly in recent years. "Internet usage has increased in popularity as evidenced by the entry of many Internet service providers (ISPs) into the market." [1]. So, potential market is also the highly competitive environment and churn customer is one of various challenges in this business world.

In recent years, our company had tried to solve this challenge in order to minimize churn rate, which accounts for approximately 2% monthly. Building long-term relationship with the customers is one of the most important objectives. For purpose of retention, Customer Service Department intuitively chose customers based on a few conditions such as the number of consecutive days of inactive service, the number of inbound calls and the number of complaints. However, this approach did not detect the probable churning users effectively. After that, a new approach using machine learning technology to predict customer churn was considered. This is a classification task that churners are minority class and non-churners are extremely major class.

The services are monthly subscriptions. In terms and conditions, they are expired at the end of each month. The valid duration to renew the services is from the expired time to next 16 days. In renewable duration, customers can still normally access to the services. After this period, users who do not renew the services are identified as churn customers and their services will be stopped from that time. The customers terminating their contract are also recognized as churn customers. Non-churn customers, on the other hand, are defined as users who renew the services or continue using the services after interruption. The status of a customer, churn or non-churn, is determined at the end of each month, regardless of the previous status.

This paper is organized as follows. Section 1 shows the challenge of loosing the customers, telecommunication industry and customer churn problem. Section 2 reviews the related papers and summarizes background knowledge required to comprehensively follow the remainder. Section 3 describes the data source and the feature selection. Section 4 presents the models construction. Section 5 evaluates and compares the performance between these models. Last section gives the conclusion.

II. Related work

There have been attempt to solve the churn customer issue such as machine-learning classifier techniques [2] [3], data mining technology [4] [5], hybrid neural networks [6], and so on. However, many issues in real life face considerable imbalanced data that these reports have not

978-1-5386-2715-0/17/$31.00 ©2017 IEEE

dealt with, in term of precision and recall. For purpose of cost estimate when taking care of the possible churning customers, precision measures would be more effective. For purpose of retaining most customers, recall of the model needs to be improved. In order to balance between these measures, the most suitable models and data are selected.

One of the challenges in this study is that the number of non-churners is much more than churners'. The ratio of non-churn customers to churn customers stood at 98:2. Therefore, reducing the imbalance of the classes before implementing the models is very important. While under-sampling can put models in danger by removing truly representative of the major class, over-sampling can lead to over-fitting by duplicating the minority class. In order to solve this challenge, the algorithm SMOTE (Synthetic Minority Over-sampling Technique) is selected. SMOTE is defined as "an over-sampling approach in which the minority class is over-sampled by creating synthetic examples rather than by over-sampling with replacement" [7]. This algorithm creates new instances of the minority class instead of duplicating the existing ones.

In the model section, various models such as AdaBoost, Extra Trees, kNN, Neural Network and XGBoost were implemented, then compared with each others. These models are the popular well-known models for classification tasks.

AdaBoost (also known as Adaptive Boosting) is an ensemble technique which constructs a strong learners as linear combination of simple weak learners. Unlike previous boosting algorithms, AdaBoost "adjusts adaptively to the errors of the weak hypotheses" [8]. The final model converges to a strong classifier. However, this algorithm is sensitive to noise data and outliers.

Extra-Trees (also known as Extremely randomized trees) is a variant of Random Forest algorithm [9], but slightly different from Random Forest. When splitting a node during the construction of the tree, Extra-Trees picks a number of randomized decision trees on entire sample instead of the best split among a random subset of the features. Therefore, Extra-Trees is computationally faster than Random Forest.

k-NN (also known as k-Nearest Neighbor) is an non-parametric lazy learning algorithms [10]. When classifying a new case, k-NN finds its k nearest neighbors from all available cases in training set based on a similarity measure such as Euclidean distance, Minkowski distance or Mahalanobis Distance. k-NN is one of the fundamental classification methods.

Neural Network (also known as An Artificial Neural Network) is defined as "...a computing system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs." [11]. It imitates the idea of human brain, which is compose of billions neurons. Neural Network includes feedforward and feedback topologies [12].

XGBoost is an open-source software library which implements an optimized distributed gradient boosting system under the Gradient Boosting framework [13]. Gradient boosting optimizes loss function based on an ensemble of weak learners, typically decision trees, to predict the targets of regression and classification problems.

III. Data and Feature Engineering

This study was investigated by an Internet Service Provider, one of the leading companies in the telecommunications industry in Vietnam. The company provided many different sources that were separated into three main groups: customer information, their usage data and service data.

- Customer information: It contains almost entirely information of contract such as registration date, termination date, location, service type, cable type, bandwidth, payment history, promotion, and so on.
- Customer usage data in telecommunications: This is user activity log data daily, including two major parts. The first part is about information of connection between user's modem and company's server such as the initial date time of connection, disconnection date time, reason for rejection, type of modem. The second part is about user's usage daily such as the amount of data downloaded and the amount of data uploaded.
- Customer service data: These data are collected from data source of customer service activities such as customer's inbound and outbound call phone history, customer satisfaction surveys, maintenance and support, and so on. They show the explicit concerns and problems of customers and are considered as the most useful data. However, unlike from two groups of data mentioned above, this kind of data is not collected frequently for each customer or collected from all customers at the same period because of the human resource cost. Therefore, they are not enough to show substantial differences between churn customers and non-churn customers. In this study, they have not been used in the models yet. In case of having complete data, it will be very helpful for customer churn prediction.

A large number of attributes from the data source in the previous section have been generated, but only stable and relevant attributes were used for modeling. Based on various statistical tests, 121 most relevant features were selected to use for training and testing purpose. These attributes were collected over one year and examined the seasonal effect. It shows that the tendency of them was stable month over month of the year, except the month of Lunar New Year - the biggest holiday in Vietnam. This special month is not included in this study since there are substantial changes in policy and user behavior. For the sake of simplicity, the study collected user's usage data in last 28 days of each month since months have different number of days. Following are the short descriptions and the visualizations of these features.

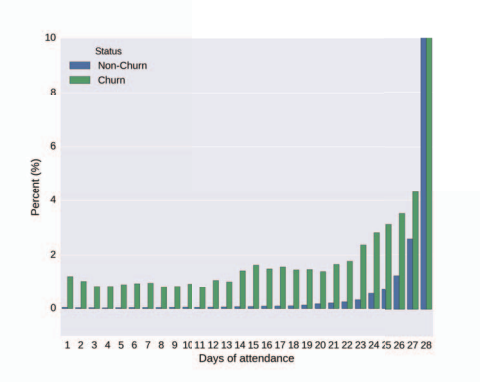

Figure 1. Attendance distribution of churn and non-churn customers.

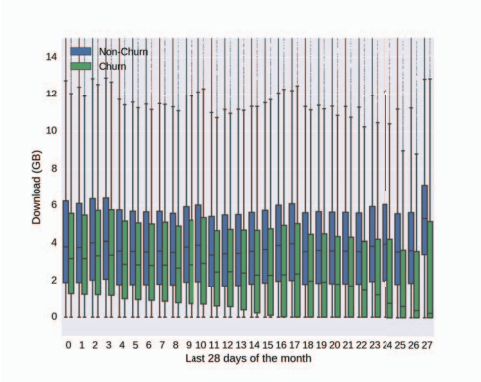

Figure 2. Download distribution of churn and non-churn customers.

A. Point of Presence

A Point of Presence (also known as POP) is a local access point for an Internet Service Provider. Each ISP has thousands POPs at different locations. Many user's modems are connected to a POP to access Internet. They could be seriously affected if the POP was inaccessible. Therefore, this feature could disclose a set of customers who were in trouble together because of their POP. It is the highest scoring variable on feature importance ranking.

B. Attendance

Attendance is the number of days of the month that existed connection between modem and server. This feature reveals a piece of information of the device's quality or Internet access's quality whether it is stable or not. The poor quality during a long time may be a high risk factor which leads customer to leave the services. The difference of attendance distribution between churners and non-churners is shown in Fig. 1.

In this figure, it is obvious to observed that churn customers often had low attendances while most non-churn customers had 28-day attendances.

C. Download and Upload

Download and Upload are the amount of data downloaded and data uploaded by user on one day, respectively. Each of the dataset contains 28 download features and 28 upload features named from 0 to 27, associated with last 28 days of a month. The bigger number is, the later day of the month is. These features present customer needs of download and upload, which involves their decision to stay or leave the service. The following figure illustrates the tendency of user's data download during last 28 days of a month.

In Fig. 2, there is a difference between the trend of customers likely to churn and customers likely to not churn. The quantity of churners' data downloaded is decreased steadily at the end of the month while the trend of non-churners was stable over the period shown.

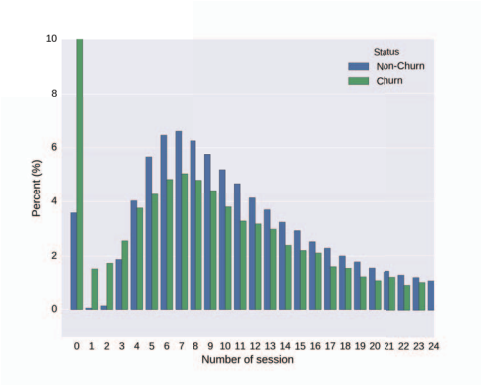

Figure 3. Session distribution of churn and non-churn customers.

D. Session

A session is a semi-permanent interactive information interchange from modem to server. It lasts between a few seconds and a few days. From session logging, the several features such as the number of sessions in a month; the average, minimum, maximum and standard deviation of session duration were created. As can be seen from the Fig. 3 that users having less session often are the probable churning users rather than non-churn users, especially users had not any session.

E. Other features

However, there is an attention that these feature not only depends on device and the Internet qualify but also is affected by the user behavior. Therefore, it is difficult to declare that how much noise of the feature is made by user behavior. In addition to these above features, dataset also contains other relevant features such as difference of download/upload between two consecutive days, service type, cable technology bandwidth range, and so on.

However, in case of having too many variables, implementing all of them consumes too much resource. It is impossible to use descriptive statistical methods for all these features. In that case, using XGBoost feature importance

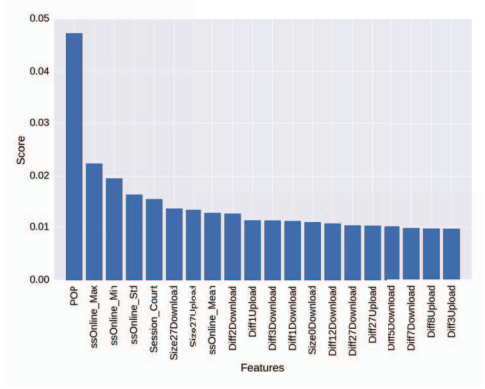

Figure 4. Top important features.

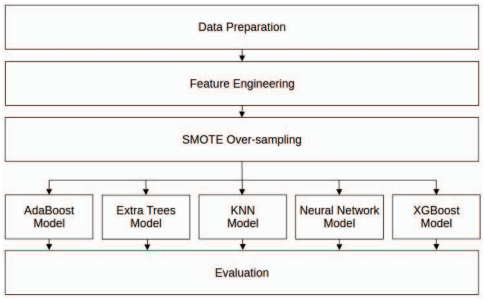

Figure 5. Flow of customer churn prediction.

ranking to extract the top relevance of output target is an appropriate alternative. The performance depends on the number of top selected features.

Fig. 4 shows the sequence of best features of this study selected by XGBoost feature important ranking approach. The top features are POP, session, type of service and difference of download between two consecutive days in turn.

IV. Model

In this section, several models were built to predict the output variable on a sample dataset. The sample consists of 373,137 customers, including 7,688 churners and 365,449 non-churners. A summary flow of prediction process is illustrated in Fig. 5.

There are five main steps in the process. In real life, raw data are often collected from a mess of things that contains missing values, unorganized structure, error records, noise, and so on. Before using it, these issues were handled in data preparation step. Understanding business and data is always required. Then, in feature engineering step, there are several tasks such as extracting useful information, creating new variables, analyzing and selecting informative features. Because of highly imbalanced data, an oversampling step is needed to reduce the imbalance between major class and minority class. After that, five different models were implemented, then evaluated and compared with performance of each others.

A. Oversampling

There are several approaches to resolve the imbalanced data such as under-sampling, over-sampling, and so on. Some works indicate that the appropriate algorithm could handle this problem is SMOTE. Using SMOTE over-sampling, parameters were set with ratio = 0.5 and random state = 42 to increase churn class.

B. Model Parameters

After oversampling, five models were applied to the over-sampling training set. This study tried testing many different sets of parameters on each model before choosing two final sets, which give the highest performance. Following is the set of parameters used on each model.

For AdaBoost model, the set of parameters was *_estimators* = 20, *max_depth* = 50, *criterion* = 'gini', *min_samples_split* = 2 and *min_samples_leaf* = 1.

On Extra-Trees model, parameters were set with *n_estimators* = 20, *max_depth* = 50, *criterion* = 'gini', *min_samples_split* = 2, *min_samples_leaf* = 1.

Similar to that, the parameters of k-Nearest Neighbors were *neighbors* = 5, *weights* = 'uniform' and *leaf_size* = 30.

The best set of parameters for Neural Network model on this dataset is as follows: *solver* = 'lbfgs' (quasi-Newton methods), *alpha* = 1e-5, *hidden_layer* = 10, *random state* = 1, *activation* = 'logistic'.

For XGBoost model, the algorithm was implemented with parameters: *learning rate* = 0.1, *number of estimators* = 20, *max depth* = 50, *min child weight* = 1, *gamma* = 0, *subsample* = 0.5, *colsample by tree* = 0.8, *scale pos weight* = 1, *objective* = 'binary:logistic'.

V. Experiments

This section evaluates the performance of these above models on the dataset having concisely 98:2 ratio of non-churners to churners.

Based on XGBoost feature importance score, the study tried using various number of top important variables in order to find the best fit between each model and these features. For example, ROC curves of XGBoost model reflecting the change of features numbers is shown in Fig. [6].

In this figure, the performance has slight change when the number of features was increased. Dataset of top 50 features is the best fit for XGBoost model. For purpose of measuring the efficiency and cost of customer retention, precision and recall were balanced using F1-score. These measures are shown in Table I. At column "F1-score" in the above table, the highest performance is 43.56%.

Similarly, the study found the most suitable top features for AdaBoost, Extra Trees, KNN and Neural Network model

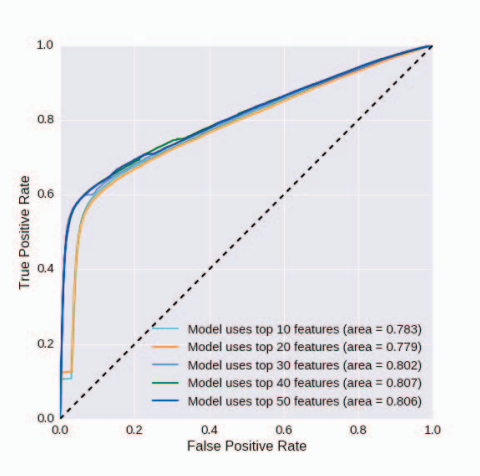

Figure 6. Flow of customer churn prediction.

Table I
XGBOOST MODEL WITH VARIETY NUMBER OF FEATURES

Number of features	Precision	Recall	F1-score
10	16.70%	40.73%	23.69%
20	17.08%	45.89%	24.90%
30	43.25%	38.88%	40.95%
40	44.91%	36.91%	40.52%
50	45.71%	42.06%	43.56%
60	17.11%	46.16%	24.97%

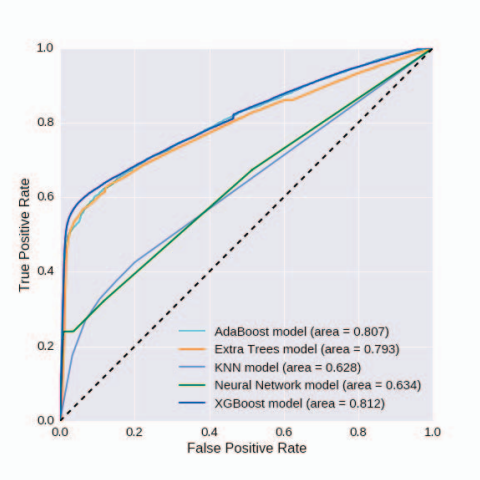

Figure 7. Flow of customer churn prediction.

before comparing their performances. Receiver Operating Characteristic curves (ROC curves) of five models are visualized together in Fig. [7].

As can be seen from this picture that 3 models giving good performances are XGBoost, AdaBoost and Extra Trees in turn. XGBoost is the best model with AUC = 81.2%. The precision, recall and F1-score of each model are shown in

Table II
THE PERFORMANCE OF MODELS

No.	Model	Precision	Recall	F1-score
1	XGBoost	45.71%	42.06%	43.56%
2	Extra trees	35.57%	42.98%	38.92%
3	AdaBoost	25.12%	51.39%	33.74%
5	Neural Networks	32.24%	23.98%	27.50%
4	k-NN	6.06%	32.75%	10.22%

Table II.

In this table, XGBoost is better than others with 43.56% of F1-score. Two models giving the lowest results are Neural Network and k-Nearest Neighbor.

In addition, it is important to consider time and resources when executing these models. Based on the experiment results on the dataset, it is an insignificant difference. Therefore, XGBoost model is recognized as the most suitable solution for this challenge.

Comparing with previous basic approach of Customer Service Department, applying machine learning algorithms to predict churn customers is more effective. Precision increased by nearly 30%, from about 15% to 45.71%, hence the business can reduce a waste of time, money and effort. Moreover, the model also gave a large number of customers, who had some problems with the services. Thus Customer Service can support them effectively and early.

VI. CONCLUSION

This paper presented a classification solution on an imbalanced dataset that identifies users at the risk of leaving the service in telecommunication industry. It mainly focused on feature engineering and modeling phases, which play the crucial roles in the process. Purpose of this study is giving a methodology dealing with the popular problem in real life – the extremely imbalanced data in customer churn prediction challenge. It also applied to other problems in which data is imbalance. Besides that, the paper provided an benchmark results to this problem for improvement purpose in the future. However, the study still has a limit. Feature engineering is based on most history usage of customer but policies. So that, it is difficult when plaining the strategy for retention. Our future researches aim to solve these aspects to complete this problem.

ACKNOWLEDGMENT

This research is sponsored by FPT Telecom churn prediction project. Many thanks to all members who have had the pleasure to work during this and related projects, each of them have important contribution. We are also thankful to the colleagues at Customer Service Department and Information System Center, who provided expertise that greatly assisted the research.

REFERENCES

[1] U.S. Commercial Service Vietnam, "Vietnam Market for Telecommunications Equipment and Services," The U.S. Commercial Service, 2014.

[2] V. Umayaparvathi and K. Iyakutti, "Applications of Data Mining Techniques in Telecom Churn Prediction," in *International Journal of Computer Applications* (0975 8887), vol. 42, no. 20, 2012, doi:10.5120/5814-8122.

[3] Niken Prasasti and Hayato Ohwada, "Applicability of Machine-Learning Techniques in Predicting Customer Defection," in *Technology Management and Emerging Technologies (ISTMET), 2014 International Symposium*, 2014, doi:10.1109/ISTMET.2014.6936498.

[4] Rahul J Jadhav and Usharani T. Pawar, "Churn Prediction in Telecommunication Using Data Mining Technology," in *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 2, Issue 2, 2011, doi:10.14569/IJACSA.2011.020204.

[5] Afaq Alam Khan, Sanjay Jamwal and M.M.Sepehri, "Applying Data Mining to Customer Churn Prediction in an Internet Service Provider,"" in *International Journal of Computer Applications* (0975 8887), vol. 9, no.7, 2010, doi:10.5120/1400-1889.

[6] Chih-Fong Tsai and Yu-Hsin Lu, "Customer churn prediction by hybrid neural networks,"" in *Expert Systems with Applications*, vol. 36, Issue 10, pp. 12547–12553, 2009, doi:10.1016/j.eswa.2009.05.032.

[7] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall and W. Philip Kegelmeyer, "SMOTE: Synthetic Minority Oversampling Technique," in *Journal of Artificial Intelligence Research*, vol. 16, pp. 312–357, 2002, doi:10.1613/jair.953.

[8] Yoav Freund and Robert E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting,"" in *Journal of Computer and System Sciences*, vol. 55, pp. 119–139, 1997, doi:10.1006/jcss.1997.1504.

[9] Pierre Geurts, Damien Ernst and Louis Wehenkel, "Extremely randomized trees," in *Machine Learning*, vol. 63, Issue 1, pp. 3–42, 2006, doi:10.1007/s10994-006-6226-1.

[10] Daniel T. Larose and Chantal D. Larose, "k-Nearest Neighbor Algorithm," in *Discovering Knowledge In Data: An Introduction to Data Mining*, John Wiley & Sons, Inc., 2005, ch. 5, pp. 90–106, doi:10.1002/9781118874059.ch7.

[11] Maureen Caudill, "Neural networks primer, part I," in *Journal AI Expert* , vol. 2, Issue 12, pp. 46–52, 1987.

[12] Richard L. Welch, Stephen M. Ruffing and Ganesh K. Venayagamoorthy, "Comparison of feedforward and feedback neural network architectures for short term wind speed prediction," in *International Joint Conference on Neural Networks*, 2009, doi:10.1109/IJCNN.2009.5179034.

[13] Tianqi Chen and Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceeding KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785-794, doi:10.1145/2939672.2939785.