

Lao Telecom Customer Churn Prediction: A Case Study in Borikhamxay Branch

Outhong Hutsady¹, Phouthone Vongpasith², Phet Sonevilay³, Chitnavanh Phonekhamma⁴, Orlady Khammanivong⁵

Department of Computer Science, Faculty of Natural Sciences,
National University of Laos Vientiane Capital, Lao P.D.R

Abstract

Customer churn practice is essential in competitive and rapidly developing in telecom sector. The process of migrating from one service provider to another telecom service provider occurs due to various of services. The prediction of customer churn has developed as an indispensable part of planning process and strategic decision making in Telecom sector. The main aim of the study is to explore the customer churn prediction in Lao Telecom Borikhamxay Branch using in machine learning algorithms. This study makes use of logistic regression, decision tree and random forest classifiers for predicting consumer churn. The dataset containing 101104 customers was used for training and testing. The experimental results show that when trained and tested on the dataset, the decision tree and random forest classifiers attained a remarkable accuracy of 93%.

Keywords: Customer churn, Lao Telecom Borikhamxay Branch, Machine Learning

Introduction

The underlying principle of customer churn prediction in terms of telecom industry is to calculate subscribers approximately who literally feel like to leave from a company they used so far and suggest solutions to prevent considerable churns. Recently, making an estimation of churners before they quit has become necessary in the environment of stiff competition amongst companies. The major role that the telecom industry plays made it all the more significant to build prediction mechanisms alongside the lines of churn prediction.

Lao Telecommunication Company Limited is a company that provides telecommunication services that has been operating since November 11, 1996 which conducting telecommunication services in various fields such as basic telephone services, mobile telephone services, digital systems, GSM, radio signal services called "Lao Link", public telephone services using cards, international telephone services, Internet services, rural long-distance telephone services, receiving and sending telegrams, faxes and other supplementary services. The customer is important and is a part of the Lao Telecommunication Company that generates the main income. If the customer cancels the service, it may cause the company to have a lack of funds and is the starting point of the problem of this chapter, which is the customer is an important part of the service that the Lao Telecommunication Company provides to the customer.

As a result, tools for developing a customer churn prediction model is critical for the Lao Telecommunication Company. Researchers believe that developing a customer churn prediction model is critical for maintaining consumers in order to create a model. Existing research shows that the primary goal is to identify the valuable churn customer using a large volume of telecom data. However, there are several limitations in existing models that impediments to solving this problem in the real world. In the telecom industry, a large volume of data is generated, and the data contains noise, resulting in poor prediction model results.

To address the churning prediction problem, machine learning algorithms have been proposed. In this paper, churn prediction models is proposed to predict the churning of customers such as logistic regression, decision tree and random forest algorithms. A large volume of Lao Telecom Borikhamxay Branch data is used for training and testing models.

Research Objectives

- To address the churning prediction problem using a large volume of Lao Telecom Borikhamxay Branch data.
- To predict the churning of customers, machine learning algorithms such as logistic regression, decision tree and random forest algorithms were evaluated.

Research Methodology

Proposed Scheme for the Prediction of Customer Churn: In this section, the proposed scheme is presented in Figure 1 with detailed description. The phases used for the proposed scheme are collection of data, Pre-Processing of data, applying machine learning algorithms, evaluations models and customer churn prediction.

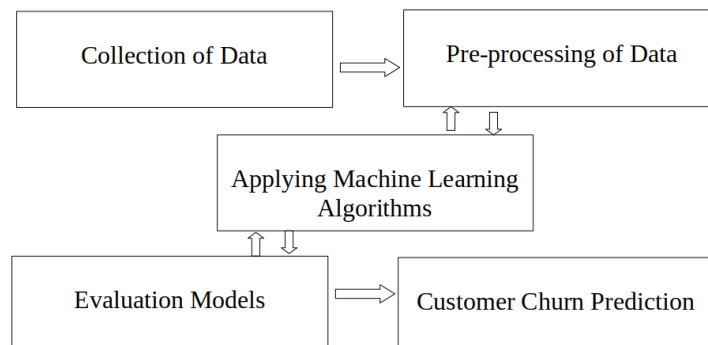


Figure 1 Phases used for proposed scheme.

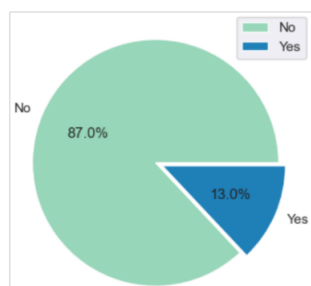
From the above fig.1, the first phase of the proposed scheme is collection of data. This is done to collect the relevant data required to create training and testing of dataset for making predictions. In this case, the data could include information such as customer demographics, usage patterns, billing history, and customer service interactions. The second phase is pre-processing, the main purpose of preparation of data is to improve the quality of data and enhance the performance of customer churn prediction. The preparation of data requires to be undertaken in a much iterative way until a conclusive result is met. This phase involves 3 steps namely data exploration, feature selection and data transformation. The different visualization techniques are applied to different types of variables, so it's helpful to differentiate between continuous and categorical variables and look at them separately. Before training of model, feature selection is one of the most essential factors that can influence the model's performance. Data transformation comprises of explanatory variables which can be transformed from binomial form into binary. So more variables were formed to estimate the transformation in usage of data. In the third first, machine learning algorithms are applied for the classification of customers to distinguish between churn and non-churn customers. In this paper, Logistic Regression, Decision Trees and Random Forest are used for customer classification and prediction.

Collection of Data: The dataset used for experiments in this paper, contains results of Lao Telcom Customer Churn dataset obtained from Lao Telecommunication Company, Borikhamxay Branch (it is also known as LTC). Each row represents a customer, each column contains attribute described on the column Metadata. It consists of 101104 customer information. Every customer has 13 features and the "Churn". The last attribute contains labeled data with two classes where 13 % of total customers are labeled as "Yes" indicating true customers i.e., categorized as churning customers and the remaining 87 % customers are labeled as "No" indicating false customers i.e., categorized as non-churning customers. The attribute selection depends on the results of techniques of feature selection that find useful, the most similar and effective attributes to predict the churning customers. A total of 87960 are non-churners and 13144 are churners. The dataset contains 8 categorical columns and 5 numeric columns. Figure 2 shows details of dataset.

```

RangeIndex: 101104 entries, 0 to 101103
Data columns (total 13 columns):
# Column Non-Null Count Dtype
-----
0 msisdn 101104 non-null int64
1 product 101104 non-null object
2 DNAME 101104 non-null object
3 voice_rev 101104 non-null int64
4 data_rev 101104 non-null int64
5 total_rev 101104 non-null int64
6 refill 101104 non-null int64
7 last_pk 101104 non-null object
8 KYC 101104 non-null object
9 Gender 101104 non-null object
10 lifetime 101104 non-null object
11 Downtime 101104 non-null object
12 Churn 101104 non-null object
dtypes: int64(5), object(8)
memory usage: 10.0+ MB

```



msisdn	product	DNAME	voice_rev	data_rev	total_rev	refill	last_pk	KYC	Gender	lifetime	Downtime	Churn
2055043986	MP	Pakkading	12087	30000	42087	25000	Topup	Yes	F	Over 2 Year	Yes	No
2055044143	MP	Borikhan	17098	0	17098	0	No	Yes	M	Over 2 Year	Yes	No
2055044208	MP	Thaphabat	34734	0	34734	30000	Cards	Yes	F	Over 2 Year	Yes	No
2055044437	MP	Vienthong	0	0	0	0	No	Yes	F	Over 2 Year	Yes	No
2055044797	MP	Khamkeut	33411	0	33411	30000	Cards	Yes	M	Over 2 Year	Yes	No
...
2052812799	MP	Borikhan	11606	0	11606	20000	Topup	Yes	M	Over 2 Year	Yes	No
2052812891	MP	Vienthong	0	30000	30000	20000	Cards	Yes	M	below1Month	Yes	No
2052812898	MP	Paixan	534	28000	28534	20000	Topup	Yes	F	Over 2 Year	Yes	No
2052813011	MP	Vienthong	3574	0	3574	0	No	Yes	M	Over 2 Year	Yes	Yes
2052813110	MP	Khamkeut	6029	98000	104029	278000	Topup	Yes	M	Over 2 Year	No	No

Figure 2 Details of dataset

Pre-processing of Data: Classification of data and performing pre-processing makes it easy to use. Because the column's data type is object type, we convert it to numeric and make a duplicate of the underlying data to manipulate and process. This conversion is performed using panda's library imported in the program and using on-hot encoding function of library, as shown in Figure 3. Next, check for duplicate values in dataset, but there are no duplicate values present in dataset. Further continue checking for unique values. So according to the feature with high number of unique values, it would better to drop those columns due to less analysis and insights. It was found the phone number of customers is not useful and contains unique value which won't affect the prediction results, so drop the column names “msisdn” which is not necessary to evaluate the results. In data exploration, the distribution of individual predictions by churn will be determined, the plot demonstrates that clients are more likely to not churn and they appear lowest churning. Fig. 3 shows details of data pre-processing.

time_Over 2 Year	lifetime_below1Month	last_pk_Cards	last_pk_No	last_pk_Topup	msisdn	voice_rev	data_rev	total_rev	refill
1	0	0	0	1	0.548989	0.021061	0.042150	0.052433	-0.074955
1	0	0	1	0	0.548989	0.220823	-0.648801	-0.503132	-0.244928
1	0	1	0	0	0.548989	0.923881	-0.648801	-0.111042	-0.040961
1	0	0	1	0	0.548989	-0.460786	-0.648801	-0.883262	-0.244928
1	0	1	0	0	0.548990	0.871139	-0.648801	-0.140455	-0.040961

Figure 3 Details of data pre-processing.

Applying Machine Learning Algorithms: The dataset was split into the training and validation set. 80 % of the data was used for training and 20 % was used for validation. There are 80883 records in the training set and 20221 records in the validation set. The dataset was used to train three machine learning algorithms. The algorithms include logistic regression, decision tree and random forest. Logistic regression predicts the output based on probability. It uses the sigmoid function. The output lies between 0 and 1. If the threshold is 0.5 then values from 0 to 0.49 are false, while values between 0.5 to 1.0 are true. It is a linear classifier. The Logistic Regression class was imported from the linear model module of the Sci-kit Learn library. The maximum number of training iterations was set to 300 while penalty was set to ‘none’. The Decision Tree is a tree-structured classifier. The internal nodes represent the features of the dataset, the branches represent the decision rules, and each leaf node represents the outcome. I import the Decision Tree Classifier class from the tree module of the Sci-kit Learn library. The criterion is “entropy”. Random Forest is based on ensemble learning, which is a process of combining multiple classifiers to solve a complex problem. This algorithm uses multiple decision trees to make predictions. It takes less training time when compared to other algorithms. Random Forest classifier class was imported from the ensemble module of the Sci-kit Learn library.

Evaluation Models: In this research, certain performance metrics like accuracy, precision, recall and f1-score were used to determine the performance of each model. Accuracy can be calculated by dividing the total number of correct predictions by the total number of predictions and then multiply by 100. Precision is the ratio of true positives and total positives predicted. The recall metric focuses on type-II errors (false negative). Although it cannot measure the existence of type-I error which is false positives. And F1- score is the combination of precision and recall. The is the harmonic mean of Precision and Recall.

Results

We performed several experiments on the proposed churn model using machine learning algorithms on the dataset. From the result obtained in training and testing models in Figure 4, Decision Tree and Random Forest outperformed Logistic Regression with mean precision of 96 % and 97% respectively which are the ability to identify only customers that are about to not churn, recall (also known as sensitivity) of 96 % and 95 % which is the fraction of relevant instances that were retrieved by Decision Tree and Random Forest. F-scores also known as F-measure of 96 % is a measure of a model's accuracy on a testing dataset. However, the Logistic Regression displays lowest accuracy of 91% in model training, followed by 93 % accuracy for Decision Tree and Random Forest classifier.

train accuracy: 0.9135912367246517 test accuracy: 0.9063844518075268					
classification report for Logistic Regression					
	precision	recall	f1-score	support	
0	0.95	0.95	0.95	17594	
1	0.64	0.63	0.64	2627	
accuracy			0.91	20221	
macro avg	0.79	0.79	0.79	20221	
weighted avg	0.91	0.91	0.91	20221	

train accuracy: 0.9389612155830026 test accuracy: 0.9321002917758766					
classification report for Decision Tree					
	precision	recall	f1-score	support	
0	0.96	0.96	0.96	17594	
1	0.75	0.72	0.73	2627	
accuracy			0.93	20221	
macro avg	0.85	0.84	0.85	20221	
weighted avg	0.93	0.93	0.93	20221	

train accuracy: 0.936772869453408 test accuracy: 0.9300232431630483					
classification report for RandomForest					
	precision	recall	f1-score	support	
0	0.97	0.95	0.96	17594	
1	0.71	0.77	0.74	2627	
accuracy			0.93	20221	
macro avg	0.84	0.86	0.85	20221	
weighted avg	0.93	0.93	0.93	20221	

Figure 4 The result obtained in training and testing models.

The confusion matrix which shows the number of true positives, false positives, true negatives, and false negatives is shown in Figure 5. The logistic regression model correctly classified 1660 customers that churned. It misclassified 926 customers as churned, whereas the customers did not leave. It correctly classified 1668 customers as retained. It misclassified 967 customers as retained, whereas they actually churned. The Decision Tree model correctly classified 1901 customers that churned. It misclassified 647 customers as churned, whereas the customers did not leave. It correctly classified 16947 customers as retained. It misclassified 726 customers as retained, whereas they actually churned. The Random Forest model correctly classified 2020 customers that churned. It misclassified 808 customers as churned, whereas the customers did not leave. It correctly classified 16786 customers as retained. It misclassified 607 customers as retained, whereas they actually churned.

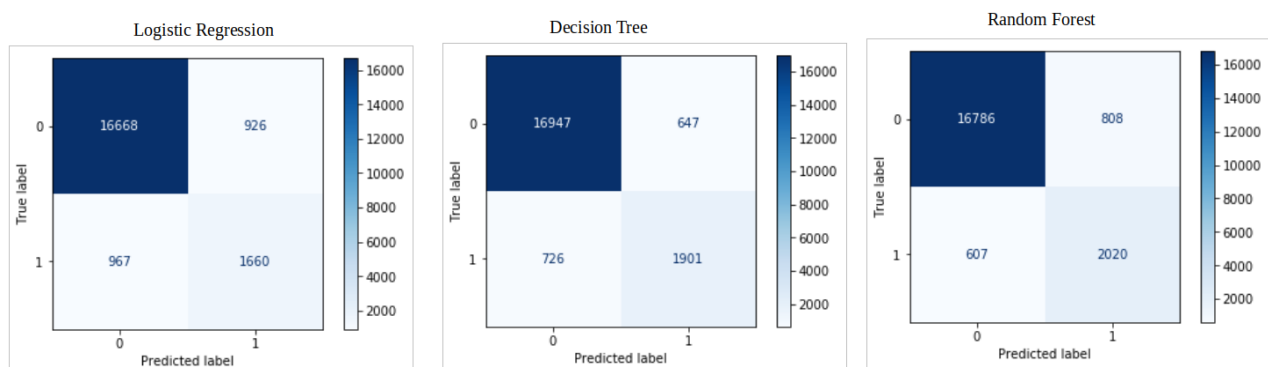


Figure 5 The testing results from the confusion matrix.

CONCLUSION

Algorithms of machine learning helps in predicting the consumer churn in the Lao Telecommunication Company, Borikhamxay Branch. Here we are using the dataset of 101104 customers to predict customer churn. This research makes use of logistic regression, decision tree and random forest for predicting the consumer churn in the Lao Telecommunication Company. From the findings of the result, it was found that accuracy rate of prediction in consumer churn is found to be 95%, 96% and 97% respectively. Further this research could be extended by adopting some algorithms to improve accuracy and prediction rate of churners in the Lao Telecommunication Company.

References

1. Ammara, and D. M. Linen, "A review and analysis of churn prediction methods for customer retention in telecom industries." In 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS), pp. 1-7. IEEE, 2017.
2. N. Scott, A. S. Gupta, W. Kamakura, J. Lu, and C. H. Mason, "Defection detection: Measuring and understanding the predictive accuracy of customer churn models." *Journal of marketing research*, vol. 43, 2006, pp. 204-211.
3. W. Xing, P. Li, M. Zhao, Y. Liu, R. G. Crespo, and E. Herrera-Viedma, "Customer churn prediction for web browsers." *Expert Systems with Applications*, 2022, vol. 209.
4. Wilcox, and Phill. "LAOS IN 2021." *Southeast Asian Affairs*, 2022, pp. 177-192.
5. Shirazi, Farid, and M. Mohammadi. "A big data analytics model for customer churn prediction in the retiree segment." *International Journal of Information Management*, 2019, vol. 48, p. 238-253.
6. Zdziebko, Tomasz, P. Sulikowski, W. Sałabun, M. Przybyła-Kasperek, and I. Bąk. "Optimizing Customer Retention in the Telecom Industry: A Fuzzy-Based Churn Modeling with Usage Data." *Electronics*, 2024, vol. 13.
7. Anawar, Syarulnaziah, N. F. Othman, S. R. Selamat, Z. Ayop, N.i Harum, and F. A. Rahim. "Security and Privacy Challenges of Big Data Adoption: A Qualitative Study in Telecommunication Industry." *International Journal of Interactive Mobile Technologies*, 2022, vol.16.
8. Huang, and L. Francis, "Alternatives to logistic regression models in experimental studies." *The Journal of Experimental Education*, 2022, vol. 90, pp. 213-228.
9. Custode, L. Leonardo, and G. Iacca. "Evolutionary learning of interpretable decision trees." *IEEE Access*, 2023, vol. 11, pp. 6169-6184.
10. Das, Sunanda, M. S. Imtiaz, N. H. Neom, N. Siddique, and H. Wang. "A hybrid approach for Bangla sign language recognition using deep transfer learning model with random forest classifier." *Expert Systems with Applications*, 2023, vol. 213.
11. Derczynski, and Leon. "Complementarity, F-score, and NLP Evaluation." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 261-266. 2016.