ISBN: 978-981-18-3959-7
2022 the 12th International Workshop on Computer Science and Engineering (WCSE 2022)
doi: 10.18178/wcse.2022.06.013

Research on Early Warning of User Churn in Social Platform based on Data Mining Technology

Wangyang Shi [1], Jonathan M. Caballero [2 +] and Jonan Rose Montana [2]

[1] Technological University of the Philippines Manila, Philippines
Anhui Technical College of Mechanical and Electrical Engineering Wuhu, China
[2] Technological University of the Philippines Manila, Philippines

Abstract. The management of user churn is essential content in the enterprise customer management of the social platform. By constructing the early warning model of user churn, we can predict the potential lost users so that enterprises can give early warning and take corresponding measures to retain users and reduce the cost of maintaining users, which has specific practical significance.

Research the application of data mining technology in predicting user churn in the social media industry. Using data mining technologies such as Logistic regression algorithm, Bayesian algorithm, and XGBoost algorithm, the churn model is constructed, and the models are compared and integrated so that the model with high prediction accuracy can be obtained. The potential churn users can be predicted, which is the key user group maintained by enterprises.

Keywords: social media users, Data mining, Loss model, voting fusion , Churn prediction

1. Introduction

Nowadays, mobile Internet development has entered a stable period, and the number of users of Internet social platforms has become saturated. Social platforms have gradually shifted from mining new users to maintaining existing old users to avoid customer churn.

The number of users is the basis for the survival of social platforms. Facing fierce market competition, maintaining the existing users is a significant problem that social platforms must face. Especially, users with many fans and high value are the fundamental objects of maintenance [1].

In the face of massive user data, digging out high-value users, focusing on maintenance, early warning, timely intervention, enhancing user viscosity, and reducing user churn has become a crucial part of enterprise customer management [2].

Using Bayesian algorithm, XGBoost, and Logistic in data mining technology to build a model can quickly and accurately understand the characteristics of user data. Especially after making a single model, using the Voting method to fuse and get the potential churn user model is a practical user churn prediction method.

2. Literature Review

There are many kinds of researches on the user churn and early warning of enterprises, including:

Duan Pei and other authors put forward that if users in the telecommunications industry have no call records this month or the consumption amount drops to a certain threshold, and they will be lost customers [3].

Qi et al. established an early warning model of user churn using alternative decision tree (ADTree) and Logistic regression and compared the results [4].

+ Corresponding author. Tel.:(+63 2)53013047
E-mail address: jonathan_caballero@tup.edu.ph.

Masand and other authors first get the order of variable importance through the genetic algorithm, then test and compare several prediction training models (decision tree, neural network, and K-nearest neighbor). Finally, get the user churn system model to predict the reasons for user churn [5].

Rosset and other authors realized the user churn prediction model and considered the problem of sample imbalance. Adjusting different weights reduced the local accuracy, and the prediction effect of the model was improved [6].

Karan et al. predicted the problem of user churn by using the recurrent neural network model of ERNN and JRNN [7].

Garcia, Dhar S, and other authors introduced cost-sensitive factors to improve neural network and support vector machine models to deal with unbalanced sample data [8] [9].

To sum up, the number of users is the basis for the survival of Internet social platforms and is an essential resource for enterprises. Through massive historical data, enterprise user management is necessary to mine users' usage habits, improve users' experience effect of using products, enhance users' satisfaction, give users more benefits, retain users, and reduce user churn.

3. Methods

3.1. Data Mining Technology

1) Overview: Data mining is to dig out the hidden helpful information through data cleaning, data integration, data conversion, and finally evaluate the excavated information and its effectiveness, and draw valuable conclusions [10], which is an essential means for the company to make decisions.
   Data mining can be generally divided into two types: predictive and descriptive data mining [11]. Through descriptive data mining, find out the potential relationship between data, and through predictive data mining, build the corresponding model, predict and infer, and predict the potential lost users.
2) Bayesian algorithm: Bayesian classification is the general name of a class of classification algorithms. Its method is simple, accurate, and fast. Bayesian theorem assumes that the influence of one attribute value on a given class is independent of the values of other attributes. Still, this assumption is often not valid in practical situations, so its classification accuracy may be reduced [12].
3) Logistic regression algorithm: Logistic regression algorithm is a probabilistic nonlinear regression model. A multivariable analysis method for studying binary output classification [13]. Through logistic regression, we can establish a relationship between the observation result Y of the dichotomy and some influencing factors [X1, X2, X3,...] to estimate and classify the probability of a specific result under certain elements and its effect is either 1 or 0.
4) XGBoost algorithm: XGBoost algorithm (eXtreme Gradient Boosting) is a gradient boosting algorithm based on a decision tree. XGBoost algorithm realizes the generation of the weak learner's loss function, uses the first derivative and second derivative value of the loss function, and dramatically improves the algorithm's performance by pre-sorting, weighted quantile, and other technologies [14].
5) Voting method: The voting method is based on several basic models and adopts a voting strategy to select the final classification with the most votes [15]. There are two methods of Voting: hard Voting and soft Voting. For the category predicted by the hard voting method, the category with the most significant number of votes is the final result of all the individual classifiers' predicted results. For example, for a social media user, using the Bayesian algorithm, Logistic regression algorithm, and XGBoost algorithm, the prediction results of user churn are respectively 0, 1, and 1(1 means churn, 0 will not churn), then the result after hard voting fusion is 1. Soft Voting is mainly to set weights for a single classifier and multiply them by the probability of a single classifier, and the highest classification probability is the final result. The formula is as follows:

$$M=n_1M_1+n_2 M_2 +\cdots+n_kM_k,\text{among:}n_1+n_2+\cdots+n_k=1 \quad (1)$$

6) Evaluation indicators:The evaluation indexes of the basic model mainly include:

a) Recall:

$$Recall=TP/(TP+FN) \quad (2)$$

b) Accuracy:

$$Accuracy=(TP+TN)/(TP+TN+FP+FN) \quad (3)$$

c) Precision:

$$Precision=TP/(TP+FP) \quad (4)$$

d) F1 value:

$$F1=(2*Precision*Recall)/(Precision+Recall) \quad (5)$$

Description: As shown in Table 1.

Table 1: Prediction matrix

Real sample	**Prediction result**	
	Positive sample	***Negative sample***
Positive sample	TP (True positive sample)	FN (False negative sample)
Negative sample	FP (False positive sample)	TN (True negative sample)

3.2. Characteristics of Lost Users

1) User age group: This paper analyzes and sorts out the data lost by social platform users in the recent two years. There are 153,221 lost users, including 83,245 lost users under 27, accounting for 54.33% of lost users. Most of these people are students who have just started work, and they are under tremendous pressure, poor stability, fast change of behavior, and relatively influenced by external factors. There are 43,562 lost users between 28 and 45 years old, accounting for 28.43% of lost users. Between 46 and 60 years old, 12,357 users lost, accounting for 8.06%, as shown in Fig.1.

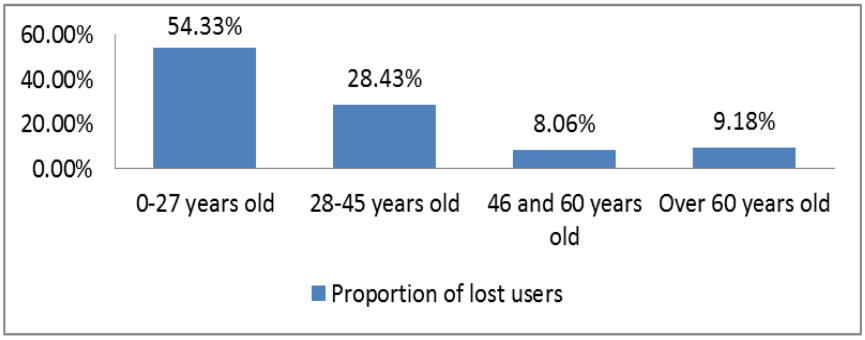

Fig. 1: Proportion of user churn in different age groups.

2) User's area: According to users' frequent usage locations, it is found that 66,783 users are lost in first-tier cities, 43,478 users in second-tier cities, 23,479 users in third-tier cities, and others are users in cities below the fourth tier. It follows that first-line and second-line users are the key maintenance groups, and the specific proportion is shown in Fig.2.

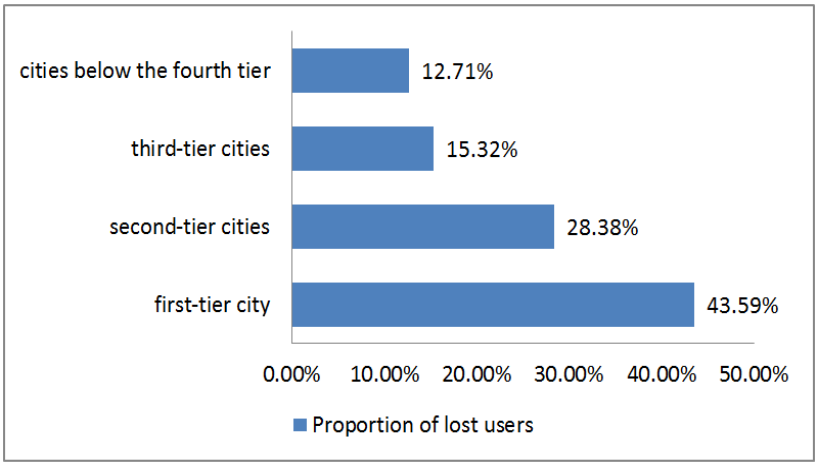

Fig. 2: Proportion of user churn in different regions.

3) User's daily activity: According to the analysis of users' daily activity, 76,548 users use the platform for less than one hour a day, 45,689 users who use the platform for 1 hour to 2 hours, and 28,654 users who use the platform for 2 hours to 3 hours a day. The number of users who use the platform for more than 3 hours is minimal, and the user activity is high, as shown in Fig.3.

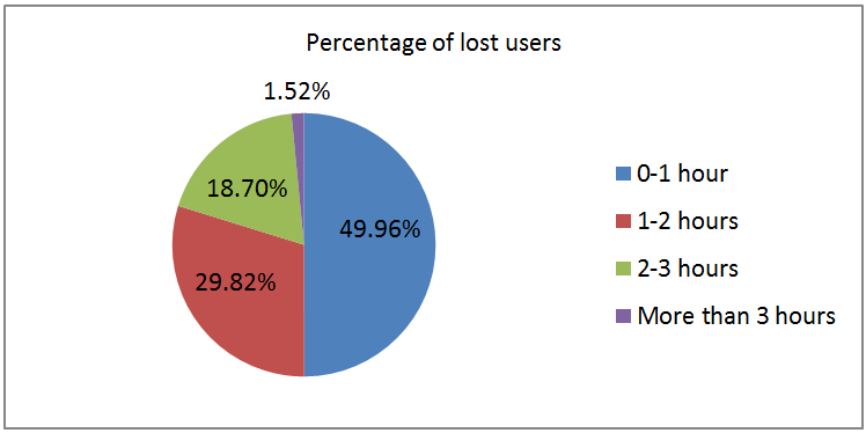

Fig. 3: Daily activity of users, the proportion of user churn.

3.3. Data Generation

This paper analyzes and sorts out the lost data of a social platform in recent two years and draws 30,000 people from lost users as positive samples and 30,000 people from non-lost users as negative samples, which constitute training samples. Through data cleaning, discretization, standardization, and other data preprocessing and feature field processing, about 31 fields are selected, including user basic information, area information, user activity information, etc. Some fields include age, gender, user level, area, online duration, traffic, etc.

4. Results And Discussion

4.1. Comparison of Basic Model Prediction

In building the user churn model, Table 2 shows the prediction effect of each basic model. Seen from the Table that the accuracy rate of the Bayesian model is the lowest 76.37% and that of the XGBoost model is the highest 90.97%, but the accuracy rates of these three models are all above 75%. The Logistic and XGBoost model's recall rate is above 89%, and the prediction effect is good. From the perspective of precision, the XGBoost model has the best prediction effect, reaching more than 92%.

Table 2: Comparison of basic model prediction and evaluation

Basic model	Accuracy	Recall	Precision	F1 value
Naive Bayesian model	76.37%	84.32%	72.76%	78.11%

Logistic regression	87.95%	89.86%	86.56%	88.18%
XGBoost model	90.97%	89.34%	92.35%	90.82%

4.2. Comparison of Voting Models

The three basic models are fused by hard Voting. The most obvious improvement of the prediction accuracy of the merged model is the combination of the Logistic model and XGBoost model, with an accuracy rate of 91.43%, which is 0.46 percentage points higher than that of the single XGBoost model. However, it is found that not any combination can improve the prediction effect.

The three basic models are fused by soft Voting. Similarly, the combination of the Logistic model and XGBoost model improves the prediction accuracy of the merged model most obviously, with the model accuracy increased by 0.69 percentage points. Compared with the model fusion by hard Voting, the prediction effect is more robust.

4.3. Analysis of Forecast Results

Two voting methods fuse the basic model, and the prediction model is better than the single model. Specifically, from the perspective of precision, hard Voting is better than soft Voting. Still, the soft voting fusion model is more effective from the statistical data of accuracy, recall, and F1 value. The specific evaluation indicators are shown in Table 3.

Table 3: Comparison of prediction and evaluation of fusion models

Fusion model	Accuracy	Recall	Precision	F1 value
Hard Voting	91.43%	90.10%	92.56%	91.31%
Soft Voting	91.66%	90.96%	92.25%	91.60%

From comparing evaluation indexes of prediction results, the fusion model under the soft voting mode has a better prediction effect, and it can be used to predict potential lost users.

4.4. Identification of Significant Value Customers

The goal of building a user churn early warning model is not to maintain all the users who are about to leave the platform but to maintain the customers whose activity is in the top 30% and user traffic is in the top 40% of the platform. These customers are of great value to the platform. Maintain these customers in advance, maintain customers and promote business according to different customers' use of the platform, improve customer stickiness, reduce churn rate, and reduce the loss of income and scale of enterprises caused by user churn in social platforms. According to the user churn warning model, there are 7,675 potential churn users, of which about 2,469 are valuable users of the platform. Enterprises mainly focus on maintaining these key customers, which greatly reduces the workload of platform customer management.

5. Conclusion

User churn warning is an essential task of social platform enterprise customer relationships. In this paper, three basic models of Bayesian, XGBoost, and Logistic are constructed, and the Voting fusion method is proposed to fuse different basic models to improve the accuracy of user churn prediction. The results indicate that the prediction model obtained by the Voting method has an obvious advantage of a more proper method compared with the single model, and the prediction model constructed by the soft Voting method is the most satisfactory.

In this paper, while predicting the potential loss of users, we can find the users who are of great value to the platform, carry out key maintenance, reduce the workload of enterprise users' maintenance, and reduce the operating costs of enterprises.

In this paper, only two basic models are fused, and three or more models are not considered for fusion testing. In the next step, we are going to make deeper research on the fusion model composed of more than three basic models, to obtain a more stable fusion model with better prediction performance.

6. References

[1] Ding Naipeng, Duan Min. Overview of the development of customer relationship management [J]. Economic Jingwei, 2005.

[2] Shao Bingjia. Customer Relationship Management (Second Edition) [M]. Beijing: Tsinghua University Publishing House, 2010,129-132.

[3] Duan Pei. Research on the construction and application of telecom customer churn early warning model [D]. Master's thesis, Zhejiang Gongshang University, 2015

[4] Qi J, Zhang Y, Zhang Y, et al. TreeLogit Model for Customer Churn Prediction[C].Proceedings of The 1st IEEE Asia-Pacific Services Computing Conference, APSCC 2006, December 12-15, 2006, Guangzhou, China. IEEE, 2006

[5] Masand B, Datta P, Mani D R, et al. CHAMP: A Prototype for Automated Cellular Churn Prediction[J]. Data Mining & Knowledge Discovery, 1999, 3(2):219-225.

[6] Rosset S, Neumann E . Integrating customer value considerations into predictive modeling[C]//Data Mining, 2003. ICDM 2003. Third IEEE International Conference on.IEEE, 2003.

[7] Kasiran Z . Customer Churn Prediction using Recurrent Neural Network with Reinforcement Learning Algorithm in Mobile Phone Users[J]. International Journal of Intelligent Information Processing, 2014.

[8] Garcia, Salvador, Herrera, et al. Cost-Sensitive back-propagation neural networks with binarization techniques address multi-class problems and non-competent classifiers[J]. Applied Soft Computing, 2017.

[9] Dhar S , Cherkassky V . Cost-Sensitive Universum-SVM. International Conference on Machine Learning & Applications, 2017.

[10] Jia Wei Han, Michelin Kamber, Jian Pei, et al. Concept and technology of data mining [M]. Machinery Industry Press, 2012.

[11] Li Deren, Wang Shuliang, Shi Wenzhong, et al. On spatial data mining and knowledge discovery [J]. Journal of Wuhan University: Information Science Edition, 2001, 26(006):491-499.

[12] Xiao Sa. Research on spam filtering system based on Naive Bayes algorithm [D]. Master's thesis, Huazhong University of Science and Technology, 2016.

[13] Liu Lizhi, Deng Jieyi and Wu Yuntao. Research on multi-classification Logistic algorithm based on HBase [J]. Research on Computer Application, 2018, 35(010):3007-3010.

[14] You Dechuang, Mo Zan. Research on bank credit evaluation based on fuzzy XGBoost algorithm [J]. Information and Communication, 2018(002):37-38.

[15] Jin Guopeng. Forecasting scheme planning of Shanghai Composite Index based on Voting fusion algorithm [D]. Master's thesis, Shanghai Normal University, 2019.