Customer churn prediction in telecom using big data analytics

Weilong Li¹*, Chujin Zhou²

¹School of Computer Science and Communication Engineering, Guangxi University of Science and Technology, Liuzhou, 545006, Guangxi, China
²School of International Education, Guangxi University of Science and Technology, Liuzhou, 545006, Guangxi, China

*Corresponding author e-mail: gxlwl@gxust.edu.cn

Abstract. Customer churn will cause huge losses to the communication company and has become a real problem. The article uses big data analysis technology to analyse user characteristics of churn customer historical information data, establish a churn prediction model, find users with a higher risk of churn in advance, develop targeted strategies, and carry out a series of retention activities to retrieve them. The paper presents a strategy of user segmentation and piecewise regression to find the highly relevant fields and divide the customers into different groups based on these fields, and then use regression analysis to establish the prediction models for different groups. Online test shows that the model can effectively identify most of the lost customers, effectively reduce the user off-network rate, and improve efficiency and effectiveness than traditional methods.

1. Introduction

With the advent of the mobile Internet era, customers have increasingly demand in online experiences, and at the same time, more and more application providers can be selected by customers. For communication companies, it is very important to protect the individual needs of all customers, especially high-value customers, and maintain a relatively stable customer base. If the telecommunications company don't pay attention to customer needs and customer relationship maintenance, some dissatisfied customers will be lost from the company. Managers want to know which users may be lost and when they will be lost. It is very necessary for the enterprise to establish a churn prediction model based on analyzing historical data and current data, and extract key data to assist decision-making, and find hidden relationships and patterns from them to predict the future. The enterprises can take measures in advance to do customer care and reduce customer churn. The use of big data analysis technology for data mining, predicting customer churn and reducing the occurrence of customer churn has become the focus of research in the telecommunications industries [1]. Studies [2] showed that the cost of retaining existing customers is much lower than acquiring new customers.

Many studies[1] have confirmed that based on the historical behavior data of customers, such as customer data, usage behavior, consumption behavior, online trajectory and other useful features to extract and combine features to accurately determine the status or tendency of customer churn, can allow companies targeting customer retention is one of the effective ways to resolve customer churn, and it is extremely important for the company's long-term development.
2. Related work

Many methods were studied to predict churn in telecom companies. Most of them used some machine learning arithmetic or data mining arithmetic, focused on applying only one method to extract feature, and the others focused on comparing several different arithmetic to predict customer churn.

Samira et al.[3] presented a predictive framework of customer churn through six stages for accurate prediction and preventing customer churn, extracted and selected the important and effective variables influencing customer behaviours, used the discriminant analysis method to predict the classes of the customers. The AUC value was over 90%, but the model needed training each period of time.

Ammar et al.[4] studied a metaheuristic-based churn prediction technique that performs churn prediction on huge telecom data, used a hybridized form of Firefly algorithm as the classifier, and obtained effective and faster results.

Arno et al.[5] presented a new hybrid algorithm named the logit leaf model (LLM) to classify data. The model consisted of a segmentation phase and a prediction phase, outperformed the decision trees or logistic regression algorithm.

Vijaya et al.[6] used Rough set theory to identify the features of telecommunication customer for churn prediction, and gotten classification accuracy of 95.13% compared to any single model.

Adnan et al.[7] presented the concept of classifier's certainty estimation using distance factor, and divided the dataset into different zones according to high or low certainty, and predicted customers exhibiting Churn and Non-churn behaviour.

Ahmad et al.[8] developed a churn prediction model which using machine learning techniques on big data platform, and built a new method of features selection.

However, similar problems are often encountered in real situations, such as how to select the most vulnerable customer groups from a large number of customers and carry out customer retention actions? How can they find as many vulnerable customers as possible When carrying out customer retention work? How to determine the minimum number of selected people to reduce costs? How to use the churn model to design different priorities and strategies for the users which have different levels probabilities of churn?

The article presents a method based on big data analyze to study the historical behavior data of lost customers to find out the key characteristics of these customers, so as to achieve the purpose of "early detection and early retention". Based on big data analysis, a high-risk churn customer early warning model is established, and the forecast threshold can be adaptively adjusted according to the needs of various departments, and the hierarchical and early warning rules with different forecast sizes can be output, which can provide a reference for subsequent customer retention. By establishing a hierarchical customer retention strategy, in practice, different retention strategies can be selected according to the status of different resources, which can significantly improve the efficiency of customer retention, reduce the cost of customer retention, and reduce interruptions to customers.

3. Customer Church Model Construction

The problem of customer churn can be generally divided into two aspects: (1) which customers (especially high-value customers) are expected to churn? (2) what are the characteristics of possible churn customers? The customer churn model is established mainly based on the characteristics of historical data and applied to predict users with a high probability of going offline. The process includes three parts: data preparation, model training and verification, and off-grid prediction.

The data used in this study came from a communications company's 1.5 million customers from March to September (including churn and non-churn users) , and divided into three parts: 1. training set data; 2. test set data; 3. test set prediction, the training set data (about 1.4 million, the last field is churn or not) is used to mine the rules of customer churn and output the relevant model. The model is used to predict the test set data (about 100,000) to test the accuracy of the model. Finally, the model is applied to the online actual system for analyzing and querying customers with a tendency to churn.
3.1. Data preparation and pre-processing

The customer's data information is derived from the database, which include 437 fields, and can be divided into two types, one is the basic information contain 11 fields, including the customer serial number, age, network access time and channel, broadband and Tariffs take effect, etc. The other is customer behavior data have 426 fields, which contains 6 months of customer behavior information, and each month's customer behavior information involves 71 fields, such as, consumer behavior, terminal bundling, deposit and other promotions Activity participation, call behavior, SMS sending and receiving behavior, traffic usage behavior, resident communication area, channel preference, complaint behavior, broadband subscription information, etc.

What variables can be used to predict churn? Customer basic information data: including customer's age, gender, network access time, etc. Customer behavior data: mainly data on customers' use of telecommunications products and services. For example, the customer's detailed call records, customer subscriptions, usage, and unsubscribed value-added services. Customer interaction data: Includes data such as customer complaints, business inquiries, and customer responses to telecommunications marketing activities, especially customer calls.

Through observation and analysis, it is found that some data information is missing. The fields of missing values are mainly concentrated on fields such as broadband activation or invalidation, terminal bundling, and arrears. When processing such numerical missing values, they are filled with a value of 0. At the same time, obviously abnormal data is removed to reduce noise interference.

If no transformation is performed on some user data fields, the predicted result will have a large error because of the random of customer's behavior. Through experimental comparisons of logarithmic, reciprocal, square root and other transformation methods, the results are found that logarithmic is beneficial to make the data into the normal distribution. In order to improve the accuracy of model prediction and reduce the value of prediction error, the target variable used for modeling should preferably conform to the normal distribution.

3.2. The customer churn model Construction

To test current data and predict future data based on historical data, various algorithm models can be used, such as SVM algorithm, random forest, linear regression, logistic regression, etc[1]. These models have their own characteristics and deficiencies. Considering that the random forest algorithm has the advantages of high accuracy, good robustness, easy to use, etc., and it is fast in training and versatile, it can be applied to many types of model tasks. The reason is that as the model scale increases, the processing speed will be significantly slowed down, and the other is that it is easy to overfit when the data is relatively noisy. Need to carry out cluster analysis based on the results later to find the key features. We use the random forest algorithm to build a customer churn model.

The general customer churn model is difficult to take into account such a large number of random customers. Seizing the characteristics of similar behaviors and similar behaviors, the target variable is transformed into a normal distribution and segmented to improve the accuracy of prediction. Customer segmentation, which reduces model parameter requirements and improves accuracy. Experiments have been performed on multiple segmentation methods, such as 10-segment, 20-segment, and 50-segment. Finally, considering performance and efficiency, 10 segments were used. For the segment after the normal distribution of the target variable, the two sides belong to extreme populations is small; the middle belongs to normal populations, and the scale is relatively concentrated.

In order to explain the model, some rules based on churn models must be outputted, such as:
"Age is larger than 60 or less than 18"
"March DOU greater than 1G & May busy traffic less than 500M & Online duration less than 20 & Campus users"

"Users with an average number of SMS messages <10 & total complaints greater than 5"

To get a more accurate churn model, some measures can also be considered:
(1) Reasonably integrate and use the given data. Considering that the existing data labels are related to some fields of the previous two months of data, directly modeling the data in July and August may cause those fields to be too relevant and cover up some features. You can combine the March, April, and May fields into one data set as the independent variable, and August churn as the dependent
variable, and filter out some data. The training set and the test set both do the same processing before modeling.

(2) The interpretability and operability of the model is to output rules, which is convenient for implementation in practical applications, but it may not necessarily get the ideal ROC curve.

(3) The feature field with obvious effects is used in the training set to establish a churn model. The training results are very good, but when tested on the test set, the effect is not very good, so the previous model may be overfitted. We use the correlation degree of the feature field to filter out the more relevant fields, and then apply the decision tree and voting mechanism to build a customer churn model to improve model accuracy, as shown in table 1.

| Field Name     | Correlation | Field Name     | Correlation |
|----------------|-------------|----------------|-------------|
| DOU_201705     | .716        | GPRS_XS_M_4    | .521        |
| DATA_4G_5      | .704        | GPRS_ARPU_5    | .516        |
| GPRS_MS_M_5    | .687        | GPRS_XS_M_3    | .486        |
| GPRS_BT_M_5    | .671        | IF_APP_PRE_5C  | .482        |
| GPRS_YW_M_5    | .644        | DOU_201703     | .468        |
| DATA_4G_4      | .632        | GPRS_ARPU_4    | .455        |
| GPRS_MS_M_4    | .617        | IF_APP_PRE_4C  | .452        |
| GPRS_BT_M_4    | .605        | ARPU_201705    | .448        |
| DATA_4G_3      | .596        | GPRS_ARPU_3    | .431        |
| DOU_201704     | .585        | IF_APP_PRE_3C  | .429        |
| GPRS_MS_M_3    | .585        | ARPU_201704    | .407        |
| GPRS_XS_M_5    | .583        | ARPU_201703    | .394        |
| GPRS_BT_M_3    | .574        | TAOCANWAI_MB_5 | .318        |
| GPRS_YW_M_4    | .571        | CHAOTAOCAN_ARPU_5 | .301  |
| GPRS_YW_M_3    | .535        | BHDU_GPRS_5    | .283        |

3.3. The customer churn model verification

For the customer churn model, there are generally four prediction results: (1) Churn customers are accurately hit by the model (referred to as hits); (2) Churn customers are incorrectly predicted by the model as non-churn (referred to as missed judgments); (3) Non-churn customers Hit accurately by the model; 4. Non-churn customers were wrongly predicted by the model as churn (referred to as misjudgment). Omission and misjudgment are two types of errors, which will have different degrees of negative impact on actual work. A good prediction model should maximize hits while reducing missed judgments and misjudgments.

An excellent customer churn early warning model should be able to adaptively achieve a compromise between the maintenance cost and the maintenance success rate, and output the optimal retention strategy based on the actual needs and constraints of the marketing. Automation must also allow marketing planners to intervene manually in conjunction with actual marketing needs.

There are often two types of results obtained through churn analysis through models: one is the description of the characteristics of the churn customers, and the other is the churn score for each customer. The characterization of churn customers can be used as a reference for marketing department business personnel to formulate retention marketing strategies, and to develop targeted retention strategies; while churn scores combined with other variables (such as customer value) can help business personnel decide which Customer retention.

From the information obtained from the user churn model, we can roughly outline the characteristics of users that are easy to churn:

(1) Older users and young people who are not married and economically independent are more likely to lose.

(2) The various network service items provided can reduce the user churn rate.

(3) The longer the contract is signed, the higher the retention rate of users.
Therefore, we can make targeted recommendations:

It is recommended that elderly users and young users use digital networks and sign a two-year or longer contract. If the relevant network services can be provided, user loyalty can be increased, so some new services can increase user loyalty and reduce churn.

4. Results

Model evaluation generally includes accuracy rate and coverage rate. The higher the accuracy rate and the greater the coverage rate, the better the model's effect. Among them: accuracy rate = the number of customers predicted to churn accurately / the number of customers predicted to churn; coverage = accuracy predicted of customers / actual customers lost.

We select the data from March to June 2017 for training and perform model prediction on the data from July to September. As shown in Figure 1, after the verification of the data from October 2017 to December 2017, we can get the prediction from July to September. Data, the precision rate is basically 80%, and the recall rate is 40%.

![Figure 1. The evaluation result of model prediction](image)

Through effective maintenance methods, it can accurately maintain the predicted churn users, reduce the number of users leaving the network, and increase the value of online users. The model was also evaluated using the new data set and tested the impact of the system on customer churn decisions. The model produced good results and was deployed into online system.

According to the in-depth mining of customer churn behavior, output the churn probability and churn characteristics of each customer, the company can formulate a hierarchical customer retention strategy, which is of great significance to the retention and value improvement of the company's existing customers. Compared with traditional carpet or empirical customer retention strategies, hierarchical customer retention strategies can significantly reduce the cost of customer retention and improve the efficiency of customer retention.

5. Conclusion

We propose a modeling idea of user segmentation and piecewise regression, find the fields that is highly relevant to customer traffic and consumption behavior, and use clustering methods to divide the total customers into different groups based on these fields, and then apply regression analysis to establish personalized regression prediction models for different groups of customers. In order to further improve the prediction accuracy, we analyze the regularity of customer information and turn the linear segmentation of the customer's original variable into a normalized transformation. The model was also evaluated using the new data set and tested the impact of the system on customer churn decisions. The model produced good results and was deployed into production.

6. Acknowledgments

This work was financially supported by the PhD Fund Project of Guangxi University of Science and Technology(16Z04).
References
[1] F. Kayaalp, Review of Customer Churn Analysis Studies in Telecommunications Industry, Karaelmas Science and Engineering Journal. 7(2017) 696-705.
[2] S.A. Qureshii, A.S. Rehman, A.M. Qamar, A. Kamal, A. Rehman. Telecommunication subscribers’ churn prediction model using machine learning, Eighth international conference on digital information management. 2013, pp. 131–6.
[3] K. Samira, Z.R. Mahmoud, Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior, Journal of Systems and Information Technology.19(2017)65-93.
[4] A.Q.A. Ammar, D. Maheswari, Churn prediction on huge telecom data using hybrid firefly based classification, Egyptian Informatics Journal. 18 (2017) 215–220.
[5] D. C. Arno, C. Kristof, W. D. B. Koen, A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees, European Journal of Operational Research. 269 (2018)760–772.
[6] J. Vijaya, E. Sivasankar, Computing efficient features using rough set theory combined with ensemble classification techniques to improve the customer churn prediction in telecommunication sector. Computing. 100(2018)839–860.
[7] A. Adnan, A.O. Feras, S. Babar, A. Awais, L. Jonathan, A. Sajid, Customer churn prediction in telecommunication industry using data certainty, Journal of Business Research. 94 (2019) 290–301.
[8] A.K. Ahmad, A. Jafar, K. Aljoumaa, Customer churn prediction in telecom using machine learning in big data platform, J Big Data. 6 (2019) 28.