Data Visualization and Prediction for Telecom Customer Churn

Pulin Yang *

Software Engineering, Guangxi University of Science and Technology, Guangxi, China

* Corresponding author email: p.yang.12@student.scu.edu.au

Abstract. With the deepening of telecom industry reform and the intensification of competition, the customer churn rate of telecom enterprises is gradually increasing. How to predict and effectively reduce customer churn is directly related to the survival and development of telecom enterprises. In order to effectively deal with unbalanced classification and improve the accuracy of high-value customer churn prediction in telecom industry, this paper uses telecom customer data set from kaggle platform to analyze people's use of telecom services, and help telecom operators find out the reasons for customer churn, and establish churn prediction model to reduce customer churn rate. In this paper, firstly, the data set is imported, and then the data visualization analysis is carried out. Then, the random forest model, SVM model and GBDT model are introduced for comparison. Experiments show that random forest has better classification performance than other methods, and improves the accuracy of high-value customer churn prediction.

Keywords: Random Forest; Support Vector Machine; Visual Analysis; Customer Loss Forecast.

1. Introduction

Telecommunications is a common customer service in the life. In modern society, anyone who makes phone call or watches TV at home must rely on the services provided by telecom operators, such as the telephone and Internet. Customer churn refers to customers who used products or services before, but stopped using them due to various reasons such as losing interest. Telecom companies, Internet service providers and insurance companies usually take customer churn analysis and customer churn rate as one of their key business indicators, because the cost of retaining old customers is far lower than the cost of acquiring new customers. Studies have shown that if a company's turnover rate is reduced by 5%, its profits will increase by 25% ~ 85% [1]. Generally speaking, the cost for a company to acquire a new user is relatively high, but sometimes it only takes a phone call or strategy to keep an existing user. At present, the existing customers in the communication industry have become the core resources of the company. How to establish a data mining model of customer churn warning, it is particularly important to accurately locate the customers with churn tendency. Moreover, in this era of fierce competition in the communication industry and basically saturated number of users, it is easier to maintain old users than to acquire new ones.

Researchers have been investigating the churn forecast over the past ten years. In this case, machine learning algorithms can be considered as an effective method due to their strong performance in many tasks [2-5]. For instance, a PCA-SVM technique is suggested for predicting telecom customer attrition based on Principal Component Analysis (PCA) and Support Vector Machine (SVM) [6]. And the improvement coefficient of accuracy is much higher than that of other prediction methods. In [7], in order to improve prediction accuracy, a support vector machine model based on cost-sensitive learning is proposed to overcome the shortcomings of SVM in dealing with unbalanced samples. Using cost-sensitive learning, this model adopts different penalty coefficients for unbalanced sample sets. Compared with standard support vector machine and neural network, the results show that this model improves the prediction accuracy, effectively solves the problem of data set imbalance, and is an effective method to predict telecom customer churn.

2. Dataset Description and Preprocessing

2.1 Dataset Description

The dataset used in this study is collected from the Kaggle [8], the data set has 21 features and 7043 items. Each item represents a customer. The detailed description for every feature can be found in Table 1.

Table 1. The detailed information about the features on the collected dataset.

No.	Field name	Data type	describe	Classification of fields
1	customer	numerical	Customer ID	Basic Information
2	gender	text	Gender (male, female)	
3	SeniorCitizen	numerical	Whether the elderly (0,1)	
4	Partner	text	Is there a partner (Yes, No)	
5	Dependents	text	Do you have family members (Yes, No)	
6	tenure	numerical	Number of registered months for customers	Service Information
7	PhoneService	text	Is the telephone service enabled (Yes,No)	
8	MultipleLines	text	Is the multi-line service (Yes, No, No phone service) enabled?	
9	InternetService	text	Is the network service (DSL. Fiber optic,No) enabled?	
10	OnlineSecurity	text	Is network security service (Yes,No, No internet service) enabled?	
11	OnilneBackup	text	Open online backup service (es.No.No internet service)	
12	DeviceProtection	text	Open online protection service (es.No.No internet service)	
13	TechSupport	text	Open technical support service (Yes, No, No internet service)	
14	StreamTV	text	Turn on streaming TV (Yes,No, No internet service)	
15	StreamMovies	text	Open streaming movie (Yes,No, No internet service)	
16	Contract	text	Term of contract (Month-to-month, One year,Two year)	Paid Information
17	paperlessBilling	text	Is there a paperless bill (Yes,No)	
18	PaymentMethod	text	Payment method (electronic check, mailedcheck, bank transfer (automatic), credit card (automatic))	
19	MontalCharges	numerical	Monthly consumption amount	

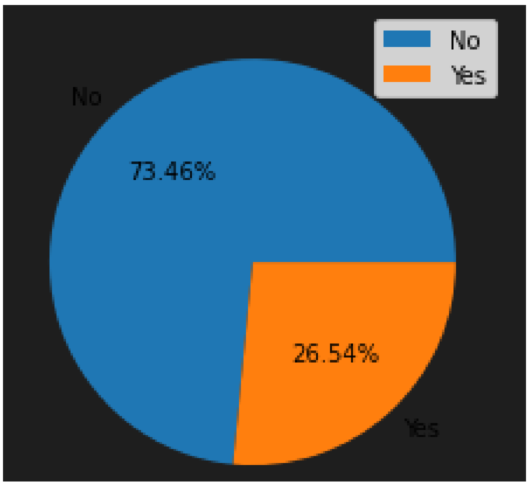

Fig 1. Total loss rate in the dataset.

By understanding the features, the features can be divided into three categories, including the basic information of customers, the opened service information and the payment information. In addition, the label is to determine whether the customer will be lost. In this data set, a total of 26.43% customers lost, that is, the total loss rate was 26%. The basic information includes gender, age (according to whether it is elderly or not), whether there are partners (which can be understood as working partners or friends) and family members. Observing service information can be divided into two basic services: telephone+network services. These two basic services have their own value-added services, such as multi-line telephone, network security, online backup, etc. In addition, there are payment information such as payment method and consumption amount.

2.2 Data Cleaning

After checking the data type, it is found that the data type of TotalCharges is Object, which means that this field contains an empty string. Therefore, the null value is removed firstly so that the model can accept the input data normally.

3. Visual Analysis

3.1 Correlation between Variables

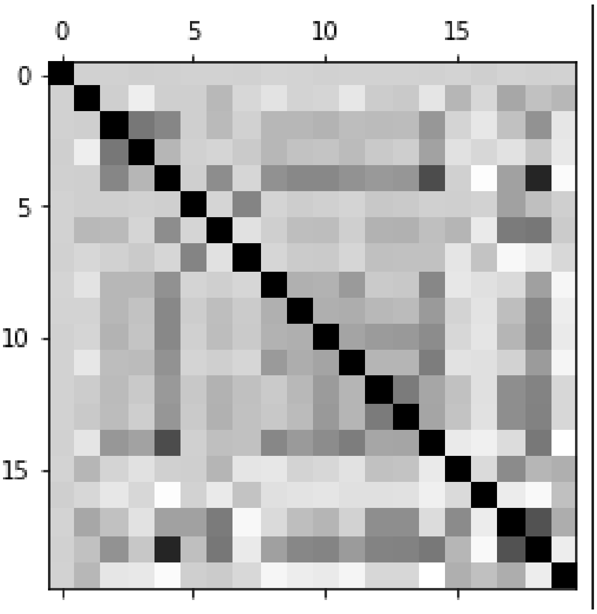

Fig 2. Variables to establish correlation matrix

Observe the correlation between variables to establish correlation matrix. Where the matrix is bright, it shows strong correlation, which is positive correlation. Internet services, network security, online backup, equipment maintenance services, technical support services, Internet TV services, and online movies are highly correlated, and they are positive correlation.

3.2 Frequency Analysis

From the above frequency distribution diagram, it can be found that: single people (partner), people without financial resources (dependents), people with Internet service = fiber optic, people with PaymentMethod=Electronic check, people without OnlineBackup, people with monthly contract, people with paperless billing = yes, and people with DeviceProtection=No are more likely to become lost customers.

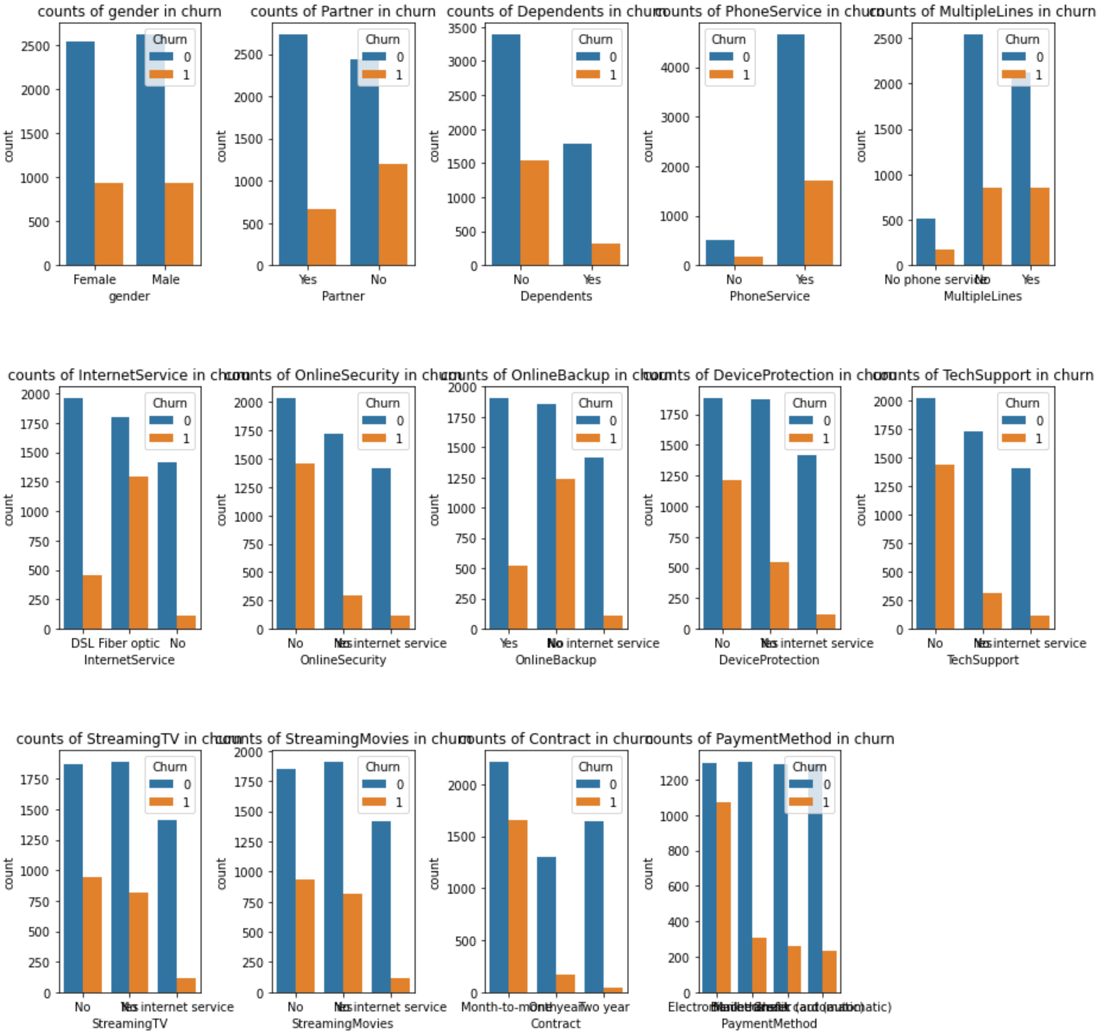

Fig 3. Frequency distribution diagram

4. Machine Learning Models

4.1 Support Vector Machine

Using binary classification, Support Vector Machines (SVM) is a model. It differs from perceptron in that its fundamental model is a linear classifier with the greatest interval determined in the feature space. SVM is a significantly nonlinear classifier since it also incorporates kernel skills. The interval maximization learning technique of SVM is comparable to the minimization of the regularized hinge loss function and is formalized as a convex quadratic programming problem. Convex quadratic programming is solved using the optimization procedure of the SVM. The core ideas of SVM algorithm are as follows: 1. If the data are linearly separable, the hyperplane is determined based on the maximum interval to ensure the global optimization and make the classifier as robust as possible; 2. If the data is linearly separable, transform the low-dimensional samples into high-dimensional samples by kernel function to make them linearly separable [6].

4.2 Random Forest

A machine learning algorithm is random forest. Ensemble Learning is a broad subset of machine learning. Combining different classifiers to create an ensemble classifier with greater prediction accuracy is the fundamental concept behind ensemble learning. The three main categories of integration algorithms are bagging, boosting, and stacking. It is possible to use random forest for both classification and regression problems. The classification task is to predict discrete values; The regression task is to predict the continuous values [9]. The forest adopts the idea of Bagging randomly, and the so-called Bagging is:

(1) Taking out n training samples from a training set every time, and forming a new training set;

(2) M sub-models are trained by using the new training set;

(3) Voting is employed for the classification problem, and the classification category of the sub model that receives the most votes become the final category. A simple average approach is used for the regression problem to get the projected value. A random forest is created by combining several decision trees, using the decision tree as its fundamental unit [10].

5. Result and Discussion

After analyzing the four dimensions of lease term, total charge, contract and monthly charge respectively, the following conclusions are drawn.

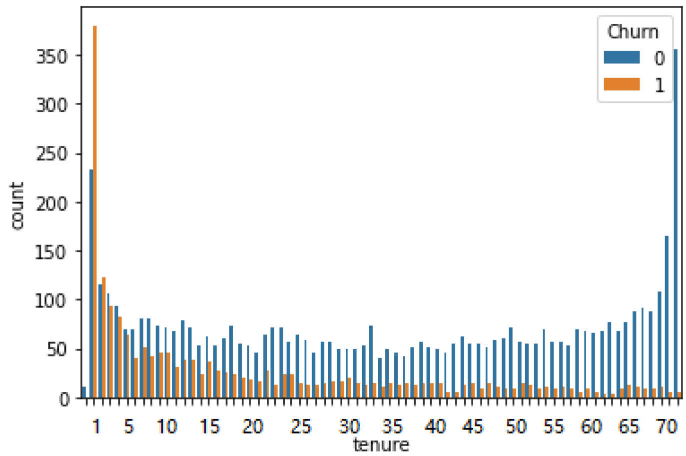

Fig 4. Visualization of drain ratio

As can be seen from the figure, the loss ratio of customers 1-5 is large, so more attention should be given to some customers 1-5. In terms of the experimental results, it can be found that the lost customers have low TotalCharge value, and more attention should be paid to the customers with low TotalCharge value. In addition, as can be seen from the Figure 4, Month-to-Month has a large customer turnover, so w more attention should be also given to it. Besides, MonthlyCharges' larger customers are easy to lose.

Table 2. Model comparison data

	Rate of accuracy	Rate of recall	precision
svm	0.614925	0.536458	0.782115
Random forest	0.60745	0.552083	0.780696
GBDT	0.669967	0.542781	0.807665

Comparing the three models, it can be found the of random forest is slightly higher and the recall rate is higher. So, the model is better.

Portrait of customers with high churn rate: single people, people with no financial resources, people with InternetService=Fiber optic, people with PaymentMethod=Electronic check, people without OnlineBackup, and people with monthly contract. People with PaperlessBilling=Yes and DeviceProtection=No are more likely to become churn customers.

Lonely people, who should be singled out as a priority customer, had an attrition rate of 50 percent, nearly double the total attrition rate of 27 percent. Such customers have high value and good quality, and their monthly consumption per capita is 78-yuan, accounting for 120% of the total monthly consumption per capita; Among the lost users, the average life expectancy is 16 months, which is 89% of the total average life expectancy of 18 months.

Fiber-optic network service is the one with the largest number of users (44%) and the highest churn rate (42%). The main reason is that the package price is expensive. 56% of the customers who have opened this service have also opened streaming movie & TV service, and the churn rate of this kind of people is also high (30%).

In view of this situation, this paper puts forward the following suggestions:1, for the elderly, no partner, no family customers, regularly launch telephone return visits, gratitude feedback, free lottery, gift cards and other caring activities, strengthen the contact between customers and products; 2. Using the churn prediction model, select a group of customers with high churn rate, launch churn prevention products through questionnaire survey, conduct AB test, and continue iterative optimization. 3. Recall lost customers through system push, SMS, etc. For returning customers, we will stimulate users' consumption by giving away the talking time, watching movies for free, and returning the package to the preferential price.

6. Conclusion

The study proves that the analysis of random forest telecom customer loss model has a wide range of application prospects and values. However, there are still many improvements in the optimization of the prediction model, such as adjusting the parameters of the decision tree, refining the feature screening, or using a variety of algorithms for model evaluation. Besides, there are still many things worth analyzing and mining in this data, such as user classification, user life cycle analysis, cross analysis among various variables, etc. In the future, this project needs to be improved in many aspects. Therefore, the next research is mainly aimed at choosing more reasonable function types and their parameters. At the same time, the application of customer loss prediction in other industries is also the future research direction.

References

[1] Gallo, A. The Value of Keeping the Right Customers, 2010.

[2] Shehab, Mohammad, et al. Machine learning in medical applications: A review of state-of-the-art methods. Computers in Biology and Medicine 145, 2022, 105458.

[3] Majumder, Madhumita Guha, Sangita Dutta Gupta, and Justin Paul. Perceived usefulness of online customer reviews: A review mining approach using machine learning & exploratory data analysis. Journal of Business Research 150, 2022, 147-164.

[4] Aziz, Saqib, et al. Machine learning in finance: A topic modeling approach. European Financial Management 28.3, 2022, 744-770.

[5] Yu, Q et al. Clustering Analysis for Silent Telecom Customers Based on K-means++. 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). Vol. 1. IEEE, 2020.

[6] Tao, K., & Qian, z. Application of improved support vector machine in Telecom customer churn prediction. Computer simulation, 28(7), 4, 2011.

[7] Guan Yiu, W., & Yong, G. Application research of Support vector machine in Telecom customer churn prediction. Computer simulation, 28(4), 5, 2011.

[8] Kaggle. Telco customer churn. https://www.kaggle.com/blastchar/telco-customer-churn, 2017.

[9] Liaw, A, & Wiener, M. Classification and regression by random forest. R News, 23(23), 2002.

[10] Shishi, H., & shexue, H. Brief analysis of Random Forest theory. Integration technology, 002(001), 1-7, 2013.