Data mining models in telecom churn prediction

Zekai Xing

School of Computer Science & Technology, Beijing Institute of Technology, Beijing, China

1120211934@bit.edu.cn

Abstract. Customer attrition has become one of the biggest problems facing the telecommunications industry due to its fast growth. According to telecom studies, acquiring new customers is more expensive than keeping current ones. Telecom companies may use the information gleaned from telecom data to understand the causes behind customer churn and take steps to maintain their current clientele. This study explores the popular data mining methods for spotting trends in loss of clients. Principal component analysis is used in the survey to reduce the dimension of the attributes. Some conclusions about the relationship between costumer usage data their churn can be summarized through the exploratory data analysis. And three prediction techniques (Logistic Regression, SVM Regression, Random Forest Regression) are applied in the customer churn prediction. This paper compares the accuracy and performance of these models. The result shows, among these models, SVM regression has the highest accuracy. It has an accuracy of 0.92 and a precision of 0.48.

Keywords: Data mining, Telecom Churn, SVM.

1. Introduction

The telecommunications industry is characterized by intense competition, rapid technological advancements, and an ever-growing customer base. In this dynamic landscape, the issue of client attrition is one of the biggest problems telecom service providers confront, where subscribers discontinue their services and switch to other providers. High customer churn has serious consequences, such as decreased income and higher client acquisition expenses, and diminished market share. Therefore, predicting and mitigating customer churn has become a critical objective for telecom companies worldwide [1].

Data mining is a subfield of artificial intelligence and machine learning, which is good at finding potential patterns between variables in large-scale data and has carried out related application research on telecom customer churn. Therefore, this study collects relevant data, applies the machine learning model to the problem of telecom customer churn, and provides a reasonable and reliable reference for relevant enterprises through the identification of customers with loss risk.

The primary goals of this study encompass the evaluation of various data mining methods and algorithms in terms of their accuracy and effectiveness in predicting customer churn. Additionally, this research aims to unravel the underlying patterns and factors that influence churn, including customer demographics, usage behavior, service quality and customer support interactions.

Throughout the subsequent sections, this paper will delve into the data mining techniques employed, the dataset utilized for analysis, and the results obtained from our predictive models. By shedding light on the application of data mining in the context of telecom customer churn prediction, this study contributes to the body of knowledge surrounding customer retention strategies and data-driven decision-making within the telecommunications sector.

2. Literature review

In recent years, extensive research has been conducted in the realm of forecasting telecom churn, demonstrating the substantial interest and scholarly attention this domain has received. For example, Ye, Cheng and Lin employed the Bayesian Network technique to predict the customer churn [2]. Wang, Chlang, and Hsu presented a decision tree algorithm as part of their discussion of a recommender system for customer attrition [3]. The study contained data from thousands of members and transactions over a three-month period. Decision tree was also used by Shuai and Huang to predict the client churn for the customer relationship management system [4]. Xu, Liu and Yao used the backpropagation neural network to build the prediction model [5]. The goal of Kamalraj and Malathi's study was to better understand churn prediction through the use of data mining techniques [6]. This kind of client retention operations may be applied by the telecommunications sector as part of their client Relationship Management initiatives. The DM approach is used by the author on client details. While the existing literature has made significant contributions to our understanding of this model, there remain several unexplored avenues for future research. These gaps present opportunities for further survey.

3. Data

Data for this paper is downloaded from kaggle. The data preparation phase includes data cleaning, feature extraction and normalisation steps. Moreover, this paper conducts appropriate exploratory analysis to extract useful insights. There are 226 attributes in the dataset. They are divided into categories and do EDA on them. Through EDA, some relationships between them can be found for reducing the data dimension, and at the same time, the relationship between these attributes and user churn can also be found through visualization. Four more important types of attributes are given next as examples.

3.1. Recharge amount related variables

This paper chooses some attributes like the total recharge amount, total recharge amount for data and maximum recharge amount as examples. There is a drop in all of them for churned customers in the 8th month (See Figure 1-3).

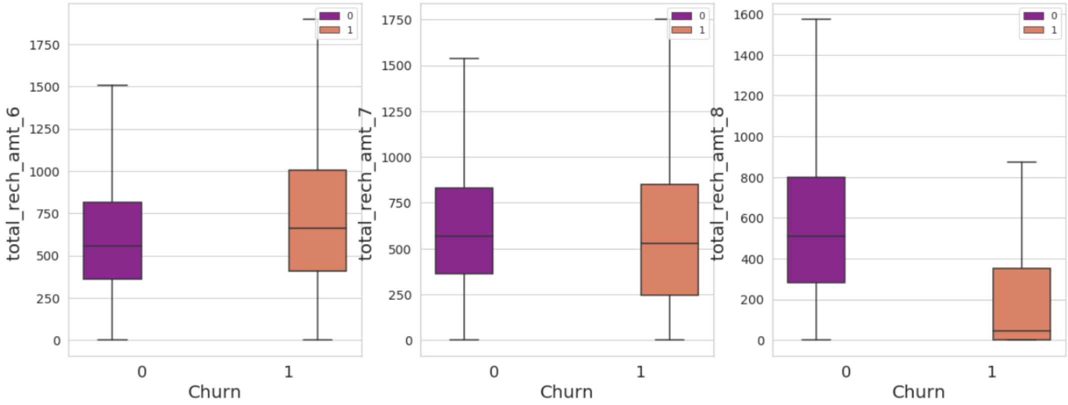

Figure 1. Box plot for the total recharge amount from June to August

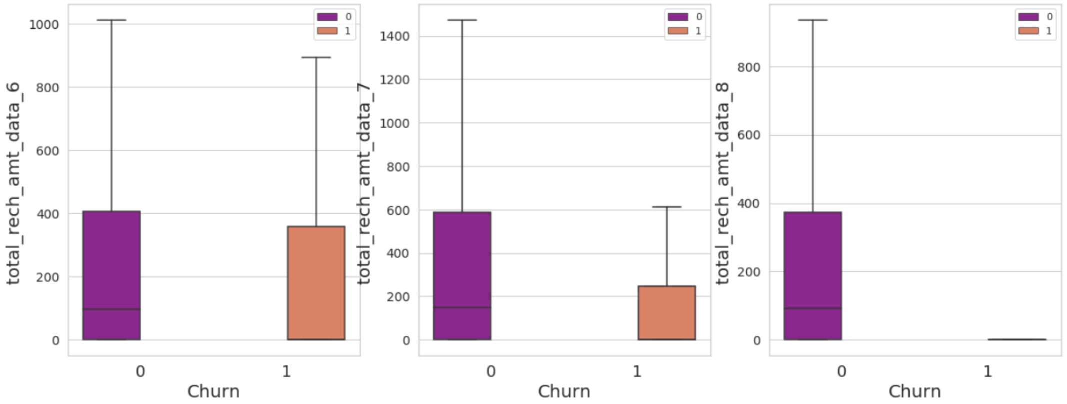

Figure 2. Box plot for the total recharge amount for data from June to August

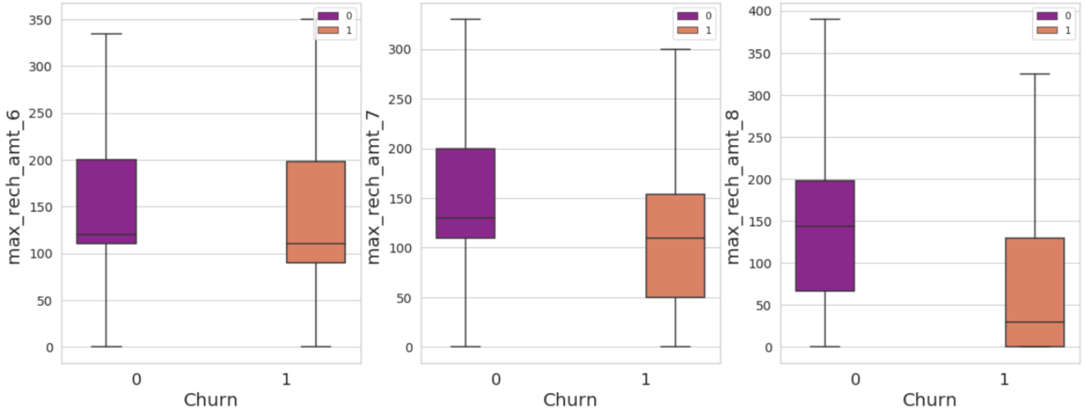

Figure 3. Box plot for the maximum recharge amount from June to August

3.2. 2G and 3G usage related attributes

Two observations can be derived from above: 1) In the eighth month, churned consumers use less 2G and 3G networks. 2) The fact that non-churned consumers use 2G and 3G more frequently suggests that churned customers may come from places where 2G and 3G service is not adequately accessible (See Figure 4-6).

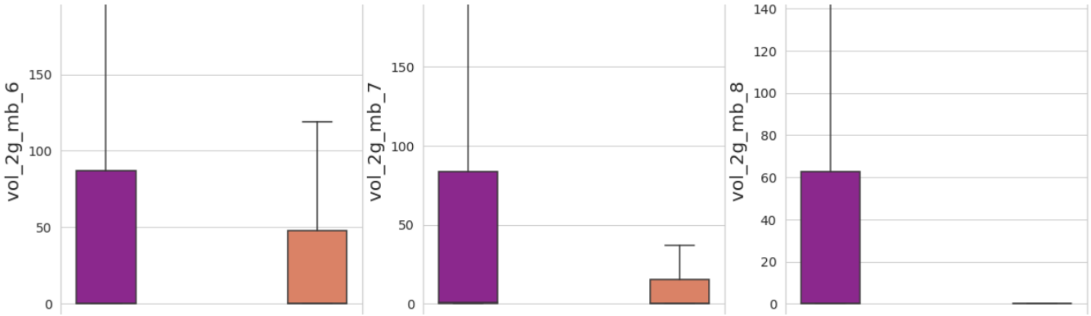

Figure 4. Box plot for the volume of 2G usage from June to August

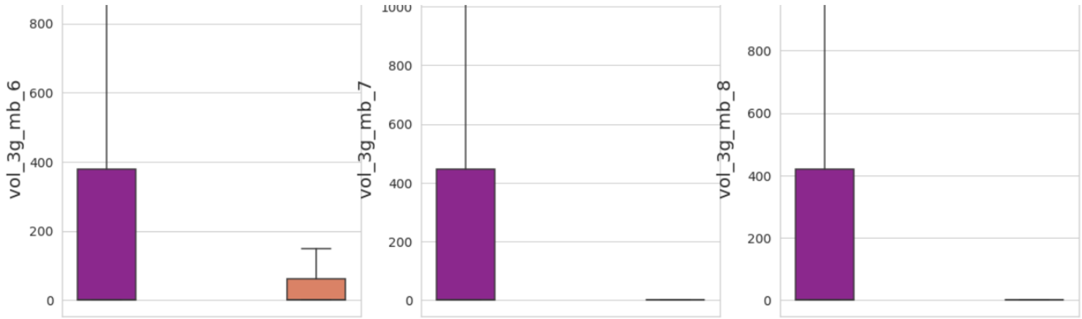

Figure 5. Box plot for the volume of 3G usage from June to August

For the 2G&3G monthly subscription, again a drop in monthly subscription for churned customers in 8th month can be seen.

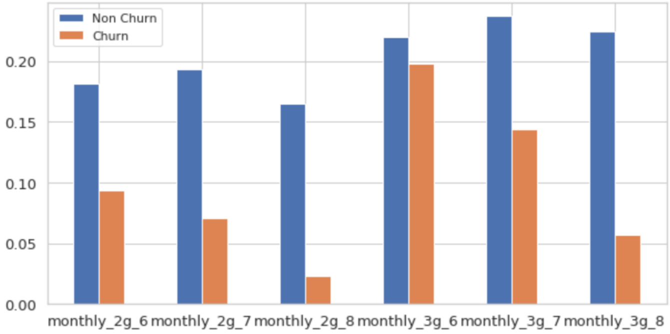

Figure 6. Bar plot for the monthly subscription from June to August

3.3. *Average revenue per user*

Huge drops can be seen for average revenue per user in 8th month for churned customers (See Figure 7).

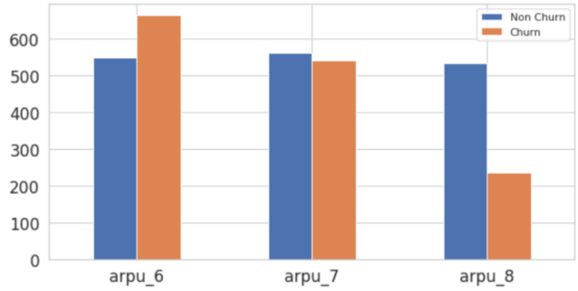

Figure 7. Bar plot for the monthly subscription from June to August

3.4. *Minutes of usage for voice calls*

There are lots of attributes related to the minutes of usage for the voice calls, the heat map can be seen in the following Figure 8.

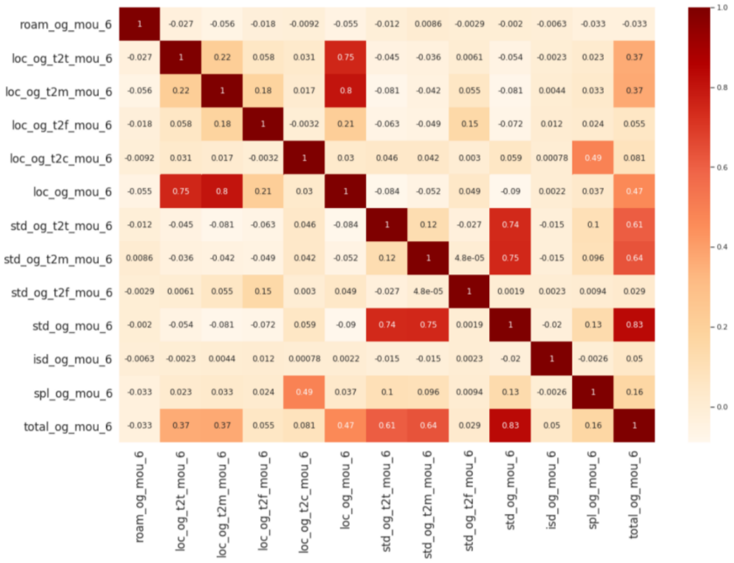

Figure 8. Heat map for the attributes related to the minutes of usage for voice calls

From the heat map, it can be seen that the inspection of some attributes like total incoming call on month 6, standard incoming call on month 6, and local incoming call on month 6 is necessary to prevent multicolinearity concerns since they appear to have a high association with other fields. And it can be found that some other variables in the dataset are combined into total incoming call on month 6, standard incoming call on month 6, and local incoming call on month 6. Thus, it is possible to delete these columns from the data for every month.

4. Methods & Results

4.1. Principle

The customer churn model is mainly based on whether telecom customers are churned within a certain period of time. The essence of this judgment is a classification problem [7], that is, to divide existing customers into two categories: customers with churn propensity and customers without churn propensity. Customer churn prediction is to use the sample database (which contains both customer data that has been lost in a period of time and customer data that has not been lost) to and analyse the factors such as customer use, business, payment, service satisfaction and whether it is off the grid for a period of time, and find out the regular knowledge that affects whether different customer groups will churn. This model can then be used to analyze the probability of churn from customer-related data tracked at a certain stage.

4.2. PCA

Principal component analysis, or PCA for short, uses the concept of dimensionality reduction to combine several indicators into just a few of combined indicators. It is a statistical approach used to reduce the complexity of a data collection. The transition is linear. Any data projection's first significant variation is located at the first coordinate, also known as the first principal component; the second significant variance is located at the second coordinate, also known as the second principal component; and so on. The data is transformed into a new coordinate system as a result. Principal component analysis is frequently used to decrease a dataset's dimensionality while preserving the characteristics that account for the majority of its volatility. The higher order principal components are disregarded in favor of the lower order principal components in order to achieve this. In this manner, the most significant parts of the data are often preserved by the lower-level components. This isn't always the case, though; it all depends on the particular use. And it looks like 60 components are enough to describe 95% of the variance in the dataset. 60 components will be chosen finally for modeling (See Figure 9).

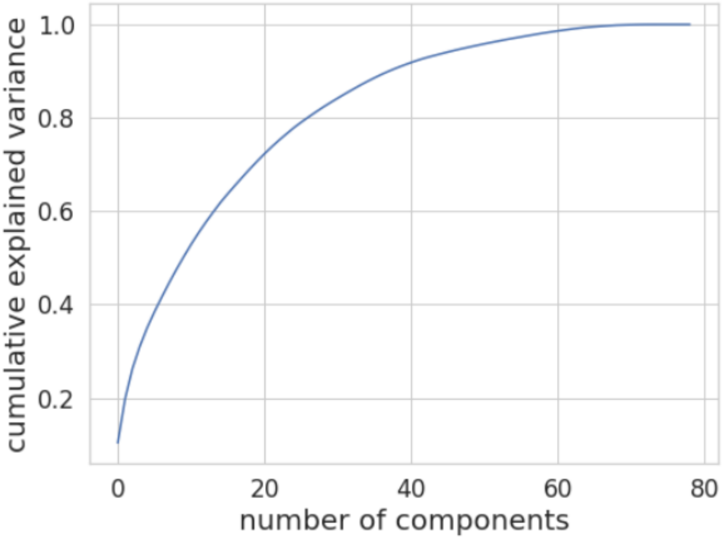

Figure 9. Scree plot

4.3. Logistic regression

The logistic function has the following form:

$$p(x) = \frac{1}{1 + e^{-(x-\mu)/s}} \tag{1}$$

Where μ is a location parameter and s is a scale parameter. Logistic regression is mainly used in classification problems [8]. Take binary classification as an example, for the given dataset it is assumed that there exists such a straight line that can complete the data linearly separable. An extra layer is needed for logistic regression in order to identify the category by comparing the probability values and discovering a direct association between the input vector x and the classification probability P(Y=1). Consider a binary classification issue, given the following data set.

$$D = (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_{Nn}), x_i \subseteq R^N, y_i \in 0, 1, i = 1, 2, \ldots, n \tag{2}$$

Given that $\omega^T x + b$ takes the continuous values, so it cannot fit discrete variables. So, probability can be used since it takes continuous values as well. The function is widely used in this model.

$$y = \frac{1}{1 + e^{-(\omega^T x + b)}} \tag{3}$$

$$\ln \frac{y}{1-y} = \omega^T x + b \tag{4}$$

Let y denote the likelihood that x is a positive sample and let 1-y represent the likelihood that x is a negative sample. If the chance of the event occurring is p, the ratio of the likelihood that it will occur to the likelihood that it won't occur is known as the odds. The formula is as follows.

$$\omega^T x + b = \ln \frac{P(Y = 1|x)}{1 - P(Y = 1|x)} \tag{5}$$

$$P(Y = 1|x) = \frac{1}{1 + e^{-(\omega^T x + b)}} \tag{6}$$

That is, the linear function of the input x decides the log chances of the outcome Y = 1. This is the logistic regression model. Through the logistic regression model, our churn prediction has an accuracy with 0.88.

4.4. SVM regression

Support vector machines are commonly used in supervised learning for analyzing data in regression and classification problems. They have corresponding learning methods in machine learning [9]. A non-probabilistic binary linear classifier is generated by an SVM training method, which builds a model that, given a set of training examples that have been labeled as belonging to one of two categories, assigns new samples to one or the other category. Training examples are mapped to points in space using SVM in order to maximize the difference between the two categories. Next, additional samples are predicted to belong to a certain category and mapped into the same space based on which side of the gap they fall into. Through a technique known as the kernel trick, SVMs may effectively perform both linear and non-linear classification by implicitly transforming their inputs into high-dimensional feature spaces (See Figure 10-12).

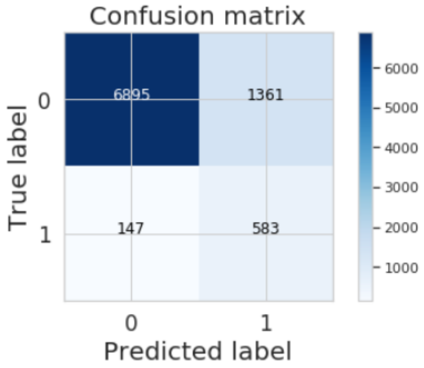

Figure 10. SVM (Default)-linear Model Stats Scores Summary

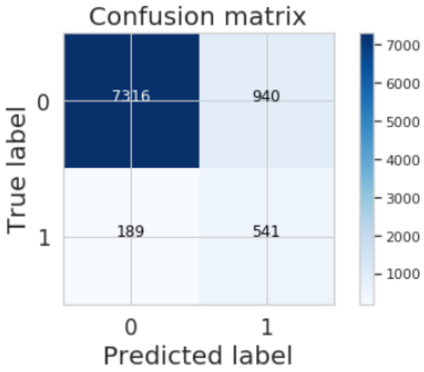

Figure 11. SVM (Default)-rbf Model Stats Scores Summary

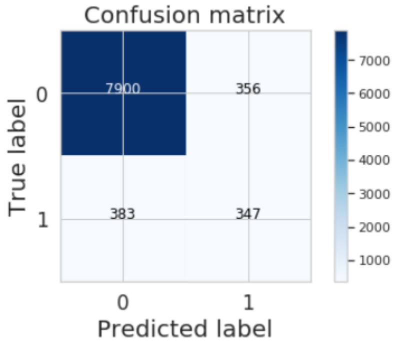

Figure 12. SVM (rfb) [Hyper] Model Stats Scores Summary

This research builds three different models for the SVM regression, the following Table 1 shows the best performance of the SVM (rfb) [Hyper] model.

Table 1. SVM (rfb) [Hyper] Model Stats Scores Summary

Model	Accuracy	Precision	Recall	AUC	F1
SVM (rfb) [Hyper]	0.92	0.48	0.49	0.72	0.49

4.5. Random forest regression

For classification, regression, and other tasks, random forest is an ensemble learning technique that builds a huge amount of decision trees in the training process [10]. For classification tasks, the random forest output is the class that most of the trees select. The average prediction result made by each single tree is returned for regression tasks. By employing this model, optimal results can be achieved in the following manner (See Figure 13 and Table 2).

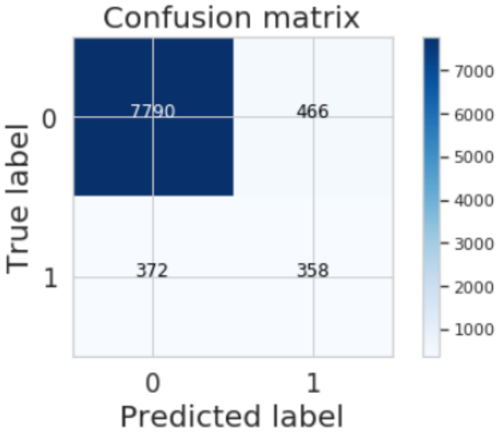

Figure 13. Random Forest (Hyper) Model Stats Scores Summary

Table 2. Random Forest (Hyper) Model Stats Scores Summary

Model	Accuracy	Precision	Recall	AUC	F1
Random forest (Hyper)	0.90	0.66	0.41	0.79	0.51

4.6. Comparison

This paper compares the accuracy and shows the results in the following Table 3. As shown, SVM with tuned hyperparameters produce best result on this dataset with 0.92 accuracy.

Table 3. Comparison of different models

Model	Logistic Regression	SVM Regression	Tree Model
Accuracy	0.88	0.92	0.91

5. Conclusions

The telecommunications industry has experienced significant customer churn rates and substantial losses due to customer attrition. While some level of business loss is inevitable, it is possible to manage and control churn to keep it at an acceptable level. To solve the issues facing the telecom sector, new and improved ways must be created in addition to developing effective ones. This paper has explored various prediction models and conducted a comparative analysis of prediction model quality measures. The findings indicate that SVM models achieve significantly higher accuracy compared to other models and it can be selected to predict churn data for future dataset or production. Besides, the results show that average revenue per user seems to be most important feature in determining churn prediction. Despite achieving some substantive results, there are still some shortcomings in this study. For example, in terms of model selection, this study only adopted partial models, and further research is needed to use other more complex models and obtain corresponding results.

References

[1] Jia, L., & Li, M. (2004). Establishment and implementation of telecom customer churn model based on data mining. Computer Engineering and Applications, (04), 185-187.

[2] Ye, J., Cheng, Z. K., & Lin, S. M. (2019). Telecom customer churn prediction analysis based on Bayesian network. Computer Engineering and Applications, (14), 212-214.

[3] Wang, Y., Chang, D., & M. Hua, S. (2009). A recommender system to avoid customer churn. Expert System with application, 36(4), 8071-8075.

[4] Ruan, S., & Sheng, Z. H. (2011). Data Mining Algorithm Based on Decision Tree Application and Research. Energy Procedia, 120-127.

[5] Xu, J. B., Liu, J. R., Yao, T. N., & Li, Y. (2023). Prediction and Big Data Impact Analysis of Telecom Churn by Backpropagation Neural Network Algorithm from the Perspective of Business Model. Big data, (5), 355-368.

[6] Kamalraj, N., & Malathi. A. (2013). Applying Data Mining Techniques in Telecom Churn Prediction. International Journal of Advanced Research in Computer Science and Software Engineering, 31, 515-524.

[7] Lemmens, A., & Croux, C. (2006). Bagging and Boosting Classification Trees to Predict Churn. Journal of Marketing Research, 43(2), 276-286.

[8] Thomas, V., Wouter, V., & Bart, B. (2013). A Novel Profit Maximizing Metric for Measuring Classification Performance of Customer Churn Prediction Models. IEEE Trans. Knowl. Data Eng, (5), 961-973.

[9] Cortes, C., Vapnik, V. (1995). Support-vector networks. Mach Learn, 20, 273–297

[10] Breiman, L. (2001). Random Forests. Machine Learning, 45, 5–32