Customer churn prediction based on LASSO and Random Forest models

Customer churn prediction based on LASSO and Random Forest models

Qiannan Zhu[1], Xinyi Yu[1], Yuankang Zhao[1], Deyi Li[1*]

[1] College of Science, Wuhan University of Science and Technology, Wuhan, Hubei, China, 430065

[*] Corresponding author：Prof, College of Science, Wuhan University of Science and Technology, Wuhan China

* Corresponding author's e-mail: 1210534766@qq.com

Abstract. Customer churn probability is influenced by many factors, due to the complexities of actual problems, high-dimensional data often exists multicollinearity, and ordinary regression model is no longer applicable, while Random Forest model without data processing will lead to a large amount of calculation and make the model become not generalizable. So we try to construct a LASSO-RF model based on the existing theories that the Random Forest model was used to predict the variables selected by LASSO model. This paper takes the member data of an airline company as an example to carry out an empirical study. The results show that compared with the LASSO model or Random Forest model alone, the LASSO-RF model constructed in this paper has a smaller amount of calculation, higher prediction accuracy and stronger generalization ability.

1. Introduction

With the vigorous development of enterprises of the same type, the competition within the industry becomes extremely fierce, and customer care and maintenance, as a rather important core competitiveness, affects the economic benefits of enterprises, so the study on customer loss is particularly important. Many industries, such as telecommunication and financial industry, have taken a series of measures for this, but some industries still pay less attention to the problem of customer loss, which is likely to become a potential hidden danger in the development of enterprises [1].

At present, the commonly used methods for customer loss include Logistic regression model, decision tree, neural network, etc. Among them, the Logistic regression model is a kind of algorithm that attempts to explore the relationship between variables through the measurement of errors, which is widely used. However, due to the tendency of multicollinearity between high-dimensional data, the Logistic regression model is no longer applicable. The decision tree does not need to consider the multicollinearity among variables, has a good capacity to contain dirt, and has a more intuitive visual presentation, but when the dimensional disaster occurs to the data, the tree will be too large and easy to produce overfitting, which is not applicable [2,3].

The LASSO model used in this paper has strong processing capacity for high-dimensional data, and can screen out variables with high correlation from a large number of variables, so as to solve dimensional disaster and multicollinearity problem. Random Forest model is an integrated algorithm

based on cart decision tree, which has the advantage of adopting voting mechanism in data analysis and effectively avoids the over-fitting problem of a single tree [4,5,6]. Therefore, on the basis of the existing theories, this paper establishes LASSO-RF model for the prediction of customer loss probability, and the empirical results show that this model has good prediction accuracy and generalization ability.

2. Theory

2.1. LASSO model.

The LASSO (The Least Absolute Shrinkage and Selection Operator) model is a solution to the dimensional disaster problem by adding The L1 regular term on the basis of the ordinary least squares and reducing the regression coefficient of some insignificant variables to 0.

Now suppose vector $X_i = (x_{i1}, x_{i2}, \cdots\cdots, x_{in})^T$ as independent variables, Y_i is the i^{th} observation values corresponding to the dependent variable, the multivariate regression model is:

$$Y_i = \alpha_i + \sum_{j=1}^{p} \beta_j x_{ij} + \mu_i, \mu_i \sim N(0, \sigma^2) \tag{2.1}$$

Ordinary least square method gives the following parameter estimation：

$$(\hat{\alpha},\ \hat{\beta}) = \min\left\{\sum_{i=1}^{n}(Y_i - \alpha_i - \sum_{j=1}^{p}\beta_j x_{ij})^2\right\} \tag{2.2}$$

LASSO regression adds L1 regular term on the basis of the above formula and gives the following parameter estimation：

$$(\hat{\alpha},\ \hat{\beta}) = \min\left\{\sum_{i=1}^{n}(Y_i - \alpha_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 + c\sum_{j=1}^{p}|\beta_j|\right\} \tag{2.3}$$

In formula (2.3), c is the adjusted parameter, and its choice will directly affect the regression results of LASSO regression. According to Robert Tibshirani's study, c can be selected by the generalized cross validation method, and the form of the generalized cross validation method is：

$$GCV(c) = \frac{1}{N} * \frac{SSE(c)}{[1 - p(c)/N]^2} \tag{2.4}$$

Where, *SSE(c)* is the sum of squares of residues, *p(c)* is the number of effective regression coefficients in LASSO regression, and the optimal adjustment parameter c should minimize the verification value of generalized crossover[7,8].

Although the LASSO model can conduct variable screening to solve the multicollinearity problem among high-dimensional data, its essence is linear regression, so the single LASSO model has the problem of low prediction accuracy when making prediction decisions.

2.2. Random Forest model.

The basic idea of the random forest algorithm is as follows: firstly, feature selection is carried out on the decision tree to make the classified data set more pure. Here, *GINI* index is selected as the purity measurement standard:

$$\Delta G = 1 - \sum_{i=1}^{q} p_i^2 - \sum_{j=1}^{k} \frac{D_j}{|D|} G_{D_j} \tag{2.5}$$

Where, G represents the GINI function, q represents the number of categories in sample D, P_i represents the proportion of category i samples to the total number of samples, and k represents that sample D is divided into k parts, that is, there are k D_j data sets. When the gain value of *GINI* index in formula (2.5) reaches the maximum, node splitting is carried out.

Secondly, the generated multiple decision trees constitute the random forest, and the simple majority voting mechanism is adopted to complete the prediction[4,5,6]. The final classification decision is shown in formula (2.6):

$$L(X) = \max \sum_{i=1}^{k} I(l_i(x) = y) \tag{2.6}$$

Where, *L(X)* represents the combined classification algorithm, l_i represents the classification algorithm of the i^{th} decision tree, and *y* is the target variable. *I(•)* is the indicative function.

The Random Forest model is an integrated model built on the basis of cart decision tree, which effectively improves the prediction accuracy and does not need to consider the multicollinearity problem among variables. However, for the high-dimensional data, the model is of high complexity and heavy computation, and the model does not have generalization ability.

2.3. Establishment of LASSO-RF model.

In this paper, LASSO-RF model will be constructed to solve the shortcomings of the above models. By combining the advantages of each model, the combined model will have the characteristics of smaller computation, higher prediction accuracy and stronger generalization ability. To this end, the following algorithm is constructed:

Data pre-processing → LASSO model for variable screening → Preliminary construction of Random Forest model → Parameter tuning → Test data → Optimal model → Prediction decision.

1) Data pre-processing: as practical problems are always complicated, involving many variables and large amount of data, it is necessary to check the data set, fill in missing values, delete invalid values, and centralize standardization. In order to facilitate subsequent comparison and verification without loss of generality, the data set can be divided into 70% training set and 30% test set.

2) LASSO model for variable screening: on the basis of pretreatment, formula (2.3) is used for variable screening. First, the corresponding LASSO model is selected according to the type of dependent variable. Second, according to the demand to choose the appropriate minimum target parameters; finally, formula (2.4) is used for cross validation to select the optimal model.

3) Preliminary construction of Random Forest model: the Random Forest model is preliminarily established for the variables selected by the above LASSO model, and the prediction decision is made according to formula (2.6).

4) Parameter tuning: the above-mentioned preliminary build a Random Forest model need to be further optimized in order to achieve the best effect, the selection of Random Forest parameters is a very difficult thing, parameters optimization generally need to artificial selection, there is no one based on optimization theory, the optimal parameters of the generated, can select the classification error, prediction accuracy indicators as the standard of parameter tuning.

5) Test data: put 30% test set into the above optimization model for verification.

6) Optimal model: if the conclusion of the training set and the test set is consistent under a certain confidence level, the model is established successfully; otherwise, repeat steps 4) and 5).

7) Prediction decision: the optimal model is used to predict the future customer loss probability, and timely adjust the customer relationship maintenance strategy.

3. Empirical analysis

3.1. Data pre-processing.

This paper takes the member data of an airline company as an example. After data processing, the data set contains 53 variables, among which the dependent variables are binary discrete variable. It can be seen from the observation of 4964 articles that the customer loss accounts for 43.01% and the non-customer loss accounts for 56.99%, that the customer loss of the company is very serious and it is urgent to solve this problem.

3.2. Variable selection based on LASSO model.

The above standardized data contains 53 variables, and to a certain extent, dimensionality disaster occurs, for which the LASSO model is used for Variable Selection and Regularization. Import the *glmnet* package in R software and call the *cv.glmnet* function. First, since the dependent variable is binary discrete variable, the LASSO model based on Logistic regression is adopted, and the parameter is *family = binomial*. Secondly, the classification error of the model is selected as the minimum target parameter, and the parameter is *type.measure = class*. Finally, after cross validation, the adjusted parameter *c=0.05* was selected in this paper. At this time, the model had excellent performance and the number of independent variables was the least. The 7 variables with regression whose coefficient is not 0 were showed as follows:

Table1 Variable selection and coefficient

id	Variable	coefficient
1	(Intercept)	-1.37
2	flight_count	-0.39
3	flight_count7	-0.79
4	flight_count8	-2.16
5	seg_km_sum	-0.04
6	avg_flight_interval	-0.82
7	l1y_flight_fount	-0.21
8	ration_l1y_flight_count	-0.13

3.3. Build LASSO-RF model

3.3.1. Parameter selection.

Based on the above analysis, 7 main variables were selected finally, and the data dimension was greatly reduced. On this basis, Random Forest model is used, *randomForest* package is imported into R software, and *randomForest* function is called to training set. In this paper, model generalization ability, classification error, prediction accuracy, and Jaccard coefficient are used as evaluation criteria for multiple experiments, and the results are shown in the following table:

Table2 Parameter selection (training set)

Model	Model 1	Model 2	Model 3
Number of variables	3	4	5
classification error	5.91%	6.32%	6.37%
prediction accuracy	0.997	0.999	0.998
Jaccard	0.993	0.998	0.996

As can be seen from the above table, with error less than 5%, the prediction accuracy and Jaccard coefficient of the three models are considered equal, while the classification error of model 1 is the minimum, and the number of variables is small, the generalization ability is strong. Therefore, the parameters *ntree=35, nPerm=5 and mtry=3* corresponding to model 1 are the optimal parameters.

3.3.2. Model evaluation.

The final LASSO-RF model was obtained after a series of parameter selection. In order to investigate the prediction effect of LASSO-RF model, the trained model was firstly used to predict the training set and the test set respectively. Secondly, the LASSO model and the Random Forest model were separately used to predict the training set and test set, and Jaccard coefficient was selected as the index to compare the three. The results are shown in Table 3.3:

Table3 Jaccard coefficient

Model	Training set	Test set
LASSO	0.758	0.772
Random Forest	0.998	0.851
LASSO-RF	0.995	0.851

As can be seen from the above table

1) The Jaccard coefficient of LASSO model is obviously lower than Random Forest model and LASSO-RF model, so the latter two models have better prediction effect.

2) With error less than 5%, the Jaccard coefficients of Random Forest and LASSO-RF models are considered equal, and the variables' number in LASSO-RF model is less than that in Random Forest model, therefore, the complexity is lower. It shows that on the customer loss prediction problem, that first using LASSO model for variable screening and then applying the Random Forest model to predict can not only guarantee the accuracy of prediction, but make the model has stronger generalization ability.

4. Summary

The probability of customer loss is affected by many factors. Due to the complexity of practical problems and the limitation of the level of researchers, it is difficult to directly find a group of independent variables that are not related to each other and have a significant impact on the dependent variables, so the multicollinearity problem is inevitable. In this paper, on the basis of existing theory to build the LASSO-RF model is applicable to the dimension disaster and multicollinearity problem, the model effectively set the LASSO model and the Random Forest model's advantage, through empirical analysis found that compared with the single use LASSO model or Random Forest model, LASSO-RF model has a smaller amount of calculation, higher prediction accuracy and stronger generalization ability. This model can be widely applied to the data with binary discrete variables as dependent variables, so it can not only be used to predict the probability of customer loss of airlines, but also be extended to other industries to solve the problem of customer loss prediction.

Acknowledgements

This article is periodical achievement of National Under-graduated Innovative Training Program (201810488006), and the Humanities and Social Science Research Project of 2018 Hubei Provincial Education Department (18Q032).

References

[1] Cui Ya-qi. Analysis of the Airline Customer Churn Based on C5.0 Algorithm [J]. Journal of Xi′an Aeronautical University. 2018,36(1):72-77.

[2] Liu Tingting,Wang Xiaoli,Ge Mingtao.Airline customer value analysis based on data mining [J].Shandong Industrial Technology.2017,(4):287-288

[3] Wu Tongshui, He Liang . Analysis of airline customer loss based on decision tree［J］. Business modernization，2006(35)：381-383.

[4] Kai Liu. Research on Adaptive Feature Selection and Parameter Optimization Algorithm for Random Forest [D].Ji Lin: Changchun university of technology,2018.

[5] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32

[6] B Gregorutti, B Michel ,P Saint-Pierre. Correlation and variable importance in random forests[J]. Statistics & Computing , 2017 , 27 (3) :659-678

[7] Shi Guoliang, Jing Zhigang, Fan Liwei. Research on the Original Oil Price Prediction Based on Lasso-Xgboost Combination Method [J]. Industrial Technology & Economy, 2018, 37(7):31-37.

[8] Efron B, Hastic T, Johnstone I, Tibshrani R. Least Angle Regression[J]. Annals Statistics, 2004, (32):407~499.