Research on Application of Stacking Technique in Telecom Churn Prediction

Junhui Gu*

School of information, Xi'an University of Finance and Economics, Xi'an, Shanxi, 710199, China

*Corresponding author: 1931053594@xaufe.edu.cn

Abstract. In today's saturated telecommunication market, operators need to fully tap the value of data, strengthen the management of existing users, and reduce customer churns. In order to further to reduce the maintenance cost of enterprises, it is necessary to apply more efficient data mining technology to the prediction of telecommunication user churns in order to improve the operating profit of enterprises. This paper took a telecommunication customer churn data set on Kaggle as the research data, and explores a high accuracy customer churn prediction model. Logistic Regression was exploited to create customer churn prediction models, and the parameters of each model were optimized. On this basis, the Stacking technique was used to fuse the models, and the best combination method is obtained. After research, this paper mainly drew the following conclusions: First, in a single model, the model built with XG Boost algorithm has the highest prediction accuracy, which is 86.20%. Secondly, combining the accuracy rate of training set, cross-validation set and test set, it can be found that XG Boost and Random Forest have the best performance under Stacking fusion algorithm. The accuracy of the above three datasets is 97.61%, 85.84%, 86.51%, respectively, which improves the prediction accuracy of the prediction model.

Keywords: Telecom customs churn, Churn prediction, Data mining, Stacking technique.

1. Introduction

1.1.Research Background

At present, the telecommunication industry has become one of the major industries in many countries. Telecom companies are producing data at an extremely fast rate. Technological advances and an increase in the number of operators have accelerated competition in the telecommunications industry. In such an environment, any customer dissatisfaction with the telecommunication company will result in customer loss, that is, customer transfer from one telecommunication service provider to another. Telecom companies aim to maximize benefits while ensuring long-term survival in the market. Research shows that it is easier to earn more revenue by extending a customer's retention period than by upsell the existing customers or acquire new customers. This is because of the costs involved in advertising, labor and preferences, which can expand to nearly five or six times the cost of existing customers [1]. On the other hand, in the Internet era, the use of traditional voice business is declining, and the value of consumers in telecommunication enterprises is being nibbled by Internet enterprises. Communications operators need to focus their attention from mining new users to maintaining the stock users. While expanding the incremental market, they should also pay close attention to the needs of the stock users to avoid customer churn.

1.2.Research Gap

First, the data used in the customer churn prediction model built in this article are according to the available data on Kaggle. There will be more missing and abnormal values in the data in the enterprise database. Although this paper fills in and modifies these data to make it close to business rules, it also affects the prediction results of the model to some extent. Secondly, the data set used in this paper is ideal and cannot completely correspond to the user information of the telecommunication company under the real business.

This paper uses Stacking model fusion method to fuse models, explores different combinatorial methods of models, improves the prediction effect of models, and enriches the research of fused

models. Secondly, the application of the model is considered from the point of view of discriminant threshold, and the selection of discriminant threshold for a single model with the largest F1 value is shown by the way of calculating.

1.3.Fill The Gap

The first part is the introduction. In this part, the article combines the background of the contemporary telecommunication industry to discuss the necessity of avoiding customer churn for telecommunication enterprises. At the same time, it also introduces the inadequacy and innovation of the article.

In the second part, this article lists the research overview and methods of customer lose in the telecommunication industry.

In the third part, the article uses a telecommunication customer history dataset downloaded on Kaggle's official network as experimental data, conducted in-depth analysis and data preprocessing.

In the fourth part, the Random Forest, XGBoost, Logistic Regression and SVM (rbf) algorithms used are optimized to build prediction models.

In the fifth part, the article shows the main results of the study. First, the results of single classifier classification (accuracy, precision, recall, F1, AUC, ROC curve) after parameter tuning are presented and analyzed. Then, each single classifier is analyzed from the perspective of discriminant threshold through visualization. Finally, the output of Stacking fusion model is shown and analyzed.

Finally, the fifth part summarizes the conclusions of this study.

2. Literature Review

Prediction model based on decision tree algorithm: The classification tree algorithm used by Cox L. A in the research of customer behavior prediction model can see the embryonic form of the decision tree algorithm [2]. Lines et al. constructed a model of customer churn prediction based on Decision Tree algorithm [3]. By comparing the representation of Logistic Regression algorithm and decision tree algorithm, the experimental results clarify that decision tree is an efficient technique for customer churn prediction. Wei Hu et al. analyze customer turnover behavior, make improvements to the traditional decision tree algorithm, prune the Decision Tree, solve the defects of the traditional Decision Tree algorithm in the multi-rule set, build a prediction model, and improve the properties of the prediction algorithm [4].

Prediction model based on Random Forest algorithm: Huang et al. use Random Forest algorithm [5]. By comparing all previous research results, this paper finds that the model built by Random Forest performs well in the prediction model of telecommunication customer churn. Wu et al. proposed using hierarchical sampling Random Forest algorithm, using SVM as a cascade classifier at each tree node, which proves to be more effective than original Random Forest, SVM and other algorithms through research [6]. Yang et al. proposed a deep Random Forest model [7]. This model was chosen to stack shallow Random Forests in order to produce more accurate prediction. This stacked deep structure has better prediction effect than traditional Random Forests. The improvement of algorithm complexity will inevitably reduce the execution efficiency of the model.

To explore the application effect of distinct machine learning algorithms in customer churn warning, some studies have compared and analyzed the effects of several algorithms. Shin. Yuan Hung et al. constructed a model using two classification methods, Decision Tree algorithm and BP artificial neural network [8]. The experimental results clarifies that the model has excellent performance. Qi et al. used AD Tree, Logistic Regression Integration Model and Tree Net to build customer churn warning models, and made a comparative analysis of the results [9]. Buckley B et al. used logistic regression, decision tree, naive Bayesian, and neural network methods to model customer churn prediction system, and compared the accuracy and efficiency of several algorithms under the same conditions, and introduced the advantages and shortcoming of various algorithms [10].

To further improve the prediction effect of the model, researchers began to apply the fused algorithm to customer churn prediction. Ma Jun et al. used Bayesian inference-based Decision Tree algorithm to model, and the accuracy and recall rates were higher than those based on a single decision tree algorithm [11]. Xu proposed a fusion model based on XGBoost Bagging method combined with mixed sampling method, made full use of the unbalanced dataset, constructed a two-classification model to predict the churn of telecommunication customers, and verified the model with multiple indicators, expecting good results of profit verification model [12].

The above research on customer churn mainly focuses on the construction or improvement of a single or a small number of models, so it is necessary to explore the application of model fusion method in prediction model construction to improve the accuracy of prediction.

3. Sample and Data Processing

3.1.Sample Source

The data researched in this paper comes from Kaggle. The data contains 7043 samples and 40 fields. Among them, 5174 positive samples (both "Joined" and "Stayed" labels are considered as positive samples) and 1869 negative samples, which are stored in 3 CSV files. One of the files contains the information of the population field corresponding to the “Zip Code” field. Before data preprocessing, the information of the population field has been added to the training dataset with reference to this file.

3.2.Field Requirements

Selection of target field: in the prediction and research of telecom customer churn, the data corresponding to the target field contains information identifying customer status (whether there is churn). The target field of the model built in this paper is "customer status". Among them, the division rules of positive and negative samples are as follows: positive samples refer to customers whose information (label) in the target field is "stayed" or "joined", and positive samples refer to customers whose information in the target field is "churched".

Selection of input field: the input field refers to the feature field required in the construction of the model, and the information of these features is related to the information of the target field. By reading File A, which stores field descriptions, all fields are classified into seven categories:

(1) Basic customer information. Including customer ID, gender, age and other information

(2) Environmental information. Including the city, longitude and latitude, zip code and population of the city where the customer is located

(3) Call information. Including whether to subscribe to multi line services, whether to handle home phone service packages and other information

(4) Network information. Includes information such as whether to provide network services to the company, whether to handle unlimited traffic, last month traffic, etc.

(5) Consumption information. Include monthly average long distance call consumption, payment method, paperless billing, etc.

(6) Value-added service information. Includes access to unlimited data downloads/uploads, advanced technical support, and device protection plans.

(7) Additional information. Include contract life, loss category, loss reason and other information

3.3.Data Preprocessing

3.3.1 Delete Field

Through reading the data set description file and preliminary judgment on the number of field value types, some fields unrelated to classification or unable to be processed were deleted. The following is described by [Irrelevant / Unable to process]:

Table 1. Description of the deleted field

Deleted fields name	Describe
Customer ID	Irrelevant
Churn Reason	Unable to process
Zip Code	Irrelevant (In the "original data description" section, its functions have been described)
City	Irrelevant
Churn Category	Unable to process

3.3.2 Classification Label Field Processing

Because the research in this paper needs to train the customer data set through various machine learning algorithms to obtain a prediction model, it requires that the sample values under the label field in the data set be dichotomized, so only the "stayed" and "churched" labels are retained. That is, change the sample whose value is "joined" in the column of the field customer status to "stayed".

3.3.3 Missing Value Processing

Because the processing methods for missing values of different fields are different, the fields of the dataset are classified before processing. All fields in the dataset can be split into two categories: 1. Classification field and 2. Numerical field. Numeric fields refer to fields whose values are numbers and do not have classification functions. Fields other than these numeric fields are classified fields. Classification fields can be divided into multiple classification fields and two classification fields. When the number of types of values in the classification field is equal to two, it is called a two classification fields, and other classification fields are multi classification fields.

Numeric field: fill with the mean value of the corresponding column of the field. Secondary classification fields: the secondary classification fields with missing values in this dataset are all caused by not opening Internet service (field). Therefore, the phenomenon of missing information itself has useful information. Therefore, assign new value none to the missing value of the column where these two classification fields are located, divide them into multi classification columns, and participate in the training of the model.

Multi category field: except for the missing value of the churn reason field and too many existing value types that cannot be processed. The processing methods and reasons of the remaining multi category fields are the same as those of the two category fields.

3.3.4 Split Multiple Classification Columns and Normalize Numerical Field Data

Convert the information in the column of the multi category field into One-Hot encoding. When there are types of values in field, perform the following operations:

$$\{x_i\}_{i=1}^{N} \Rightarrow \{(x_i^1, x_i^2, \cdots, x_i^p)\}_{i=1}^{N}$$

x_i^p : : Represents the data under the numerical value type after the split of multi category field x .

The Offer field is converted to One-Hot encoding

	Offer
0	Offer A
1	Offer B
2	None

→

	Offer _ Offer A	Offer _ Offer B	Offer _ None
0	1	0	0
1	0	1	0
2	0	0	1

Figure 1. The Offer field is converted to One-Hot encoding

Normalization of numerical field data:

Normalize the value according to the data with *Min – Max scaling* method:

$$X_{norm} = (X - X_{\min}) / (X_{\max} - X_{\min}) \quad (2)$$

3.4.Dataset Segmentation

It is necessary to randomly extract 1 / 4 of the samples from the original data set to form a test data set, and the remaining data form a training data set. At this time, the dimension of the training dataset is: 5282 rows × 54 columns; The dimension of the test data set is: 1761 rows × 54 columns.

4. Model Parameter Optimization

There are great divergences in the performance of models under different parameters. In order to realize the best prediction effect of each model, it is essential to adjust the parameters of a single model through grid search.

4.1.Optimization of Logistic Regression Model

When the Logistic Regression algorithm is used to build the model, the penalty term parameter "penalty" and the regularization parameter "C" are mainly adjusted. Wherein, the penalty term parameter has two values of "L1" regularization and "L2" regularization; The value of regularization parameter C represents the intensity of regularization. The smaller the value of this parameter, the greater the intensity of regularization. In this paper, when the grid search method is used for parameter optimization, it is found that the AUC value of the model when using the L1 regular term is higher than that when using the L2 regular term. Therefore, the L1 regular term is selected as the penalty term parameter, and the grid search is used for optimization after setting the value range of C. The optimal parameter setting of logistic regression is shown in Table 2.

Table 2. Parameters of Logistic Regression model

Parameter name	Effect	Optimal parameter value
penalty	Regularization method	L1
C	Reciprocal of regularization intensity	3

4.2.Parameter Optimization of Random Forest Model

As the grid search method is exploited to adjust the model parameters, it needs to traverse the different combinations of various parameters to obtain the optimal parameter combination, and the amount of data used in this paper is large. Using this method to find the optimal parameter combination will greatly increase the time cost, so this paper only adjusts the important parameters. The parameters that need to be adjusted in the Random Forest algorithm are mainly n_ estimators, max_ depth, min_ samples_ leaf, min_ samples_ Split and Max_ features. When n_ If the value of estimators is too small, the model will be under-fitting; if the value is too large, the model will be easily over-fitting; max_ The depth parameter can limit the maximum depth of the Decision Tree to reduce overfitting; min_ samples_ leaf, min_ samples_ Split and Max_ Features can serve the purpose

of limiting the size of the decision tree. The important parameter values of Random Forest are set as shown in Table 3.

Table 3. Random Forest parameters

Parameter name	Effect	Optimal parameter value
n_estimators	The number of sub decision trees set artificially, the higher the value, the more accurate the prediction is, but the operation efficiency is low	180
max_depth	Limit the maximum depth of the decision tree to reduce the impact of overfitting	9
min_samples_leaf	Control leaf growth and decision tree size	8
min_samples_split	Control leaf growth and decision tree size	16
max_features	Limit the size of the tree	30

4.3.Parameter Optimization of SVM (rbf) Model

When building a model using the SVM (rbf) algorithm, the selected kernel function is RBF (Radial Basis Function), and the main adjustment parameters are C and gamma. The higher the C, the lower the toleration of error and the easier the model to fit. The smaller C, the easier the model will be to underfit. Whether the C value is too large or too small, it will directly result in a poor generalization ability of the model. The gamma parameter is a parameter attached to the RBF function when it is selected as the kernel function. The value of gamma (only two options: auto, scale) determines the number of support vectors. Parameter C is optimized using grid search with different gamma values. Finally, the optimal parameter settings for the support vector machine model are shown in Table 4.

Table 4. SVM (rbf) parameters

Parameter name	Effect	Optimal parameter value
C	C is the regularization parameter. The intensity of regularization is inversely proportional to the value of C. The value must be positive. The type of penalty is the square L2 penalty.	3.5
gamma	Coefficient of kernel function	0.1

4.4.Parameter Optimization of XG Boost Model

The parameters in XGBoost library can be divided into three categories: macro parameters, booster parameters and learning target parameters. This article only adjusts the booster parameters. First, set other parameters to the default state, and adjust the maximum number of iterations n_ Estimators and learning rate_ rate, the optimal iteration times and the learning rate are determined to be 300 and 0.05 respectively through grid search; Second, adjust max_ Depth parameter, whose value is generally set between 3 and 10; Finally, adjust subsample and gamma parameters, which are important to reduce overfitting. The values of some booster parameters are shown in Table 5.

Table 5. XGBoost model parameters

Parameter name	Effect	Optimal parameter value
n_estimators	For the number of sub decision trees set artificially, the higher the value, the more accurate the prediction is, but the operation efficiency decreases	350
Learning_rate	Learning rate	0.05
max_depth	Limit the Maximum depth of the Decision Tree to reduce the impact of overfitting	7
subsample	Proportion of random sampling of decision tree	0.85
gamma	The lowest value of loss function decline when nodes are split	0

5. Result & Discussion

After adjusting the parameters of the underlying model, the optimal parameter values of each model are brought into the model, and the communication user loss early warning model is constructed. Multiple evaluation indicators are utilized to compare and evaluate the modeling effect of each model. The information shown in Table 6.

Table 6. comparison of effects of various models

Evaluation index	Logistic Regression	Random Forest	SVM(rbf)	XGBoost
accuracy	83.33%	85.46%	83.40%	86.20%
precision	87.61%	87.30%	86.72%	88.88%
recall	89.67%	93.58%	89.44%	92.57%
F1	88.63%	90.33%	88.06%	90.69%
AUC	78.08%	79.15%	78.87%	79.66%

In terms of accuracy, the prediction accuracy of XGBoost model is the highest, reaching 86.20%, while the prediction accuracy of Logistic Regression is the lowest, only 83.33%. However, in general, the prediction accuracy of Random Forest, XGBoost, Logistic Regression and SVM (rbf) models are all over 80%; From the perspective of precision, the prediction results of XGBoost model are the best, reaching more than 88%. From the perspective of recall rate, the Random Forest model has the highest score. Generally speaking, the two indicators of accuracy rate and recall rate often restrict each other. The two indicators can be considered comprehensively, that is, F1 value. From the perspective of this indicator, XGBoost model also has the best prediction effect.

Next, the ROC curves of each model are compared, as shown in Figure 2:

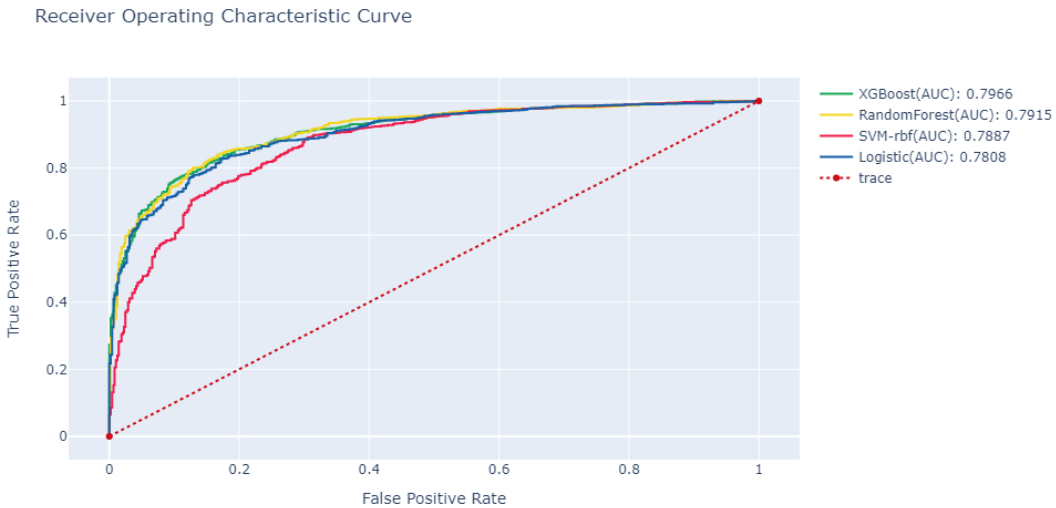

Figure 2. ROC curve and AUC value of each model

From the ROC curves of each model, the ROC curves of XGBoost model have better effect than the other three models, and their corresponding AUC value is 0.7966. The AUC values of Random Forests and SVM (rbf) models are very similar, 0.7915 and 0.7887, respectively. By contrast, the logistic regression AUC value was 0.7808, which was the lowest in all models. The results are shown in Figure 3.

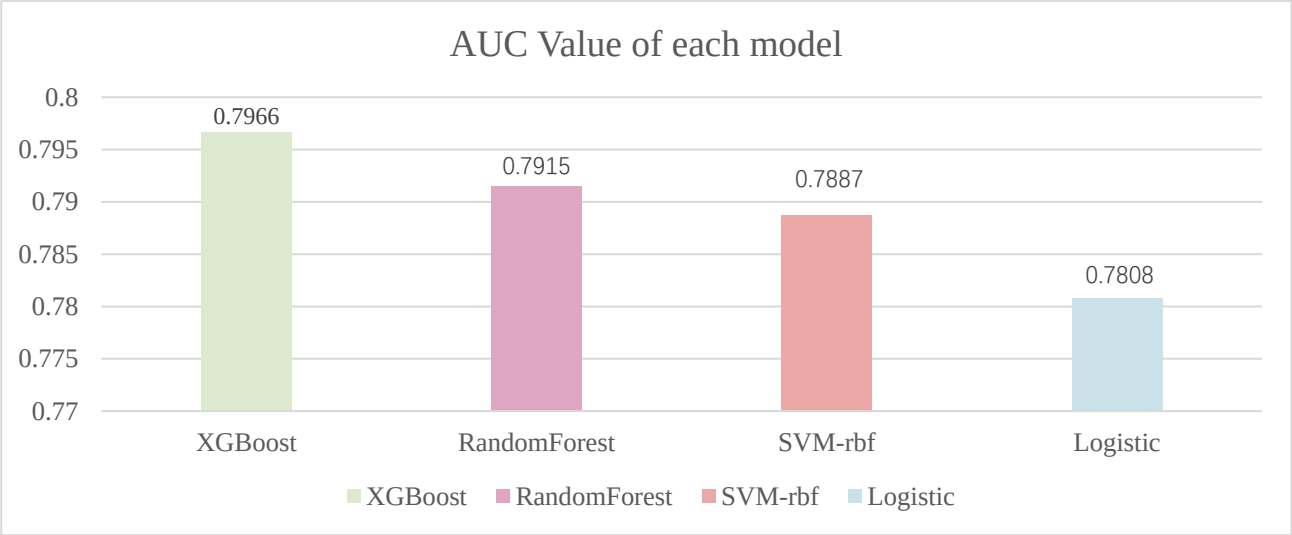

Figure 3. AUC Value of each model

Consider the application of the model from the perspective of discrimination thresholds. Visualize four base classifiers as discrimination thresholds increase and four indicators (precision, recall, f1, queue rate) change, as shown in Figure 4.

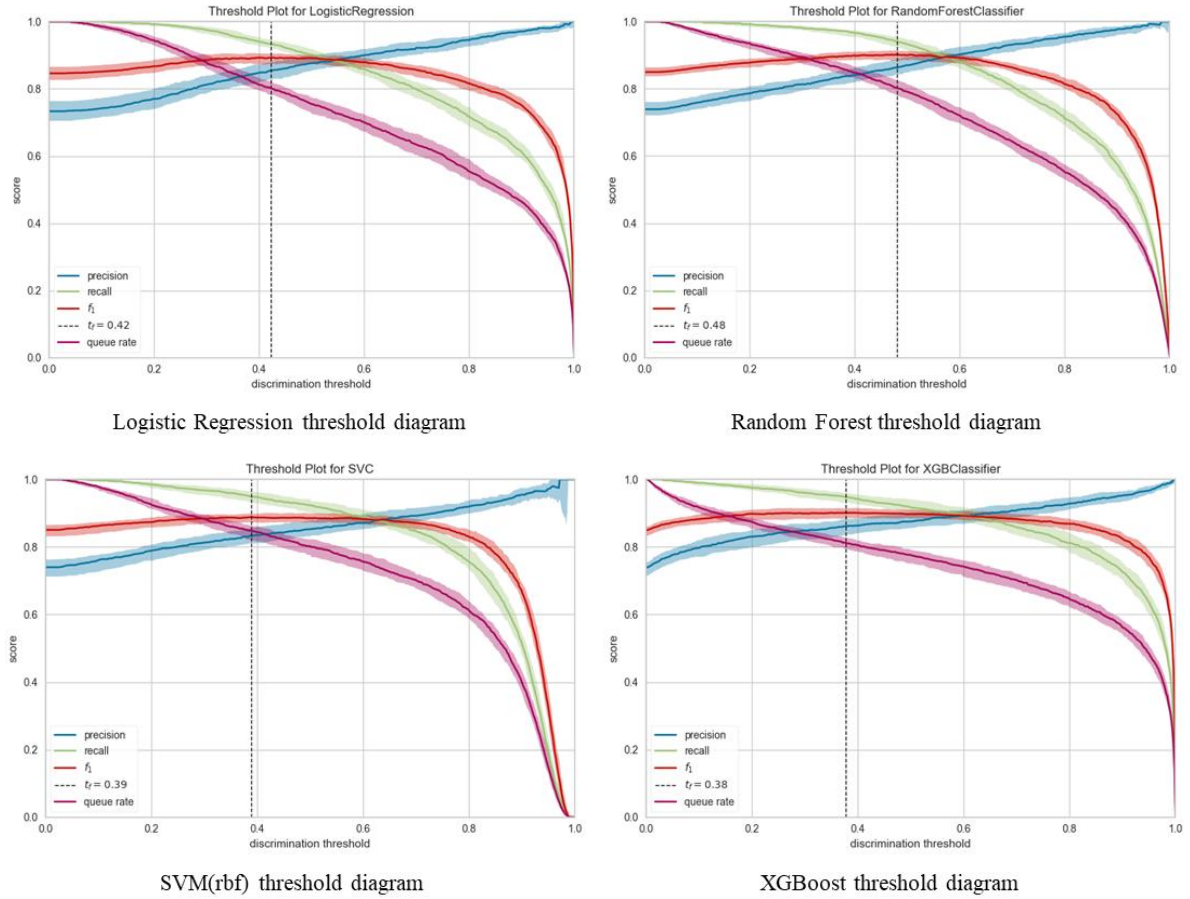

Figure 4. Threshold diagram of base classifier

The F1 value should be considered when evaluating model performance because of the unbalanced proportion of positive and negative samples in this paper's dataset (the proportion of positive samples to the total sample can be determined from the precision initial value on the y-axis). The visualizer has identified the discrimination threshold that should be set when F1 is at its maximum, as shown in Table 7.

Table 7. Corresponding discrimination threshold when F1 value is the maximum

Model	Logistic Regression	Random Forest	SVM(rbf)	XGBoost
Discrimination threshold	0.42	0.48	0.39	0.38

On the premise of maximizing the accuracy of the model, the basic learners in the fusion model are adjusted according to the following rules. 1. Increase the diversity of basic learners while excluding those with significantly lower accuracy and other learners. 2. Meta-learner selection of more complex learners 3. Prevent over-fitting. After many experiments, Compare the accuracy of 5-fold cross validation and test dataset, confirms that the model incorporating the two most accurate basic learners (XGBoost, Random Forest) has the best accuracy. The structure is shown in Table 8. (Because there are too many combinations, only the best combinations are shown).

Table 8. Stacking fusion model information

Base Classifier (Cross validation accuracy / Test accuracy)	Meta-Classifier	Fusion Model (Cross validation accuracy / Test accuracy)
XGBoost (0.8513/0.8640) Random Forest (0.8446/0.8445) XGBoost2 (0.8502/0.8636)	Random Forest	0.8584/0.8651

After fusing the above structure with overlay method, a new prediction model is obtained, which improves the prediction accuracy slightly compared with the basic learner in cross-validation and test set, 0.8584 and 0.8651.

6. Conclusion

The significance of this study is to predict customer turnover for telecommunication companies, which can bring huge profits for telecommunication companies. Therefore, the goal of this study is to establish a high accuracy telecommunication customer churn prediction model. Using a telecommunication customer churn data set from Kaggle website as sample data, the sample data is preprocessed. Then, 70% of the samples are randomly extracted from the sample data to form the training set, and the remaining sample data are to form the test set. The Random Forest, XGBoost, SVM and Logistic Regression algorithms are used to build the prediction model, and the Grid Search method is utilized to optimize the model hyper-parameters. Then the AUC, accuracy and other indicators of the models built by each single algorithm are compared to confirm that the prediction model built by XGBoost is the best of the four models, and its corresponding AUC value is 0.7966. The AUCs of the models built by Random Forests and SVM (rbf) are 0.7915 and 0.7887. Logistic Regression has the lowest AUC of 0.7808. On this basis, in order to optimize the accuracy of the model, the Stacking model fusion method is used to fuse the base classifier (the algorithm mentioned above), and then 5-fold cross-validation and test set validation is exploited to evaluate the accuracy of the fusion model. Retain the hyper-parameters of the model built by the optimized base classifier, compare the accuracy of cross-validation and test set validation under various combinations, and finally select the base classifier composed of Random Forest and XGBoost algorithm trained under the action of two different random tree seeds. The meta-classifier also consists of a fusion model composed of Random Forest. The accuracy of the fusion model is 0.8584 and 0.8651 under both validation methods, which is higher than that of any single base classifier algorithm. The use of this

integration strategy can provide new ideas of telecommunications companies to predict customer churns.

References

[1] Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. Expert systems with applications, 38(3), 2354-2364..

[2] Cox, L. A. (2002). Data mining and causal modeling of customer behaviors. Telecommunication Systems, 21(2), 349-381.

[3] Farquad, M. A. H., Ravi, V., & Raju, S. B. (2014). Churn prediction using comprehensible support vector machine: An analytical CRM application. Applied Soft Computing, 19, 31-40.

[4] Wei Hu (2014, May). Customer Churn Analysis in Mold Industry Based on Data Mining.

[5] Huang, Y., Zhu, F., Yuan, M., Deng, K., Li, Y., Ni, B., ... & Zeng, J. (2015, May). Telco churn prediction with big data. In Proceedings of the 2015 ACM SIGMOD international conference on management of data (pp. 607-618).

[6] Wu, Q., Ye, Y., Zhang, H., Ng, M. K., & Ho, S. S. (2014). ForesTexter: an efficient random forest algorithm for imbalanced text categorization. Knowledge-Based Systems, 67, 105-116.

[7] Yang Xiaofeng (2016). Multi-layer Random Forest for Telco Churn Prediction.

[8] Hung, S. Y., Yen, D. C., & Wang, H. Y. (2006). Applying data mining to telecom churn management. Expert Systems with Applications, 31(3), 515-524.

[9] Qi, J., Zhang, Y., Zhang, Y., & Shi, S. (2006, December). TreeLogit model for customer churn prediction. In 2006 IEEE Asia-Pacific Conference on Services Computing (APSCC'06) (pp. 70-75). IEEE.

[10] Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. Expert Systems with Applications, 39(1), 1414-1425.

[11] YIN Ting, MA Jun, QIN Xizhong, et al. Bayesian decision tree applying in forecasting customer churn. Computer Engineering and Applications, 2014, 50 (7): 125-128.

[12] Xu Shuqiao (2019). A Research on Bagging of Xgboost Classifiers for Prediction Churn in Telecom