2023 2nd International Conference on Data Analytics, Computing and Artificial Intelligence (ICDACAI) | 979-8-3503-3976-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICDACAI59742.2023.00076

Prediction of Customers' Behaviors Based on XGBoost Model

Yibin Zhang#,*
Tianjin College
University of Science and Technology
Beijing
Tianjin, China
ayx2276@163.com

Chunyan Shao#
Business School
Jianghan University
Wuhan, China
cy_shao2022@163.com

Chen Zou#
School of Mathematics and Science
Wenzhou University
Wenzhou, China
987261450@qq.com

#These authors contributed equally, *Corresponding author

***Abstract*—Useing XGBoost model to predict customer behavior is the main contribution of this paper. The data set describing customer behavior includes 37 characteristic variables and one response variable. We divided the 37 characteristic variables into numerical variables and classification variables, conducted data preprocessing respectively. The maximum and minimum normalization was applied to standardize the numerical variables, and the digital coding processing technology was used to digitize the classification variables. In the feature selection part, numerical variables are selected by variance filtering, and the chi-square test was used to choose categorical variables. In addition, Fisher score was used to determine the merging characteristics, and 14 variables were retained in the scoring results. According to the ratio of 7:3, the data set corresponding to the features filtered by Fisher's score is divided into training set and test set, and XGBoost model was established. Four significant variables that significantly affect the prediction results were found, and the accuracy of our model was verified to be 0.63, which indicated that the prediction has certain reliability and was very helpful for the company to predict customer behavior. This shows that the prediction has a certain reliability, which is very helpful for the company to predict customer behavior.**

Keywords—XGBoost, Variance Filtering, Chi-square Test, Fisher Score, ROC Curve

I. INTRODUCTION

For most companies, their customers' behaviors influence the companies' profits. In this paper, our work is to predict that if the customers would cheat the companies and make these companies loss their profits. Our data contains 700 customers' informations, 37 features and a goal variable of fraud or not.

In the use of the algorithm, the XGBoost algorithm is selected, and algorithms such as variance selection method, chi-square test, and Fisher score feature screening are added to assist in data processing, and more highly characteristic samples are created for data analysis. Look for indicators that play a more critical role in identifying fraud state.

II. DATA PREPROCESSING AND ANALYSIS

A. Data Introduction and Missing Value Processing

The data set used in this article comes from the Alibaba Cloud Tianchi Big Data Learning Competition. The data set contains 37 variables and 1 feature variable, the training set has 700 records, and the test set has 300 records. Among them, the 37 variables, there are 26 categorical variables, and 14 numerical variables the characteristic variable is "fraud", "1" means fraud, and "0" means not fraudulent. Observation shows that in the training set, there are abnormal data in the three columns of "collision_type" (collision type), "property_damage" (whether there is property damage), and "police_report_available" (whether there is a police record report), and the proportions of the abnormal data are respectively 0.176, 0.370, 0.353. These three variables are all categorical variables, and the number of categories is 2-3. If the filling method is used, large errors may be caused by inaccurate filling. If the deletion method is used, a large amount of data will be lost, which will seriously affect the effect of model training. Here, we choose to single out outliers as a category.

B. Exploring the Correlation between Variables

The correlation study of variables helps us to screen variables. If the correlation between multiple variables is high, it may cause interference between groups and affect the discrimination of eigenvalues; A strong similarity indicates that some variables are redundant and can be deleted. We use the functions in the pandas_profiling library for the analysis of the betrayal of the correlation. For categorical variables can be measured by Phil (φk) correlation coefficient or Cramér's V, Phil (φk) is a new, practical correlation coefficient that works consistently between categorical, ordinal, and interval variables, capturing nonlinear correlations and reverts to the Pearson correlation coefficient in the case of bivariate normal input distributions; Cramér's V is a measure of association for nominal random variables. This coefficient ranges from 0 to 1, with 0 indicating independence and 1 indicating complete association.

The Pearson coefficient has the characteristics of decentralization and normalization and has an excellent performance in reflecting the correlation between the target value and the eigenvalue [1]. Pearson's correlation is one of the most common indicators for measuring linear dependence. The Pearson correlation coefficient p is a measure of the linear dependence between two random variables. Its focus is on inferring the unknown p from actual observed data [2]. Therefore, for numerical variables, we use the Pearson correlation coefficient to measure the correlation, and the results are shown in Figure 1. It can be seen from Figure 1 that there is a strong correlation between age and the length of time you have been a customer. There is a strong correlation between the overall claim amount, injury claim amount, property claim amount, and auto claim amount.

979-8-3503-3976-5/23/$31.00 ©2023 IEEE
DOI 10.1109/ICDACAI59742.2023.00076

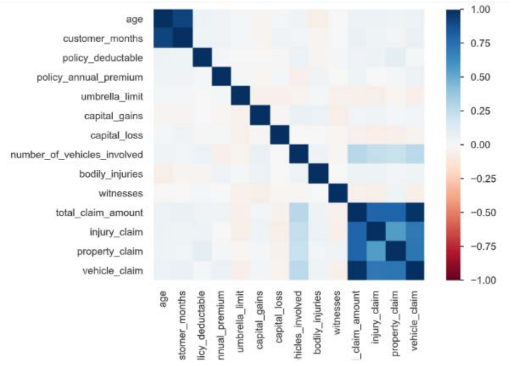

Figure 1: Heat map of the Pearson correlation coefficient of numerical variables

We use the Phil (φk) correlation coefficient for categorical variables to calculate and get the results shown in Figure 2. It can be seen from Figure 2 that there is a strong correlation between the four variables of the type of accident, the type of collision, the severity of the accident, and which local agency was contacted. The hour of the accident, the number of vehicles involved, and the type of accident, type of collision, there was also a strong correlation between accident severity and which local agency was contacted. The car model has a strong correlation with the car brand.

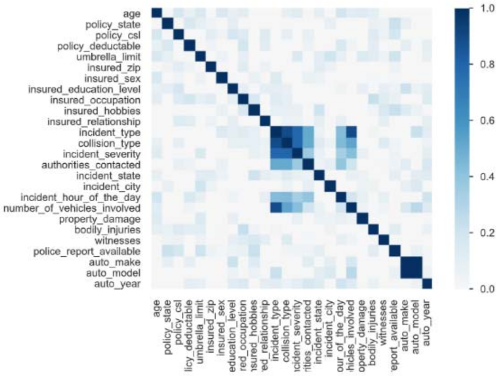

Figure 2: Heat map of correlation coefficient of categorical variable Phil (φk)

For categorical variables, we can also evaluate with Cramér V's empirical estimation, the results are shown in Figure 3. It can be seen from Figure 3 that there is a strong correlation between which local agency was contacted, the number of vehicles involved, the type of collision, the date of the accident, and the type of occurrence. The type of collision also has a strong correlation with the severity of the accident and the type of accident. There is also a strong correlation between the severity of the accident and the type of accident.

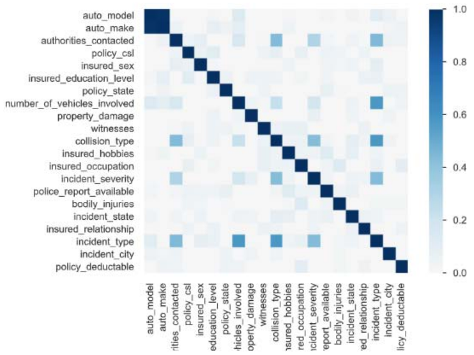

Figure 3: Cramér V heat map for categorical variables

From the 38 feature sets, 14 features that can be used for numerical feature screening are extracted. This sample contains data such as the age of the insurance client, insurance deductions, and annual premiums. The gap between these numerical data is too large. Therefore, to eliminate the impact of the range between variables, MIN-MAX standardization is used to preprocess the data.

C. Min-Max Data Standardization

The Min-Max data standardization method is the method to eliminate the influence of variable dimension and variation range. Specific method: Find the minimum and maximum values of each attribute, and map an original value x to a value X in the interval [0,1] through Min-Max standardization formula.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

X_{min} is the minimum value and X_{max} is the maximum value in the attribute.

The original data of 14 labels, after processing, the numerical variation range of each variable satisfies 0≤X'≤1, and both the positive index and the inverse index are transformed into positive indicators, and the direction of action is the same.

D. Data Preprocessing of Categorical Variables

Extract 26 classification features from 38 feature sets, and save these 26 feature sets in a new table file for further analysis. Many feature labels about date and classification that have no obvious correlation with insurance fraud prediction results are excluded from the total dataset.

LabelEncoder is to number discontinuous numbers or texts. For example, four different categories will be given different labels 1, 2, 3, and 4. Before performing pure categorical variable data feature screening, this method is often used to process data. To better deal with these 26 classification features in the future, the label encoding method in sklearn will be used to process each feature.

III. ATA INITIAL SCREENING

Definition: Measures the degree of deviation between a random variable and its mathematical expectation

Formula:

$$\sigma^2 = \frac{\sum(X-\mu)^2}{N} \tag{2}$$

X is the number in a set of data, μ is the mean value of the data, and N is the number of numbers in the data.

Variance is a measure of the degree of dispersion of a variable (that is, the degree to which the data deviates from the average value). The greater the variance of a variable, the greater its degree of dispersion, which means the greater the contribution and effect of this variable to the model, so variables with large variance should be retained, and on the contrary, meaningless features should be eliminated.

To select a suitable number of numerical feature sets, the selected variance threshold is 0.062. That is, the following numerical variable features are selected: customer_months, policy_deductable, capital_gains, capital_loss, number_of_vehicles_involved, bodily_injuries, witnesses, injury_claim, vehicle_claim.

Classification feature screening based on chi-square test

Karl Pearson's family of chi-square tests represents one of the most commonly used statistical analyzes to answer questions about associations or differences between categorical variables. The chi-square test is a nonparametric statistic, also known as a distribution-free test. The categorical variables in the sample data conform to the non-parametric test in the chi-square test, that is, the measurement levels of all variables are nominal or ordinal [3].

Calculate the chi-square statistic formula:

$$X^2 = \sum \frac{(f_0-f_e)^2}{f_e} \tag{3}$$

f_0 is the actual value and f_e is the theoretical value.$X^2$is used to measure the difference between the actual value and the theoretical value.

Substituting the categorical data into the chi-square test model, the following analysis diagram is obtained:

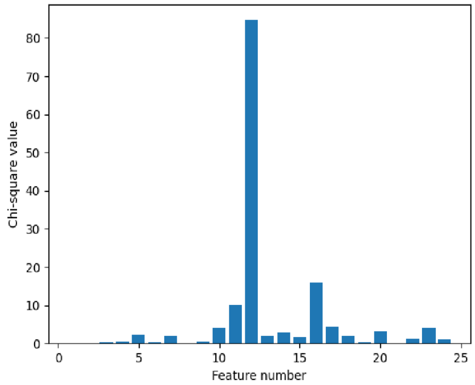

Figure 4: Chi-square values

In Figure 4, the ordinate represents the chi-square value of the classification feature and the target feature in the model, and the abscissa is the position code of the 26 features in the table, representing 26 classification features. The chi-square test assumes that "two events are independent of each other", that is, when the calculated chi-square value is larger, the probability of two events being independent is smaller and the correlation is greater. According to the chi-square value, the following relevant labels are extracted from it: incident_type, collision_type, incident_severity, incident_hour_of_the_day, number_of_vehicles_involved, witnesses.

IV. Secondary screening of data

A. One Hot Encoding

One-hot encoding, that is, One-Hot encoding, also known as effective encoding, uses N-bit state registers to encode N states, each state has its independent register bit, and at any time, only one valid. In a thermal effect feature, we created a new variable. Each category uses a binary variable to control 0 or 1. Here, 0 means that it does not exist, and 1 means that the category exists [4]. After reintegrating the numerical variable and the categorical variable, the data needs to be preprocessed twice. One-hot encoding will be used to process the categorical data with a small amount of classification, that is, the data with the following labels: incident_type, collision_type.

B. Fisher Scores

Fisher scores are a filter-based supervised feature selection method with feature weights. As a feature correlation criterion, the Fisher scoring model has many advantages, such as reduced calculation, higher precision, and stronger operability, compared with using supervised learning for feature selection and can effectively reduce the time and space complexity [5]. In general, the larger the Fisher score, the better the selected features. Like the central limit theorem, Fisher's method is also a rule that a distribution is transformed into another distribution through a certain operation, but the original distribution here is no longer random but must be a random distribution independent of each other. The distribution is a chi-square distribution, and the formula is as follows:

$$X_{2k}^2 \sim -2\sum_{i=i}^{k} \ln(p_i) \tag{4}$$

Take K samples from the independent random distribution, multiply the logarithm by -2 and add the sum respectively, and the new random variable follows the chi-square distribution of 2k degrees of freedom

The effect of each feature in the training set is determined by the Fisher Score algorithm. Fisher Score creates a weighting problem based on the classification effectiveness of each feature. A subset of the training data is obtained according to the k features with the largest weights [6]. The following figure can be obtained from the Fisher fraction model:

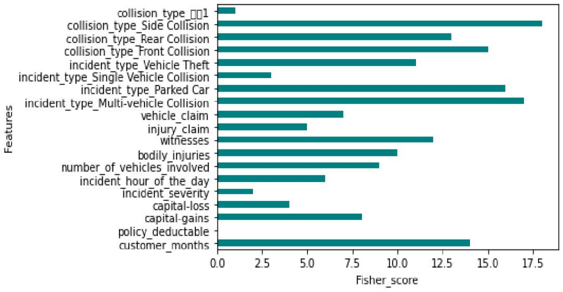

Figure 5: Fisher Scores

According to the data in Figure 5, the following label data with small scores will be excluded: collision_type_unknown1, incident_type_Single Vehicle Collision, incident_severity, capital_loss, policy_deductable

B. Divide test set and training set

The original data set is directly divided into two mutually exclusive data sets by a certain ratio, namely the training set and the test set. In this data experiment, the training set and the test set will be divided in a ratio of 7:3, and finally, 489 samples are selected from 700 samples as the training set and 209 samples are used as the test set.

C. Random Oversampling

Another way to solve the problem of unbalanced black-and-white samples is called oversampling [7]. There are only 489 pieces of data in the training set table, and the result of whether the insurance constitutes fraud, yes or no, the two data samples are not balanced. The problem of unbalanced data set is a special kind of classification problem. Random oversampling augments the number of minority class data points in the training set by randomly duplicating existing minority class members. Although an oversimplification, random oversampling has performed well in empirical studies [8]. So far, we have adopted a random oversampling method to filter the data set. For the statistics of fraud results in the training set, non-fraud is the majority class sample with a sample size of 365; fraudulent class is the minority class sample with a sample size of 125. By the method of random oversampling, the number of samples of both categories is converted to 365.

V. Model construction

A. XGBoost Algorithm

XGBoost is a training model with remarkable effects in dealing with classification problems. It is not only efficient in training samples but also very flexible and can be directly transplanted, which can well solve large-scale data problems in all walks of life. The XGBoost model uses an additive training method to optimize the objective function. This means that the optimization process in the latter step is eased. In this training, we used several parameters that have a great influence on particle swarm optimization: learning rate (learning_rate), number of trees (n_estimators), depth of the tree (max_depth), minimum weight of leaf nodes (min_child_weight), Gamma value (gamma), build a decision tree (subsample). The impact of this system is widely recognized in many machine learning and data mining challenges. Taking the challenge hosted by the machine learning competition website Kaggle as an example, among the 29 challenge-winning solutions published on the Kaggle blog in 2015, 17 solutions used the XGBoost model. Among these solutions, 8 use XGBoost alone to train the model, which fully demonstrates the superiority of the XGBoost model [9].

The principle of the XGBoost algorithm is to divide the original data set into multiple sub-datasets, randomly assign each sub-data set to the base classifier for prediction, and then calculate the results of the weak classification according to a certain weight, and predict the final result once. . More generally, the principle of XGBoost is like a relay race. Each runner is selected from many runners who are not so fast, and these selected runners are generally the fastest. , and finally, win the game together [10]. The XGBoost algorithm introduces the balance between the performance of the model and the operation speed into the objective function, and performs second-order Taylor expansion on it when solving the objective function, to speed up the solution and reduce the running time of the model; at the same time, it introduces regularization to control the complexity of the model to avoid excessive Fitting [11].

B. ROC Curve

The ROC curve uses positive and negative samples to measure the recognition accuracy, and it is relatively easy to find out the recognition ability of performance when any threshold value is used. The curve can present the data simply and intuitively, and judge the data better. Combined with the above Fisher score feature screening, the 14 feature label-related training sets are substituted into the XGBoost model for training, and then the data is predicted and compared in the test set, and the following prominent feature importance is obtained and the ROC curve.

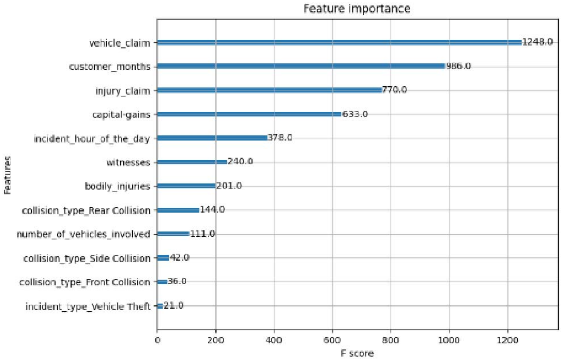

Figure 6: Feature Importance of XGBoost

As shown in Figure 6, in the feature importance graph, the car claim amount score is 1248, the customer duration score is 986, the injury claim amount score is 770, and the capital gain score is 633. The importance scores of these four variables are much higher than other variables.

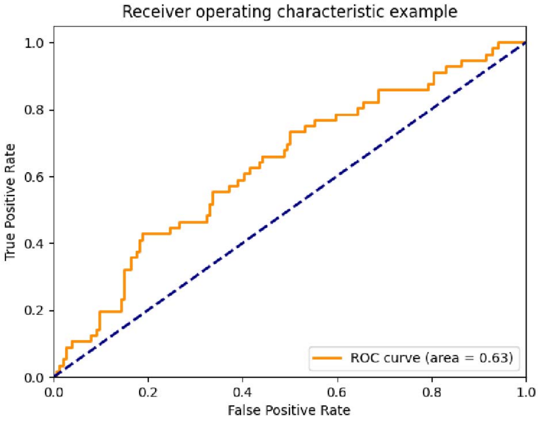

Figure 7: ROC curve

According to the importance of each feature shown in Figure 6, the four items of injury claim amount, auto claim amount, the length of time the party has been an insurance customer, and capital gains have a greater correlation with whether this insurance constitutes fraud. ROC plots are great for organizing classifiers and visualizations. An advantage of such graphs is that they can highlight visualization and organize classifier performance without regard to distribution or error costs [12]. The ROC curve finally obtained by the XGBoost model is shown in Figure 7, where AUC, the area under the ROC curve, is 0.63. The model achieved an accuracy of 0.63. That is to say, the model can be used as a predictive analysis of customers' behaviors, and its predictive results still have certain reference values.

VI. Summary

From the feature importance diagram made in the XGBoost model, it can be concluded that the four characteristic indicators of the number of injury claims, the number of auto claims, the length of time the parties have been customers, and capital gains have the greatest impact on whether fraud happens.

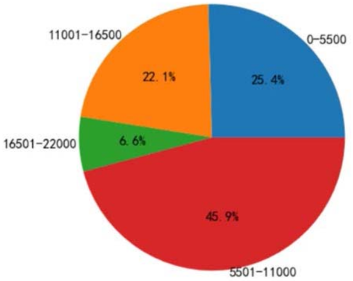

Figure 8: Amount of Injury Claims vs Number of Fraudulent

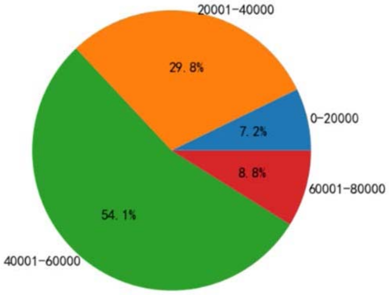

Figure 9: Auto Claim Amount vs Fraudulent Number

From Figures 8 and 9, the number of injury claims and the number of auto claims have a greater impact on whether a customer's default is constituted. If the customer misrepresents the number of injury claims and car claims, that is, exaggerates the authenticity of the data, it may lead to an abnormal default amount. Among the fraudsters, 71.3% of the fraudsters reported the damage claim amount in the range of 0-11,000; 83.9% of the fraudsters reported the car claim amount as more than 40,000. Therefore, the company should pay close attention to the people who fill in the lower amount of the injury claim and the higher amount of the car claim, because the probability of fraud is higher in this group of people.

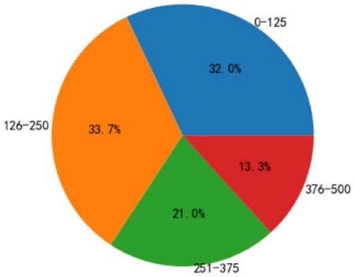

Figure 10: Becoming a customer and the number of fraudsters

It can be seen from Figure 10 that among the fraud related to the length of time the parties have been customers, the number of people who have been customers for less than 250 months accounted for 65.7% of the fraudulent population. It is often in the news that opportunists create artificial disasters for short-term profit. Adding fraud prevention measures to customers who have been insured for less than a certain number of months will help reduce the occurrence of fraud.

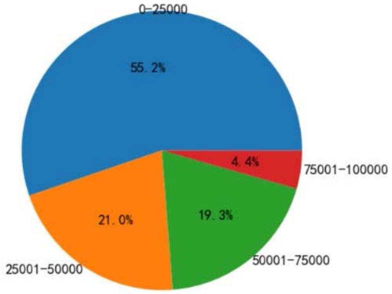

Figure 11: Capital Gains vs Number of Fraudsters

Capital gains, also known as capital profits, are the portion of an asset that people sell for more than they paid for it. As shown in Figure 11, 76.2% of fraudsters have capital gains of less than 50,000, indicating that the defaulting customer will not ask for high compensation.

It can be seen from Figure 10 and Figure 11 that people who have been customers for a short time and whose capital gains are not high need to pay special attention, and fraudulent behavior occurs more frequently among these groups.

VII. Conclusion

The main research results of this paper are as follows:

1) The XGBoost algorithm had a good performance, and the feasibility of the algorithm in the application is verified.

2) Among the 37 features about the customers, we used the feature importance map generated by the XGBoost model to screen out the most important four features. They are the amount of the injury claim, the amount of the auto claim, time as a client, and capital gain. And it can be found that fraud is more likely to occur when the number of injury claims is in the middle and low range; fraud is more likely to occur when the amount of automobile claims is in the middle range; Fraud is more likely to occur in the low-to-middle range. Overall, it can be shown that fraudulent insurance behaviors are more likely to occur when the insurance application period is shorter, personal injuries are minor, car injuries are

moderate, and capital gains are at the low-to-medium end.

REFERENCES

[1] Sanqiang Chang, Chuiri Zhou. Personal Credit Prediction Based on Feature Optimization and Boosting Algorithm [J/OL]. Computer System Applications:1-8[2022-11-26]. DOI: 10. 15888/ j. cnki. csa. 008959.

[2] Ly, Alexander, Maarten Marsman, and Eric-Jan Wagenmakers. "Analytic posteriors for Pearson's correlation coefficient." Statistica Neerlandica 72.1 (2018): 4-13.

[3] McHugh, Mary L. "The chi-square test of independence." Biochemia Medica 23.2 (2013): 143-149.

[4] Daoud, Mwamba Kasongo, and Inwhee Joe. "A deep-learned embedding technique for categorical features encoding." IEEE Access 9 (2021): 114381-114391.

[5] Sun, Lin, et al. "Feature selection using Fisher score and multilabel neighborhood rough sets for multilabel classification." Information Sciences 578 (2021): 887-912.

[6] Aksu, Doğukan, et al. "Intrusion detection with comparative analysis of supervised learning techniques and fisher score feature selection algorithm." International symposium on computer and information sciences. Springer, Cham, 2018.

[7] Dan Sun, Weili Shi, Lanxiang Rao, Shasha Meng, Xiaoming Guo, Yilun Li. Credit Card Fraud Detection Method Based on Improved Mixed Sampling and XGBoost Algorithm [J]. Computer and Modernization,2022(09):111-118.

[8] Liu, Alexander, Joydeep Ghosh, and Cheryl Martin. "Generative Oversampling for Mining Imbalanced Datasets." DMIN. 2007.

[9] Chen, T. and Guestrin, C. (2016)XGBoost: a scalable tree boosting system.ArXiv.1603 arXiv:1603.02754.

[10] Xiang Li. Multi-factor quantitative stock selection scheme planning based on XGBoost algorithm [D]. Shanghai Normal University, 2017.

[11] Ziwei Shang. Based on the SMOTE+ENN and Application research of ECG assisted diagnosis and treatment with random forest [D]. Shanghai: Donghua University, 2019.

[12] Fawcett, Tom. "An introduction to ROC analysis." Pattern recognition letters 27.8 (2006): 861-874.