2022 2nd International Signal Processing, Communications and Engineering Management Conference (ISPCEM) | 978-1-6654-9271-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISPCEM57418.2022.00061

Customer Churn Prediction in Bank Based on Different Machine Learning Models

Xiaofeng Li, Zhongwei Chen*
School of Information Engineering
Guangxi University of Foreign Languages
Nanning, 530222, Guangxi, China
Corresponding Author*: 306567550@qq.com

***Abstract*—With the rise of internet finance, the competition in the banking industry has become increasingly fierce. Preventing the loss of customers and retaining old customers has become an important concern of major banks. Firstly, according to an existing open dataset, this paper makes a descriptive statistical analysis of each feature, and uses Logistic, Random Forest and Scv models to predict customer churn, such as AUC curves, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Error (MSE). The better-performing models listed above are then chosen. Finally, Svm is selected as the best performance model, and according to descriptive statistics and feature importance, some suggestions are put forward for banks to retain customers.**

***Index Terms*—customer churn, Logistic, random forest, Svm, AUC curve**

I. Introduction

Customer churn refers to the fact that the customers of the bank no longer participate in the original business, purchase repeatedly or terminate the original products or services. The sudden emergence of internet finance and the intensification of competition in traditional banking have made it increasingly important for banks to tap into their own potential, attract high-quality customers, and prevent customer churn. Studies have found that the cost of acquiring a new customer can be up to 5 to 6 times higher than the cost of maintaining an existing one. Thus, if an enterprise loses customers, the cost of acquiring new ones can be significant, and the profits generated by these new customers may not be as good as those of the old ones. Therefore, no matter which industry, more and more attention is paid to customer churn management. Predicting the potential loss of customers, effectively retaining and caring for customers is one of the important issues concerned by all enterprises.

To ensure that users do not have to worry about these issues, Zhang, W.H [1] proposed a solution for modeling and forecasting mobile customer churn by considering the changing characteristics of mobile customer churn. To accomplish this, they collected historical data on mobile customer churn and applied the model to specific mobile customer churn forecasting. The results show that the modeling speed of this model is faster than other mobile customer churn prediction models, and it can obtain better mobile customer churn prediction results. Yang, S. and Yue, J.J [2] In the customer churn prediction system of banks, unknown customer data are often used to predict customer service information, so as to provide the basis for the bank's future business strategy. In the prediction of customers, they often need to classify the mining attributes of some classification rules [3]. The training data of bank customer churn analysis based on random forest is applied to various models. A range of techniques was used, including random forest, decision tree, support vector machine (SVM), logistic regression, and extreme learning machine (ELM). In this process, we found the best function of each model by using gradient descent and adjusting network parameters (such as basic learning rate). Then, the test data are also applied to each model we study to test the accuracy of each model. Meng X. and Cai S. and Du K. and Kou J [4] used data mining and statistical techniques to analyze and identify the factors that affect customer churn of commercial banks, and established a customer churn prediction model. The empirical results show that the disturbance prediction technique used has achieved satisfactory prediction results. Zhang Y [5] Customer churn is a major problem faced by mobile communication operators. Modeling and customer churn prediction analysis through data mining can not only retain potential churn customers, but also enable operators to optimize products and services and improve customer management strategies [6]–[10].

Many scholars have used different methods to do in-depth research. The data mining of bank customer churn is analyzed and studied in this study using the logistic, random forest, and SVM models. We evaluate and mine the customer data set in accordance with the features of each user's data piece to investigate why consumers lose their characteristics. Through data analysis and machine learning model construction, The experimental outcomes demonstrate that, when compared with Logistic and Random Forest approaches, our method offers banks a means for accurately and dependably predicting customer turnover by modeling a sizable quantity of data.

II. Concept and theoretical basis

A. Accurately prevent customer loss

According to studies, "the profit produced by an existing customer is more than 10 times that of a new customer, while the expense of establishing a new client is 3-8 times that of keeping an old customer." For every 5% drop in

978-1-6654-9271-3/22/$31.00 ©2022 IEEE
DOI 10.1109/ISPCEM57418.2022.00061

user engagement, the profit of the enterprise will drop by 25%. The price of obtaining new users is far more expensive than the price of keeping current customers, and the price of gaining lost consumers is considerably more expensive. In fact, after a series of tests and studies, it has been proved that the loss of users is the biggest damage to the company's profits.

Any bank's resources are limited. In order to get high returns while maintaining low costs, banks should tilt the limited resources to key customers. The premise of realizing these strategies is to analyze the value of customers and understand their value differences. According to the customer value and its proportion in all customers, customers can be divided into VIP customers, major customers, ordinary customers and small customers in the value dimension. Differentiated service strategies are adopted for different customer segments to reduce the customer churn rate.

B. Random forest model

Random forest classifiers in machine learning are made up of many decision trees, with the categories produced by individual trees serving as the basis for the output categories. This approach was developed by Leo Breiman and Adele Cutler, who also named it "Random Forests." It combines Tin Kam Ho's "random subspace technique" from Bell Laboratories with Breiman's "bootstrap aggregating" concept to create a collection of decision trees.

C. Logistic regression model

Given that logistic regression is an extended linear model, multiple linear regression analysis and logistic regression share many characteristics. Although binary is more popular and simpler to understand, logistic regression may process many categories if the dependent variable is multi-classified. Multiple categories can also be handled using the soft-max approach. Actually, binary logistic regression is the approach that is most frequently utilized.

Directly deducing the linear regression model into the logistic regression results in a generic non-linear connection and various value ranges for the two sides of the equation. Since the dependent variable in the logistic equation is a binary variable, the estimated value range for a given probability as the dependent variable is 0-1, but the estimated value range on the right side of the equation is infinite or infinitesimal. To address this issue, regression analysis using logs was developed.

D. Support vector machine model

Support Vector Machines (SVMs) are a type of supervised learning classifier that use the maximum-margin hyperplane as their decision boundary for learning samples. SVMs calculate empirical risk using the hinge loss function and optimize structural risk by adding a regularization term to the solution system. They are known for being sparse and reliable classifiers, and are capable of performing kernel-based nonlinear classification. SVMs were first developed by Soviet-era academics Alexander Y. Lerner and Vladimir N. Vapnik using the generalized portrait method, and later refined by Vapnik and Alexey Y. Chervonenkis, who created a linear SVM with strict margins. Since then, SVMs have been gradually integrated into statistical learning theory through the investigation of the maximum edge distance decision boundary in pattern recognition, the development of technology for solving planning problems with slack variables, and the introduction of the VC dimension (Vapnik-Chervonenkis dimension).

III. Problems to be studied

A. Description of data set

The data used in this paper comes from the bank data platform, which mainly affects the related factors of customer churn, with a total of 3,033 pieces of data. The data sets include RowNumber, CustomerId, CreditScore, geographical location, gender, age, tenure, balance, the number of products purchased by customers through the bank, whether customers have credit cards, less possibility of active customers leaving the bank, salary estimation, and whether customers leave the bank. As can be seen in Table 1, various field data is employed as the study object for predictions. We apply the same strategy to handle the data in order to correctly conduct the future experiment and compare the models. First, we must determine whether the data collection has any missing values and then zero out such values. Table I displays the specific field data, while Figure 1 displays the correlation coefficient between these fields.

TABLE I
Data information

type	data1	data2	.....	data3033
RowNumber	1	2	......	3033
CustomerId	15647311	15619304	......	15701354
Surname	Hill	Onio	......	Boni
CreditScore	608	502	......	699
Geography	Spain	France	......	France
Gender	Female	Male	......	Male
Age	41	18	......	42
Tenure	2	1	......	1
Balance	159660.8	535867	......	59660.8
NumOfProducts	1	3	......	2
HasCrCard	0	1	......	0
IsActiveMember	1	0	......	0
EstimatedSalary	112542.58	113931.57	......	93826.63
Exited	1	0	......	0

B. Experimental process of Logistic regression model

Logistic regression (binomial Logistic regression model) is a classification model, which is represented by$P(Y \mid X)$of conditional probability distribution, and its form is parameterized logistic distribution. Here, the random variable x is a real number, and the random variable y is 1 or 0. We estimate the model parameters by supervised learning.

The conditional probability distribution shown below represents the logistic regression model. In this study, we used logistic regression to classify data. The logistic regression model is defined by Equations 1 and 2, where x is the input, Y is the output, w and b are the parameters, w is

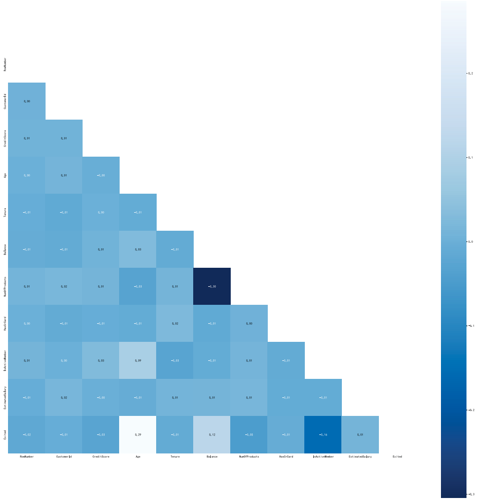

Fig. 1. correlation coefficient

the weight vector, b is the offset, and $w \cdot x$ is the dot product of w and x. Given an input example x, $P(Y = 1 \mid x)$ and $P(Y = 0 \mid x)$ can be calculated using Equations 1 and 2. Logistic regression then compares these two conditional probability values and classifies the example x into the category with the larger probability value.

For convenience, the weight vector and input vector can be expanded and denoted as w and x, respectively, such that $w = \left(w^{(1)}, w^{(2)}, \cdots, w^{(n)}, b\right)^{\mathrm{T}}$, $x = \left(x^{(1)}, x^{(2)}, \cdots, x^{(n)}, 1\right)^{\mathrm{T}}$. The current form of the logistic regression model is given by Equations 3 and 4. These equations determine the category by comparing the size of $P(Y = 1 \mid x)$ and $P(Y = 0 \mid x)$ (similar to the Naive Bayes classifier). In this form, b can be thought of as $w_0 x_0$, where $x_0 = 1$. This is essentially a binomial distribution, which follows the distribution law of the binomial distribution.

The odds of an event, or the probability of it occurring divided by the probability of it not occurring, is a characteristic of the logistic regression model. The likelihood of an event occurring with probability p is $\frac{p}{1-p}$, and the log odds or logit function is given by Equation 5. For logistic regression, this can be obtained from Equations 3 and 4 as $\log \frac{P(Y=1|x)}{1-P(Y=1|x)} = w \cdot x$. This indicates that the likelihood of the output Y being equal to 1 in the logistic regression model is a linear function of the input x. Alternatively, the logistic regression model can be seen as representing the logarithmic likelihood of output $Y = 1$ as a linear function of input x.

The experimental results of using our data set with the logistic regression model can be found in table II-III and figure 2.

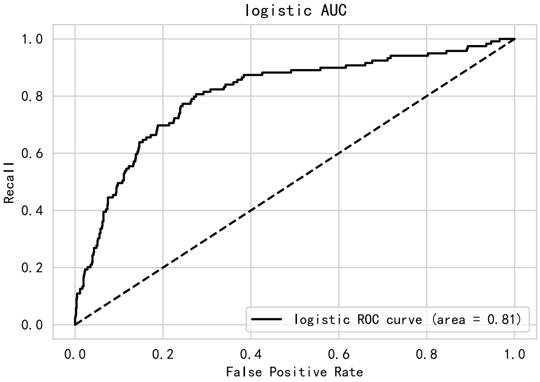

Fig. 2. Logistic AUC

TABLE II
Logistic Model Evaluation Form

type	precision	recall	f1-score	support
False	0.910	0.936	0.93	883
True	0.414	0.243	0.30	119
accuracy			0.87	1002
macro avg	0.627	0.519	0.61	1002
weighted avg	0.835	0.837	0.86	1002

TABLE III
Logistic index

MAE	MSE	RMSE	accuracy
0.135	0.192	0.105	0.87

C. Experimental process of stochastic forest model

Two anticipated σ^2 values from the same population, each with a variance of $\widehat{T}^{*(b)}\left(X_0\right)$ in the random forest's fundamental theory.Because they are independent of each other, the variance of their mean$\hat{f}_{bag}^{*}\left(\boldsymbol{X_0}\right)$decreases to σ^2/R

It can be challenging to become independent. The variance of the expected value is equal to the product of the correlation coefficient and the variance, plus the product of the difference between 1 and the correlation coefficient, and the variance divided by the number of iterations (B) when the correlation coefficient of the expected values is equal to P. As B increases, the second term approaches zero, leaving only the first term.

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \tag{1}$$

With the increase of b, the second term tends to zero, leaving only the first term.

$$\operatorname{var}(\bar{Z}) = \operatorname{var}\left(\frac{1}{N}\sum_{i=1}^{N} Z_i\right) \tag{2}$$

$$\operatorname{var}(\bar{Z}) = \frac{1}{N^2}\left[N\sigma^2 + N(N-1)\rho\sigma^2\right] \tag{3}$$

$$\operatorname{var}(\bar{Z}) = \frac{\sigma^2 + (N-1)\rho\sigma^2}{N} \tag{4}$$

$$\operatorname{var}(\bar{Z})=\rho\sigma^2+\frac{1-\rho}{N}\sigma^2 \tag{5}$$

To create a random forest with B trees, a bootstrapped sample (Sb*) is created by resampling the data set with a sample size of N. A regression tree or classification tree is then created based on the bootstrapped sample (Sb*), and it is constantly developed from the root node of the decision tree until the pre-pruning requirements are satisfied. A random subset of input variables is created by randomly selecting M input variables from P input variables, with M equal to the square root of P or the logarithm of P to the base 2. The ”best” grouping variables are then chosen from the subset, and two sub-nodes are created to produce the random forest with the B-tree (Tb*). The prediction result for regression is the average of the B trees in the random forest, and the classification result is the class with the highest number of votes $m=[\sqrt{p}]$ or $m=[\log_2 p]$.

Find the ”best” grouping variables currently available from B, create two sub-nodes, and produce the random forest with the B-tree $\left\{\widehat{T}^*(b)\right\}^B$.

The regression prediction result is$\hat{f}_{rf}^*\left(\boldsymbol{X}_{\mathbf{0}}\right)=\frac{1}{B}\sum_{b=1}^B \widehat{T}^{*(b)}\left(\boldsymbol{X}_{\mathbf{0}}\right)$The classification result is the class with the highest number of votes.

The experimental results of applying the random forest model to our data set can be seen in Table IV-V and Figure 3.

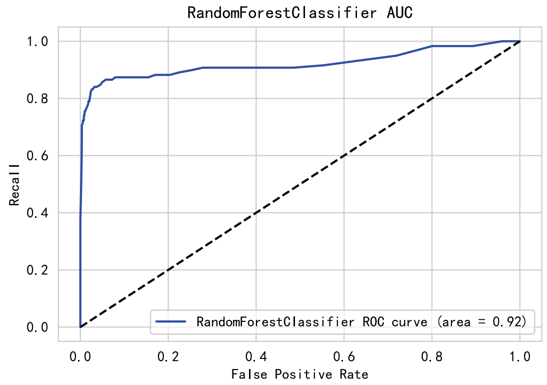

Fig. 3. RandomForestClassifier AUC

TABLE IV
RANDOMFORESTCLASSIFIER MODEL EVALUATION FORM

type	precision	recall	f1-score	support
False	0.95	0.99	0.97	883
True	0.94	0.63	0.75	119
accuracy			0.95	1002
macro avg	0.94	0.81	0.86	1002
weighted avg	95	0.95	0.95	1002

TABLE V
RANDOMFORESTCLASSIFIER INDEX

MAE	MSE	RMSE	accuracy
0.106	0.093	0.122	0.95

D. Experimental results of support vector machine model

Support Vector Machine (SVM) is a popular binary classification method that has shown significant benefits due to the improved stability of the discovered hyperplane. In the binary classifier dataset $D=\left(x^{(n)},y^{(n)}\right)n-1^N$, where $\boldsymbol{y}n\in+1,-1$, if the two types of samples are linearly separable by the equation $w^{\top}\boldsymbol{x}+b=0$, the samples can be separated by the inequality $y^{(n)}\left(\boldsymbol{w}^{\top}\boldsymbol{x}^{(n)}+b\right)>0$. The distance of each sample in the dataset D from the hyperplane of separation is given by:

$$\gamma^{(n)}=\frac{\left|\boldsymbol{w}^{\top}\boldsymbol{x}^{(n)}+b\right|}{|\boldsymbol{w}|}=\frac{\boldsymbol{y}^{(n)}\left(\boldsymbol{w}^{\top}\boldsymbol{x}^{(n)}+b\right)}{|\boldsymbol{w}|} \tag{6}$$

The goal function of formula (14) is expressed as a convex optimization problem that aims to find the largest interval division hyperplane.

$$1-y^{(n)}\left(\boldsymbol{w}^{\top}\boldsymbol{x}^{(n)}+b\right)\leq 0,\quad \forall n\in\{1,\cdots,N\} \tag{7}$$

The Lagrange function of formula is obtained through the Lagrange multiplier technique.

$$\Lambda(\boldsymbol{w},b,\lambda)=\frac{1}{2}\|\boldsymbol{w}\|^2+\sum^N\lambda_n\left(1-y^{(n)}\left(\boldsymbol{w}^{\top}\boldsymbol{x}^{(n)}+b\right)\right) \tag{8}$$

Where $\lambda_1\geq 0,\cdots,\lambda_N\geq 0$ is Lagrange multiplier.Calculate $\Lambda(\boldsymbol{w},b,\lambda)$ derivatives about $\boldsymbol{w}_{\text{and}}$ $\boldsymbol{ib}$, and make them equal to 0. Gets the Lagrange dual function.

$$\Gamma(\lambda)=-\frac{1}{2}\sum_{n=1}^N\sum_{m=1}^N\lambda_m\lambda_n y^{(m)}y^{(n)}\left(\boldsymbol{x}^{(m)}\right)^{\top}\boldsymbol{x}^{(n)}+\sum_{n=1}^N\lambda_n \tag{9}$$

Consequently, Table VI-VII and figures 4 & 5 shows the experimental findings produced by feeding our accomplishment data into the support vector machine model.

TABLE VI
SVM MODEL EVALUATION FORM

type	precision	recall	f1-score	support
False	0.98	1.00	0.99	883
True	0.99	0.84	0.91	119
accuracy			0.98	1002
macro avg	0.98	0.92	0.95	1002
weighted avg	98	0.98	0.98	1002

TABLE VII
SVM INDEX

MAE	MSE	RMSE	accuracy
0.095	0.072	0.95	0.98

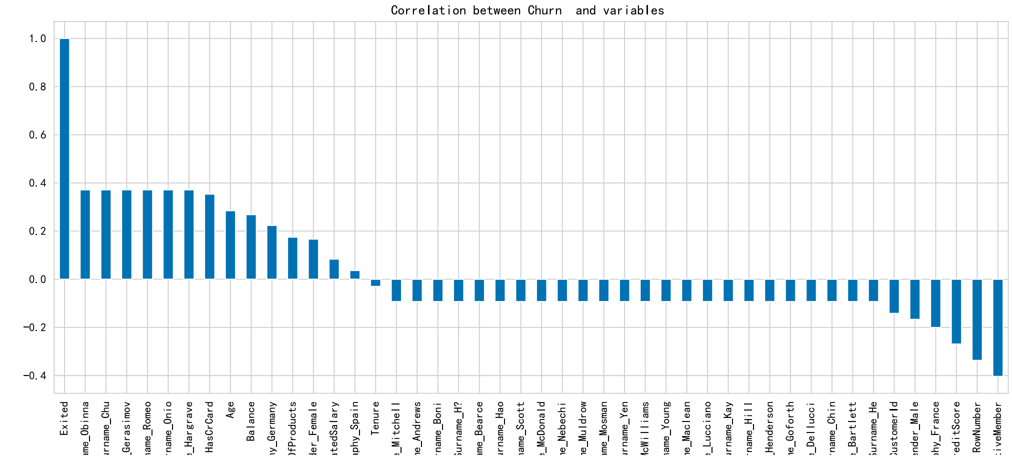

Fig. 4. Correlation between Churn and variables

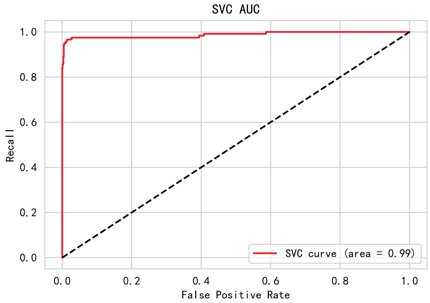

Fig. 5. SVM AUC

E. Comparison of experimental results of models

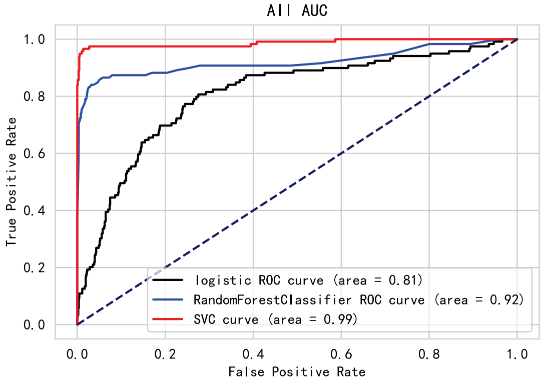

Fig. 6. All AUC

TABLE IX
MODEL PERFORMANCE COMPARISON RESULTS

	Logistic	Random forest	Svm
MAE	0.135	0.106	0.095
MSE	0.192	0.093	0.072
RMSE	0.105	0.122	0.095
accuracy	0.871	0.95	0.98

The results from table VIII-IX show that Svc outperforms random forest and Logistic in terms of MAE and MSE, as well as RMSE. The AUC figure 6 demonstrates the effectiveness of the models based on their true positive and false positive rates. The ROC curve for an excellent classifier is close to the top left corner of the unit square, and a high true positive rate near the Y axis indicates strong model performance. The ROC curve's points being close to the Y axis indicates minimal classification error and a successful threshold. AUC ranges from [0.5,1], with a higher value indicating better model performance. The AUC index for logistic regression is 0.87, while it is 0.95 for random forest and 0.98 for SVC.

Our analysis shows that the Svm model performs exceptionally well in predicting customer churn. The Svc model demonstrates the highest levels of RMSE, correct rate, MAE and MSE.

F. Comparison of experimental results of models

The connection diagrams between customer turnover and other dimensions shown in Figures 4 and 6 allow us to evaluate the relevant elements that impact customer churn in addition to using the model to forecast customer churn. Age and Balance had connection values of 0.29 and 0.12, making them the most important variables for predicting customer attrition, according to the correlation study.

IV. CONCLUSION

The prediction and analysis of customer churn is rooted in market competition. By establishing a churn prediction

TABLE VIII
EVALUATION FORM OF MODEL COMPARISON RESULTS

type	precis-Log	recall-Log	f1-sc-Log	precis-Rfo	recall-Rfo	f1-sc-Rfo	precis-Svm	recall-Svm	f1-sc-Svm	support
False	0.90	0.96	0.93	0.95	0.99	0.97	0.98	1.00	0.99	883
True	0.44	0.23	0.30	0.94	63	0.75	0.99	0.84	0.91	119
accuracy			0.87			0.95			0.98	1002
macro avg	0.67	0.59	0.61	0.94	0.81	0.86	0.98	0.92	0.95	1002
weighted avg	0.85	0.87	0.86	0.95	0.95	0.95	0.98	0.98	0.98	1002

model, banks can control customer churn from the source and prevent it before it happens, thus effectively preventing customer churn. In the increasingly fierce market competition, preventing customer churn is not a passive behavior in bank management, but a marketing strategy throughout bank management. Preventing customer churn is as important as developing new markets and new customers, and even considering marketing efficiency, preventing customer churn is more economical than developing new customers.

To sum up, this paper studies customer churn. SVM model executes more quickly and provides greater model impact. The Support Vector Machine (SVM) model demonstrates exceptional performance in predicting customer churn, as indicated by its high scores in terms of area under the curve (AUC), accuracy, and other evaluation metrics. Its superior prediction capabilities and stability make it a valuable tool for identifying potential consumers.

ACKNOWLEDGEMENT

The 2020 Guangxi University Young and Middle-aged Teachers' Basic Research Ability Improvement Project "Cross-platform Research and Development of Panoramic Campus Roaming System" (2020KY63022).

REFERENCES

[1] W. H. Zhang, "Forecast model of mobile customer churn based on data mining," *Journal of Inner Mongolia Normal University(Natural Science Edition)*, 2016.

[2] S. Yang and J. J. Yue, "Bank customer churn decision tree prediction algorithm under data mining technology," *Computer Knowledge and Technology*, 2014.

[3] M. Rahman and V. Kumar, "Machine learning based customer churn prediction in banking," in *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. IEEE, 2020, pp. 1196–1201.

[4] M. X., C. S., D. K., and K. J., "Research on customer churn prediction model of commercial banks," *Systems Engineering*, vol. 22, no. 12, p. 5, 2004.

[5] Z. Y, "Research and application of mobile communication customer churn prediction model based on data mining," *Modern Information Technology*, vol. 6, no. 11, p. 6, 2022.

[6] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu, "Applications of support vector machine (svm) learning in cancer genomics," *Cancer genomics & proteomics*, vol. 15, no. 1, pp. 41–51, 2018.

[7] M. Mohammadi, T. A. Rashid, S. H. T. Karim, A. H. M. Aldalwie, Q. T. Tho, M. Bidaki, A. M. Rahmani, and M. Hosseinzadeh, "A comprehensive survey and taxonomy of the svm-based intrusion detection systems," *Journal of Network and Computer Applications*, vol. 178, p. 102983, 2021.

[8] W. Huang, H. Liu, Y. Zhang, R. Mi, C. Tong, W. Xiao, and B. Shuai, "Railway dangerous goods transportation system risk identification: comparisons among svm, pso-svm, ga-svm and gs-svm," *Applied Soft Computing*, vol. 109, p. 107541, 2021.

[9] K. S. Sahoo, B. K. Tripathy, K. Naik, S. Ramasubbareddy, B. Balusamy, M. Khari, and D. Burgos, "An evolutionary svm model for ddos attack detection in software defined networks," *IEEE Access*, vol. 8, pp. 132 502–132 513, 2020.

[10] R. Sharma, A. Sungheetha *et al.*, "An efficient dimension reduction based fusion of cnn and svm model for detection of abnormal incident in video surveillance," *Journal of Soft Computing Paradigm (JSCP)*, vol. 3, no. 02, pp. 55–69, 2021.