scientific reports

Check for updates

OPEN

A novel classification algorithm for customer churn prediction based on hybrid Ensemble-Fusion model

Chenggang He[1,3]✉ & Chris H. Q. Ding[2,3]

Nowadays, customer churn issues are becoming more and more important, which is one of the most important metrics for evaluating the health of a business it is difficult to measure success without measuring customer churn metrics. However, it has become a challenge for the industry to predict when customers are churning or preparing to churn and to take the necessary action at the critical time before they do. At the same time, how to keep the place of deep research on the 17 machine learning algorithms in 9 major classes of machine learning classics production is the first problem we are facing. Through customer churn deep research, we mentioned the Ensemble-Fusion model based on machine learning and introduced a smart intelligent system to help reduce the actual customer churn about the production. Comparing with most popular predictive models, such as the Support vector machine algorithm, Random Forest algorithm, K-Nearest-Neighbor algorithm, Gradient boosting algorithm, Logistic regression algorithm, Bayesian algorithm, Decision tree algorithm, and Neural network algorithm are applied to check the effect on accuracy, AUC, and F1-score. By comparing with 17 algorithms in 9 categories of machine learning classics, the data prediction accuracy of the Ensemble-Fusion model reaches 95.35%, AUC score reaches 91% and F1-Score reaches 96.96%. The experimental results show that the data prediction accuracy of the Ensemble-Fusion model outperforms that of other benchmark algorithms.

Keywords Customer churn, Machine learning, Ensemble-Fusion model, Smart intelligent system

Customer churn is one of the key factors affecting the benign development of industries and enterprises, and at the same time, it is a very challenging research topic in both academia and industry[1–3], especially for those information industries relying on the subscription model and the order purchase operation model, customer churn, especially the churn of key customers, can be fatal to their impact. Reducing **5%** of customer loss rate can increase profits by 25–125%[2]**.** Unfortunately, this always requires lots of manual efforts to analyze data, and it is often too late to take actions to retain them. In order to retain more existing old customers, especially some key customers, many companies have made many attempts to differentiate between churned and non-churned customers, so as to achieve the purpose of retaining churned customers, but the actual effect is very poor. As we all know, the loss of old customers not only affects revenue, but also affects the attraction of new customers. In addition, the cost of developing a new customer is often much higher (almost 5–6 times) than the cost of retaining an old customer[4,5]. So, is it possible to research efficient customer churn prediction models for customer churn prediction by using machine learning-related algorithms in conjunction with the actual needs of the industry? At the same time, in order to help those decision makers who do not have the theoretical foundation of algorithms to make decisions quickly and efficiently, is it possible to develop an intelligent, convenient, efficient and intelligent early warning system that can detect or predict the existing customer churn in a timely manner to help the industry, and then the enterprises can take relevant actions to retain customers when they find that there is a risk of churning key customers, so as to minimize the losses of the enterprises? In part of the related work the theoretical basis of Gradient Boosting Algorithm[6,7], Bayesian Algorithm[8,9], Support Vector Machine Algorithm[10–15], Random Forest Algorithm[16], K Neighborhood Algorithm[17,18], Logistic Regression Algorith[19,20],

[1]School of Public Safety and Emergency Management, Anhui University of Science and Technology, No.15 Fengxia Road, Hefei 230041, Anhui, China. [2]School Department of Computer Science and Engineering, University of Texas at Arlington, 701 S. Nedderman Drive, Arlington, TX 76019, USA. [3]School of Computer Science and Technology, Anhui University, 111 Jiulong Road, Hefei 230039, Anhui, China. ✉email: hechenggang@aust.edu.cn

Decision Tree Algorithm[21–24] and Neural Network Algorithms[25–29] are described and the research on application of these algorithms in customer churn prediction is discussed. The literature related to the above algorithms is restating the superiority of the single algorithm they use, and after analyzing them, it can be concluded that these algorithms are affected by the characteristics of the dataset, and there is a strong dependency between their algorithms and the dataset, and then there is no such thing as being able to use one algorithm alone to solve all the problems in any practical application scenarios. Based on the shortcomings of the traditional algorithms analyzed above, this paper proposes a model based on Ensemble-Fusion (Integrated Learning Fusion), in order to meet the universality of various complex scenarios through the model, and expects to be able to provide academia and industry with a pervasive and efficient customer churn prediction solution. So in this paper, we first propose a customer churn prediction algorithm based on the Ensemble-Fusion model. Then it proposes an efficient churn solution based on the Ensemble-Fusion model. Finally, in order to help the information industry make efficient customer churn decisions, a real-time intelligent early warning system for customer churn is developed through theory-guided practice, which can monitor customer dynamics in real-time, help enterprises to identify potential lost customers in advance, and provide early warning at the first moment to remind the sales team or the Customer success management team (CSM) to take proactive action to retain lost customers, thus reducing the risk of fatal blow to the enterprise because of customer churn.

Given the above purposes, this paper conducts research on customer churn prediction through machine learning related theories and algorithms, firstly gives a solution to deal with the huge and complex datasets in the industry, then proposes the Ensemble-Fusion (Integrated Learning Fusion) prediction model for customer churn, and finally, in order to further guide the theory to practice, facilitate the enterprises to take actions quickly and efficiently to retain customers, especially the key customers, in order to improve customer retention. Especially the retention of key customers. Combined with my many years of experience in the industry, I have developed an end-to-end real-time intelligent early warning system for customer churn, which not only predicts customer churn in an organization's production environment, but also sends out early warnings to alert the relevant personnel such as the sales team and the customer success team, so that the relevant teams can take effective action to retain the customers who are about to be lost in the first time. The system not only predicts customer churn in an organization's production environment, but also sends out early warnings to alert relevant personnel such as sales and customer success teams so that they can take immediate action to retain lost customers. In order to solve the above problems, we must first deal with the problems encountered in the research, specifically in the research work encountered in the actual research and development of the very difficult problems are as follows: First, the real structure of the production data is very complex and the relevant data are often distributed in different regions of the world in different departments and data structure of different databases, the collection of data is very difficult, and due to the restriction of some sensitive information and the relevant agreements, it is very difficult to collect all the relevant data. It is also difficult to collect all the relevant data due to sensitive information and related protocol issues. Therefore, the problem of customer churn data collection becomes how to construct an effective model with a limited data set. Secondly, in the collected relevant data, there is still a lot of noise in the data, which is very imbalance[30–38] due to the actual impact of business complexity and there are no labels to mark whether a customer is churned or not, which requires that a lot of prior work and business knowledge should be involved before proceeding with the collection and processing of the data. In order to address the above issues in customer churn data prediction, this paper's main contributions of the work are as follows:

(1) This paper proposes a novel model named Ensemble-Fusion based on ML (Machine Learning) related theories and algorithms to predict customer churn in SAAS[36] (Software-as-a-Service, SAAS is a cloud-based software delivery model in which the cloud provider develops and maintains cloud application software) production environments, which focuses on the exceptionally complex data collection, processing and application in the actual production line, and organizes a detailed customer churn prediction data processing architecture diagram is shown(detailed in Sect. "Customer churn prediction solution based on Ensemble-Fusion model"), and finally the solution proposed in this paper is used in the actual production environment to achieve good results.

(2) This paper combines machine learning theories and algorithms, such as support vector machine algorithms, random forest algorithms, K-neighborhood algorithms, gradient boosting algorithms, logistic regression algorithms, Bayesian algorithms, deci- sion tree algorithms and neural network algorithms, and other 9 categories of 17 machine learning algorithms as a baseline classifiers to propose the "customer churn data processing architecture based on the integration of learning fusion (Ensemble- Fusion)". Fusion-based customer churn prediction model and verified the high accuracy and effectiveness of the churn prediction model by evaluating the key indexes of the machine learning model, such as precision, recall, accuracy, AUC [37](Area under the ROC[38] Curve, AUC measures the entire two-dimensional area underneath the entire ROC curve. AUC provides an aggregate measure of performance across all possible classification thresholds.) and F1-score[39,40](F1-score is an important evaluation metric that is commonly used in classification task to evaluate the performance of a model. F1-score is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model's precision and recall).

(3) In order to further improve the productivity of the industry efficiently, by linking theory to practice, this paper also designs and develops an intelligent early warning system based on the Ensemble-Fusion model to help enterprises predict customer churn, especially the churn of important customers, quickly and effectively, so as to help them retain churned customers and reduce the churn that brings. The system is designed to help companies retain lost customers and minimize the fatal blow to the company due to customer churn. The intelligent system can not only present important customers with high probability

> of churn, but also automatically provide relevant information based on the prediction results to remind relevant personnel to take proactive actions to retain important customers that are about to be churned, so as to reduce losses.

This paper not only provides specific solutions to the important problem of cus- tomer churn from theory, but also translates the theory into a specific intelligent early warning system, which can efficiently help enterprises, especially those who don't know the background knowledge of machine learning and other relevant leadership decision- making personnel to easily make effective decisions about customer churn, so as to be able to retain key customers and increase the competitiveness of the enterprise. The system can be used to retain key customers and increase the competitiveness of an organization.

The rest of this paper is organized as follows, in Section "A research approach to customer churn prediction based on Ensemble-Fusion model", it mainly introduces the theory and methodology, solution, and overall architectural design of the machine learning-based customer churn intelligent system and introduces the customer churn prediction algorithm based on the Ensemble-Fusion model proposed in this paper. In Section "Experiment and result", the proposed customer churn prediction algorithm is validated and the high accuracy and effectiveness of the churn prediction model are verified by the key metrics of machine learning model evaluation, such as precision, recall, accuracy, AUC , and F1-score[37–40]. Section "Intelligent early warning system for customer churn prediction based on Ensemble-Fusion model" describes the main functions of the intelligent early warning system for customer churn prediction, and also provides a detailed description of the User Cases associated with this intelligent system. A review of relevant customer churn research is presented in Section "Related work". Finally, relevant conclusions and outlook are summarized in Section "Conclusions and future work".

A research approach to customer churn prediction based on Ensemble-Fusion model

This part proposes a solution for customer churn prediction based on the Ensemble- Fusion model: firstly, it comprehensively outlines the specific scenarios to be solved for customer churn, and gives the ideas and feasible solutions to solve the problem from top to bottom. Then the specific design and implementation of an end-to-end customer churn intelligent prediction system is proposed: specifically including the collection and processing of complex datasets, the construction of prediction models, and the intelligent system platform in three parts, each of which contains a detailed process. Then this paper provides an in-depth analysis of the machine learning model for customer churn prediction, and finally this paper proposes a new customer churn prediction model and gives a specific implementation algorithm.

Customer churn prediction solution based on Ensemble-Fusion model

This part proposes a solution based on the Ensemble-Fusion model to predict customer churn and help organizations reduce customer churn. The detailed process of the solution is depicted in Fig. 1, as shown in Fig. 1, the solution consists of two main parts: the offline training part and the online inference part. During offline training, data preprocessing[30–33] s first required to clean and label the input data, the annotation is done by labeling the data with churn or non-churn. Then, the relevant features of the data are extracted based on the business knowledge, such as the feature "Trend of meetings compared to last year" which is used to describe the

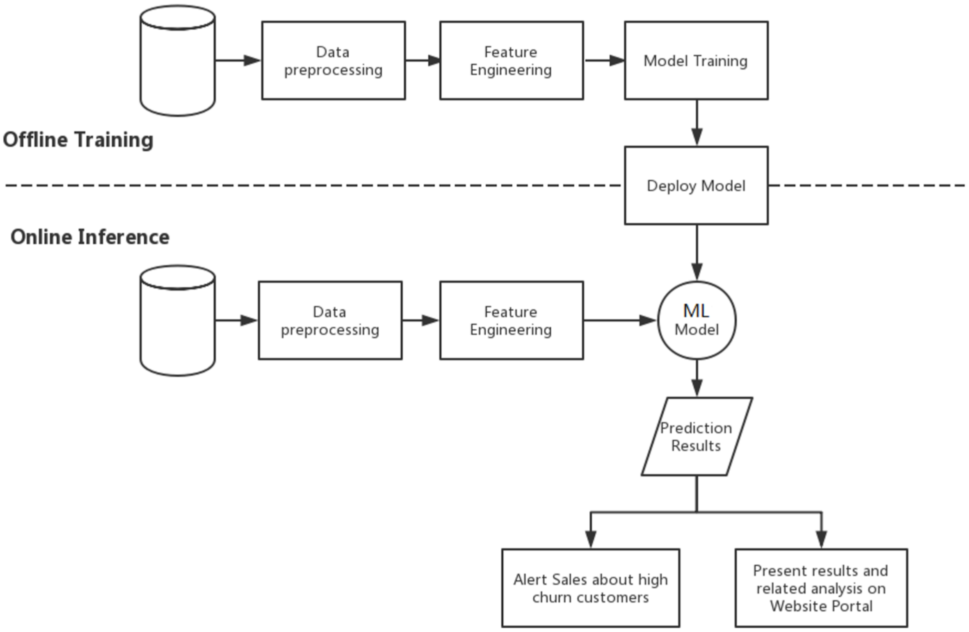


Fig. 1. Customer Churn Solution Flowchart.

number of meetings booked by customers in the current year compared to the number of meetings booked by customers in the previous year, and the number of meetings booked also reflects the trend of imminent churn of customers. The feature "Trend in meeting duration compared to last year" can be used to characterize the total duration of meetings in the current year compared to the total duration of meetings in the previous year, which can be used to predict the trend of customer churn. These extracted features can effectively reflect the trend of imminent or significant customer churn. Specific model features are described in Table 1, where model training data information is used from actual production line usage data.

The process of customer churn prediction processing and the logical relationship between data transfers are detailed in Fig. 2. In addition, since there are only a few churned (noisy) data, data balancing-related processes must be performed before training. These features can then be used to iteratively train and validate the machine learning model until the model is validated well enough to be deployed directly to a production environment.

Feature	Descriptions
MTG CORRELATION INDEX	Meeting correlation trend versus last year
MINS CORRELATION INDEX	Mins correlation trend versus last year
MTG GROUTH	Meeting growth percentage versus last period
MINS GROUTH	Mins growth percentage versus last period
MINS GROUTH YEAR	Mins growth percentage versus last year
HOST CORRELATION INDEX	Host number correlation trend versus last year
HOST GROUTH	Host number growth percentage versus last period
YEAR USAGE	The usage in 1 year
STD USAGE	Variance for usage in the past year
PLATFORM	Billing Platform Type (SAAS or Native)
EFFECTIVE FROM	Service effective from
EFFECTIVE TO	Service effective to
TEL CORRELATION INDEX	Tel correlation trend versus last year
TEL GROUTH	Tel growth percentage versus last period
TEL GROUTH YEAR	Tel growth percentage versus last year
YEAR AMOUNT LOCAL	Annual local currency amount in the past year
YEAR AMOUNT USD	Annual USD currency amount in the past year
CURRENCY	Currency
RENEW TERM	How long will renew, when old ser- vice is expired
INITIAL TERM	The first contract period

Table 1. Detailed description of characteristics related to customer churn.

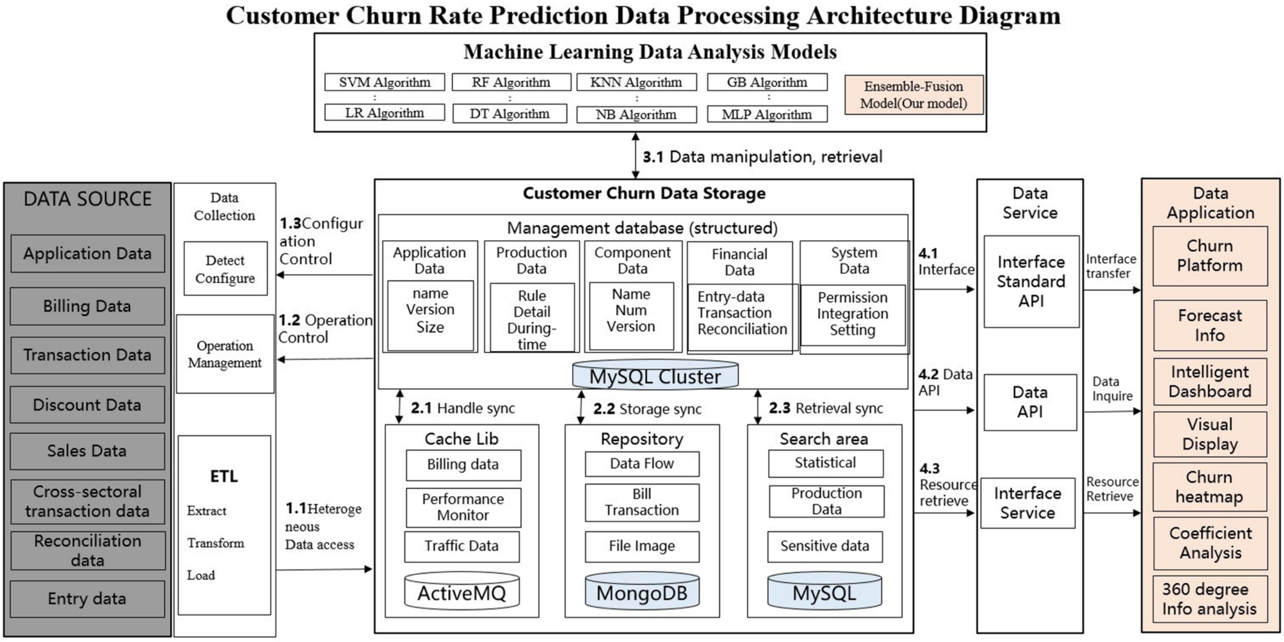


Fig. 2. Architecture diagram of customer churn prediction data processing.

Finally, the rigorously validated model can be deployed in a production environment to predict the likelihood of customer churn in real time.

For the online inference component, data cleaning and feature engineering[35–37] are also required to construct the training dataset. The dataset here does not contain labeled data, mainly because the goal to be predicted is whether customers will churn in the following months, which has not occurred in the previous inference process. After obtaining the trained model, test data also needs to be fed into the machine learning model to infer the final prediction. Finally, information about the high churn customers predicted by the validated machine learning model will be displayed on the intelligent churn prediction system. Information about the churn prediction will be notified to the project stakeholders in real-time via email, instant messaging, and other messaging channels so that they can proactively take action to minimize the risk of churn losses.

Customer churn data prediction algorithm based on Ensemble-Fusion model

In order to better carry out the research on customer churn rate, this paper focuses on the theoretical basis of the Support Vector Machine algorithm, Random Forest algorithm, K-neighborhood algorithm, Gradient Boosting algorithm, Logistic Regression algorithm, Bayesian algorithm, Decision Tree algorithm, and Neural Networks algorithm in Section "Related work" and discusses the research on the application of these algorithms in the prediction of customer churn rate. The literature related to the above algorithms restates the superiority of the single algorithm they use, and after analyzing them, it can be concluded that these algorithms are affected by the characteristics of the dataset, and there is a strong dependency between their algorithms and the dataset, and then there is no such thing as being able to use one algorithm alone to solve all the problems in any practical application scenarios. Based on the shortcomings of the traditional algorithms analyzed above, this paper proposes a model based on Ensemble-Fusion (Integrated Learning Fusion), in order to meet the universality of various complex scenarios through the model, and expects to be able to provide academia and industry with a pervasive and efficient customer churn prediction solution.

This subsection focuses on the detailed construction process of the customer churn prediction method based on the Ensemble-Fusion model, which is described in detail in Algorithm 1, and compared with the experimental results of 17 machine learning algorithms through the model in the experimental part of Section "Experiment and result", so as to validate that the model has a high accuracy rate, strong robustness, and ease of scalability.

End-to-end customer churn prediction real-time intelligent early warning system design

To further help organizations reduce customer churn, this subsection designs and develops a customer churn intelligent prediction system. The system consists of three main parts, the first part is mainly the collection and processing of different business-related data set and detailed processing, which mainly includes four major processes, of which the first major process includes the access of heterogeneous data, due to the unusual complexity of the source of data in the real production environment, which mainly includes the system application data, Billing (financial billing) customer data, prod- uct transaction data, Product discount data, product sales data, cross-departmental transaction data, reconciliation data and posting data. In a large multinational group.

```
Require: Training data $X = \{x_i, y_i\}$ $x$ is the customer-related data, and $y$ is its
    corresponding label.
Ensure: prediction$X_{test}$ usingensemble-fusionmodel
Step 1: Data Aggregation and De-drying
Select$\{x_i, y_i\}$ from$X$
Heterogeneous data source processing via ETL (Extra, Transform,Load)
Default value processing and Sparse coding $x' \Leftarrow x$
Using PCA (Principal Component Analysis) for feature extraction from $x'$
Step 2:   Ensemble Fusion Model Training
$0 \Leftarrow t$
$ALGO_{set}$ = {SVM,RF,KNN,Gboost,LR,Bayes,DT,Adaboost,ET,MLP}
$METRICS_{set}$ ={Precision,Recall,Accuracy,F1-Score}
for $t = 0$ to $e_m$ do
    if $t = 0$ then
        $Model_{ensemble} = Model_0$
    else
        for $i = 0$ to $Len(ALGO_{set})$ do
            Predictions $y_i$ are obtained in the $ALGO_{set}$
        end for
        for $i = 0$ to $Len(ALGO_{set})$ do
            Majority vote fusion of baseline model predictions
            if $Metrics_{fusion} > Metrics_i$ then
                $Model_{ensemble} = Model_{fusion}$
            end if
        end for
    end if
end for
```

Algorithm 1 Customer ChurnPrediction Algorithm Based on Ensemble Fusion Model

of companies, due to the different technical architectures of each system, the data for- mat is not the same, generally JSON, XML, plain text files and other formats. To process the data, it is necessary to unify the data format here, from different hetero- generous databases through ETL (Extra, Transform, Load) to achieve from different types of databases (e.g., MySQL, Oracle, MongoDB, and Redis) to get the data, and finally unified storage in the MySQL database. The second major process is to structure the data by managing the database to construct training and testing datasets for the next machine learning models. The third major process is to perform the construction of the machine learning model for customer churn prediction through the formatted and unified dataset acquired in the previous step (details will be elaborated in Sect. "AUC results and analysis"). The fourth major part is the transfer of business logic through the standardized API interface (Restful API), and ultimately display of relevant information on the front-end page, which mainly includes the display of customer churn information, the display of customer churn heat map, the customer churn management platform, and the analysis of customer churn 360-degree related information, which is elaborated in detail in Fig. 2(Customer Churn Prediction Data Processing Architecture Diagram). The second part is the ML (Machine Learning) modeling system, which includes data acquisition, feature engineering, and model training, and this part is elaborated in subsection 2.3. The third part is the visualization and presentation plat- form which will display the information related to customer churn, and this relevant part will be described in detail in Section "Experiment and result". The details of the system architecture are described in detail in Fig. 3, as shown in Fig. 3, the system mainly consists of the following parts, the first part is the collection of data, for the Fortune 500 multi- national corporations, their various businesses are spread all over the world, and the collection of data is a very complex and time-consuming work. The second part is the data processing such as feature engineering on the data collected in the first part, then the training and validation of the machine learning model, and finally obtaining a machine learning model with the highest accuracy rate to be used in the customer churn prediction system. The third part is the platform display part, which mainly displays multi-dimensional warning information and real-time forecasts for specific customer churn information, and the specific related information and functions will be elaborated in Section "Experiment and result".

Specific user usage examples of this intelligent system are described in detail in Fig. 4. As shown in Fig. 4, the sales layer and the leadership layer are two important key target roles that are important in the platform. At the sales level, the intelligent system displays customers with high churn risk on the platform and provides relevant details. The platform also sends out regular alert emails, timely messages, and other early warning information to notify the relevant project stakeholders to take proactive action to intervene in the impending churn. Additionally, salespeople can send feedback about forecasts to help continuously improve and optimize the proposed machine learning model. For leadership, it is even more important to keep track of global customer churn rather than individual customer churn. To solve this problem, the intelligent real-time alert system is

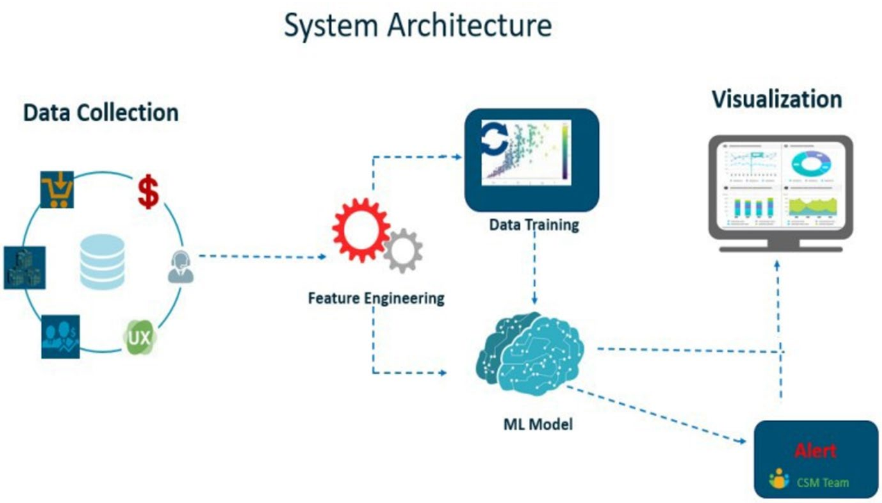


Fig. 3. Architecture diagram of customer churn intelligent early warning system.

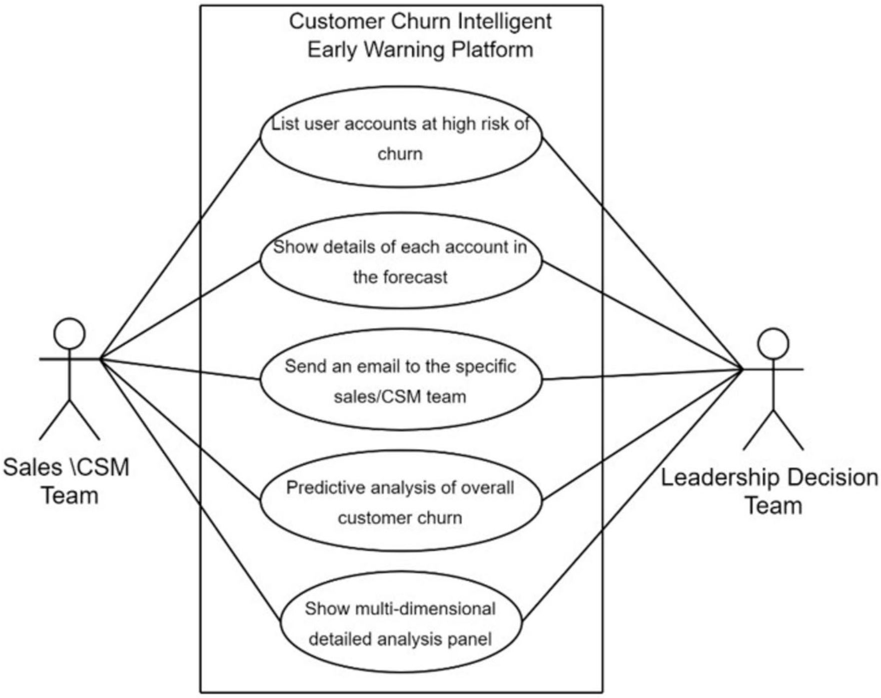

Customer Churn Intelligent
Early Warning Platform
List user accounts at high risk of churn
Show details of each account in the forecast
Send an email to the specific sales/CSM team
Predictive analysis of overall customer churn
Show multi-dimensional detailed analysis panel
Sales \CSM Team
Leadership Decision Team

Fig. 4. Use case diagram for a customer churn platform.

designed with a dashboard module for leadership managers to show the overall churn trend from a global perspective, thus facilitating decision-makers to make efficient decisions at the first time.

Experiment and result

This section focuses on the comparison of the experimental results of the proposed Ensemble-Fusion model-based machine learning for customer churn prediction and the classical machine learning 9 categories and 17 algorithms for customer churn predic- tion. Here, a private dataset of the customer production line system of the Company from 2015 to 2022 is used, where 80% of the data is used for training and 20% of the data is used for testing, in which K-fold cross-validation is used to test the accuracy of the model.

Model evaluation indicators

In order to evaluate the performance of machine learning models, relevant metrics recognized in the field of machine learning are usually used, namely precision, recall, accuracy and F1-score[38–41]. These metrics represent the performance of predictive models for customer churn prediction. The meanings of the metrics are explained here in a relevant way, with true positives and false positives denoted as TP and FP, respectively[42], and true negatives and false negatives denoted as TN and FN, respectively[43].TP stands for the number of customers whose actual labels are churned (predict label is churn), FP stands for the customers whose actual customers are labeled as not churned but whose predicted customer labels are churned number, FN represents the number of customers whose actual label is churn but whose predicted label is not churn, and TN represents the number of customers whose actual label is not churn and whose predicted label is not churn. Thus, precision, recall, accuracy, and F1 score can be described as follows:

$$Prection = \frac{TP}{TP + FP} \; Recall = \frac{TP}{TP + FN} \tag{1}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \; F1 - Score = \frac{2Prection * Recall}{Prection + Recall} \tag{2}$$

Results of model indicators related to customer churn prediction

To evaluate the performance of the customer churn prediction algorithm based on the Ensemble-Fusion model proposed in this paper, the customer churn prediction is performed by the model proposed in this paper and 17 machine learning algorithms in 9 major categories of machine learning classics respectively. The performance metrics of precision, recall, accuracy, and F1-score[38–41] are compared, and the detailed results of the specific comparison can be found in Table 2. Among the 17 machine learning algorithms in 9 major classes of machine learning classics, the accuracy of gradient boosting classifiers and random forests are 95.32% and 94.29%, respectively, and the F1-score of the gradient boosting classifier is up to 96.3%, which is better than other machine learning classic algorithmic classifiers, while the integrated learning fusion model proposed in this paper achieves an accuracy rate of 95.35%, and the F1-Score reaches 96.96% significantly better than other machine learning classic benchmark classifier algorithms. The results of Precision, Recall, Accuracy, and F1-Score of 17 machine learning algorithms in 9 categories of machine learning classics are shown in detail in Figs. 5, 6, 7 and 8 for comparison.

AUC results and analysis

To further evaluate the performance of the model, this section also uses AUC[13] curve for evaluating the machine learning model. A higher AUC score represents better performance of the model. Here, fivefold cross-validation[14] is used to calculate the ROC, and the highest AUC is obtained for the integrated learning-based fusion model proposed in this paper, the detailed results of the specific comparison can be found in Table 3, and the ROC[15] results for the related machine algorithms are shown in Figs. 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26 and 27.

No	ML algorithm	Precision	Recall	Accuracy	F1-score
1	Random forests[1,2]	0.94660846	0.989123	0.942994	0.967399
2	K-nearest neighbors[5]	0.8984902	0.981404	0.889289	0.938118
3	Gradient boosting classifier [6,7]	0.95816327	0.98842	0.9532	0.96306
4	Logistic regression[3]	0.87862377	0.967719	0.858086	0.921022
5	MLPClassifier(activation = 'logistic')[1]	0.94264507	0.980351	0.932193	0.961128
6	MLPClassifier(activation = 'tanh')[1]	0.93855503	0.975439	0.924392	0.956641
7	MultinomialNB classifier[8,9]	0.86323214	0.987719	0.855686	0.921289
8	BernouiliNB classifier[8,10]	0.85508551	1	0.855086	0.921883
9	GaussianNB classifier[8,9]	0.85508551	1	0.855086	0.921883
10	DecisionTreeClassifier (CART)[3]	0.95308642	0.94807	0.915692	0.950572
11	DecisionTreeClassifier (ID3)[3]	0.95149385	0.949825	0.915692	0.950658
12	SVM classifer (Linear)[10–12]	0.85508551	1	0.855086	0.921883
13	SVM classifer (Poly)[10–12]	0.92019704	0.983158	0.912691	0.950636
14	SVM classifer (RBF)[10–12]	0.92028749	0.988421	0.916892	0.953138
15	SVM classifer(sigmoid)[10–12]	0.8604878	0.928421	0.810081	0.893165
16	Adaboost classifier[6,7]	0.94765282	0.984561	0.940294	0.965755
17	ExtraTreesClassifier[5]	0.92671706	0.989474	0.924092	0.957068
18	**Our model***	**0.960088**	0.989013	**0.953533**	**0.969631**

Table 2. Comparison of results of customer churn prediction algorithm metrics. Significant values are in bold. An asterisk means that our proposed model achieves the optimal result.

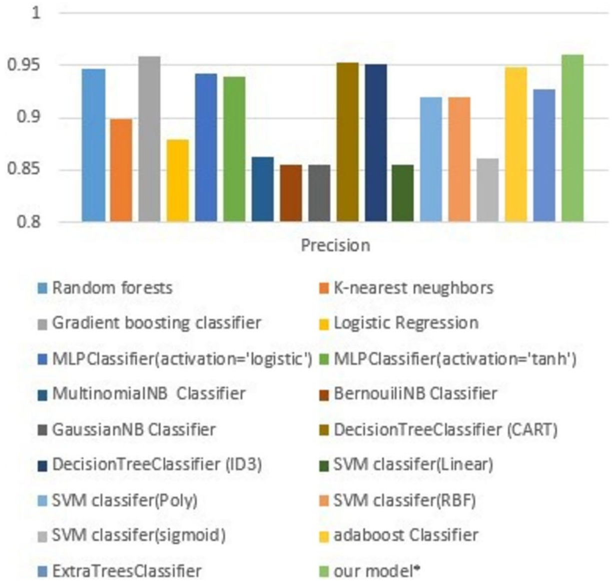


Fig. 5. Comparison of algorithm precision.

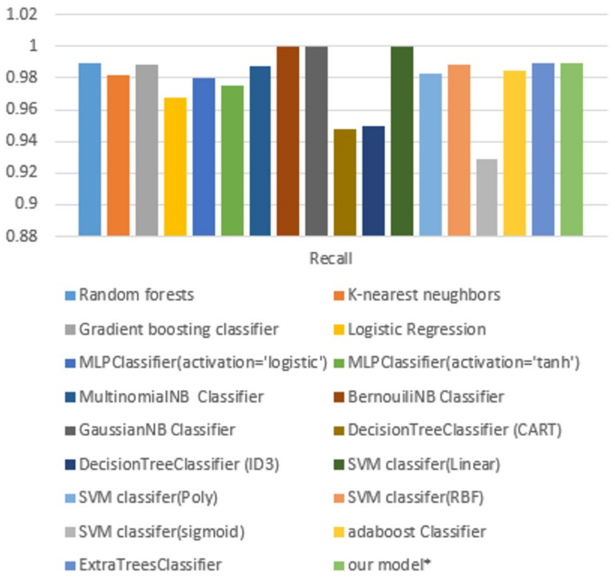


Fig. 6. Comparison of algorithm recall.

Intelligent early warning system for customer churn prediction based on Ensemble-Fusion model

In this section, the main functions of the real-time intelligent early warning system for customer churn data prediction based on the Ensemble-Fusion model will be elaborated in detail, and the relevant descriptions of the main functions are described as follows.

Information relevant to predicting customer churn

Figure 28 shows the top five of the "Top 100" accounts with high churn risk, as shown in Fig. 28, with detailed information (e.g., account name, account ID, etc.) displayed in the table. If the prediction is incorrect, the user can give feedback by clicking on the relevant action, and then feedback through the system. Of course, it is also possible to click on the Account ID to enter the detailed prediction page, which will be analyzed in detail in Section "Demonstration of the intelligent system of customer churn prediction".

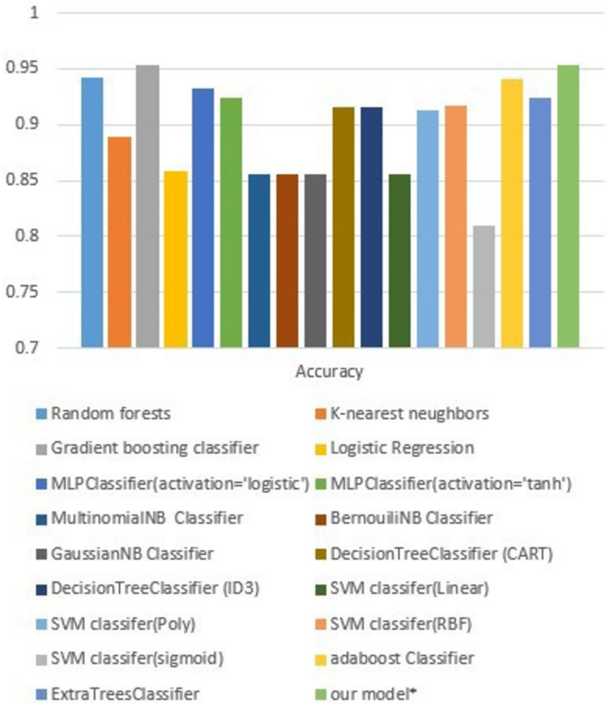


Fig. 7. Algorithm Accuracy comparison chart.

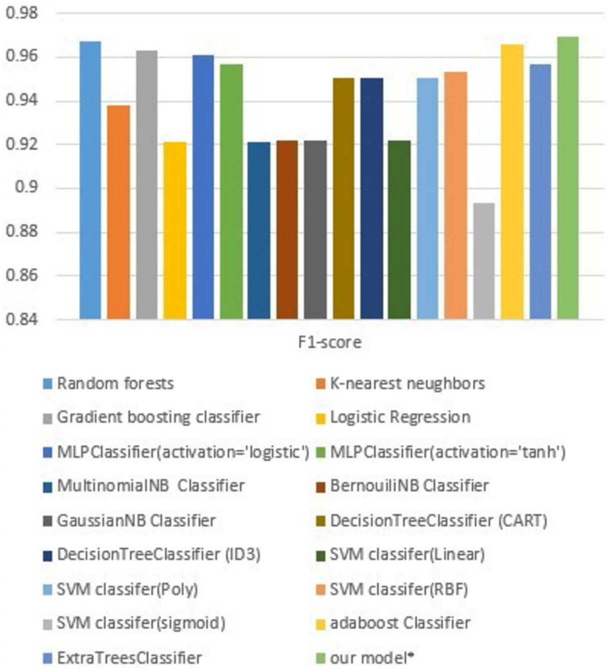


Fig. 8. Algorithm F1-Score Comparison chart.

Demonstration of the intelligent system of customer churn prediction

In Figs. 29 and 30, detailed information of a detailed page of a real-time intelligent prediction system for customer churn is described, which consists of two parts, wherein the upper half of the page displays the basic information of the current churned customer data prediction, which specifically includes information such as the user's ID, name, and the type of platform. In the second half, the reasons for the churn are provided and a multi-dimensional analysis of the specific reasons is provided to help the relevant stakeholders and personnel in the

No	ML algorithm	ROC fold0	ROC fold1	ROC fold2	ROC fold3	ROC fold4	Mean ROC
1	Random forests[1,2]	0.9	0.91	0.93	0.91	0.91	0.91
2	K-nearest neighbors[5]	0.84	0.82	0.86	0.87	0.85	0.85
3	Gradient boosting classifier[6,7]	0.92	0.92	0.91	0.89	0.93	0.91
4	Logistic regression classifier[3]	0.82	0.8	0.81	0.8	0.86	0.82
5	MultinomialNB classifier[8,9]	0.76	0.79	0.79	0.79	0.78	0.78
6	BernouiliNB classifier[8,9]	0.72	0.71	0.72	0.74	0.7	0.72
7	GaussianNB classifier[8,9]	0.85	0.85	0.8	0.86	0.85	0.84
8	DecisionTreeClassifier (CART)[3]	0.84	0.84	0.83	0.81	0.86	0.83
9	DecisionTreeClassifier (ID3)[3]	0.85	0.84	0.84	0.83	0.85	0.84
10	SVM classifer(Linear)[10–12]	0.72	0.73	0.59	0.73	0.65	0.68
11	SVM classifer(Poly)[10–12]	0.88	0.9	0.89	0.89	0.92	0.89
12	SVM classifer(RBF)[10–12]	0.93	0.9	0.91	0.88	0.9	0.9
13	SVM classifer(sigmoid)[10–12]	0.51	0.52	0.46	0.53	0.48	0.5
14	Adaboost classifier[6,7]	0.92	0.88	0.91	0.91	0.9	0.9
15	ExtraTreesClassifier[5]	0.87	0.92	0.88	0.92	0.92	0.9
16	MLPClassifier(activation = 'logistic')[1]	0.91	0.91	0.92	0.89	0.9	0.9
17	MLPClassifier(activation = 'tanh')[1]	0.89	0.89	0.93	0.9	0.85	0.89
18	Our model*	0.91	0.92	0.92	0.9	0.91	**0.91**

Table 3. Comparison of AUC score results of customer churn data prediction algotihms AUC score. Significant values are in bold. An asterisk means that our proposed model achieves the optimal result.

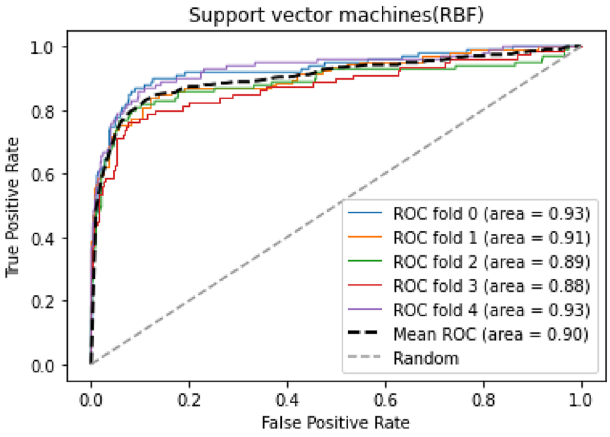


Fig. 9. SVM(RBF)algorithm ROC and AUC.

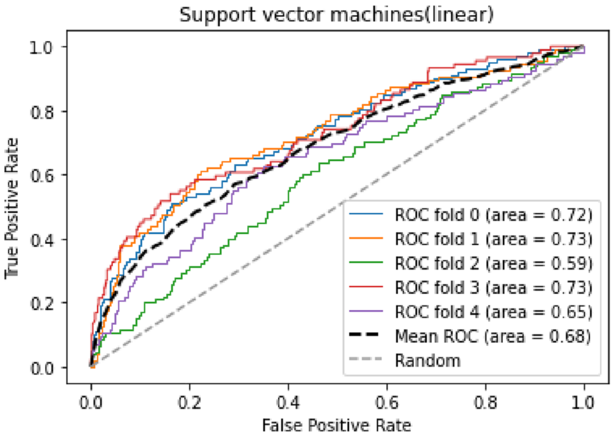


Fig. 10. SVM(RBF)algorithm ROC and AUC.

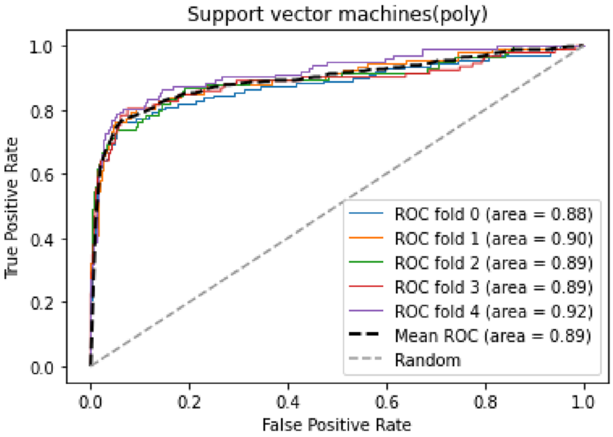


Fig. 11. SVM(Poly) algorithm AUC.

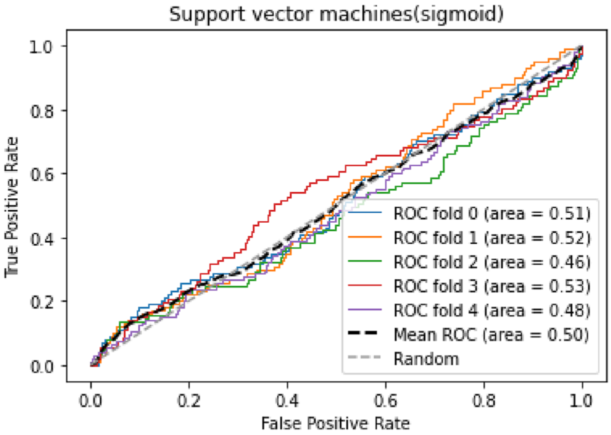


Fig. 12. SVM (Sigmoid) algorithm AUC.

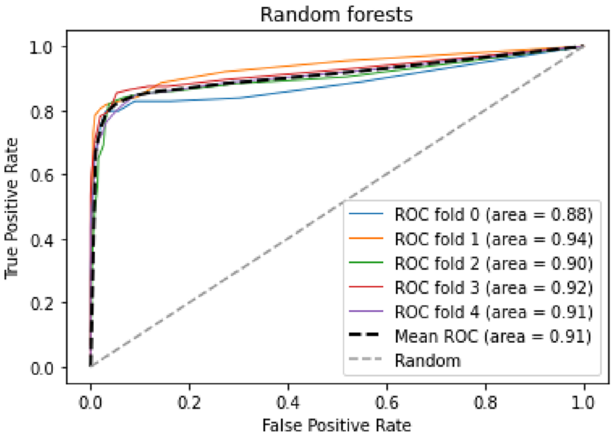


Fig. 13. Random Forest algorithm AUC.

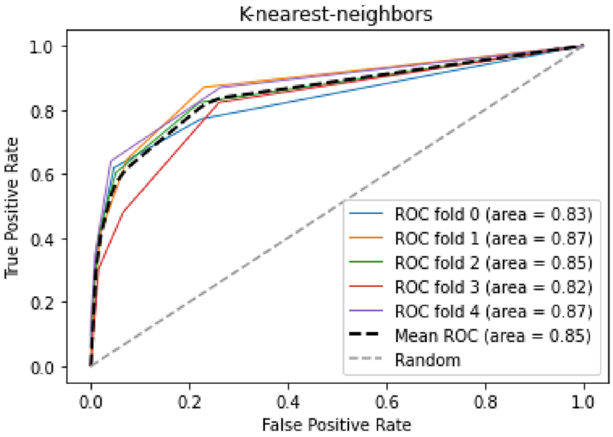


Fig. 14. KNN algorithm AUC.

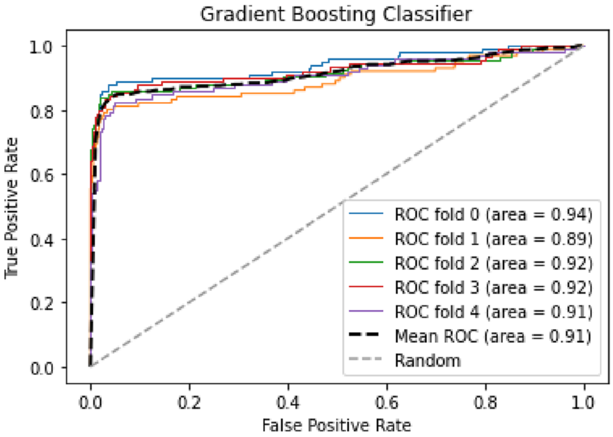


Fig. 15. Random Forest algorithm AUC.

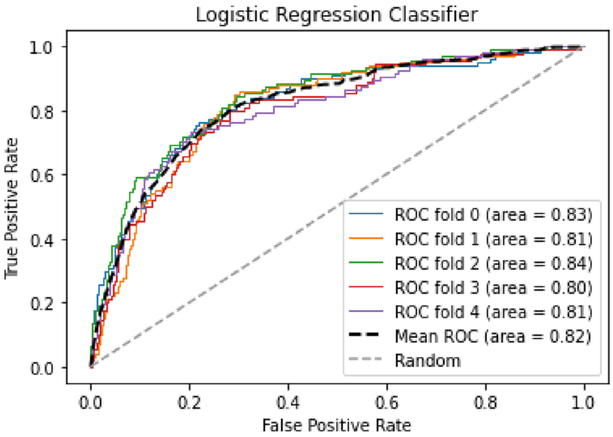


Fig. 16. LR algorithm AUC.

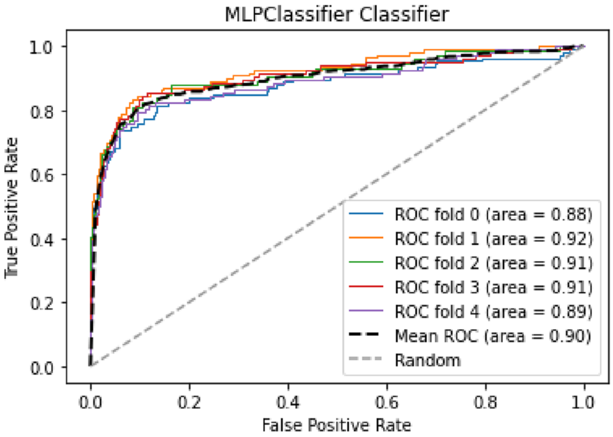


Fig. 17. MLP (Algorithm 16) AUC.

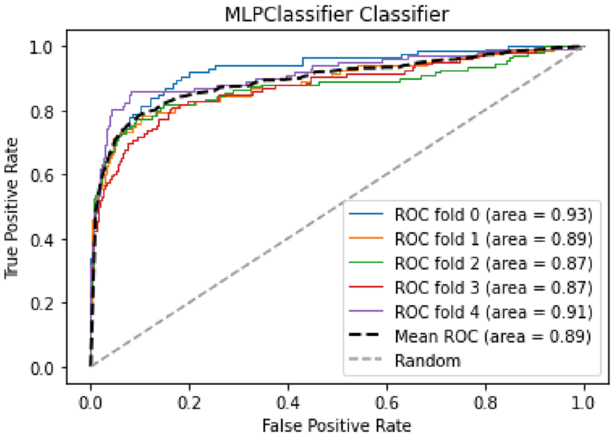


Fig. 18. MLP (Algorithm 17) AUC.

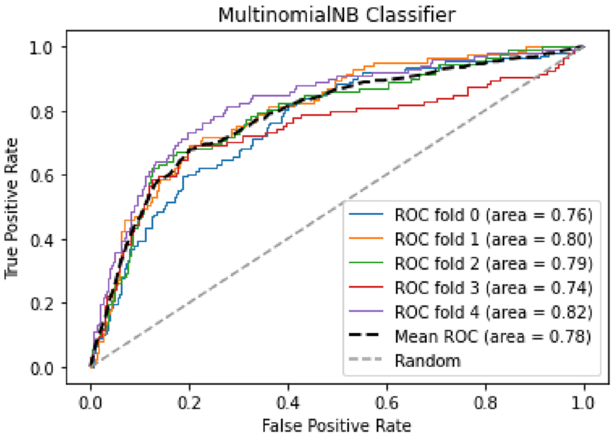


Fig. 19. MultinomialNB algorithm AUC.

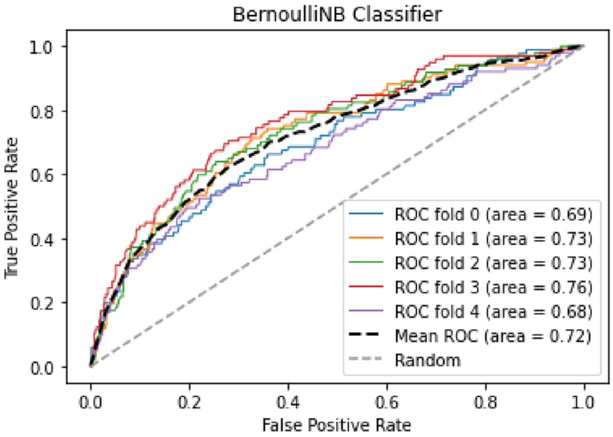


Fig. 20. BernouiliNB algorithm AUC.

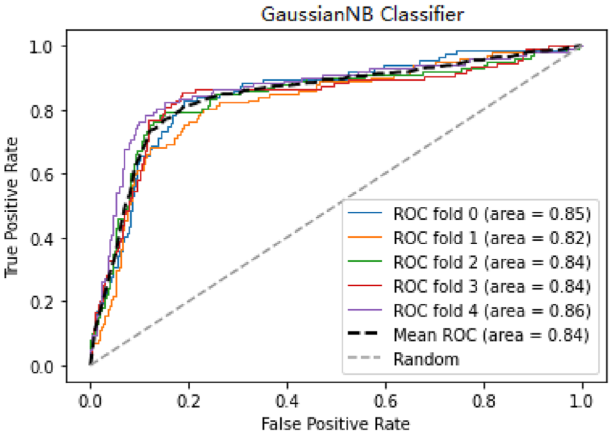


Fig. 21. GaussianNB algorithm AUC.

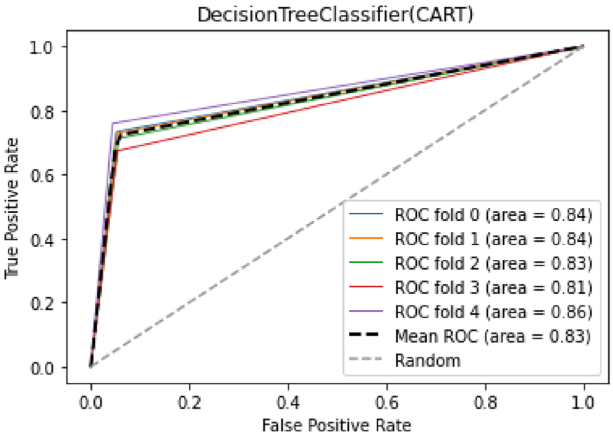


Fig. 22. DT(CART) algorithm AUC.

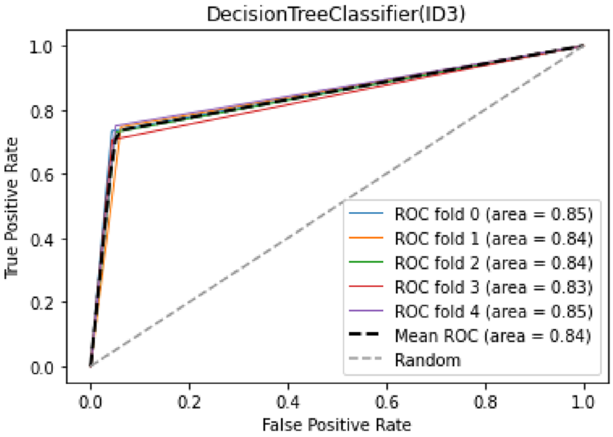


Fig. 23. ID3 algorithm AUC.

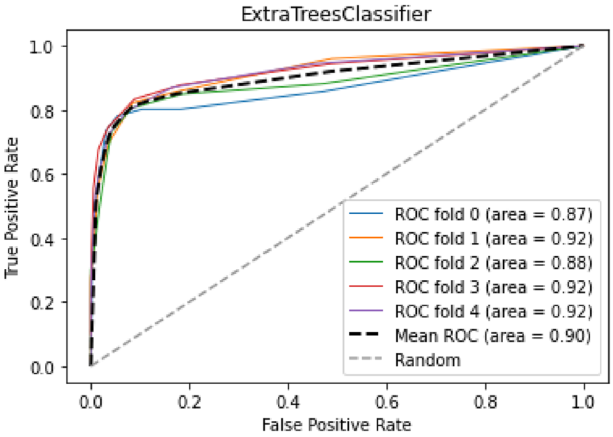


Fig. 24. ExtraTrees algorithm AUC.

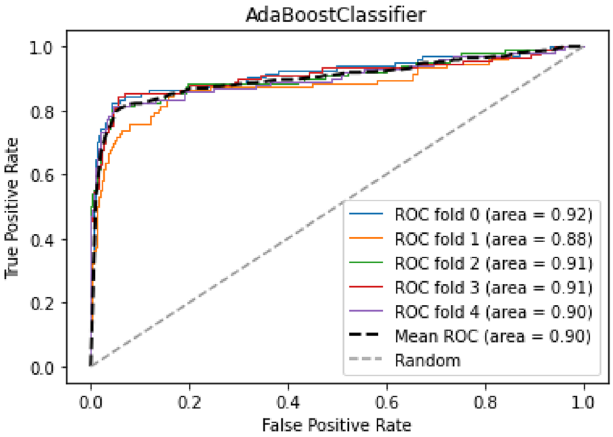


Fig. 25. AdaBoost algorithm AUC.

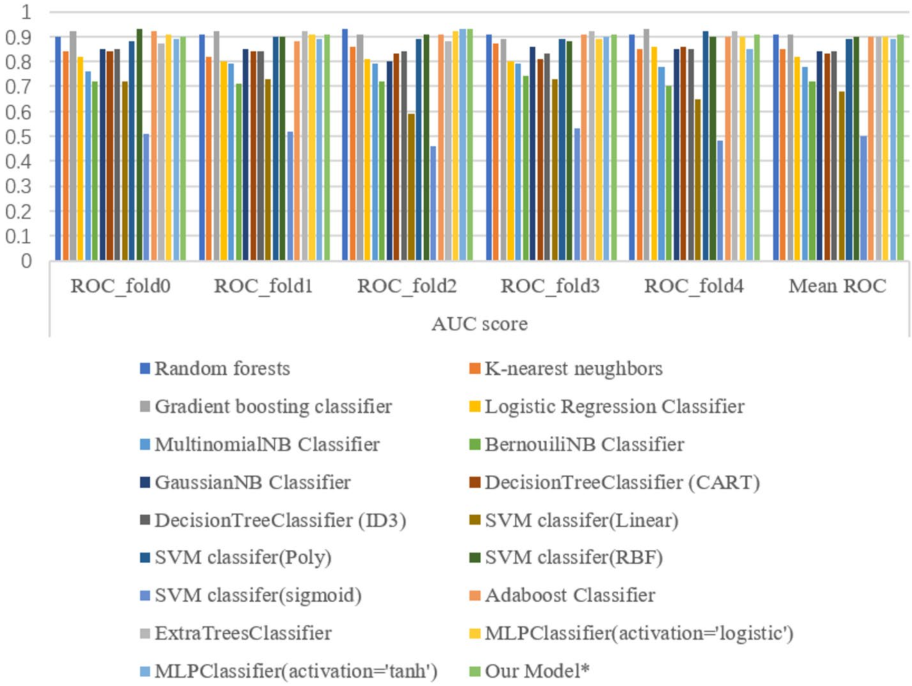


Fig. 26. Comparison of K-fold AUC for each algorithm.

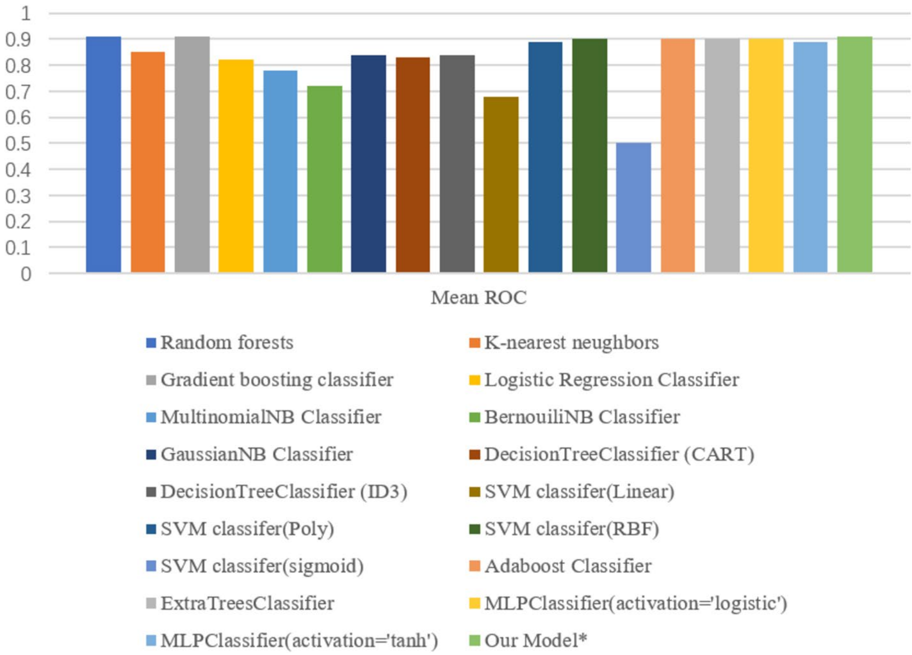
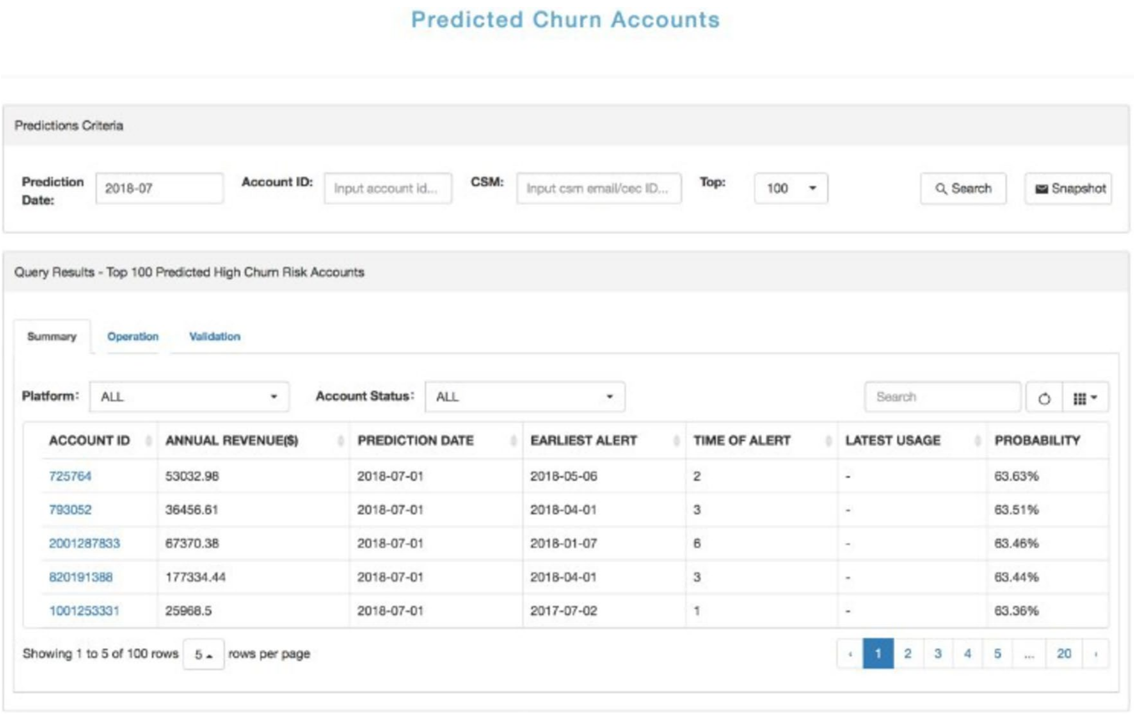


Fig. 27. Comparison of average AUC by algorithm.

relevant departments in the industry to analyze the current billing and usage trends of the account so as to identify the churn trends in time to take effective action.

Dashboard for an intelligent system for customer churn prediction

For dashboards designed for leadership decision makers, specific information about the results of predictive analysis of relevant customer churn data is presented in Figs. 31, 32, 33 and 34. The Real-Time Intelligent Alerts dashboard consists of a total of five sections. The first section is the overall trend in customer churn, which includes three parts: average churn rate, fully renewed accounts, and new onboarding contracts. The second section is Customer churn as a key driver for leading decision-making teams to make decisions. The third section



Predicted Churn Accounts

Predictions Criteria

Prediction Date: 2018-07 Account ID: Input account id... CSM: Input csm email/cec ID... Top: 100

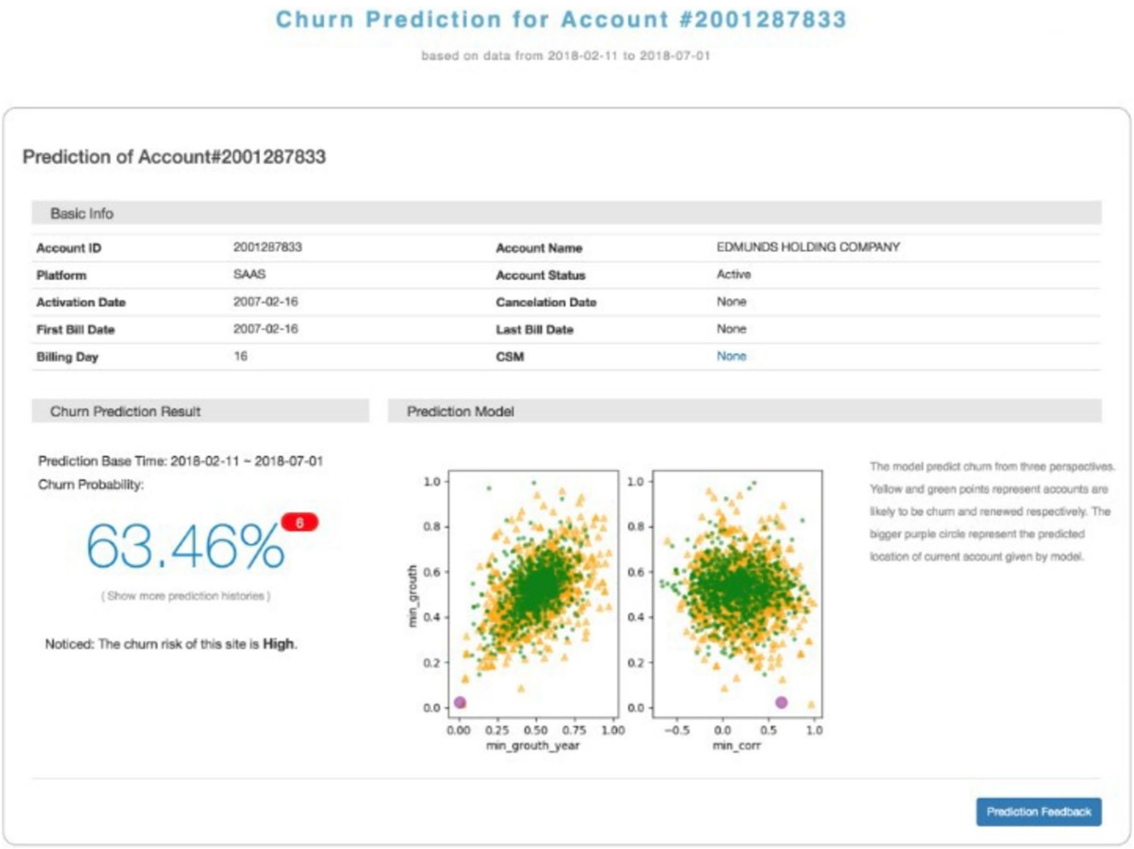
Query Results - Top 100 Predicted High Churn Risk Accounts

Platform: ALL Account Status: ALL

ACCOUNT ID	ANNUAL REVENUE($)	PREDICTION DATE	EARLIEST ALERT	TIME OF ALERT	LATEST USAGE	PROBABILITY
725764	53032.98	2018-07-01	2018-05-06	2	-	63.63%
793052	36456.61	2018-07-01	2018-04-01	3	-	63.51%
2001287833	67370.38	2018-07-01	2018-01-07	6	-	63.46%
820191388	177334.44	2018-07-01	2018-04-01	3	-	63.44%
1001253331	25968.5	2018-07-01	2017-07-02	1	-	63.36%

Showing 1 to 5 of 100 rows 5 rows per page

Fig. 28. Example display of customer churn information.



Churn Prediction for Account #2001287833

based on data from 2018-02-11 to 2018-07-01

Prediction of Account#2001287833

Basic Info

Account ID	2001287833	**Account Name**	EDMUNDS HOLDING COMPANY
Platform	SAAS	**Account Status**	Active
Activation Date	2007-02-16	**Cancelation Date**	None
First Bill Date	2007-02-16	**Last Bill Date**	None
Billing Day	16	**CSM**	None

Churn Prediction Result

Prediction Base Time: 2018-02-11 ~ 2018-07-01

Churn Probability:

63.46% 6

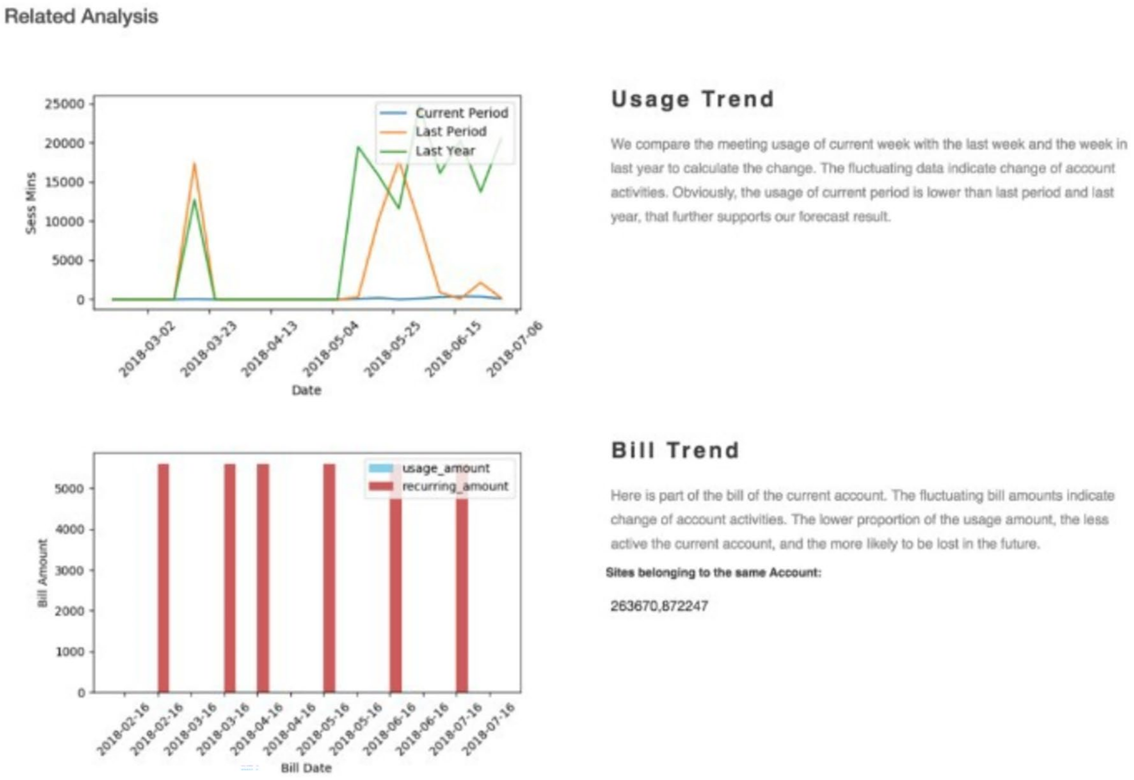
(Show more prediction histories)

Noticed: The churn risk of this site is **High**.

Prediction Model

The model predict churn from three perspectives. Yellow and green points represent accounts are likely to be churn and renewed respectively. The bigger purple circle represent the predicted location of current account given by model.

Fig. 29. Example display of lost customer details.


Related Analysis

Usage Trend

We compare the meeting usage of current week with the last week and the week in last year to calculate the change. The fluctuating data indicate change of account activities. Obviously, the usage of current period is lower than last period and last year, that further supports our forecast result.

Bill Trend

Here is part of the bill of the current account. The fluctuating bill amounts indicate change of account activities. The lower proportion of the usage amount, the less active the current account, and the more likely to be lost in the future.

Sites belonging to the same Account:

263670,872247

Fig. 30. Example display of user and account trends.

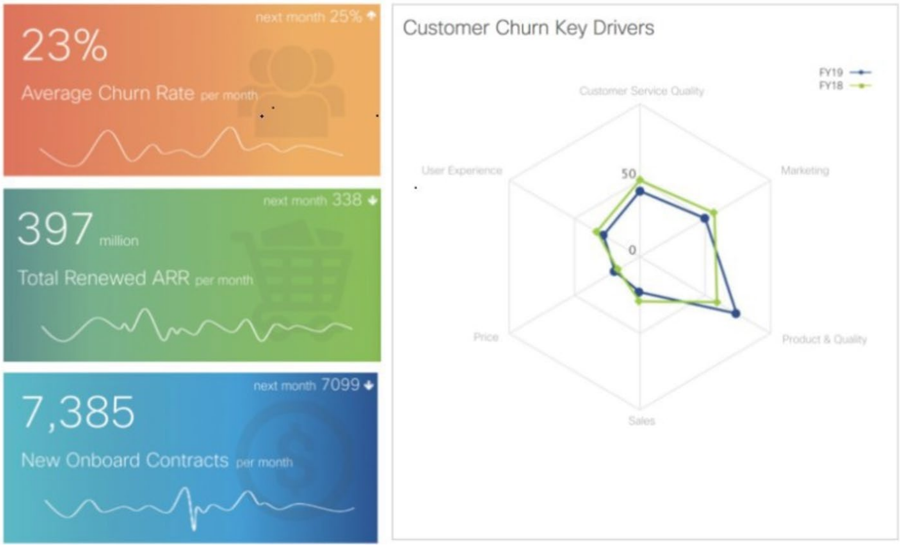


Fig. 31. Leadership Decision Panel Design—Generalized Information.

is the Churn heatmap (Churn Heatmap Description), which displays churn rates for selected regions and also provides a top correlation analysis and top correlation forecast for the next six months.

Customer churn prediction intelligent system evaluation module

In order to evaluate the performance of the model in the intelligent early warning system for customer churn based on the Ensemble-Fusion model, this subsection tests the 2018 production line production data. Figure 35 demonstrates the specific results of the evaluation, and the accuracy of the model is obtained by testing and validation to be above 95.8%, which achieves a high level of accuracy prediction. Higher accuracy means that more predicted churned customers are indeed likely to actually churn in the future, which does reduce the churn rate and retention of customers thus reducing the risk of fatalities to the organization due to customer churn.

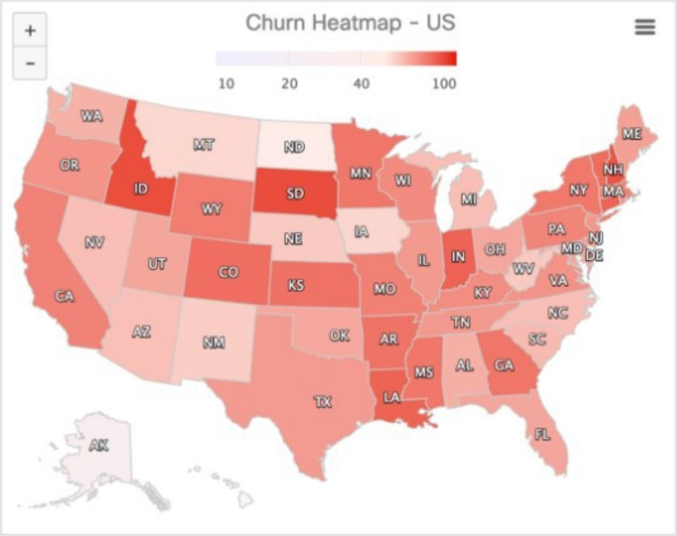
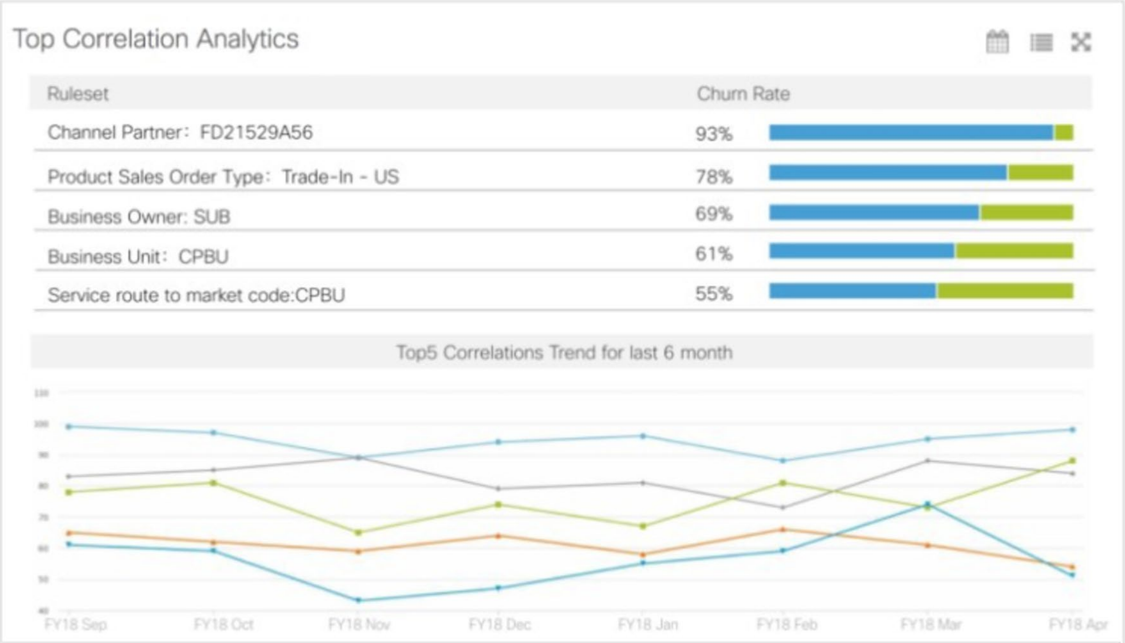


Fig. 32. Leadership Decision Panel Design—Churn Heat Map[44](We developed a customer churn intelligent early warning system using open source pyecharts, https://github.com/pyecharts/pyecharts).


Top Correlation Analytics

Ruleset	Churn Rate
Channel Partner: FD21529A56	93%
Product Sales Order Type: Trade-In - US	78%
Business Owner: SUB	69%
Business Unit: CPBU	61%
Service route to market code:CPBU	55%

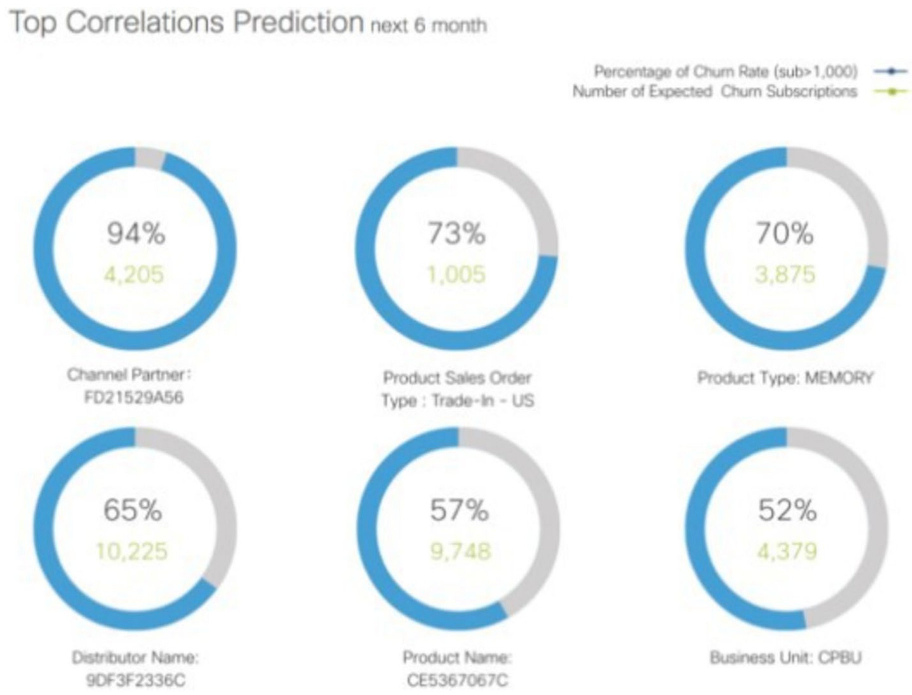
Fig. 33. Leadership decision panel design—correlation coefficient analysis.

Related work

To obtain the best model for customer churn prediction, this section will conduct a theoretical analysis of related machine learning algorithms and models. First, 9 categories and 17 algorithms related to machine-learning are expounded, and then in the third part, a prediction model of customer churn rate based on an ensemble-fusion model is proposed, and 17 sets of experiments are carried out to verify that the model has strong performance. Robust and easy to extend.

Support vector machines

Support vector machines(SVM)[10,11] are a set of supervised learning methods used for classification, regression, and outlier detection[12]. The advantages of support vector machines are effective in high dimensional spaces. Still effective in cases where the number of dimensions is greater than the number of samples. The objective function:



Top Correlations Prediction next 6 month

Percentage of Churn Rate (sub>1,000)
Number of Expected Churn Subscriptions

94%
4,205
Channel Partner: FD21529A56

73%
1,005
Product Sales Order Type : Trade-In - US

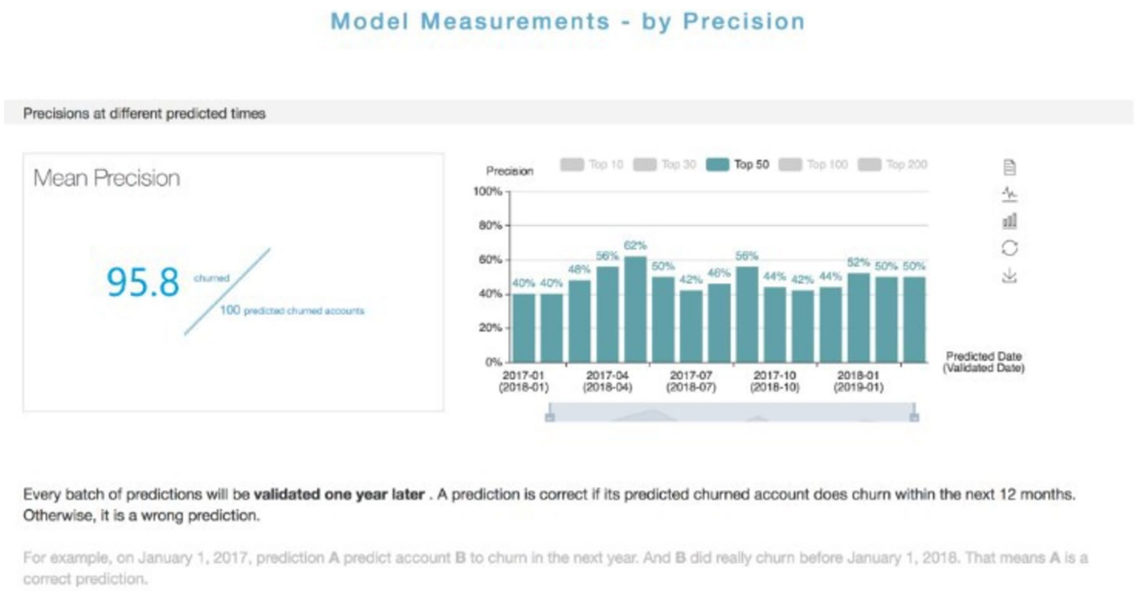
70%
3,875
Product Type: MEMORY

65%
10,225
Distributor Name: 9DF3F2336C

57%
9,748
Product Name: CE5367067C

52%
4,379
Business Unit: CPBU

Fig. 34. Leadership decision panel design—360 degree information analysis presentation.



Model Measurements - by Precision

Precisions at different predicted times

Mean Precision

95.8 churned / 100 predicted churned accounts

Every batch of predictions will be **validated one year later** . A prediction is correct if its predicted churned account does churn within the next 12 months. Otherwise, it is a wrong prediction.

For example, on January 1, 2017, prediction A predict account B to churn in the next year. And B did really churn before January 1, 2018. That means A is a correct prediction.

Fig. 35. Customer churn prediction model evaluation page.

$$L(w, b, alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{n} \alpha_i y_i (wx_i + b) + \sum_{i=1}^{n} \alpha_i \tag{3}$$

SVM is a supervised learning models that analyze data used for classification and regression analysis. In the customer churn prediction, SVM divides the result of prediction into two parts, such as positive is customer churn while negative is customer non-churn. The kernel of SVM is used like linear, poly and RBF.

Random forests

Random forests are constructed by several trees[16] and each decision tree is trained by random samples. A random forest is a data construct applied to machine learning that develops large numbers of random decision trees analyzing sets of variables. This type of algorithm helps to enhance the ways that technologies analyze complex data. The Random Forest algorithm is one of the best algorithms for classification. RF can classify large data with accuracy. It is a learning method in which the number of decision trees is constructed at the time of training and outputs of the modal predicted by the individual trees. RF acts as a tree predictor where every tree depends on the ran- dom vector values. The basic concept behind this is that a group of "weak learners" may come together to build a "strong learner". Random forest models are machine learning models that make output predictions by combining outcomes from a sequence of regression decision trees. Each tree is constructed independently and depends on a random vector sampled from the input data, with all the trees in the forest having the same distribution. The predictions from the forests are averaged using bootstrap aggregation and random feature selection. RF models have been demonstrated to be robust predictors for both small sample sizes and high dimensional data. RF clas- sification models were constructed that directly classified bioreactor runs as having sufficient or insufficient cardiomyopathy content.

K-nearest-neighbors

K-nearest-neighbors algorithm (KNN) is a non-parametric classification method first developed by Evelyn Fix and Joseph Hodges in 1951[17]. It is used for classification and regression. In both cases, the input consists of the k closest training examples in the data set. The output depends on whether KNN[18] is used for classification or regression. The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples. The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. The principle behind nearest neighbor methods is to find a predefined number of training samples closest in distance to the new point and predict the label from these. The number of samples can be a user-defined constant (k-nearest neighbor learning) or vary based on the local density of points (radius-based neighbor learning). The distance can, in general, be any metric measure: standard Euclidean distance is the most common choice. Neighbors- based methods are known as non-generalizing machine learning methods since they simply "remember" all of their training data. KNN is a non-parametric algorithm, which means it does not make any assumptions on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the data set and at the time of classification, it performs an action on the data set. KNN algorithm at the training phase just stores the data set and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Gradient boosting classifier

Gradient boosting[34, 35]produces a model in the form of an ensemble of the prediction model, usually there using decision trees. Gradient boosting classifier has a lot of advantages, such as high prediction rate, dealing with non-linear data, and flex- ible handling of various types of data. Predictions are made by the majority vote of the weak learners' predictions, weighted by their individual accuracy. Gradient boosting machines are an extremely popular machine learning algorithm that has proven successful across many domains. A simple GBM model contains two categories of hyper-parameters: boosting hyper-parameters and tree-specific hyper-parameters. Gradient boosting re-defines boosting as a numerical optimization problem where the objective is to minimize the loss function of the model by adding weak learners using gradient descent. Gradient descent is a first-order iterative optimization algorithm for finding a local minimum of a differentiable function. As gradient boosting is based on minimizing a loss function, different types of loss functions can be used resulting in a flexible technique that can be applied to regression, multi-class classification.

Theoretical analysis of customer churn rate prediction based on logistic regression

Logistic regression[19, 20]is a generalized linear regression analysis model, which is divided from the classification of machine learning. It belongs to the classification algorithm in supervised learning. Due to the good performance of logistic regression[19], it can often be used for binary classification. or multi-classification problems. In the research on the prediction of customer churn rate, logistic regression can be abstracted here to deal with the binary classification problem, such as the label of marking customer churn as 0, and the label of non-churn as 1. At this time, for each set of input data, according to the Sigmoid function[20]

$$g(z) = \frac{1}{1 + e(-z)} \tag{4}$$

in logistic regression, the predicted value can be mapped to between [0, 1]. If $y \geq 0.5$, it is recorded as 0 category is the loss, and similarly, it is 1 category that is not lost.

Theoretical analysis of customer churn rate prediction based on Bayesian theory

The research on customer churn prediction is currently limited to the application stage of Naive Bayes[8]. The basic idea of the Naive Bayes algorithm[9]: for a given category to be classified, solve the problem under the condition that this category appears. The probability of occurrence of each category, which category has the highest probability of occurrence, is considered to be the category to which the item to be classified belongs.

Theoretical analysis of customer churn rate prediction based on decision tree

In the research on customer churn prediction, a few pieces of literature use a decision tree algorithm[21]. A decision tree is also called a decision tree in some literature[22]. This kind of algorithm belongs to supervised learning in machine learning, which can be used to solve classification and regression problems. The decision tree algorithm is a top-down divide-and-conquer strategy, a recursive algorithm from the root node to the leaf node,

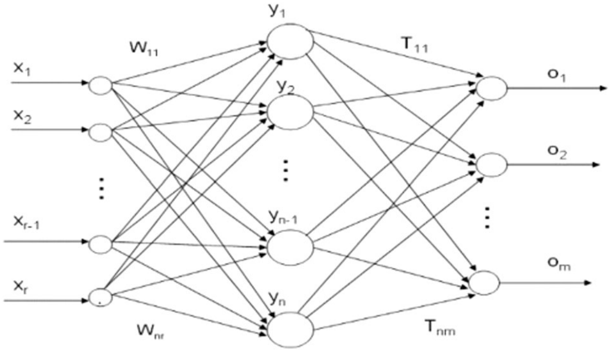


Fig. 36. The structure of three-layer BP neural network.

where the leaf nodes are divided according to different division methods, generally according to information gain, gain rate, and Gini index[23].The decision tree is divided the algorithms are ID3 algorithm, C4.5 algorithm and CART algorithm[24].

Theoretical analysis of customer churn rate prediction based on neural network

In recent years, deep learning has been widely used to solve some complex problems, and it is also used in the prediction of customer churn rate[25]. The BP neural network was proposed by a group of scientists led by Rumelhart and McCelland in the book "Parallel Distributed Processing" in 1986, which detailed the error back-propagation algorithm for multilayer perceptions with nonlinear continuous transformation functions. The analysis of, realizes Minsky's vision of multi-layer network[26]. The structure of BP neural network[26]is a backpropagation (Back Propagation) neural network, referred to as the BP neural network. The standard BP neural network is divided into three layers, namely the input layer, the hidden layer and the output layer, as shown in Fig. 36.

The principle of the neural network algorithm mainly includes two stages: (1)FP (forward propagation) data is input from the input layer, then input through the hidden layer under the mapping of the relevant activation function, and finally reaches the output layer for output, and then according to the error between the expected output and the actual output is used to construct the cost function (loss function) for the second stage (2) BP (backpropagation) from the output layer through each hidden layer to correct the weight and bias of the hidden layer by layer, and finally correct the weights and biases from the hidden layer to the input layer, and finally get the neural network model. Neural networks can approximate any nonlinear function arbitrarily. Because of their simple structure and easy implementation, they have been widely used in time series analysis and nonlinear function regression estimation. However, the development of such networks is limited due to the difficulty of determining the network structure, the existence of over-learning, and the tendency to fall into local extreme values. This paper expects to use it in the research of customer churn prediction to get good results.

Conclusions and future work

In this paper, we proposed a novel model named Ensemble-Fusion that utilized 9 categories of 17 machine learning algorithms as baseline classifiers. Through experiment proves that the Ensemble-Fusion model**(Our model)** reaches 95.35%, AUC score reaches 91% and F1-Score reaches 96.96%, and the experimental results show that the data prediction accuracy of Ensemble-Fusion model outperforms that of other benchmark algorithms. This paper first elaborates on the important role of research in today's information industry and gives important contributions, then this paper focuses on the research of customer churn prediction based on an integrated learning fusion model, mainly from the customer churn prediction solution based on the integrated learning fusion model, the design of real-time intelligent early warning system of customer churn, the machine learning algorithm of customer churn prediction and this paper. The newly proposed customer churn prediction model is compared and the specific implementation algorithm based on the integrated learning fusion model is given. Then this paper validates the proposed churn prediction algorithm experimentally and evaluates the robustness of the algorithm by using evaluation metrics such as precision, recall, accuracy, F1-score, and AUC. Finally, this paper provides a detailed description of the main functions of the theoretically and practically developed customer churn intelligent early warning system, in order to efficiently help the information industry improve its productivity and to be able to excel in today's globally competitive environment.The study presented in this paper is not free of limitations. Firstly, it is challenging to gather all relevant data on customer churn due to sensitive information and related protocol issues. Therefore, how to construct an effective model using the limited dataset becomes a bottleneck in customer churn prediction research. The other limitation of the study is that there is still a lot of noise and no labels to mark customer churn in the collected data, which requires a lot of time to organize and learn relevant business knowledge before data collection and processing. Finally, customer churn is a multidisciplinary issue involving a variety of fields such as psychology, sociology, and economics, but current research may lack an interdisciplinary perspective and approach.Concerning future research, we intend to develop a similar ensemble-fusion classification algorithm that substitutes the baseline classifiers with reinforcement learning model-related algorithms. The primary aim here is to construct an ensemble classifier that can more easily be used in complex data structures such as multisource isomerization. In order to study customer churn in more depth in the future, there are several potential directions for further research. The first direction is to obtain more data from industry, e.g., combining different feature data. Another interesting direction is to relax strict algorithmic constraints to support compact and dense feature representations, which can be explored in areas such as fast symmetric decomposition techniques.

Data availability

Availability of data and materials： Data is available on request from the author (Chenggang He).

Received: 27 January 2024; Accepted: 26 August 2024
Published online: 30 August 2024

References

1. Fujo, S. W. *et al.* Customer churn prediction in telecommunication industry using deep learning. *Inf. Sci. Lett.* **11**(1), 24 (2022).
2. Xie, Y., Li, X., Ngai, E. & Ying, W. Customer churn prediction using improved balanced random forests. *Expert Syst. Appl.* **36**(3), 5445–5449 (2009).
3. De Caigny, A., Coussement, K. & De Bock, K. W. A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *Eur. J. Oper. Res.* **269**(2), 760–772 (2018).
4. Ahmad, A. K., Jafar, A. & Aljoumaa, K. Customer churn prediction in telecom using machine learning in big data platform. *J. Big Data* **6**(1), 1–24 (2019).
5. He, C., Ding, C.H., Chen, S., Luo, B. Intelligent machine learning system for predicting customer churn. In: *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 522–527 (2021). IEEE.
6. Estran, R., Souchaud, A. & Abitbol, D. Using a genetic algorithm to optimize an expert credit rating model. *Expert Syst. Appl.* **203**, 117506 (2022).
7. Deng, Z., Huang, Z.-H. & Miao, X. Sufficient conditions for judging quasi-strictly diagonally dominant tensors. *Comput. Appl. Math.* **42**(1), 63 (2023).
8. Chen, Y., Matsubara, T., Yaguchi, T. Kam theory meets statistical learning theory: Hamiltonian neural networks with non-zero training loss. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp 6322–6332 (2022).
9. Bhavan, A. *et al.* Bagged support vector machines for emotion recognition from speech. *Knowl.-Based Syst.* **184**, 104886 (2019).
10. Sadohara, R. *et al.* Seed coat color genetics and genotype× environment effects in yellow beans via machine- learning and genome-wide association. *Plant Genom* **15**(1), 20173 (2022).
11. Varshney, R. K. *et al.* Resequencing of 429 chickpea accessions from 45 countries provides insights into genome diversity, domestication and agronomic traits. *Nat. Genet.* **51**(5), 857–864 (2019).
12. Bergström, A. *et al.* Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**(6484), 5012 (2020).
13. Keramati, A., Ghaneei, H. & Mirmohammadi, S. M. Developing a prediction model for customer churn from electronic banking services using data mining. *Financ. Innov.* **2**, 1–13 (2016).
14. Hudaib, A. *et al.* Hybrid data mining models for predicting customer churn. *Int. J. Commun. Netw. Syst. Sci.* **8**(05), 91 (2015).
15. Li, H., Yang, D., Yang, L., Lu, Y., Lin, X. Supervised massive data analysis for telecommunication customer churn prediction. In: 2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Comput- ing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom)(BDCloud-SocialCom-SustainCom), pp 163–169 (2016). IEEE.
16. Deng, Q. & Söffker, D. A review of hmm-based approaches of driving behaviors recognition and prediction. *IEEE Trans. Intell. Vehic.* **7**(1), 21–31 (2021).
17. Graham, S. E. *et al.* The power of genetic diversity in genome-wide association studies of lipids. *Nature* **600**(7890), 675–679 (2021).
18. Shen, J. *et al.* Identification of a novel gene signature for the prediction of recurrence in hcc patients by machine learning of genome-wide databases. *Sci. Rep.* **10**(1), 4435 (2020).
19. Devriendt, F., Berrevoets, J. & Verbeke, W. Why you should stop predicting customer churn and start using uplift models. *Inf. Sci.* **548**, 497–515 (2021).
20. Wang, Q.-F., Xu, M. & Hussain, A. Large-scale ensemble model for customer churn prediction in search ads. *Cognit. Comput.* **11**, 262–270 (2019).
21. Alboukaey, N., Joukhadar, A. & Ghneim, N. Dynamic behavior based churn prediction in mobile telecom. *Expert Syst. Appl.* **162**, 113779 (2020).
22. Wang, S., Cao, J. & Philip, S. Y. Deep learning for spatio-temporal data mining: A survey. *IEEE Trans. Knowl. Data Eng.* **34**(8), 3681–3700 (2020).
23. Zdravevski, E., Lameski, P., Apanowicz, C. & Ślęzak, D. From big data to business analytics: The case study of churn prediction. *Appl. Soft Comput.* **90**, 106164 (2020).
24. Vo, N. N., Liu, S., Li, X. & Xu, G. Leveraging unstructured call log data for customer churn prediction. *Knowl.-Based Syst.* **212**, 106586 (2021).
25. Goecks, J., Jalili, V., Heiser, L. M. & Gray, J. W. How machine learning will transform biomedicine. *Cell* **181**(1), 92–101 (2020).
26. Aria, M., Cuccurullo, C. & Gnasso, A. A comparison among interpretative proposals for random forests. *Mach. Learn. Appl.* **6**, 100094 (2021).
27. Alotaibi, M. Z. & Haq, M. A. Customer churn prediction for telecommunication companies using machine learning and ensemble methods. *Eng. Technol. Appl. Sci. Res.* **14**, 14572–14578 (2024).
28. Alabdulwahab, A., Haq, M. A. & Alshehri, M. Cyberbullying detection using machine learning and deep learning. *Int. J. Adv. Comput. Sci. Appl.* **14**, 10 (2023).
29. Haq, M. A., Khan, M. A. & Alshehri, M. Insider threat detection based on NLP word embedding and machine learning. *Intell. Autom. Soft Comput.* **33**, 619–635 (2022).
30. Oksuz, K., Cam, B. C., Kalkan, S. & Akbas, E. Imbalance problems in object detec- tion: A review. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(10), 3388–3415 (2020).
31. Zidan, M. A. *et al.* A general memristor-based partial differential equation solver. *Nat. Electron.* **1**(7), 411–420 (2018).
32. Devriendt, F., Berrevoets, J. & Verbeke, W. Why you should stop predicting cus- tomer churn and start using uplift models. *Inf. Sci.* **548**, 497–515 (2021).
33. Shirazi, F. & Mohammadi, M. A big data analytics model for customer churn prediction in the retiree segment. *Int. J. Inf. Manag.* **48**, 238–253 (2019).
34. Amin, A. *et al.* Cross-company customer churn prediction in telecommunication: A comparison of data transformation methods. *Int. J. Inf. Manag.* **46**, 304–319 (2019).
35. Stripling, E., Broucke, S., Antonio, K., Baesens, B. & Snoeck, M. Profit maximizing logistic model for customer churn prediction using genetic algorithms. *Swarm Evolut. Comput.* **40**, 116–130 (2018).

36. Liu, Z. *et al.* Extreme gradient boosting trees with efficient Bayesian optimization for profit-driven customer churn prediction. *Technol. Forecast. Soc. Change* **198**, 122945 (2024).
37. Chicco, D. & Jurman, G. The advantages of the matthews correlation coeffi- cient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genom.* **21**(1), 1–13 (2020).
38. Carvalho, D. V., Pereira, E. M. & Cardoso, J. S. Machine learning interpretability: A survey on methods and metrics. *Electronics* **8**(8), 832 (2019).
39. Chou, J.-S. & Nguyen, T.-K. Forward forecast of stock price using sliding- window metaheuristic-optimized machine-learning regression. *IEEE Trans. Ind. Inform.* **14**(7), 3132–3142 (2018).
40. Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G. & Chatzisavvas, K. C. A com- parison of machine learning techniques for customer churn prediction. *Simul. Model. Pract. Theory* **55**, 1–9 (2015).
41. Ismail, M. R., Awang, M. K., Rahman, M. N. A. & Makhtar, M. A multi-layer percep- tron approach for customer churn prediction. *Int. J. Multimed. Ubiquitous Eng.* **10**(7), 213–222 (2015).
42. Riedmiller, M., Lernen, A. Multilayer perceptron. Machine learning lab special lecture, University of Freiburg **24** (2014).
43. Jain, H., Khunteta, A. & Srivastava, S. Churn prediction in telecommunication using logistic regression and logit boost. *Procedia Comput. Sci.* **167**, 101–112 (2020).
44. Pyecharts, https://github.com/pyecharts/pyecharts.

Acknowledgements

Thanks for Authors Contributions: "Conceptualization, Chenggang He; methodology, Chenggang He; validation, Chenggang He,Chris H.Q.Ding; investigation, Chenggang He; writing—original draft preparation, Chenggang He; writing—review and editing, Chenggang He; supervision, Chris H.Q.Ding.

Author contributions

Conceptualization, C.H.; methodology, C.H.; validation, C.H.; investigation, C.H.; writing—original draft preparation, C.H.; writing—review and editing, C.H., C. H. Q. D.; supervision, C. H. Q. D.. All authors reviewed the manuscript.

Funding

This research was funded by the Scientific Research Foundation for High- level Talents of Anhui University of Science and Technology(2023yjrc120), Anhui Quality Engineering Project(2023cyts013), NSFC Key Project of International (Regional) Cooperation and Exchanges (61860206004), Natural Science Foundation of China (61976004,61572030).

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to C.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2024