Loan Default Prediction with Machine Learning Techniques

Lili Lai
College of Information Engineering, China Jiliang University, Hangzhou, 310018, China

***Abstract*—Loan business is one of the major income sources for bank. However, loan default problem is a major issue for loan business. With the rise of big data era and the development of machine learning techniques, nowadays we have more options for classifying and predicting loan default, other than manual processing. With a real-world dataset from a prestigious international bank, we demonstrate that the AdaBoost model can achieve a 100% accuracy for predicting loan default, outperforming other models including XGBoost, random forest, k nearest neighbors, and multilayer perceptrons. Our result shows the promising application of machine learning techniques in the financial industry.**

Keywords-loan default; machine learning; XGBoost; AdaBoost

I. INTRODUCTION

Credit lending business is an innovation in banking loan business, and powerfully promotes the growth of consumption and economy. However, with the popularity of credit lending business and the gradual expansion of its scale, loan default problem exposes to the public. Loan default problem plays a vital role in both traditional banking industry and emerging Internet financial industry, loan default will cause damage to banks supporting country's economy, what is worse, may even result in economic crisis. Therefore, it is necessary to establish and perfect credit lending risk management system, and predict loan default problem to reduce default risk in the meantime.

In the past, human screening was the major method to predict loan default. It always cost a huge amount of time and workforce with a lot of trouble due to the considerable data. Using machine learning to predict loan default probability is time-saving and workload-saving, which can heavily increase the effectiveness and accuracy.

The driven factors behind the use of machine learning methods for financial applications are two-fold. With the popularity of electronic transactions, online and mobile banks, the development of third-party mobile and online payment platforms, banks are collecting more data about their customers from both internal and external approaches, which become the basis of predicting loan default in a view of big data. The rapid development and success of applying machine learning and deep learning models in various problems (e.g., image classification [1, 2], stock market prediction [3, 4] and traffic forecasting [5, 6]) also inspires the idea of applying them in a series of financial applications, which were previously dominated by manual operations.

In this paper, based on a real-world data set provided by Xiamen International Bank, we predict the happening of loan default or not and compare the performance of five machine learning models from different families, including XGBoost, random forest (RF), AdaBoost, k nearest neighbors (kNN), multilayer perceptrons(MLP). Besides analyzing the prediction result, we also present our preprocessing process in this paper. Our result indicates that AdaBoost achieves an accuracy of 100% in this dataset, thus showing a promising application of AdaBoost model in the real scenario.

Our contributions in this paper are summarized as follows:

(1) We study the loan default prediction problem with a real-world dataset, which is also publicly accessible.

(2) We conduct experiments with five sophisticated machine learning models, i.e., XGBoost, RF, AdaBoost, kNN, and MLP.

(3) We manage to achieve a 100% accuracy with our feature extraction techniques and the AdaBoost model.

The rest of this paper is organized as follows. Section 2 gives a short review of related work in the area of Fintech, which inspires our approach of using machine learning methods for loan default prediction. In Section 3, we describe the characteristics of the data set we use, including its initial features and what information it contains. Section 4 presents our experiment process on the data set with machine learning models and the results. We draw our conclusion in the final section.

II. RELATED WORK

In this section, we give a short review of the related work of applying machine learning techniques to financial applications, i.e., corporate credit risk, consumer credit risk, and credit risk of P2P lending.

A. Corporate Credit Risk

Corporations are the major clients to the loan business of a bank. Due to the high average loan of corporations as clients, banks will be hugely impacted once default occurs. Therefore, there was a long time in the past that abundant researches about corporate credit risk were conducted, some of which used machine learning.

In [7], the author applies support vector machines (SVM) to corporate credit rating problem, using a grid search technique with 5-fold cross validation to figure out the optimal parameter value of RBF kernel function of SVM. In addition, the author compares SVM's performance with multiple discriminant analysis (MDA), case-based reasoning (CBR) and three-lay fully connected back-propagation neural networks (BPNs) in order to evaluate the prediction accuracy of SVM. The results show that SVM is superior to other existing methods and can be considered as a promising alternative method.

978-1-7281-4349-1/20/$31.00 ©2020 IEEE
DOI 10.1109/CCNS50731.2020.00009

In [8], the authors conduct a comprehensive experimental comparison study on the effectiveness of four learning algorithms (i.e., single hidden layer feedforward networks (SLFN), backpropagation (BP), extreme learning machine (ELM), incremental extreme learning machine (I-ELM) and SVM) for a data set composed of real financial data of corporate credit ratings. The results show that SLFN based approaches are more reliable than SVMs, while SVMs outperform other methods on output distributions.

In [9], the authors propose a new MSVM classifier termed ordinal multi-class support vector machine (OMSVM), which aims to extend binary SVM through applying ordinal pairwise partitioning (OPP). It can cope with multiple ordinal classes effectively. To validate the effectiveness of this new classifier, the authors apply it in a real bond rating case in Korea, comparing the results of OMSVM to results of conventional MSVM and other artificial intelligence models including MDA, MLOGIT, CBR and ANN. The conclusion is that OMSVM achieves better classification performance and needs less computing resources than other typical multiple-class classification techniques.

In [10], the authors propose a new method for multi-classification based on support vector domain and fuzzy clustering algorithm to complete enterprise credit rating. By using a fuzzy clustering algorithm to preprocess the data, only the boundary data points are selected as training samples to complete the designation of the support vector field, thereby reducing the calculation cost and obtaining better performance. In order to verify the proposed method, this paper uses actual cases for experiments and compares the results with traditional multi-class support vector machine methods and other artificial intelligence techniques. The results show that the proposed model improves the performance of corporate credit ratings with less computational consumption.

B. Consumer Credit Risk

Consumer credit scoring has been a challenging issue for a bank because of the changeable personal situations of individual users. Different loans, such as consumer loans and mortgage loans, require different models to predict and control risk. With the emergence of the big data era, the digital information recorded on social networks and mobile applications can also be used for consumer credit risk related research.

In [11], authors apply convolutional neural networks to consumer transaction data to predict mortgage defaults. For each consumer, apart from the number of daily transactions on the checking account and the money transferred to the checking account, this paper also uses the balance of the checking account, savings account and credit card. In the absence of other information about each consumer, the convolutional neural network model can obtain a ROC AUC of 0.98, and the random forest classifier model can obtain a ROC AUC of 0.926.

In [12], the authors do some research about behaviors of default prediction models based on credit scoring methods and computational techniques with machine learning algorithms. The authors compare the prediction performance of different models with the data of "My Home, My Life" program, and the results indicate that: (1) the accuracy of models improves with the number of days overdue increasing; (2) traditional ensemble techniques, including bagging (BG), random forest (RF), and Boosting, obtained the best prediction results; (3) a negative impact was caused on all standards when a smaller number of observations was used, especially on the type II error.

In [13], the authors propose a nonparametric ensemble tree model named gradient boosting survival tree (GBST), which extends survival tree with the gradient boosting algorithm and learns survival tree ensemble by minimizing the negative log-likelihood in an additive manner, in order to process highly heterogeneous industry data collected in the Chinese consumer financial market. GBST optimizes the survival probability simultaneously for each time period, which can reduce the overall error significantly.

In [14], the authors provide a common framework to assess individual consumer's credit risk through applying machine learning methods. In this paper, both machine learning methods and optimized logic regression are applied to analyze the dataset of complete payment histories of short-termed installment credits, and describe an algorithm used to adjust the size of terminal nodes for probability assessment. The results show that the machine learning method of regression random forests is superior to optimized logic regression especially when are applied to a large credit scoring data set.

C. Credit Risk of P2P Lending

Peer to Peer (P2P) lending is a significant innovation application in financial field for some time past. However, the incomplete personal information as well as neglect of risk control and management during the process of rapidly growing business make P2P lending one of the fields where defaults occur most easily. To help P2P blossom healthily and reduce default rates, there were considerable studies trying to utilize machine learning models to conduct studies relevant to Credit Risk of P2P Lending.

In [15], the authors utilize the loan data on Australia P2P platform and compare the traditional machine learning algorithm with CatBoost algorithm. The analysis results of the AUC value and accuracy of these algorithm as evaluation indicators demonstrate that CatBoost outperforms traditional machine learning algorithms with higher accuracy in credit scoring.

In [16], the authors apply modern machine learning algorithm LightGBM and XGBoost to real P2P transaction data of Lending Club where creatively forecast the loan default risk, and compare the results of diverse methods. The multiple observational data set classification prediction results show that LightGBM is the best.

In [17], the authors propose a credit scoring model with artificial neural networks to divide P2P loan application into groups of default and non-default, the output results demonstrate that the credit scoring model based on neural networks can screen default effectively.

III. Dataset Description

In this paper, we try and compare five sophisticated machine learning models (i.e., XGBoost, RF, AdaBoost, kNN, and MLP), using the loan dataset of actual business occasions in Xiamen

International Bank. The dataset totally includes 132,029 records which can be divided into three types of input information, which are:

(1) User's basic attributes: containing the user's ID number, gender, age, region, education, work type, ethnicity, highest education, and the start and expiration date of the ID card.

(2) Lending related information: including the user's loan product type, loan for margin transactions, basic rating level,

TABLE I. USER'S BASIC ATTRIBUTES

Field names	Attributes	Field names	Attributes
id	User's unique number	edu	User's education
target	Default sign: 0 or 1	job	Type of user's job
certId	User's credit ID number	ethic	User's ethnicity
gender	Female of male	highestEdu	User's highest education
age	User's age	certValidBegin	Credit start date
dist	User's region	certValidStop	Credit invalid date

TABLE II. LENDING RELATED INFORMATION

Field names	Attributes	Field names	Attributes
loanProduct	Loan product type	residentAddr	Resident address
lmt	loan for margin transactions	setupHour	Application hour for the loan
basicLevel	Basic rating level	weekday	Application date
bankCard	Bank card number	isNew	Whether it is a new data

bank card number, resident address, application hour and application date, etc.

(3) Information related to user credit reporting: This part of the data involves sensitive third-party data, and Xiamen International Bank did not provide further instructions.

In Fig.1, the distribution of loan of margin transactions is given as an example of input characteristics, with the variable of lmt refers to loan of margin transaction.

Our prediction target is a categorical variable, namely, default (0) and non-default (1). According to Fig.2, this is a highly misbalanced problem, thus we will use AUC instead of accuracy as our evaluation metric.

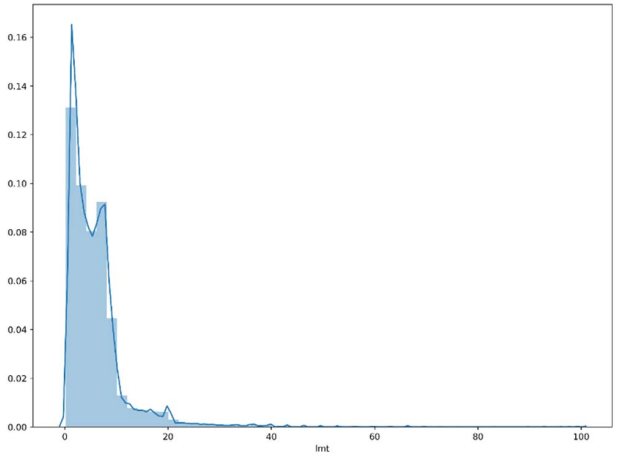

Figure 1. Distribution of loan of margin transactions

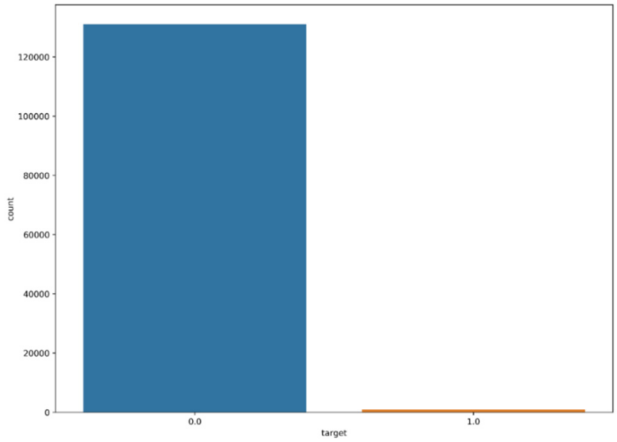

Figure 2. Distribution of prediction target

Through the process of experiment, we randomly pick 80 percent of the whole data as the training set and the rest as the validation set.

IV. EXPERIMENT

A. Feature Extraction

Our feature extraction section consists of the following steps:

Remove duplicate features: there are many duplicate or meaningless data, for example, X_0—X_9 in the user credit related information. Hence, we delete these data and the final data was reduced by 34 columns. Furthermore, we also fill in some missing feature values, the retained data is more streamlined and valuable.

Classify the rest data which is worth as features into numeric data and category data, and one-hot encode the category data and convert it into dummy variables. Different features are grouped and their median and mean values are used as new features.

Extraction of fine-grained address information: we pick up the users' address information, count the user ratings and education level of regional loan users, and take the average levels of users in different regions as new features.

Extraction of bank card information: extract information such as the validity period of bank cards reserved by users.

B. Evaluation Metric

In this paper, we mainly use AUC value as our evaluation indicator. The binary classification we are going to deal with is a common problem in machine learning. Receiver operating characteristic (ROC) and area under the curve (AUC) we are going to use are common tools to evaluate the strengths and weaknesses of a binary classifier.

A receiver operating characteristic (ROC), is a coordinate graph analysis curve representing the diagnostic capability of the binary classifier system when discrimination threshold changes. Its abscissa is specificity, which is false positive rate (FPR), illustrating the proportion of samples that are predicted to be positive but in fact negative in all negative examples; the ordinate is sensitivity, which is true positive rate (TPR), representing the proportion of samples that are predicted to be positive and actually positive.

AUC is the area of under the ROC curve, thus obviously it is not bigger than 1. Since the ROC curve is usually above the straight line of y=x, the value range of AUC is generally between 0.5 and 1. The AUC value is used as the evaluation criterion because the ROC curve does not clearly indicate which classifier has better effect, and as a value, the classifier corresponding to a larger AUC has a better effect.

The criteria of AUC based prediction classifier:

AUC = 1: it is a perfect classifier with 100% accuracy.

AUC = [0.85, 0.95]: work well.

AUC = [0.7, 0.85]: with general effect.

AUC = [0.5, 0.7]: limited effect, but useful in loan default prediction.

AUC = 0.5: as valueless as random guesses.

AUC <0.5: worse than random guesses.

C. Machine Learning Models

In this paper, we use XGBoost, random forest, AdaBoost, and MLP models to analyze, while comparing their performance in prediction accuracy.

The base model of XGBoost, random forest and AdaBoost is a decision tree. A decision tree contains a root node, several internal nodes and leaf nodes. Leaf nodes correspond to the decision results, and the other nodes correspond to an attribute test. The sample set contained in each node is divided into child nodes, according to the result of the attribute test. However, the learning ability of a single decision tree is limited, so we will use an ensemble learning method. Ensemble learning completes the learning task by constructing and combining multiple learners which often perform weaker.

There are two typical ensemble learning methods, namely, boosting and bagging.

In boosting, we firstly train a base learner from the initial training set, and then adjust the distribution of training samples according to the performance of the base learner, so that the training samples that the previous base learner made mistakes will receive more attention in the future. Based on the distribution of the adjusted samples, we train the next base learner. We repeat this process until the number of base learners reaches the value T specified in advance, and finally the T base learners are weighted and combined. XGBoost and AdaBoost are both boosting methods.

In bagging, we can sample and get T samplers with m training samples through m times random sampling with replacement by T round. Then we train a base learner based on each sampling set. Finally, we combine these base learners. Random forest is a kind of bagging. For each node of a base decision tree, a subset of k attributes is randomly selected from the node's attribute set, and then an optimal attribute is selected from this subset for partitioning.

K-nearest neighbor (kNN) refers to the method that finds k training samples that are closest to the given training sample in the training set based on a certain distance metric, and then makes predictions based on the k “neighbors” information. In the k-nearest neighbor method, when we choose a different parameter k, the result will be different.

Multilayer perceptron (MLP) is a kind of feedforward artificial neural network, mapping a set of input vectors to another set of output vectors. Each neuron has a series of parameters that can be learned, and uses a nonlinear function as the activation function. By introducing nonlinear functions, MLP can obtain a strong learning ability and perform well in many machine learning problems.

TABLE III. HYPER PARAMETER SEARCH SPACE FOR DIFFERENT MODELS.

Model	Hyper Parameters
RF	'max_depth': [5, 10, 15, 20, 25], 'n_estimators': [20, 50, 100, 200]
AdaBoost	'max_depth': [5, 10, 15, 20, 25], 'n_estimators': [20, 50, 100, 200]
XGBoost	'max_depth': [5, 10, 15, 20, 25], 'n_estimators': [20, 50, 100, 200]
kNN	'n_neighbors': [3, 5, 10]
MLP	'hidden_layer_sizes': [(50,50,50), (50,100,50), (100,)], 'activation': ['tanh', 'relu']

D. Results and Analysis

In Table 3, we list the different parameter search spaces for different models. The parameters are optimized by a 5-fold cross validation.

Then we apply the models after hyper parameter search into testing set and get their performance, indicated by AUC, of different models, as shown in Table 4.

TABLE IV. PERFORMANCE OF DIFFERENT MODELS.

Model	AUC
RF	0.5010
AdaBoost	1.0000
XGBoost	0.7166
kNN	0.5036
MLP	0.5000

According to the result, we find that:

RF, kNN and MLP perform weaker than other models because they overfit the training set and predict that the users always don't default, so their AUC values are equal to or slightly larger than 0.5, which indicates that the accuracy of these models is very low, and even shows no reference value.

Boosting performs well, the AUC values of both AdaBoost and XGBoost are significantly larger than 0.5, which shows the better prediction effect.

A good news is that AdaBoost gets the highest AUC value of 1 which shows that the accuracy of AUC is 100%, the corresponding optimal parameter is {'base_estimator__max_depth': 20, 'n_estimators': 100}.

V. CONCLUSION

Loan default prediction is an important and challenging problem. In this paper, we compare the performance of five machine learning models which are AdaBoost, XGBoost, random forest, k nearest neighbors, and multi-layer perception, using ROC and AUC as evaluation metrics. The outcomes show that boosting models perform better and AdaBoost even achieves a 100% accuracy. Based on our results, we believe that machine learning methods have a huge potentiality to be applied to process the loan default problem.

ACKNOWLEDGEMENT

The author(s) received no financial support for the research, authorship, and/or publication of this paper. There is no conflict of interest in this paper.

REFERENCE

[1] Jiang W, Zhang L. Edge-SiamNet and Edge-TripleNet: New Deep Learning Models for Handwritten Numeral Recognition[J]. IEICE Transactions on Information and Systems, 2020, 103(3): 720-723.

[2] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

[3] Jiang W. Time series classification: nearest neighbor versus deep learning models[J]. SN Applied Sciences, 2020, 2(4): 1-17.

[4] Jiang W. Applications of deep learning in stock market prediction: recent progress[J]. arXiv preprint arXiv:2003.01859, 2020.

[5] Lv Y, Duan Y, Kang W, et al. Traffic flow prediction with big data: a deep learning approach[J]. IEEE Transactions on Intelligent Transportation Systems, 2014, 16(2): 865-873.

[6] Jiang W, Zhang L. Geospatial data to images: A deep-learning framework for traffic forecasting[J]. Tsinghua Science and Technology, 2018, 24(1): 52-64.

[7] Lee Y C. Application of support vector machines to corporate credit rating prediction[J]. Expert Systems with Applications, 2007, 33(1): 67-74.

[8] Zhong H, Miao C, Shen Z, et al. Comparing the learning effectiveness of BP, ELM, I-ELM, and SVM for corporate credit ratings[J]. Neurocomputing, 2014, 128: 285-295.

[9] Kim K, Ahn H. A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach[J]. Computers & Operations Research, 2012, 39(8): 1800-1811.

[10] Guo X, Zhu Z, Shi J. A corporate credit rating model using support vector domain combined with fuzzy clustering algorithm[J]. Mathematical Problems in Engineering, 2012, 2012.

[11] Kvamme H, Sellereite N, Aas K, et al. Predicting mortgage default using convolutional neural networks[J]. Expert Systems with Applications, 2018, 102: 207-217.

[12] de Castro Vieira J R, Barboza F, Sobreiro V A, et al. Machine learning models for credit analysis improvements: Predicting low-income families' default[J]. Applied Soft Computing, 2019, 83: 105640.

[13] Bai M, Zheng Y, Shen Y. Gradient Boosting Survival Tree with Applications in Credit Scoring[J]. arXiv preprint arXiv:1908.03385, 2019.

[14] Kruppa J, Schwarz A, Arminger G, et al. Consumer credit risk: Individual probability estimates using machine learning[J]. Expert Systems with Applications, 2013, 40(13): 5125-5131.

[15] Li H, Huang H, Zheng Z. Research on Credit Risk of P2P Lending Based on CatBoost Algorithm[J]. 2019.

[16] Ma X, Sha J, Wang D, et al. Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning[J]. Electronic Commerce Research and Applications, 2018, 31: 24-39.

[17] Byanjankar A, Heikkilä M, Mezei J. Predicting credit risk in peer-to-peer lending: A neural network approach[C]//2015 IEEE Symposium Series on Computational Intelligence. IEEE, 2015: 719-725.