Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 174 (2020) 141–149

Procedia Computer Science

www.elsevier.com/locate/procedia

2019 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI2019)

A Comparative Assessment of Credit Risk Model Based on Machine Learning

——a case study of bank loan data

Yuelin Wang[a], Yihan Zhang[a], Yan Lu[a], Xinran Yu[a]

Wuhan University, Wuhan, Hubei, China
wangyuelin@whu.edu.cn
China University of Petroleum, Beijing, China
1035050755@qq.com
Xiamen University Malaysia, selangor Darul ehsan, Malaysia
744140970@qq.com
China University of Geosciences，Wuhan, Wuhan, Hubei. China
yuxinran777@qq.com

These authors are contributed equally to this work

Abstract

Recently some techniques (such as statistical techniques and machine learning techniques) have been developed for evaluating individual credit information to decide whether the person meets the criteria of credit financing, and the process is known as credit scoring. This paper mainly focuses on the comparative assessment of the performances of five popular classifiers involved in machine learning used for credit scoring: Naive Bayesian Model, Logistic Regression Analysis, Random Forest, Decision Tree, and K-Nearest Neighbor Classifier. Each classifier has its own strength and weakness, it is assertive to say which one is the best. However, the results of this experiment pinpoint that Random Forest performs better than others in terms of precision, recall, AUC (area under curve) and accuracy.

Keywords: Credit Risk, Random Forest, K-Nearest Neighbor, Decision Tree, Naive Bayes, Logistic Regression

1. Introduction

In recent years, with the continuous development of the Internet, the integration of technology and finance has deepened, which has led to tremendous changes in the financial industry. With the stimulation of consumer finance, the demand for various credit businesses is growing, and a reasonable and reliable risk assessment model needs to be established.

In the past, when commercial banks conducted credit risk assessments on loan users, they often relied on the risk control personnel to rely on the 5C classification method to subjectively judge, and judge the loan users from five factors: their personal character, credit limit, solvency, and market economy. And weighed, as a reference for whether to lend to the user, decide whether to issue loans, this method of relying on subjective judgment is obviously inefficient, and the evaluation is very dependent on the subjective judgment ability of the risk control personnel, from the perspective of internal control of the

10.1016/j.procs.2020.06.069

company In view of this, there is even the possibility of internal cheating of risk control personnel, unable to adapt to the rapid development of the market economy, to meet the needs of loan users, and to meet the needs of risk management of online lending platforms. With the development of big data on the Internet, under the new data portraits and business scenarios, the traditional credit scoring model application is severely limited, and the original business logic framework is lost. In the face of tens of thousands or even hundreds of thousands of users applying for loans, the online lending platform needs to adopt various machine learning methods to reduce the manual participation in the monitoring and testing process, and use automated methods to improve lending. The accuracy and efficiency of the review.

In recent years, scholars have studied the application of a variety of machine learning methods in credit evaluation, such as decision trees, neural networks, support vector machines and integration algorithms.[1] Aledanjro (2016) found that the traditional logistic regression did not perform as well as the previous ones when the various characteristic variables exhibited complicated nonlinear relationships. Though the Logistic Regression model may not be as good as the machine learning model in terms of prediction accuracy, it has a strong advantage in terms of variable interpretability and stability. Therefore, some scholars have improved the Logistic regression and applied it to the borrower's default behavior prediction.[2] Cai Wenxue (2019) et al. applied the combination of GBDT and Logistic regression models to personal credit risk assessment. By extracting valid combination features from the original data through the GBDT model, the original data can be fully utilized to avoid ignoring the importance of the original data. information. Bank credit data has many indicators and complex noise characteristics. The personal credit risk assessment model using GBDT and Logistic regression is used to greatly improve the accuracy of prediction.[3] In 2006, Li Xusheng and Guo Yaohuang first proposed a personal credit evaluation model based on the naive Bayesian classifier, which was tested on German and Australian credit data sets and compared with five neural network models, indicating that the naive Bayes classifier Has a lower classification error.[4] In 2006, Li Xusheng and Guo Yaohuang proposed an extended tree-enhanced naive Bayesian classifier, which removes the condition that the variables must be discrete. The mixed variables can be manipulated in the tree-enhanced naive Bayes classifier. Empirical research shows that the method The classification accuracy is better.[5] In 1985, for the first time, Makowski used the decision tree approach in the field of personal credit assessment.[6] Carter proved in 1987 that the decision tree approach has a high classification accuracy in the field of personal credit assessment.[7] Wang Maoguang et al. (2016) also found that the decision tree model has the advantages of adaptability, high precision and strong interpretability in explaining the reasons for loan default, dividing credit rating and reducing default rate.[8] Li Xin et al. (2018) applied BP neural network model to empirically evaluate the credit evaluation of P2P online loan borrowers. The results show that the model has good feature extraction and knowledge discovery ability. When there is virtual information index in the borrower evaluation index system It can still make a more accurate judgment on the credit risk of borrowers and has a strong ability to evaluate and predict.[9] Jiang Cuiqing and other scholars (2017) analyzed the relationship between different soft information variables and borrowing defaults, and then combined the screening of soft and hard information, introduced a random forest algorithm to construct a default prediction model incorporating soft information, and combined the real data of P2P for empirical analysis. The results show that the inclusion of valuable soft information in the default prediction model of P2P borrowing can improve the prediction accuracy.[10]

2. Overview

Machine learning is a method of data analysis and a part of artificial intelligence on the basis that system is able to learn from previous data, identify patterns or distributions of datasets, and make decisions. Generally speaking, it is a system that automates analytical model building with minimal human intervention.

2.1 KNN Classifier

KNN, a non-parametric and lazy learning method, is explored for regression and classification. The model structure is depended on the dataset, and there is no assumption for data distributions, therefore, it is non-parametric. This property is useful because real world datasets do not always adhere to theoretical assumptions. Lazy algorithm implies that all training datasets are used in the testing phase.

This classifier is simple, and intuitive to understand. Consequently, it’s widely used for classification because of low calculation time and easy of interpretation. KNN used in various fields in the industry such as finance, healthcare, handwriting detection. In Credit scoring, financial and banking institutes would evaluate the credit information of their customers and decide whether the customers meet the criteria of credit financing.

KNN has the following working mechanism: compute distance, find the closest neighbors, and vote for labels. In this method, K is the number of nearest neighbors, and it is a controlling variable for this method. Each dataset has its own optimal number of K. If K is small, the noise would greatly influence the result, whereas if K is large, it would be expensive to compute. Research reveals that a small K is more flexible with high variance and low bias while a large K is not flexible with lower variance but higher bias.

The pros of the method is that the training phase of KNN is much faster than other classifiers, and it is pretty intuitive and simple since it doesn't contain any assumptions. KNN can be useful in case of conducting multi-class problem and regression problem. The weakness of this method is that the testing phase of K-nearest neighbor classification is costlier and slower since it requires large memory for storing the entire training datasets. What's more, KNN is sensitive to magnitudes and outliers because it directly choose the neighbors on the basis of the distance (Genesis, 2018). High magnitude features will weigh more than low magnitude features. It is also not suitable for large dimensional data, and dimensions require to be reduced to upgrade performance (Avinash, 2018). Additionally, this method is unable to deal with missing value problem as well.

2.2 Decision Tree

Decision Trees are intuitive to visualize the decision. It is a type of non-parametric supervised learning method and decision support tool used for regression and classification. Based on the dataset, it can identify and derive a sine curve with a series of if-then-else decision rules. The model is fitter if the tree is deeper since there are more complex if-then-else decision rules.

It has a flowchart structure that mimics the activity of human thinking. This algorithm uses the tree representation to solve the problem, internal nodes in the tree to represent an attribute and leaf nodes to represent class labels. We start from the root of the tree, separating the sample into groups of homogeneous sets according to most essential splitters of input variables. Repeat this process until leaf nodes in all the branches of the tree are found. Therefore, it is easy to visualize, understand, and interpret (Rajesh, 2018).

This method is helpful in variable screening and feature selection, and it has the capability of accurately handle with high dimensional data with requiring little users' data preparation. Compare to other methods, this method requires little data cleaning since it is not affected by outliers. Furthermore, this method can deal with both quantitative and qualitative data, and this property is superior to other methods because other classifiers always require single attribute of the data. However, there are weaknesses exist, such as over fitting problem, which can be eliminated by using random forest. The reason the decision tree is likely to have the problem of over-fitting when we don't set a limitation of maximum depth is because the tree can keep growing until it has exactly one leaf node for every single observation, perfectly classifying all of them, so the tree is over-complex that it doesn't classifies the datasets well (Avinash, 2018). This method doesn't have stability because even small variation in the datasets might lead to a completely different decision tree being generated. Finally, Calculations will be complex if there are many class labels.

2.3 Random Forest

Random forest is the most popular algorithm, namely, it assembles a large amounts of decision trees from training dataset, and it also uses a tool called bagging to perform classification and regression tasks (Krishni, 2018). Each decision tree represents a class prediction, this method collects the votes from these decision trees and the class with most votes is considered as the final class (Savan, 2017).

While training, each base model is independently created by learning from different random subsamples of the data. The samples are drawn by the process called bagging or boot strapping, meaning that some samples might be used many times in one single decision tree. By using different subsamples to train each decision tree, the whole forest will have low variance but high bias despite each tree has high variance in terms of a particular set of the training dataset. At test time, predictions of each decision tree are averaged to derive the overall predictions, and the process is known as bagging.

The underlying reason this method outperforms than decision tree is that many uncorrelated decision trees can protect each other from individual errors to derive the ensemble predictions, thus over fitting problem can be reduced and the prediction results are unexcelled in terms of accuracy among the current algorithms (Tony, 2016). It also has other strengths such as: the model can run efficiently on large number of datasets, and it can estimate which variables are significant in the classification as well (Saimadhu, 2017). Random forest outperforms than linear models because it can catch non-linear relationship between the object and the features. The flaw of the method is that it can't work with sparse features because the decision trees are building blocks. So we need to pre-process the inputs to suit the model.

2.4 Naive Bayesian Model

Naive Bayes classifier is the reliable, fast, and accurate linear classifier, and it focuses on the conditional probability. Classifier technique is developed on the basis of Bayesian theorem and it is suitable for high dimension inputs because Bayes ' theorem assumes the features and attributes are independent. This hypothesis simplifies computation, so it is considered as naive. Despite its simplicity, it often outperforms alternative classifiers, especially for small sample sizes. This classifier is used in many fields. However, practically, independence is often violated, so violations of this assumption and non-linear problems can result in bad performance of this classifier (Sebastian, 2014).

The basic working steps are as follows: to commence with, calculate the prior probability for given class labels, and find likelihood probability for each class, then put these values in Bayes formula and compute posterior probability, finally see

which class gets the highest probability, given the input belongs to the higher probability class. Naive Bayes classifiers are able to train very effectively in a supervised learning environment. This model always use maximum likelihood method to estimate parameter.

An advantage of naive Bayes is that it only requires less training data to estimate the necessary classification parameters. However, the weakness associated with it is its independent parameters assumption because in real world it is hard to fulfill this assumption since datasets usually have related features (Rohith, 2018). In this case, it would greatly reduce the effectiveness of the classification. Another disadvantage is that it needs to know prior probability which is based on assumption, and the process requires a lot of subjective choices among assumption models, contributing to bad prediction results. Naive Bayes is considered as a bad estimator. If in test dataset, categorical variable has a category, which wasn't found in training dataset, the model will assign a zero probability and thus can't make a prediction, and this is called zero frequency. In order to solve this problem, smoothing technique must be used, such as Laplace estimation (Gaurav, 2018).

2.5 Logistic Regression Analysis

Logistic Regression is a powerful regression and statistical method for the classification problem, which is simple to conduct and can be utilized as the basis for binary classification problems. Additionally, describing and estimating the relationship between one dependent binary variable and independent variables makes this model instructive in deep learning. These assumptions are likely the same as the those made in linear regression.

From training data, we must estimate the coefficients of the logistic regression equation, and it is usually done by maximum-likelihood estimation, which is commonly used by machine learning despite it has to make assumptions about the data distribution. The optimized coefficients will lead to a model which would predict the value of the default class close to one, and the value for the other classes would be zero. Therefore, researchers pursue the value of the coefficients that could minimize the error in the predicted probabilities. Predictions are made by plugging in numbers into the equation just created and computing the result.

This model has one big advantage over other models: it is not only a classification model, but also provides probabilities. However, this model has its restrictiveness. Interpretation is difficult since the weights are not additive but multiplicative. In addition, this model suffer from complete separation. If a feature can separate the two classes perfectly, this model can never be trained because the weight for that feature will not converge and the optimal weight will be infinite. This is thorny because this feature might be very useful. In order to solve this problem, we can define a prior probability distribution of weights (Christoph, 2019).

3. Experimental setup

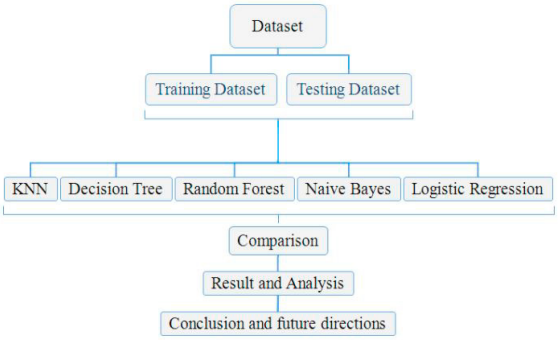

Fig.1

The data comes from the loan information statistics of a commercial bank. In order to protect personal information, the data is desensitized in advance, and the feature names have been replaced by characters.

The process of data preprocessing is:

1. Delete the irrelevant data column. If an index value has a small deviation in different sample values, delete the column.
2. If there is too much wrong data for the indicator value of a sample, delete the sample.
3. Replace the error data with the default value using the mean.

3.1 K-Nearest Neighbor (KNN)

Since the sample data is sufficient and the dimension is high, it is convenient to use the Euclidean distance calculation formula to calculate the distance between the test data and the training data, so the KNN algorithm is adopted. This set of data characteristics satisfies the intersection of the class domains and overlaps more. Therefore, the classification decision can determine the category to which the sample to be divided belongs based only on the category of the nearest neighbor or samples. The labels we set are only 1 and 0, that is, whether the bank is loaned, which is convenient for classification and high prediction accuracy.

3.2 Decision Tree

The data dimension is higher, the features are more, and it is structured data. The root node and the child nodes are convenient to construct, which is more suitable for the decision tree classification. The type is a two-category problem, and the prediction result is simple, and it is easier to train a model with higher accuracy.

3.3 Random Forest

Due to the large amount of data, the data contains more dimensions, more similar samples, and features do not need to be reduced in dimensionality, and the default value is more, so the use of random forest can be better handled. In addition, the data samples are not balanced, and random forests can better balance errors. In the process of building the model, we set the number of decision trees to 200, that is, to generate 200 decision trees and construct the final model.

3.4 Naive Bayes

In Bayesian statistical inference theory, we can assume that the different metrics of this set of data are independent of each other and calculate their conditional probability. After that, some evidence or background related to the event is also taken into account when the conditional probability is taken into account, the posterior probability is calculated, and the probability of the random event is finally calculated, that is, whether the bank is loaned, that is, the 1 and 0 of the narrative of the naive Bayes classification.

Therefore, we preprocess the samples and use feature screening. Then set the alpha parameter value to 0.01 and perform Laplacian smoothing on the data. Finally, the naive Bayes classifier is used to train and get the model.

3.5 Logistic Regression

The problem of this paper is the standard two-class problem. The categorical variables of this set of data are two-category variables, and they are numerical variables, that is, label is 1 or 0. Each observation object is independent of each other and does not interfere with each other. The feature is structured data, which is very convenient to establish a mathematical model and construct a loss function. In the process of building the model, in order to solve the over-fitting problem, we set the penalty to l2 regularization. Finally, a logistic regression model is constructed.

4.Experimental results and analysis

Table1:

Definition:

	Positive	**Negative**
True	TP	TN
False	FP	FN

Receiver Operating Characteristic: ROC
Area Under the Curve: AUC

$$\Pr ecision = \frac{TP}{TP + FP}$$

$$\mathrm{Re}\, call = \frac{TP}{TP + FN} \quad Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Method	DT=0.2				DT=0.3			
	AUC	precision	recall	accuracy	AUC	precision	recall	accuracy
KNN	0.63	61.36%	30.65%	82.27%	0.62	57.75%	28.21%	81.51%
Decision Tree	0.92	85.93%	87.80%	94.68%	0.88	79.65%	81.31%	92.11%
Random Forest	0.92	97.16%	85.16%	96.53%	0.88	95.29%	76.66%	94.57%
Naive Bayes	0.50	36.00%	0.09%	79.99%	0.50	36.00%	0.09%	79.99%
Logistic Regression	0.56	60.81%	14.77%	81.05%	0.56	61.16%	13.84%	81.01%

$$\beta = \frac{\text{the number of testing set}}{\text{the number of whole set}}$$

Table2

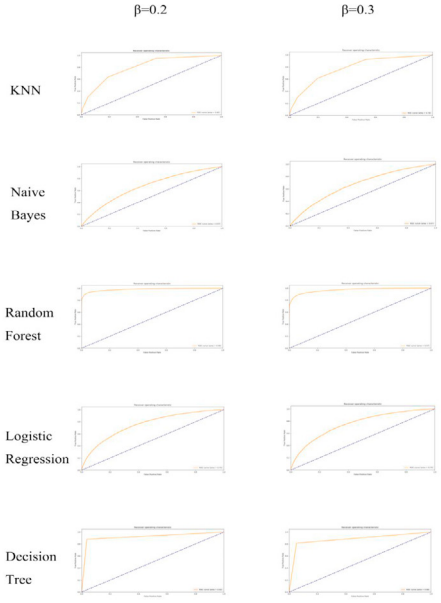

Fig.2

Random Forest can be used for classification problems as well as regression problems. The sample is random, the characteristics are random, the tree of the random forest is enough, the performance is good on the test set, and the classifier is able to avoid overfiting model.
The random forest classifier can handle missing values and be modeled with categorical values. The integrated algorithm is adopted, and its accuracy is better than most single algorithms, so the accuracy is high.
Due to the combination of trees, random forests, which are themselves nonlinear classification (fitting) models, can process nonlinear data. It can process data with very high-dimensional data (features), without making feature selection. Additionally, it can process both discrete data and continuous data, and the data set does not need to be standardized, which makes it have strong adaptability to data sets.
Training speed is fast and can be applied to large-scale data sets. The default value can be processed (alone as a class) without additional processing.Out-of-bag data (OOB) can obtain an unbiased estimate of the true error during model generation without losing the amount of training data.

5.Conclusions and future directions

Traditional commercial bank credit business practices, after the user applies for a loan, there will be a special credit approval personnel to approve the historical data of all aspects of the user, such as the user's income level, the number of family members, whether there is a car, etc., but This approach is inefficient and cannot meet the large-scale business needs of modern Internet big data lending. Secondly, this method is too subjective and prone to internal cheating.
In order to expand the scale of business and reduce the collection cost caused by user default, the Internet financial platform uses a personal credit risk assessment model built by machine learning algorithms. In the current era of big data, by collecting data such as personal characteristics and social records of users, it captures the characteristics of credit risks involved in it, and establishes a predictive classification model for users' future default risk, as a loan to a user. Decision basis. From the perspective of the user, the development and implementation of this personal credit evaluation model will also promote users to concern about the maintenance of personal credit records. From the perspective of the government, after the data in various fields are opened, the big data system and the credit risk prediction model are conducive to the further development of China's personal credit information system. In summary, the use of machine learning algorithm-based personal credit risk prediction model has expanded the scale of business, promoted the development and improvement of China's credit information system, and urged users to maintain their personal credit records, innovative financial products and services.

Acknowledgements

The completion of the thesis is attributed to many people's support and encouragement. Without of the help of many people,it could never be completed, thus here we would like to express my sincere gratitude towards them.
First and foremost, we owe our heartfelt thanks to my distinguished and cordial supervisor,Professor Yingjie Tian, who influenced us with his insightful ideas and meaningful inspirations, guided us with practical academic advice and feasible instructions, and enlightened us while we was confused during the writing procedure. His thought-provoking comments and patiently encouragements are indispensable for my accomplishment of this thesis.
Ultimately, thanks go to our parents who are our mentor and guardian from the very beginning in primary school. Without their refined education and care,we could never grow up in such a joyous and cozy environment nor have the courage to confront any obstacles on our way to success.

References

[1]Correa Bahnsen. A. (2016)Feature Engineering Strategies for Credit Card Fraud Detection. Expert Systems with Applications,51,134-142.

[2]蔡文学,罗永豪,张冠湘,钟慧玲.基于 GBDT 与 Logistic 回归融合的个人信贷风险评估模型及实证分析[J].管理现代化,2017,37(02):1-4.

[3]Wiginton J C. A note on the comparison of logit and discriminant models of consumer credit behavior. Journal of Financial and Quantitative Analysis, 1980,15(03):757-770.

[4]李旭升，郭耀煌. 基于朴素贝叶斯分类器的个人信用评估模型[J].计算机工程与应用,2006,42(30):197-201.

[5]李旭升，郭耀煌. 扩展的树增强朴素贝叶斯分类器. 模式识别与人工智能，2006,19(4):469-474.

[6]Makowsi P. Credit scoring branches out. Credit World，1985,75(1):30-37.

[7]Carter C, Catlett J. Assessing credit card applications using machine learning. IEEE expert, 1987,2(3):71-79

[8]王茂光,葛蕾蕾,赵江平.基于 C5.0 算法的小额网贷平台的风险监控研究[J].中国管理科学,2016,24(S1):345-352.

[9]李昕,戴一成.基于 BP 神经网络的 P2P 网贷借款人信用风险评估研究[J].武汉金融,2018(02):33-37.

[10]蒋翠清,王睿雅,丁勇.融入软信息的 P2P 网络借贷违约预测方法[J].中国管理科学,2017,25(11):12-21.

[11]Krishni Hewa. Nov 27, 2018. A Beginners Guide to Random Forest Regression.

[12]Avinash Navlani, December 29th, 2018. Decision Tree Classification in Python.

[13]Avinash Navlani, September 8th, 2018. Understanding Logistic Regression in Python.

[14]Chirag Sehra. Jan 20, 2018. Deicsions Trees Explained Easily.

[15]Park Hyeoun-Ae, April 2013. J Korean Acad Nurs Vol.43 No.2, 154-164.

[16]Rohith Gandi, May 6, 2018. Naïve Bayes Classifier.

Tony Yiu, Jun 12, 2016. Understanding Random Forest.

[17]Genesis. Septemeber 25, 2018. Pros and Cons of K-Nearest Neighbors.

[18]Rajesh. Oct 26, 2018. Machine Learning.

[19]Savan Patel. May 18, 2017. Random Forest Classifier.

[20]Avinash Navlani, Augest 3rd, 2018. KNN Classification And Build KNN Classifier Using Python Scikit-learn pacakage.

[21]Saimadhu Polamuri. May 22, 2017. How the Random Forest Algorithms Works In Machine Learning.

[22]Sebastian Raschka. Oct 4, 2014. Naïve Bayes and Text Classification.

[23]Gaurav Chauhan. Oct 8, 2018. All about Naïve Bayes.

[24]Christoph Molnar. Aug 16, 2019. A Guide For Making Black Box Models Explainab