Determining the Predictive Accuracy of Loan Defaulters Using R

Mugdha Sharma, Aarnav Madan, Akshat Shakarwal, Abhay Pratap Singh, and Nitin Kumar

Abstract The banking sector shares a great contribution in maintaining the economy of the country. Default in bank loans shares vital role in risk management of various bank institutions. Bank helps the customer by giving many loans, credit cards, investment, mortgage, and others. Bank loan is also an important part of banking institutions which define as the probability of returning money to the bank by its users. Although, with increase in the participant of banking loan user, the number of defaulters is increasing with the minute and that is leading to huge losses for the banking industry. Machine learning has been used to tackle this issue. This research paper proposes a more accurate way to predict loan defaulters. The model proposed in this paper predicts defaulters using a logistic regression area under Roc curve of 77% which beats the earlier accuracy of 75%. Similarly, in this paper, decision trees, random forest, and the SVM models have been able to achieve better accuracy than the models proposed in the past.

Keywords Exploratory data analysis · ML · Data science

M. Sharma (✉) · A. Madan · A. Shakarwal · A. P. Singh · N. Kumar
Department of Computer Science and Engineering, Bhagwan Parshuram Institute of Technology, Delhi, India
e-mail: mugdha.sharma145@gmail.com

A. Madan
e-mail: Madanm.aarnav@gmail.com

A. Shakarwal
e-mail: akshatshekhnew@gmail.com

A. P. Singh
e-mail: abhaytrekk@gmail.com

N. Kumar
e-mail: starknitin@gmail.com

A. Abraham et al. (eds.), *Proceedings of 3rd International Conference on Computing Informatics and Networks*, Lecture Notes in Networks and Systems 167,
https://doi.org/10.1007/978-981-15-9712-1_27

Abbreviations

ROC curve	Receiver Operating Characteristic curve
SVM	Support Vector Machine
ML	Machine Learning
K-NN	K-Nearest Neighbors

1 Introduction

With the increase in the amount of data being generated day by day, machine learning algorithms have got stronger, and data analytics have become an integral part of the industry. Various machine learning methods are being applied to solve serious business problems [6].

This research paper is designed to pay attention to one of the greatest challenges that the banking sector is facing currently. Non-repayment of loans has caused major losses to the banking sector. This is a major concern due to which banks have started to invest more and more in developing bank loan risk models that help them in reducing the risk factor of providing the loan to the customer. This is done using machine learning and predictive modeling.

Machine learning techniques that this research paper is using to find the loan prediction defaulter are logistic regression, Rpart decision tree, random forest, and SVM. Now, let us talk about logistic regression—statistic logistic model is used to find the probability of particular class or we can say event. For example, passed/failed, win the event/loss the event, alive/dead or healthy/sick. This may be applied to several event to find whether an image contains cow, bird, or any animal. Each object is detected in the image would be assigned a probability of between 0 and 1 and sum adding to one.

The decision tree is the most efficient and most favored tool which used to classify and predict dataset. A decision tree is like tree structure where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal) holds a class label [6].

Random forest or can be said as random decision forest is an assemble learning method for classification, regression, and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class, i.e., mode of classes (classification) or mean prediction (regression) of individual trees.

The input/contribution of this research paper.

The overall research process design for the proposed study is shown in Fig. 1.The objectives of the study are as follows:

To get to know which factors affect the person to default on his payment

To conduct in-depth exploratory data analysis to get insights into the data available

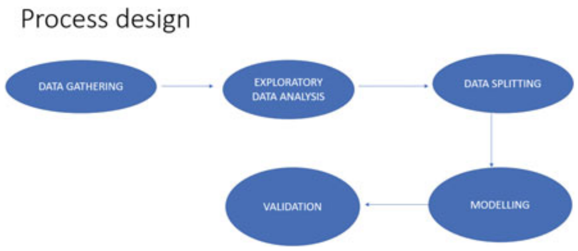


Fig. 1 Research process

To build a predictive model that accurately predicts whether the person would default or not
To improve the accuracy of the models implemented in the past work.

2 Machine Learning

2.1 Logistic Regression

Logistic regression is a classification technique used to solve a classification problem that involves predictions of a factor variable. It comes under a supervised learning algorithm that means the target variable should be known beforehand to use this algorithm.

2.2 Decision Trees

The decision tree is an unsupervised machine learning algorithm used to predict a variable by finding out the most important variables and then creating a tree-based structure using them [1]. It can be used for both regressions as well as classification problems. It can cause a problem of overfitting, but the ease of its implementation is a big factor of its being used in the industry.

2.3 Random Forest

Random forest is an unsupervised machine learning technique that uses an ensemble method to create multiple decision trees and come up with the best model using those. A random forest can also be used for both classification and regression problems. The random forest takes a lot of time to train since the generation of multiple decision trees takes time.

2.4 Support Vector Machine (SVM)

Support vector machine is used to solve regression and classification problems. In this, each data item is plotted in an *n*-dimensional space. Vectors are used to uniquely identify each group distinctly.

3 Literature Review

Based on the past literature, we have seen many different types of machine learning techniques have been used like logistic regression, decision trees, random forest, and K-NN [2, 11, 12].

The most used technique that we observed was that of decision trees. This is because of the ease of which it can be implemented. It is a technique where we find the most significant variables and make a tree concerned about that [3]. Radom forest is another such technique that was used quite often. It generates many decision trees and ensembles them to create a model with the best accuracy. The best accuracy was found out to be of Sayjdhah [6]. Nowadays, the banking sector uses efficient use of machine learning techniques with several classification techniques to split up the customers to predict the trends [8, 10]. They want to keep the all details of the customers to understand the behavior of payment data which is added to the loan scoring literature to anticipate their defaults [4]. Some researcher used the Bayesian network used for the graphical representation model showing the probability of correlation of variables [7]. Few researchers have proposed the hybrid approaches also such as merging genetic algorithms with neural network approaches to detect the financial frauds [5].

4 Methodology

4.1 Datasets

In this project, the datasets are used, and it is generated by the banking loan operations by the user. The datasets consist of 25 variables with 30,000 samples. This dataset has been used in various researches previously too [9]. So, it is not unsigned yet, the dataset includes a binary variable of Yes equal to 1 and No is equal to 0, for example, default payment outcome [6]. Table 1 shows the basic description of dataset.

Default_Payment Next_Month → This is the target variable that has to be predicted. It tells whether the person would default or not. It is a factor variable with 2 levels "Yes" and "No" as shown in Fig. 2.

Table 1 Explanation of dataset

Attribute-name	Descriptions
ID	UserID
Limit__Bal	Amount of the given credit (NT dollar)
Sex	Gender (1 = male, 2 = female)
Education	Education (1 = graduate school, 2 = university, 3 = others)
Marriage	Martial status (1 is married, 2 is unmarried)
Age	Age (year)
Pay-1_Pay-5	Status of repayment from April to September 2005
Bill_Amt1 to Bill_Amt-6	Bill statement amount from April to September 2005
pay_Amt1 to pay_Amt6	Amount paid or compensate in September 2005
default_payment_Next_Month	Amount which has to be compensated in next month

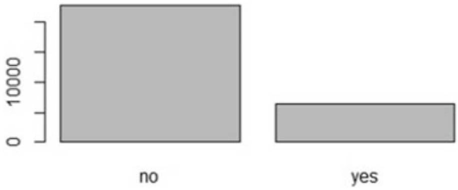


Fig. 2 Snapshot of the dependent variable

4.2 Dataset Pre-Processing and Feature Selection

Firstly, the data was cleaned to build the model. Then, all the NA values were removed. This is done so that the model can run smoothly. The next step is feature selection where only the important variables are kept and all the obsolete variables are removed. In this paper, multicollinearity and correlation have been considered to observe feature importance. Then, the outliers were treated. The interquartile range has been used to remove outliers. Outliers are values that do not follow the pattern, and these values make the model deviate from the correct predictions. Figure 3 shows the variable of bill_amt2 before and after the treatment of outlier. Feature scaling technique is also used to scale the features to a certain range to make the logistic model work fast and efficiently. *Z* score normalization was used to scale the features.

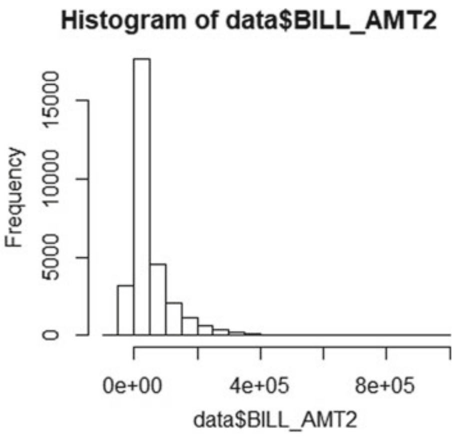


Fig. 3 Before outlier treatment

4.3 Data Visualization

This section provides interesting insights into the data that helps in understanding the relationship between different variables, all the nitty-gritty of the data are understood using different visualization techniques. Also, the demographics about the people using the credit card can be figured out from Fig. 4. Figure 5 shows the split of our dependent variable gender-wise.

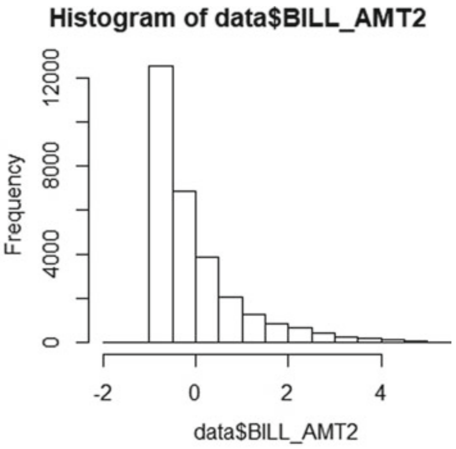


Fig. 4 After outlier treatment

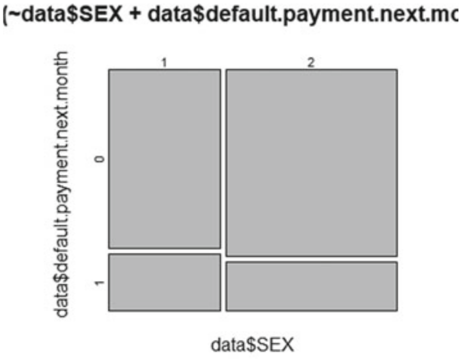


Fig. 5 Gender demographics

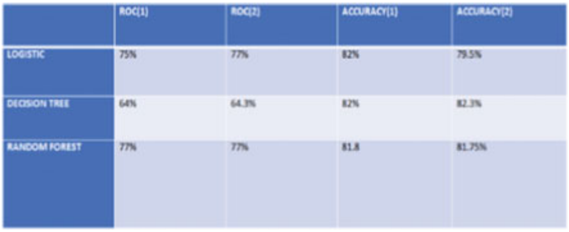
4.4 *Data Partitioning*

Data was split into two parts, train and test data. This is done so that first we can train the model on train data and then we can do the error analysis on test data. This helps us prevent overfitting and also makes the model more flexible to new data points. The data is split in the ratio of 70:30 to ensure we have sufficient data points in the test data to train the model well.

5 Performance and Evaluation

As seen in Fig. 6, it can be observed that the model used in this paper gives better results from the model being used in previous papers. Even though the accuracy of logistics reduces, but the gain in the ROC curve area shows that the model being used in the paper is much more stable and equipped to handle new incoming data.

Fig. 6 Model outcome. 1 Yashna Sayjdhah, 2 model used in this paper



	ROC(1)	ROC(2)	ACCURACY(1)	ACCURACY(2)
LOGISTIC	75%	77%	82%	79.5%
DECISION TREE	64%	64.3%	82%	82.3%
RANDOM FOREST	77%	77%	81.8	81.75%

1->Yashna Sayjdhah
2->model used in this paper

6 Model Comparison and Discussion

Four models have been used in this paper. Out of the four, random forest provides the best accuracy and a comparable area under the curve for our ROC curve making it the most stable and best-equipped model to predict the defaulters.

Decision tree was also considered, but since its area under the curve for the ROC curve is way too less, so we choose random forest as the best and most suited model for this paper.

7 Conclusion

At this stage, according to the predicting model, we have concluded that almost half of the population is married. Most of the people are graduate and from the university. Male and female have equal percentages. Bill amounts are skewed which need to be treated. Most of the people are aged between 20 and 60. More men tend to default than women in terms of ratio. Married people and others tend to default more than single. People having school or university education tend to default more. The techniques which are used in this model are random forest, decision tree, and logistic regression. Random forest has the best accuracy in this model with 81.75%. Based model had (Yashna Sayjhda 2018) the accuracy 81%. Main objective of this model is to detect the defaulters who take the loan from the bank and refuse/fail to pay within the given time which was provided by the bank itself. This paper checks on different parameters which customers likely to default more.

References

1. Rising credit card delinquencies to add to U.S. banks' worries. Accessed 13 Nov 2017
2. Venkatesh A, Gracia S (2016) Prediction of credit-card defaulters: a comparative study on performance of classifiers. Int J Comput Appl 145(7):36–41
3. AghaeiRad A, Chen N, Ribeiro B (2016) Improve credit scoring using transfer of learned knowledge from self-organizing map. Neural Comput Appl 28(6):1329–1342
4. Bakoben M, Bellotti T, Adams N (2017) Identification of credit risk based on cluster analysis of account behavior. Department of Mathematics, Imperial College London, London SW72AZ, United Kingdom
5. Azimi A, Hosseini M (2017) The hybrid approach based on genetic algorithm and neural network to predict financial fraud in banks. Int J Inf Secur Syst Manage 6(1):657–667
6. Yashna S, Kasmiran KA, Hashem IAT, Alotaibi F (2018) Credit card default prediction using ML techniques, Malaysia
7. Xia Y, Liu C, Li Y, Liu N (2017) A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Syst Appl 78:225–241
8. Yap B, Ong S, Husain N (2011) Using data mining to improve assessment of creditworthiness via credit scoring models. Expert Syst Appl 38(10):13274–13283

9. Yeh I (2017) UCI machine learning repository: default credit card clients data set. Online Archive.ics.edu. Available at. https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients. Accessed 13 Nov 2017
10. Ghasemi A, Motahari AS, Khandani AK (2010) Interference alignment for the K user MIMO interference channel. In: IEEE International Symposium on Information theory proceedings (ISIT), pp 360–364
11. Bellotti T, Crook J (2013) Forecasting and stress testing credit card default using dynamic models. Int J Forecast 29(4):563–574
12. Harrell F (2015) Regression modeling strategies: with application to linear models, logistic and ordinal regression, and survival analysis. Springer International Publishing, Berlin