Bank Customer's Credit Score Prediction Using Feature Selection and Data Mining Algorithm

Durgesh Kumar Singh and Noopur Goel

Abstract Data mining techniques of classification and prediction model are used for analyzing the customer of bank through their real data, customer is a performing as asset or non-performing asset for the bank. Every year each bank faces a similar problem that most of them customers do not refunding loan installment on time and many precisely become defaulter for the bank. So every bank needs a system that can predict in future any customer in future will be profitable or not profitable asset for the bank. If any customer found as non-performing asset means, it has bad credit score. In such case if a customer further requests for new loan, bank can easily identify him as defaulter and reject his/her request on the basis of our new proposed models. In this way, bank extracts their non-performing assets. Besides, bank can also identify their new customers for having good credit score and may offer them other services for the beneficial of bank. There are many as such models that are used for prediction. This paper compares different classification and prediction algorithms to developed best suitable model to analyze loan requesting customer's data set using their credit score. Mainly in this paper, there is a comparison between random forest algorithms and logistic regression which is best suitable for prediction of credit scores of customers with apply *k* best feature selection on data set. The aim of proposed model is to reduce bankruptcy, non-performing assets and the losses of bank.

Keywords Data mining · Classification models · F1-score · Bank credit · Feature selection

D. K. Singh (✉) · N. Goel
VBS Purvanchal University, Jaunpur, India
e-mail: durgeshsingh111@gmail.com

N. Goel
e-mail: noopurt11@gmail.com

C. R. Panigrahi et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1299,
https://doi.org/10.1007/978-981-33-4299-6_73

1 Introduction

Credit enables individuals or businesses to "eager to pay" or "buy ahead of ability." Banks give facility them to adequate loans based on costumers requirement in the corporate organization, agricultural and business sectors. In addition, at what time customer uses intrinsic intelligence just in the case of credit, it leads to financial enhancement. In recent scenario, government of country in bank decision that sometime becomes problematic for the banks and it is one the main reasons of bank's non-profitable assets (NPA). To reduce this bankruptcy, NPA and the losses of bank, we have applied some data mining algorithm. These algorithms use as a tools that will help to determine the ability of customers to repay credit loans in a timely manner or not and categorize the customers as "good credits" or "bad credit." A credit score is a numerical assessment that a bank uses with your credit report to assess the risk of providing you with a loan or providing credit to you [1].

One of the main purposes for organizing a bank is to lay out loans. But to maintain functioning, the bank issues these loans to those who are able to pay back, thereby minimizing the risk of outstanding lending. Nevertheless, risk management knowing who is notable of credit remains a continuing challenge in the banking industry. Ability to identify risk levels, customers based on characteristics such as job, age, marital status, salary span/net asset value, solvency status, etc. are key values that banks must examine before providing loan to customers. With the value of credit risk scores, you can help the bank to decide considerable interest to charge on the loan. Although, these risk elements from time to time fail to make an informed conclusion about the customer's creditworthiness.

Data mining algorithms will be used to analyze the loan approval data and will find out a patterns that will help to predict non-profitable customer and thus help banks make better decisions in the future. Data set from different sources will be used for creating a framework, and then different data mining algorithms will be applied to extract the patterns and get the results with maximum accuracy.

2 Related Work

Many research already have been discussed on this problem with data mining and machine learning algorithm in the area of bank and loaning. Some examples are as such-

Hsu et al. [2] for predicting the customer credit applied support vector machine on data set firstly without feature selection and after feature selection then compare their results and found after feature selection results of accuracy rate has been improved.

Turksen et al. [3] predict credit related to loaning of bank customer using supervised and unsupervised learning algorithm; different algorithms give different accuracy rate.

The model proposed by Jafarpour et al. [4] on Iranian bank's data set to predict loaning accuracy of customers and also establishes a relation between customer's requirement and banks and formulates a equation through which bank predict credibility of customers in concern of loaning.

Hassan et al. applied neural network algorithm for loaning of customer in bank. Firstly create a supervised model of bank credit data set. This model is very useful for bank to analyzing that any customer will return their sanction loan in future or not.

Jin et al. [1] employed data mining algorithm to found non-profitable assets of the bank in concern loaning and they also do a comparative study on different data mining algorithm and they conclude that support vector machine algorithm performed best.

Morco et al. [5] proposed a data mining approach for the prediction of customer's credit of Portuguese retail bank in telemarketing. They analyze the result of accuracy of different data mining model and found that neural network data mining algorithm performance is best.

A data mining approach conducted by Li et al. [6] on using data set of customers to predict risk using attribute bagging method. They found that the performance is outstanding using two credit databases.

3 Proposed Model

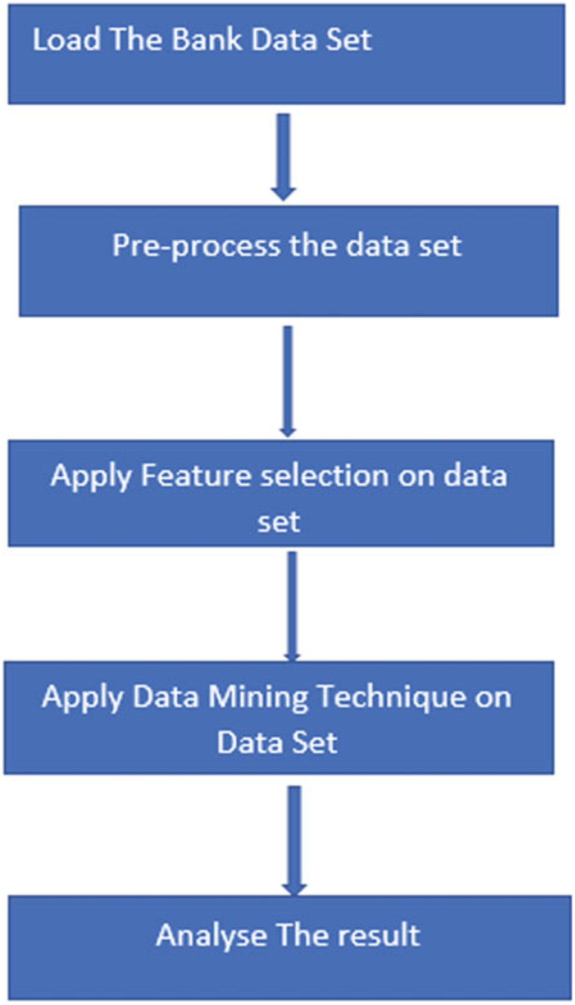
Figure 1 shows the flow of our paper. To process all the step, Python code is used on Jupiter notebook interface. The very first step is to load the data set then pre-processed to all attributes of data set. Then using very important technique that helps to improve the accuracy result of algorithm is feature selection through which redundant values attribute is eliminated. The next step after feature selection on pre-process data set apply data mining approaches to find the result and in the last analyzes the result.

4 Data Description

4.1 Data Source

The data set is used of direct marketing campaigns of a Portuguese banking institution. The data set has been available publicly at UCI machine learning repository.

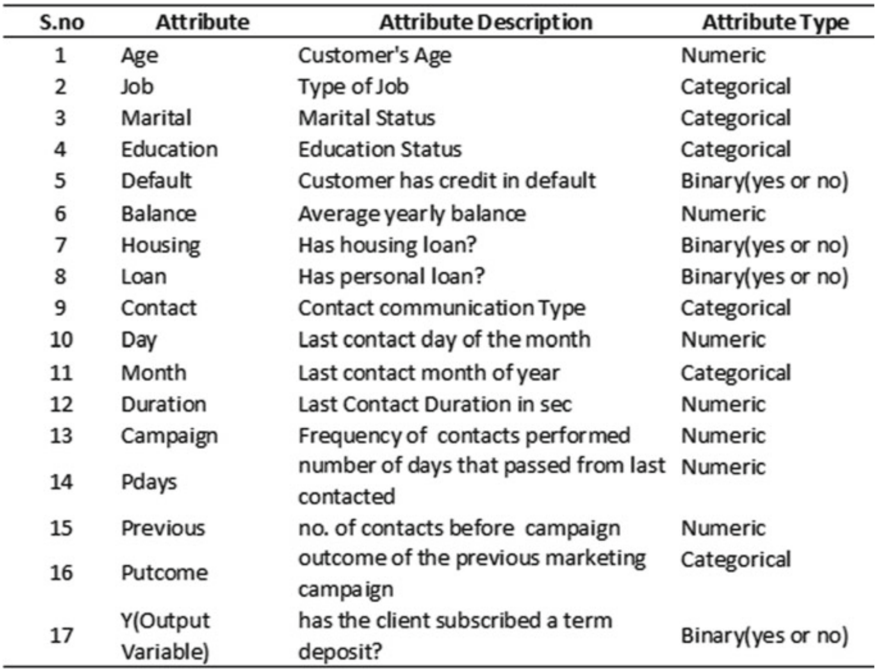
Fig. 1 Proposed model


Load The Bank Data Set
Pre-process the data set
Apply Feature selection on data set
Apply Data Mining Technique on Data Set
Analyse The result

4.2 Data Description

Number of Instances: 45,211 for bank-full.csv. Number of Attributes: 16 + output attribute. Data set is separated in 70–30%. 70% for training data set and 30% for testing data set on each particular model.

Figure 2 explains the dat set attribute description and their type.



S.no	Attribute	Attribute Description	Attribute Type
1	Age	Customer's Age	Numeric
2	Job	Type of Job	Categorical
3	Marital	Marital Status	Categorical
4	Education	Education Status	Categorical
5	Default	Customer has credit in default	Binary(yes or no)
6	Balance	Average yearly balance	Numeric
7	Housing	Has housing loan?	Binary(yes or no)
8	Loan	Has personal loan?	Binary(yes or no)
9	Contact	Contact communication Type	Categorical
10	Day	Last contact day of the month	Numeric
11	Month	Last contact month of year	Categorical
12	Duration	Last Contact Duration in sec	Numeric
13	Campaign	Frequency of contacts performed	Numeric
14	Pdays	number of days that passed from last contacted	Numeric
15	Previous	no. of contacts before campaign	Numeric
16	Putcome	outcome of the previous marketing campaign	Categorical
17	Y(Output Variable)	has the client subscribed a term deposit?	Binary(yes or no)

Fig. 2 Data set description

5 Feature Selection

Feature selection is the process of selecting best attributes from any given data set. The best attribute means that value matters most in output variable for the best prediction of accuracy.

It also reduces over fitting means to eliminate those attributes from data set that not play major role in making of decision. Due to Reduces Training time of algorithm, data set gets trained in less time. This way feature selection plays a major role in finding accuracy rate of different algorithm on different data set.

There are different feature selection method on which we used select *k* best feature selection using Python code on Jupiter notebook.

The value of attributes in which higher value attribute will play major role in data mining algorithm to finding accuracy and less value attribute will not so much affect in the accuracy result.

In Fig. 3, the least play role, in accuracy of algorithm, attribute is eliminated. Now we are using only these attribute in our data set to predict accuracy of customers.

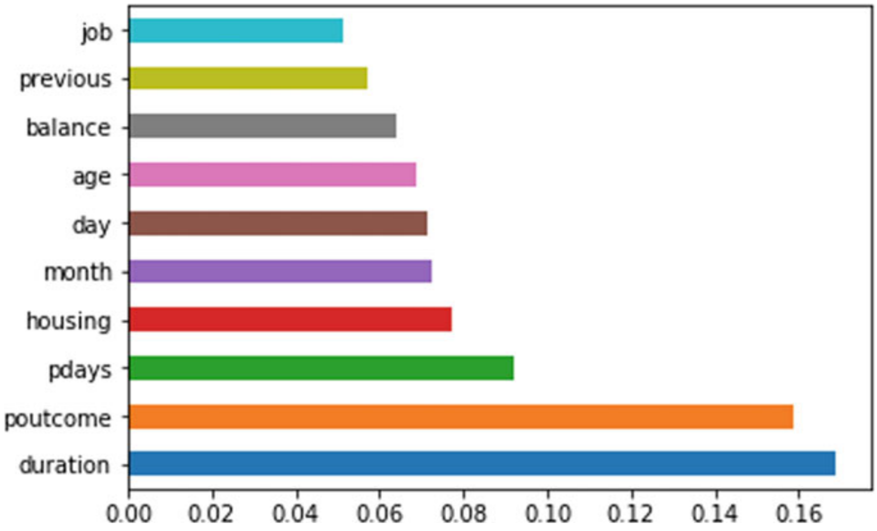


Fig. 3 Selected attribute after feature selection

6 Model Used

In this paper, two data mining models are used for prediction of bank customer behavior regarding to refund the loan. Different mining classification model gives accuracy result different on the same data set. Then create a comparison between these different mining models and suggest to bank which model will be best for evaluating the credit score of customer.

Following algorithms are used.

- Random forest algorithm
- Logistic regression algorithm.

6.1 Random Forest Algorithm

Random forest or random decision forest is algorithm that works on creating many decision tree in time of training phase. It is basically an ensemble learning method for classification, regression that operate on multiple decision tree at training time and the determination of the majority of the trees is selected by the random forest as the last finding. Random forest deals in the best way from the over fitting, that negatively impacts on the accomplishment of the model on new data. For the large data set, random forest model produces high accuracy; it runs very efficiently on large database. It also maintains its accuracy in case of missing data in large number [7].

6.2 Logistic Regression Algorithm

Logistic regression algorithm is a technique that can be used in traditional statistic as well as in data mining. Logistic regression algorithm is much similar to linear regression except that logistic regression algorithm predicts whether something true or false instead of something continuous like size.

Logistic regression is a regression analysis used to predict the results of a classification dependent variable based on one or more predictors. In logistic regression, a *S*-shaped curve used instead of straight line like in linear regression. The formula for a univariate logic curve is $p = \frac{\mathrm{e}(c0+c1\times 1)}{1+\mathrm{e}(c0+c1\times 1)}$.

To perform the logarithmic function can be applied to obtain the logistic function $\log_{\mathrm{e}} \frac{p}{p-1} = c0 + c1 \times 1$.

Here p is probability in one class and $p-1$ is on another class. Logistic regression is easy to implement and simple and used for wide variety of problem with good performance.

7 Discussion of Result

Applied data mining algorithm on the data set and calculated their training and test case accuracy on different field to find bad customer and good customer for bank. The accuracy result is following in the Table 1.

On the basis of result, shown in Table 1, the accuracy of mining algorithm the logistic regression algorithm's test case accuracy result is less than the random forest classifier's test case accuracy result. Hence, random forest algorithm is the most suitable algorithm for the data set than logistic regression. F1 score result also show that, logistic regression accuracy is less than the random forest classifier. Random forest is the best for analyzing data set of bank and on that basis, bank can easily predict which customer retention is profitable for bank. Figure 4 shows the result of accuracy using graph.

Table 1 Summarized view of results of data mining classifier

Algorithm	Training case accuracy	Test case accuracy	Precision	Recall	F1-score
Logistic regression	0.825	0.810	0.84	0.90	0.80
Random forest classifier	0.988	0.855	0.86	0.93	0.89

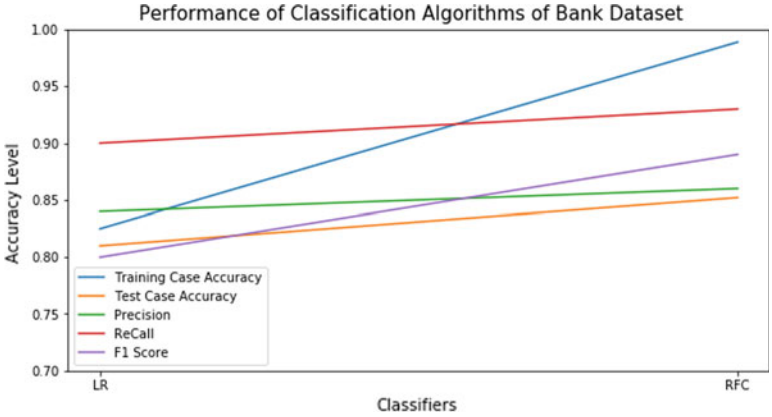


Fig. 4 Performance analysis of classification algorithm

8 Conclusion

In this paper, we used different data mining algorithm to build a model for bank on which basis bank can decide any customer as a bad customer or good customer. Then banking system can decide their future decision regarding on that particular customer. Good customer means whose credential is good and vice versa. Data mining algorithm predicts that customer is valuable or profitable asset on regarding different record as such their age, occupation, marital status, education, credit is in default, balance, loan type (as personal), last contact duration, put come (outcome of the previous campaign).

In this paper, data mining algorithm is used to create a model using Python languages with their different packages to calculate accuracy in which random forest algorithm result is best in comparison to logistic regression.

References

1. Jin, Y., Zhu, Y.: A data-driven approach to predict default risk of loan for online Peer-to-Peer (P2P) lending. School of Information, Zhejiang University of Finance and Economics, 310018 Hangzhou, China
2. Hsu, C.F., Hung, H.F.: Classification methods of credit rating—a comparative analysis on SVM, MDA and RST. In: 2009 International Conference on Computational Intelligence and Software Engineering, pp. 1–4. (2009)
3. Turkson, R.E, Baagyere, E.Y., Wenya, G.E.: A machine learning approach for predicting bank credit worthiness. In: 2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR), pp. 1–7. Lodz (2016). https://doi.org/10.1109/ICAIPR.2016.7585216
4. Jafarpour, H., Sheikholeslami Garvandani, H.: New model of customer relationship management in Iranian banks. Icbme.Yasar.Edu.Tr, pp. 1–12, (2012)
5. Moro, S., Cortez, P., Rita, P.: A data-driven approach to predict the success of bank telemarketing. Decis. Support Syst. **62**, 22–31 (2014)

6. Li, J., Wei, H., Hao, W.: Weight-selected attribute bagging for credit scoring. Math. Probl. Eng. **2013**, (2013)
7. Day, E.A., Hendershott, P.H.: Household Demand for Policy Loans. J. Risk Insur. **44**, 411–423 (1977)