Predictive Analysis of Credit Score for Credit Card Defaulters

Nupura Torvekar, Pravin S. Game

Abstract: ***Risk management has always been an important aspect of the financial institutions. Apart from the consumer frauds that cause huge losses, one more source of credit risk is nothing but the loan defaulters. Appropriate loan granting decisions therefore play an important role in avoiding these losses. Credit score and credit scoring which depends upon the credit history of a customer is one among the many factors that contribute to the loan granting decisions. Prediction of the loan defaulters in advance can help the financial institutions in undertaking some preventive measures to avoid granting loans to customers with potential risk and thereby reducing the amount of bad loans. Various machine learning techniques can play an important role in the identification of loan defaulters. The proposed work aims to identify and distinguish the good customers from bad customers by using different machine learning techniques. Two different tools Waikato Environment for Knowledge Analysis (WEKA) and KNIME (Konstanz Information Miner) are used for analyzing the performance of the classifiers. The main focus of this work is the prediction of credit card defaulters and hence two data sets relating to the credit card data of customers have been used for the purpose of this study. The results obtained from the proposed work can help the financial institutions in the identification and control of credit risk.***

Keywords: ***Classification; machine learning techniques; risk management***

I. INTRODUCTION

Banking sector is one of the most volatile and vulnerable sectors in the world with its ever-increasing risk factors. Credit risk continues to be an integral factor for the financial institutions suffering losses of the order of hundreds of millions of dollars due to the inability to recover the money granted to the customers. The prediction or the assessment of whether an individual would be capable of repaying the loan has become crucial in the financial sector. The loan granting decisions consider various factors which include the past credit history of an individual, the total amount asked for the loan, the economic conditions, the age, occupation and so on. The borrower should also be capable of paying back the loan within the given time. Among the various types of loans granted to individuals, the focus of this study is the credit card customers. Credit cards are a type of loan, and they are considered as revolving credits as the amount to be paid changes every month depending upon the expenditure of the individual. A pre-approved credit limit is assigned to the customer based on his credit history and various other factors.

The proposed work performs the classification task by distinguishing the potential defaulters from the non-defaulters and thereby trying to reduce the bad debt. Two different data sets available from the UCI machine learning repository have been used and the classification task is performed using algorithms like Naive Bayes, Logistic Regression, SVM and Random forest. The above algorithms are evaluated using Waikato Environment for Knowledge Analysis (WEKA) and KNIME (Konstanz Information Miner)

II. MOTIVATION

The Indian banking sector has faced huge losses in the recent times due to borrowers unwilling to repay their money leading to bad debts. One of the fundamental reasons for this is inappropriate decision-making and granting of loans to un-qualified applicants. Risk management and the identification of credit risk of customers require handling of substantial amount of data with broad variety of aspects. Bank managers today, therefore, experience a significant challenge in the identification of potential defaulters. Issuing of credit cards to unqualified applicants has led to huge losses for the banks. Moreover, customers with repaying ability but accumulating heavy credit and over consumption of the credit limit also can lead to severe losses for the banks. Credit cards being revolving credits, the amount to be paid changes every month. Hence, effective monitoring of these aspects can help the banks in tracking the accounts that have the probability of default and thereby take the necessary action. Incorporating machine learning techniques for the prediction of defaulters can be useful for the banks in designing preventive measures and thereby avoiding losses.

III. REVIEW OF LITERATURE

Prediction of credit risk requires the use of various machine learning techniques. Some of the work done by various researchers is summarized below:

Ref.[1] proposes model for providing the measures for loss probabilities as well as the evaluation of credit risk. The data used for this purpose consists of account level data from six different banks. Three different machine learning techniques including the decision trees, random forests and logistic regression are evaluated in the proposed model. A credit card amount that is not recovered for a period of more than 90 days is considered as non-recoverable or a default.

The prediction is therefore performed for this type of dependent variable. The accounts are then classified as bad accounts and good accounts. After evaluating the three techniques, it is observed that random forests and decision trees perform better than logistic regression and are said to be suitable for these applications. The main observation worth noting is that a single model is incapable of handling the heterogeneous data and hence a customized solution needs to be formed for different banks.

In [2] Multilayer Perceptron (MLP) and k-Nearest-Neighbor (kNN) algorithms are used for the estimation of the payment status of credit card customers.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The default of credit card clients data set is used for the estimation and the problem is solved as a binary classification problem. WEKA tool is used which provides us with the Mean Absolute error(MAE) and Root Mean Squared Error (RMSE) values which helps to identify the best performance of MLP as well as kNN algorithms for the respective values of k and the respective neurons in the hidden layer.

Ref.[3] makes use of various algorithms which include BayesNet, Meta-stacking, Naive Bayes, Random forest, SMO and ZeroR for the prediction of defaulters. The credit card default data set is used which consists of 24 different attributes. Feature selection also plays a major role in the entire process which involves the identification of the best possible features. Class-feature-centroid (CFC) and Information gain (IG) are the two algorithms used for feature selection. After selecting the relevant features, classification task is performed using six different algorithms. Precision and recall are the two important performance measures used for evaluation. Random tree and random forest give higher accuracy. The feature selection algorithms used in this case contribute significantly in improving the performance of the classifiers.

The work proposed in [4] tries to predict the probability of default. As compared to other models or techniques that perform binary classification of good or bad customers, this method of trying to provide a probability value is important. A technique known as the 'Sorting Smoothing method' is used to first identify the real probability. Six different methods such as k-nearest neighbor, logistic regression, discriminant analysis and so on are used. Artificial neural networks are found to be suitable for the proposed model as it outperforms other five techniques in terms classification accuracy.

In [5], the authors have used Extreme learning machine (ELM) for the evaluation of credit risk. The problem of credit risk evaluation is seen as a classification problem, with two different classes' bad creditors and good creditors. The classification is performed on Australian credit data set as well as German credit data set and the instances are divided as positive for good applications and negative for bad applications. Along with ELM, Support Vector Machine (SVM) is also applied on the two data sets and the corresponding results are then compared. As both the classifiers are kernel based, three different kernels Gaussian kernel, polynomial kernel and hyperbolic kernel are used. Apart from the kernel functions, performance of the classifiers is also evaluated using confusion matrix. Three different parameters overall accuracy, good accuracy and bad accuracy are obtained from the confusion matrix. After evaluating the results, it is found that ELM performs better on both the data sets and classification tasks as compared to the kernel based SVM.

Support vector based classifiers are used in [6] and [8] for the evaluation of credit risk. There are various SVM based classifiers which include LibSVM, Core Vector machines(CVM), Ball Vector machines(SVM) and so on. Applying these SVM classifiers for credit risk evaluation provides good results and proves to be suitable alternative for this model. The performance can also be optimized by using some Genetic Algorithms which can help in the selection of parameters.

Some of the work based on credit scoring is also described in [10]-[13].

IV. PROPOSED SYSTEM

4.1System Flow Diagram

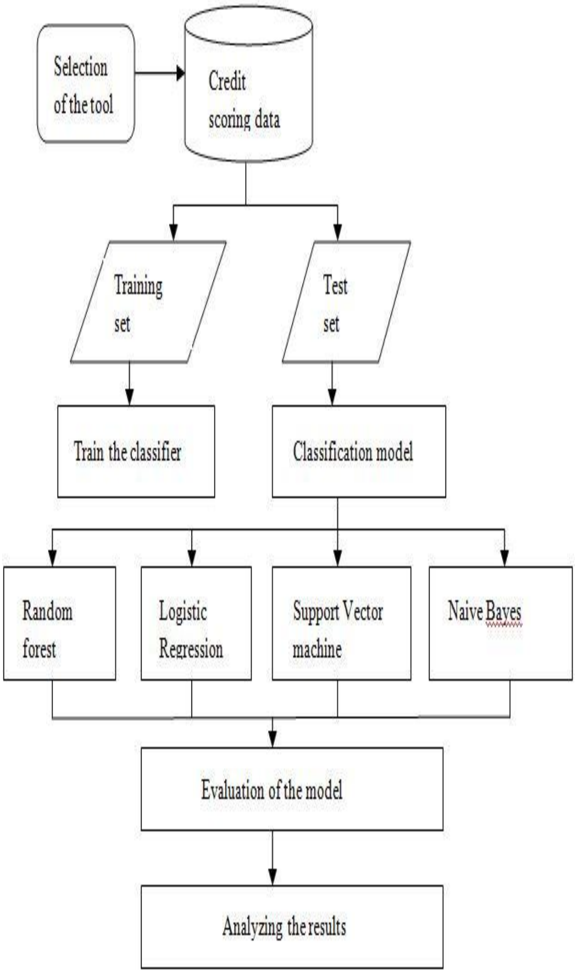

Figure 1: System Flow Diagram

The proposed system uses two different tools WEKA(Waikato Environment for Knowledge Analysis) and KNIME(Konstanz Information Miner) for the analysis of four different classification techniques. The credit scoring data set is divided into two parts- the training set as well as the test set. The classification model is trained using the training set and the remaining observations are given to perform the prediction task using different techniques.

A brief introduction of these techniques is as follows[16]:

Naive Bayes: Naive Bayes classifier is one of the supervised learning algorithms which is based on Bayes theorem and makes use of the probabilistic features. It assumes the independence between all the features for a particular model. Naive Bayes classifier is simple to implement as it assumes conditional independence.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The conditional probabilities are identified for the attributes and

classification is performed such that the class with the most probable hypothesis or maximum a posteriori (MAP) is assigned to the given element.

Logistic Regression: Logistic regression model tries to identify the correlation between the dependent and the independent variables.

It is mainly used for binary classification where the target variable is binary and one or more independent variables can be continuous or binary. This model uses the logistic function to determine the probability of the output with respect to the input. The classification is performed such that a threshold is provided, and all the probability values greater than a certain threshold are assigned one class and the values less than the threshold are assigned the other class.

Support Vector Machine: Support vector machine is a supervised learning method which can be used for classification as well as regression problems. SVM regression aims to find an approximation to a non-linear function that maps the input data into high dimensional space. Here, a hyper-plane is constructed in a way that it separates the data points with maximal margin with linear regression.

Random Forest: Random forest is also a type of supervised learning algorithm which makes use of ensemble method. It is a combination or an ensemble of decision trees. The attributes for the split are determined randomly and used for generating the individual trees. At the time of classification, the votes of all the decision trees is taken into consideration and the class with the majority vote is assigned as the class for the given value.

V. EXPERIMENTAL RESULTS

5.1Dataset Description

Two different credit scoring data sets Australian data set and the default of credit card clients data set are used which are available from the UCI machine learning repository[17].

1) Statlog (Australian Credit Approval) Data Set: Number of instances: 690

Number of attributes : 14 Class Distribution:

Class I: 383

Class II: 317

2) Default of credit card clients Data Set: Number of in-stances:30000

Number of attributes: 24 Class Distribution:

Class I: 23364

Class II: 6636

5.2Evaluation of Results

Binary classification is performed by the classifiers on both the data sets. The positive class denotes that the person will default in the next month while the negative class denotes that the person will not default.

The following tables show the accuracy values for both the data sets using WEKA and KNIME tools.

Table 1: Accuracy values for Statlog Data set

Sr. No.	Classification Algorithm	Accuracy in Weka tool (%)	Accuracy in KNIME tool (%)
1.	Naive Bayes	81.642	83.7
2.	Logistic Regression	85.507	85.6
3.	Support Vector Machine	85.507	84.6
4.	Random forest	86.47	87.5

Table 2: Accuracy values for Default of CCC Data set

Sr. No.	Classification Algorithm	Accuracy in Weka tool (%)	Accuracy in KNIME tool (%)
1.	Naive Bayes	62.42	76.6
2.	Logistic Regression	80.83	81.7
3.	Support Vector Machine	80.69	79.4
4.	Random forest	81.58	82.7

The above tables describe the classification accuracies of four different classifiers for the two data sets. The technique with the highest accuracy for both the data sets is the random forest. The performance of logistic regression and support vector machines is equally well. However, Naive Bayes classifier gave a relatively low performance with respect to the credit scoring data sets. The number of instances for the second data set is large and hence the performance of logistic regression, SVM and random forest is slightly affected with the increase in the number of instances.

VI. CONCLUSION

The use of machine learning techniques for the prediction of credit card defaulters is essential for the identification of credit risk. This can help the financial institutions in designing their future strategies. The proposed system uses Naive Bayes, logistic regression, SVM and random forest on two different credit scoring data sets. The performance of these classifiers is evaluated using the accuracy of its prediction. The classifier with the highest accuracy is found to be random forest. It is followed by logistic regression and support vector machine. Naive Bayes gives less accuracy as compared to the other classifiers. The performance of classifiers is slightly reduced when large number of instances are given for classification.

REFERENCES

1. Zhou H, Lan Y, Soh Y, Huang G and Zhang R (2012), "Credit risk evaluation with extreme learning machine", IEEE International Conference on Systems, Man, and Cybernetic(SMC), Seoul, 2012, pp. 1064-1069.
2. Butaru F,Chen Q,Clark B, Das S, Loc AW, Siddique A, (2016), "Risk and risk management in the credit card industry", Journal of Banking Finance, Vol.72, pp.218-239.
3. Koklu M, Sabanci K (2016), "Estimation of Credit Card Customers Payment Status by Using kNN and MLP", International Journal of Intelligent Systems and Applications in Engineering,Vol.4 (Special Issue) , pp.249-251 .
4. Venkatesh A, Gracia A, Shomona (2016), "Prediction of Credit-Card Defaulters: A Comparative Study on Performance of Classifiers", International Journal of Computer Applications,145, pp.36-41.
5. Yeh, Ivy Lien, Che-Hui, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients", Expert Systems with Applications. 36,pp. 2473-2480, 10.1016/j.eswa.2007.12.02

6. Paulius D, Gintautas G (2012), "Credit risk evaluation modeling using evolutionary linear SVM classifiers and sliding window approach", Proceedings of the International conference on computational science ,ICCS 2012 Book Series: Procedia Computer Science
7. Mulhim Al, Beyrouti B (2014), "Credit Scoring Model Based on Back Propagation Neural Network Using Various Activation and Error Function", International Journal of Computer Science and Network Security, Vol.14 No.3, pp. 16-24.
8. Paulius D, Gintautas G, Gudas (2011), "Credit Risk Evaluation Model Development Using Sup- port Vector Based Classifiers", Proceedings of the International conference on Computational Science ICCS 2011, Vol.4, pp.1699-1707.
9. Yao J, Lian C (2016), "A New Ensemble Model based Support Vector Machine for Credit Assessing", International Journal of Grid and Distributed Computing, Vol. 9, No.6, pp.159-168.
10. Hamid AJ, Ahmed TM (2016), "Developing prediction model of loan risk in banks using data mining", Machine Learning and Applications: An International Journal (MLAIJ), Vol.3, No
11. Brown I, Mues C (2012), "An experimental comparison of classification algorithms for imbalanced credit scoring data sets", Expert Systems with Applications, Vol. 39,Issue 3,pp. 3446-3453.
12. Tripathi D, Edla DR, Kuppili V, Bablani A (2018), "Credit Scoring Model based on Weighted Voting and Cluster based Feature Selection", Procedia Computer Science, Vol.132, pp.22-31.
13. Lessmann S, Baesens B, Seow V,Thomas L(2015),
14. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research", European Journal of Operational Research, Vol. 247, Issue 1, pp. 124-136.
15. Weka; Machine Learning Group, Waikato University, New Zealand; http://www.cs.waikato.ac.nz/ml/weka/
16. KNIME; KNIME.com AG, Germany; http://www.kni me.org/
17. J. K. Han, M. Pei, Jian, "Data Mining: Concepts and Techniques", Elsevier Publishers, Third Edition.
18. UCI – Datasets Repository, Machine Learning Center from California University; http://archive.ics.uci.edu/ml/