Prediction of Loan Status in Commercial Bank using Machine Learning Classifier

G. Arutjothi
Department of Computer Applications
Government Arts College (Autonomous)
Salem, India
garutjothi@gmail.com

Dr. C. Senthamarai
Department of Computer Applications
Government Arts College (Autonomous)
Salem, India
senthamaraiksrct@gmail.com

Abstract— **Banking Industry always needs a more accurate predictive modeling system for many issues. Predicting credit defaulters is a difficult task for the banking industry. The loan status is one of the quality indicators of the loan. It doesn't show everything immediately, but it is a first step of the loan lending process. The loan status is used for creating a credit scoring model. The credit scoring model is used for accurate analysis of credit data to find defaulters and valid customers. The objective of this paper is to create a credit scoring model for credit data. Various machine learning techniques are used to develop the financial credit scoring model. In this paper, we propose a machine learning classifier based analysis model for credit data. We use the combination of Min-Max normalization and K-Nearest Neighbor (K-NN) classifier. The objective is implemented using the software package R tool. This proposed model provides the important information with the highest accuracy. It is used to predict the loan status in commercial banks using machine learning classifier**.

Keywords: Credit Scoring; K-NN; Loan status; Loan Lending Process; Min-Max Normalization;

I. Introduction

In commercial loan lending, scoring of borrowers' creditworthiness is one of the most important problems to be addressed in the Banking Industry. Credit risk is defined as the risk that borrowers will fail to meet their loan obligations [1]. The Credit scoring system is used to predict the credit risk and to reduce the illegal activities [8]. This credit scoring systems are used to make decisions under information about the borrowers [3]. In order to make loan decisions, lenders want to minimize the risk of default of each lending decision, and realize the return that compensates for the risk [7].

In general, Banking Industry success and failure is based on their credit risk. The credit amount couldn't collect properly, then the bank will be loss. So, bank profit is correlated to their credit risk. Credit risk is a crucial challenge and a complex task to manage and evaluate [18]. Credit scoring tasks can be divided into two groups such as, application scoring and behavioral scoring. Application Scoring is to classify the credit applicant into 'good' and 'bad' risk groups. Behavioral scoring task is to classify the existing customers based on their payment history and personal information [16].

Commercial loans have always been an important part of the banking industry and lenders are always measured to minimize their credit risk. To solve this credit risk problem is too difficult [5]. Credit risk evaluation model is used to find the credit risk [2]. To evaluate the existing customer and to classify the new potential customer using credit evaluation models [9].

Data mining is the process to discover useful information from large dataset [3]. It consists of classification, clustering and association rule mining. Classification is a main function of data mining process. There are many classification techniques are available nowadays. Techniques are decision tree, support vector machine, neural network, k-nearest neighbor and logistic regression, etc. All classification techniques are already used and evaluated for this problem. But, still now couldn't find which technique is suitable for which type of dataset. Therefore, the objective of this paper is to apply a K - NN classifier to develop a credit scoring model for a commercial loan. Furthermore, the aim of this study is to classify loan applications into default customer and non-default customer group which is used for credit lenders. The results from the study would be very helpful for lenders make loan decisions.

The rest of the paper is structured as follows. In section 2 describes the basic concepts of machine learning and normalization followed by the literature survey on credit risk, K-NN in section 3. In section 4 discuss on methodology and data to this work followed by result and discussions are made in section 5. In section 6 discuss the conclusion and future work.

II. Machine Learning

A. K-Nearest Neighbor Classifier

In general, machine learning is a field of computer science that gives computers the ability to learn without being explicitly programmed. Machine learning tasks are classified into three major groups such as, supervised learning,

978-1-5386-1959-9/17/$31.00 ©2017 IEEE

unsupervised learning and semi supervised learning [14]. Supervised machine learning techniques can be used for class label known datasets and unsupervised learning techniques can be used for unknown class label datasets. In currently most of them used and focused only on classification techniques. Decision tree, support vector machine, neural network, K-NN and logistic regression are most widely using machine learning techniques.

K-NN is a vital role in the machine learning process. It can be used for both classification and regression problems [5]. This technique is mainly used for industry prediction problem and also applied for many classification problems. A distance measure is used by the K-NN classifier to make predictions without building a model [17]. Henley [11] used KNN for credit scoring and compare with other methods.

B. Min-Max Normalization

The normalization is a process of decomposing the attribute values so that they are within a specified range of smaller size [10]. It is useful to transform a large database into the simplest one and also remove the outliers. Min-Max normalization method is widely used one. This technique is responsible for accomplishing a linear transformation of actual data to the simplest one. This is done before the analysis technique is used. The normalization is a preprocessing tool [13].

Min-max normalization maps a value d of P to d′ in the range [new_min(p), new_max(p)]. The min-max normalization is calculated by the following formula:

The normalization is a process of decomposing the attribute values so that they are within a specified range of smaller size [10]. It is useful to transform a large database into simplest one and also remove the outliers. Min-Max normalization method is widely used one. This technique is responsible for accomplishing a linear transformation of actual data to simplest one. This is done before the analysis technique is used. The normalization is a preprocessing tool [13].

Min-max normalization maps a value d of P to d′ in the range [new_min(p), new_max(p)]. The min-max normalization is calculated by the following formula:

$$d' = \frac{[d\_min(p)]*[new\_max(p)-new\_min(p)]}{[max(p)-min(p)]} + new\_min(p) \qquad (1)$$

Where
min(p) = minimum value of attribute
Max(p) = maximum value of attribute

In our case min-max normalization maps a value d of P to d′ in the range [0,1], so put new_min(p) =0 and new_max(p) =1 in the above equation (6). Now we get the simplified formula of min-max normalization. Min max normalization preserves the relationship among the original data values [13].

$$d' = \frac{d - \min(p)}{\max(p) - \min(p)} \qquad (2)$$

III. Related work

We studied various articles regarding performance evaluation of data mining algorithms on different tools; some of them are described here. Loan default risk assessment is one of the crucial issues in the financial organizations. Credit scoring is widely analyzed using classification techniques. Credit scoring is a typical data mining, classification problem. Feature selection techniques are used to remove the irrelevant attributes. Feature selection based classification techniques provide empirical results than other models [2]. Abddmoula [1] applied K-NN classifier on the Tunisian commercial loan dataset. This work used on 924 credit records and they got 88.63% of classification rate.

Arutjothi [3] proposed a new credit scoring model, which is based on the hybrid feature selection method and C4.5 classifier. This relief based hybrid system not only has a strong mathematical basis, but also has higher accuracy and effectiveness. Pejic Bach [6] focused on loan decision making systems with several feature selection techniques and classifiers. The cfsSubsetEvaluator based system provides highest accuracy than other feature selection techniques.

Ajay Byanjankar [7] focused on peer-to-peer lending with a neural network classifier. The Neural network provides higher accuracy than other classifier. This result is used for lenders to make a decision on a new loan applicant. Devi [8] aimed at developing a credit scoring system which is based on ensemble learning. The ensemble learning system gives a more accurate classification rate than single classifiers. Arutjothi [9] discuss a few classification techniques with credit dataset. Compare these techniques and find the best model for credit scoring problem. Henley [11] and Feng Chia Li [17] studied about the credit scoring system with K-NN. The K-NN based system is suitable for credit scoring model.

Finally, even if there is a hundreds of research, models and methods, it is still hard to say which model is the best or which classifier or which data mining technique is the best. Each model depends on a particular data set or attributes set, so it is very important to develop a flexible model which is adaptable to every dataset or attribute set. Effective data processing and variable selection methods are also lacking.

IV. Methodology and data

A. Dataset

Data for the study has been retrieved from a publicly available data set of lending club repository loan data. The retrieved data is a pool of both defaulted and non-defaulted loans. The data comprises of demographic, financial information of borrowers and loan payment history. The dataset contains eight lakhs records and 74 attributes. We take only ten thousand records for sampling of this work. The sample data contains the default and non-default customer groups. The dataset having 73 independent attributes and one dependent attribute. The name of the dependent attribute is loan_status. This attribute defines three class labels such as, fully paid, current and bad. The class labels of the loan_status are shown in the fig 1.

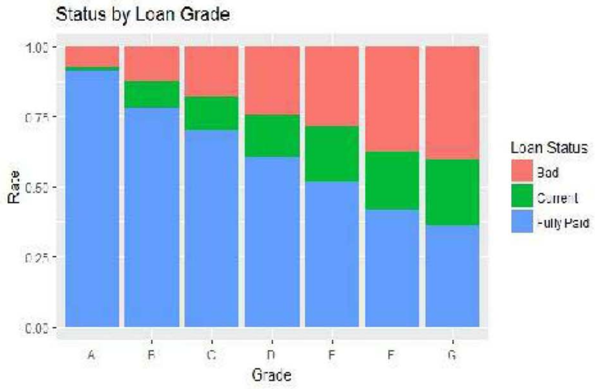


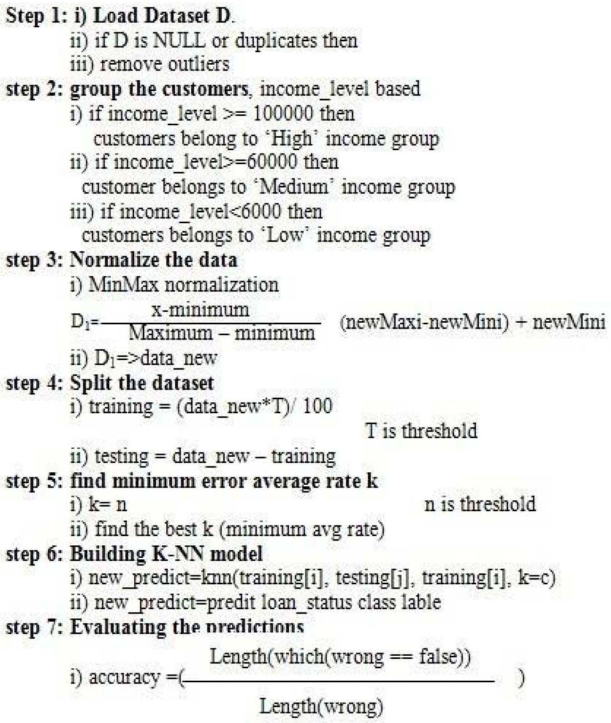
Fig 1. A simplified three class label

B. Research Methodology

The K-NN credit scoring model was built with the programming language R as it has been extensively used for academic purpose in the field of data mining. The packages of R applied in developing the model are knn, ggplot2, caTools, DescTools. Euclidean distance based K-NN is applied to build the credit scoring model. The data were partitioned into two subsets: 70% of the observations were used for training and 30% was used for model testing. In addition, the partition was performed randomly in a way that both the training and testing sets contain default cases of approximately to the same percentage. Before going to split the dataset we use Min-Max normalization to normalize the dataset.

The steps involved in this model building methodology are presented as below in fig 2. The numeric or non-numeric format of the data is loaded into R software and a set of preprocessing steps are executed before the same is used to build the classification model. The min-max normalization technique is used to transform or normalize the dataset before the classification step. The datasets have many attributes that define the credibility of the customers seeking for several types of loan. The values for these attributes can have outliers that do not fit into the regular range of data. To group the customers based on their loan_status, income_level and home_ownership attributes, in this way to remove the outliers. Some outliers are filled with Null values. The inconsistencies in the data like unbalanced dataset have to be balanced before building the classification model.

Many real time datasets have this problem and hence need to be rectified for better results. We use a random sampling method to solve this unbalanced issue. Now the training and testing dataset having the default customers records in balanced ways. The partition of the dataset which will be used in the percentage is 50%:50% (50 % of dataset is training and 50% of dataset is test dataset). The below table 1 shows the data partition summary.



```
Step 1: i) Load Dataset D.
        ii) if D is NULL or duplicates then
        iii) remove outliers
step 2: group the customers, income_level based
        i) if income_level >= 100000 then
            customers belong to 'High' income group
        ii) if income_level>=60000 then
          customer belongs to 'Medium' income group
        iii) if income_level<6000 then
          customers belongs to 'Low' income group
step 3: Normalize the data
        i) MinMax normalization
        $D_1=\frac{x\text{-minimum}}{\text{Maximum} - \text{minimum}}$ (newMaxi-newMini) + newMini
        ii) $D_1$=>data_new
step 4: Split the dataset
        i) training = (data_new*T)/ 100
                                              T is threshold
        ii) testing = data_new – training
step 5: find minimum error average rate k
        i) k= n                               n is threshold
        ii) find the best k (minimum avg rate)
step 6: Building K-NN model
        i) new_predict=knn(training[i], testing[j], training[i], k=c)
        ii) new_predict=predit loan_status class lable
step 7: Evaluating the predictions
        i) accuracy =($\frac{\text{Length(which(wrong == false))}}{\text{Length(wrong)}}$ )
```

Fig. 2. Loan_status Model.

Next step we set the iteration level (k) threshold value and find the minimum error rate based training dataset. After finding the minimum k value we can apply the knn package for R software. This knn package is defines with the k-nearest neighbor functions. This package is used to build the knn based credit scoring model. After generate the model we can evaluate this model using test dataset. Now we get the prediction dataset and it is having the predicted column. The final step is to calculate the accuracy, Root Mean Squared Error and Correlation which all of this metrics are used to evaluate the prediction dataset.

TABLE 1. DATA PARTITION SUMMARY

Samples	No. of observations	Percentage%
Training	5000	50%
Testing	5000	50%
Total	10000	100%

V. Result and Discussion

The K-NN credit scoring model was successful in classifying default and non-default loans. Hence, the commercial lenders can reduce the risk of investment failure by selecting profitable borrowers after processing the loan applications through the model. This model correctly classified the default and no-default is 75.08% in the test dataset. Table 2 presents the classification results of the model.

Table 2 shows the result for classification algorithms by using min_max normalization. We set the iteration level is 30 which is used to find the minimum error rate of the training set. The dataset is partitioned in different ways, we get the highest accuracy is 50% : 50% of the dataset.

TABLE 2. CLASSIFICATION RESULTS

S. No	K-Nearest Neighbor model (k=30)				
	Percentage Split	*Accuracy*	*RMSE*	*Correlation*	*Error Rate*
1.	80% :20%	74.5	2.04	0.99	0.256
2.	70% :30%	73.4	1.91	0.98	0.259
3.	60% :40%	74.9	2.29	0.92	0.25
4.	50% :50%	75.08	2.41	-0.26	0.24

Fig 3 shows the minimum average error rate. The graph is drawn with the iteration level on x-axis and percentage of error rate on y-axis.

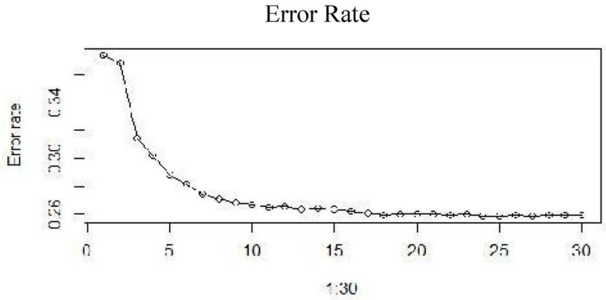


Fig. 3. Error rate retrieved from package R.

The applied k-nearest neighbor credit scoring model successfully demonstrates the applicability of K-NN in credit scoring for classification and prediction of the loan status. Data mining techniques are used to develop the credit scoring systems. The K-NN based credit scoring system provides higher accuracy than other classifiers. The data is used to develop K-NN credit scoring model with different iteration levels. The comparison study is made between the iterations, and 30 iterations based K-NN system provides the highest accuracy. Fig 3 shows the minimum error rate of training data. If the iteration level is increased then the error rate is gradually decreased. This minimum error rate based credit scoring model provides significant results for predicting the loan status in the commercial banks. This proposed model presented in this study can be effectively used by commercial loan lenders to predict the loan applicant. Lenders can use this model to predict the loan status of the loan applicant.

VI. CONCLUSION

In this paper, we have proposed a loan status model to predict the loan applicant as a valid customer or default customer. The proposed model shows 75.08% accuracy result in classifying credit applicant using R package. The credit lenders can use this model to make a loan decision on loan proposals. Further, the comparison study has been made with different levels of iterations. The iteration level is 30 based k-NN model gives significant accuracy than other levels. This model can be used to avoid the huge loss of commercial banks.

References

[1] Abdelmoula, Aida Krichene. "Bank credit risk analysis with k-nearest-neighbor classifier: Case of Tunisian banks." Accounting and Management Information Systems 14.1 (2015): 79.

[2] Arutjothi, G., Dr. C. Senthamarai. "Comparison of Feature Selection Methods for Credit Risk Assessment", International Journal of Computer Science, Volume 5, Issue 1, No 5, 2017.

[3] Arutjothi,G.,Dr.C.Senthamarai. "Credit Risk Evaluation using Hybrid Feature Selection Method." Software Engineering and Technology 9.2 (2017): 23-26.

[4] Attig, Anja, and Petra Perner. "The Problem of Normalization and a Normalized Similarity Measure by Online Data." Tran. CBR 4.1 (2011): 3-17.

[5] Babu, Ram, and A. Rama Satish. "Improved of K-Nearest Neighbor Techniques in Credit Scoring." International Journal For Development of Computer Science & Technology 1 (2013).

[6] Bach, Mirjana Pejić, et al. "Selection of Variables for Credit Risk Data Mining Models: Preliminary research." MIPRO 2017-40 th Jubilee International Convention. 2017.

[7] Byanjankar, Ajay, Markku Heikkilä, and Jozsef Mezei. "Predicting credit risk in peer-to-peer lending: A neural network approach." Computational Intelligence, 2015 IEEE Symposium Series on. IEEE, 2015.

[8] Devi, CR Durga, and R. Manicka Chezian. "A relative evaluation of the performance of ensemble learning in credit scoring." Advances in Computer Applications (ICACA), IEEE International Conference on. IEEE, 2016.

[9] G.Arutjothi, Dr.C.Senthamarai, “Effective Analysis of Financial Data using Knowledge Discovery Database”, International Journal of Computational Intelligence and Informatics, Vol. 6: No. 2, September 2016

[10] Goel, Dr Himani, and Gurbhej Singh. "Evaluation of Expectation Maximization based Clustering Approach for Reusability Prediction of Function based Software Systems." International Journal of Computer Applications (0975–8887) Volume (2010).

[11] Henley, W. E., and David J. Hand. "A k-nearest-neighbour classifier for assessing consumer credit risk." The statistician(1996): 77-95.

[12] Huang, Cheng-Lung, Mu-Chen Chen, and Chieh-Jen Wang. "Credit scoring with a data mining approach based on support vector machines." Expert systems with applications 33.4 (2007): 847-856.

[13] Jain, Y. Kumar, and Santosh Kumar Bhandare. "Min max normalization based data perturbation method for privacy protection." International Journal of Computer & Communication Technology 2.8 (2011): 45-50.

[14] Khandani, Amir E., Adlar J. Kim, and Andrew W. Lo. "Consumer credit-risk models via machine-learning algorithms." Journal of Banking & Finance 34.11 (2010): 2767-2787.

[15] Kumar, Vipin, and Sonajharia Minz. "Feature Selection: A literature review." SmartCR4.3 (2014): 211-229.

[16] Laha, Arijit. "Building contextual classifiers by integrating fuzzy rule based classification technique and k-nn method for credit scoring." Advanced Engineering Informatics 21.3 (2007): 281-291.

[17] Li, Feng-Chia. "The hybrid credit scoring strategies based on knn classifier." Fuzzy Systems and Knowledge Discovery, 2009. FSKD'09. Sixth International Conference on. Vol. 1. IEEE, 2009.

[18] Sudhamathy, G., and C. Jothi Venkateswaran. "Analytics using R for predicting credit defaulters." Advances in Computer Applications (ICACA), IEEE International Conference on. IEEE, 2016.