Survey For Credit Scoring Model using Data Mining Algorithms

Pratiksha Pawar[1], **Pranali Rajput**[2], **Shreya Shejwalkar**[3], **Pradnya Borse**[4] , **Prof. S. M. Malao**[5]

[1] Student, Computer Engineering, K.K.Wagh Nashik, Maharashtra, India
[2] Student, Computer Engineering, K.K.Wagh Nashik, Maharashtra, India
[3] Student, Computer Engineering, K.K.Wagh Nashik,. Maharashtra, India
[4] Student, Computer Engineering, K.K.Wagh Nashik Maharashtra, India
[5] Professor, Computer Engineering, K.K.Wagh Nashik, Maharashtra, India

ABSTRACT

Credit scoring means applying a statistical model to assign a risk score to a credit application.Credit scoring techniques assess the risk in lending to a particular client. They not only identify good applications and bad applications on an individual basis, but also they forecast the probability that an applicant with any given score will be good or bad.Although credit scoring systems are being implemented and used by most banks nowadays, they do face a number of limitations.The availability of high-quality data is a very important prerequisite for building good credit scoring models. However, the data need not only be of high quality, but it should be predictive as well, in the sense that the captured characteristics are related to the customer defaulting or not.The statistical techniques used in developing credit scoring models typically assume a data set of sufficient size containing enough details. This may not always be the case for specific types of portfolios where only limited data is available, or only a low number of defaults is observed. It is observed that Credit Scoring Model is much more accurate and efficient when it is executed on data that has been carefully prepared and pre-processed. Data mining could be applied in the process of Credit Scoring that is used to predict default clients n order to decide whether to grant them a credit especially by using classification algorithms. Also, data pre-processing can be used on imbalance credit data for improving risk prediction.

Keyword : *Information retrieval, Data Structures,Information Integration, Data Cleaning, Wrappers*.

1. INTRODUCTION

There are many existing systems which perform credit scoring model. But if we talk about the accuracy, most of them had failed. There are some customers be have negatively after their application are approved for loan.To prevent this situation, banks have to find some methods to predict customers behaviours.Data prepossessing techniques have a pretty good performance on this purpose for increasing the accuracy.Data quality issues can be difficult to detect without specific domain knowlegde,but have an important impact on the scorecard development and resulting risk measures. Hence, Credit Score Model is much more accurate and efficient when it is executed on data that has been carefully prepared and pre-processed.

The availability of high quality data is a very important prerequisite for building good credit scoring models. However, the data need not only be of high quality, but it should be predictive as well, in the sense that the captured characteristics are related to the customer defaulting or not.The statistical techniques used in developing credit scoring models typically assume a data set of sufficient size containing enough details. This may not always be the case for specific types of portfolios where only limited data is available, or only a low number of defaults is observed. It is observed that Credit Scoring Model is much more accurate and efficient when it is executed on data that has been carefully prepared and pre-processed. Data mining could be applied in the process of Credit Scoring that is used to predict default clients in order to decide whether to grant them a credit especially by using classification algorithms. Also, data pre-processing can be used on imbalance credit data for improving risk prediction.

2. LITERATURE SURVEY

In [1] Credit scoring using predictive models can help in the process of assessing credit worthiness during the credit evaluation process. The objective of credit scoring models is to assign credit risk score to determine if a customer is likely to default on the financial obligation. Construction of credit scoring models requires data mining techniques. Using historical data on payments, demographic characteristics and statistical techniques, credit scoring models can help identify the important demographic characteristics related to credit risk and provide a score for each customer. In[1] author illustrate the construction and comparison of three credit scoring models: logistic regression (LR) model, classification and regression tree (CART) model and neural network (NN) model to discriminate between rejected and accepted credit card applicants of a bank. Results show that Neural Network model has a slightly higher validation predictive accuracy rate (LR = 74.56, NN = 76.46, CART = 73.66).

In [3],A decision tree is an important classification technique in data mining classification. Decision trees have proved to be valuable tools for the classication, description, and generalization of data. J48 is a decision tree algorithm which is used to create classification model. J48 is an open source Java implementation of the C4.5 algorithm in the Weka data mining tool. In [3] author present the method of improving accuracy for decision tree mining with data preprocessing. We applied the supervised filter discretization on J48 algorithm to construct a decision tree.Compared the results with the J48 without discretization. The results obtained from experiments show that accuracy of J48 after discretization is better than J48 before discretization.

In [4], Analogy-based software effort estimation is a method to estimate the project cost of an unseen project based on analogies against previous projects sharing selected features. The validity of the selected features depends on many factors, and one of most crucial factors is the effectiveness of the data-preprocessing techniques applied to the datasets of the previous projects. In [4] author report the first controlled experiment that studies the class of three-stage data-preprocessing techniques with stages of missing data imputation, data normalization, and feature selection for analogy-based effort estimation. We conducted our investigation on the ISBSG data. The experimental results show that three-stage data-preprocessing techniques have significant impacts on the resultant effort estimation accuracy. The results also indicate that the combined use of Z-Score normalization, kNN imputation and mutual information based feature weighting can be an effective choice for analogy-based effort estimation.

In [5], Imbalanced credit data sets refer to databases in which the class of defaulters is heavily under-represented in comparison to the class of non-defaulters. This is a very common situation in real-life credit scoring applications, but it has still received little attention. This paper investigates whether data resampling can be used to improve the performance of learners built from imbalanced credit data sets, and whether the effectiveness of resampling is related to the type of classifier. Experimental results demonstrate that learning with the resampled sets consistently outperforms the use of the original imbalanced credit data, independently of the classifier used.

In[6], The credit scoring has been regarded as a critical topic and its related departments make efforts to collect huge amount of data to avoid wrong decision. An effective classificatory model will objectively help managers instead of intuitive experience. [6] proposes five approaches combining with the back-propagation neural network (BPN) classifier for features selection that retains sufficient information for classification purpose. Different credit scoring models are constructed by selecting attributes with five approaches. Two UCI (University of California, Irvine) data sets are chosen to evaluate the accuracy of various hybrid-BPN models. BPN classifier combines with conventional statistical LDA, Decision tree, Rough sets theory, F-score and Gray relation approaches as features preprocessing step to optimize feature space by removing both irrelevant and redundant features. In [6], the procedure of the proposed approaches will be described and then evaluated by their performances. The results are compared in combination with BPN classifier and nonparametric Wilcoxon signed rank test will be held to show if there is any significant difference between these models. The result in [6] suggests that hybrid credit scoring approach is mostly robust and effective in finding optimal subsets and is a promising method to the fields of data mining.

3. SYSTEM ARCHITECTURE

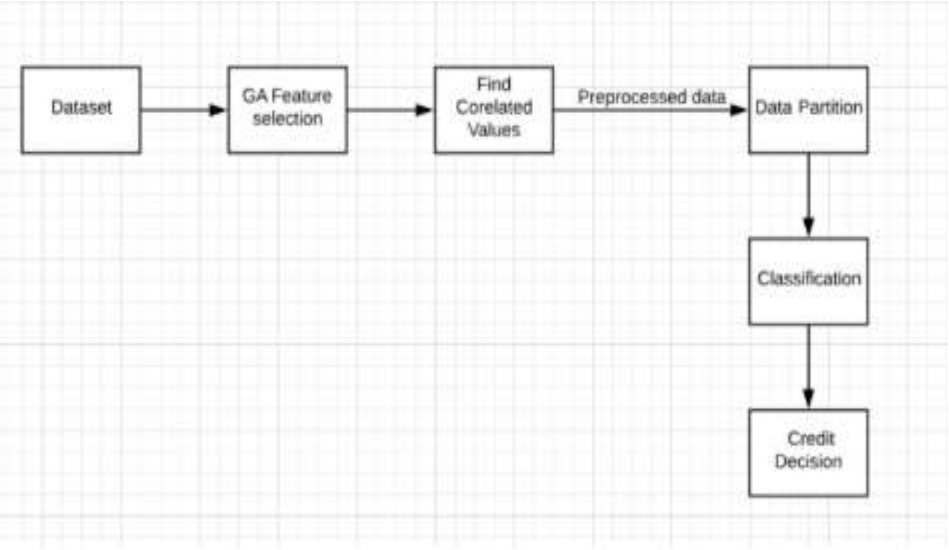

Figure: System Architecture

1.Dataset: A micro- financing institution from Bosnia and Herzegovina. Which involved data about their clients,loans and repayment history.The original dataset consisted of 23615 records described with 33 attributes.

2. GA Feature Selection : In feature selection, th function to optimize is the generalization performance of a predictive model. More specifically, we want to minimize the error of the model on an independent dataset not used to create the model. This function is called the selection error. The design variables are the presence (1) or absence (0) of every possible feature in the model.

3. Find Corelated Value : Corelation is often used as a preliminary technique to discover relationship between variables.More precisely, Corelation is a measure of the linear relationship between two variables.

4. Data Partitioning : Dataset is partitioned into a training set and testing set.Training set is used towards learning a model and the test set is then used towards evaluating the performance of the model learned from the training set.Dataset is randomly splited into approximately 70% for training and 30% for testing.

5. Classification :

• GLM : GLM is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution.

• DT : DT builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.

• SVM : SVM are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.

6. Credit Decision : Credit scoring model will give decision whether to approve loan request or not.

4. CONCLUSIONS

Credit Scoring Model can be much more accurate and efficient when it is executed on data that has been carefully prepared and pre-processed. Data mining could be apply in the process of Credit Scoring that can be used to predict default clients in order to decide whether to grant them a credit especially by using classification algorithms. Also, data pre-processing can be used on imbalance credit data for improving risk prediction. Apart from improvement in accuracy, and in other showed measurs running time of applied algorithms has also shown great improvement in algorithm execution time. Hence, Credit Scoring Model becomes more accurate and efficient by using data mining algorithms and techniques.

6. ACKNOWLEDGEMENT

It gives us great pleasure in presenting the preliminary project report on ‘Credit Scoring Model using Data Mining Algorithms’.

We would like to take this opportunity to thank our internal guide Prof. S. M. Malao for giving us all the help and guidance we needed. We are really grateful to them for their kind support. Their valuable suggestions were very helpful.

We are also grateful to Prof. Dr. S. S. Sane, Head of Computer Engineering Department, K. K. Wagh Institute Of Engineering Education & Research for his indispensable support, suggestions.

In the end our special thanks to Staff Members of the Department for providing various resources such as laboratory with all needed software platforms, continuous Internet connection, for our Project.

7. REFERENCES

[1]Y. B. Wah, I. R. Ibrahim,”Using Data Mining Predictive Models to Classify Credit Card Applicants”, 6th International Conference Advanced Information Management and Service (IMS), pp. 394- 398, 2010

[2] P. Chandrasekar, K. Qian, H. Shahriar, P. Bhattacharya, ”Improving the Prediction Accuracy of Decision Tree Mining with Data Preprocessing”, IEEE 41st Annual Computer Software and Applications Conference, pp. 481-484, 2017.

[3] J. Huang, Y. Li, J. W. Keung, Y. T. Yu, W.K. Chan, ”An Empirical Analysis of Three-stage Data-Preprocessing for Analogy-based Software Effort Estimation on the ISBSG Data”, IEEE International Conference on Software Quality, Reliability and Security, pp. 442-449, 2017

[4] V. Garcia, A. I. Marques, J. S. Sanchez, ”Improving Risk Predictions by Preprocessing Imbalanced Credit Data”, International Conference on Neural Information Processing, pp. 68-75, 2012

[5] L. Feng-Chia, W. Peng-Kai, Y. Li-Lon, ”Diversity of Feature Selection Approaches combined with Distinct Classifiers”, IEEE International Conference on Industrial Engineering and Engineering Management, 2010.

[6] R. P. Bunker, M. A. Naeem, W. Zhang, Improving a Credit Scoring Model by Incorporating Bank Statement Derived Features, October 2016

[7] F. Louzada, A. A. Guilherme, B. Fernandes, ”Classification methods applied to credit scoring: Systematic review and overall comparison”, February 2016.

[8] J. Hariharakrishnan, S. Mohanavalli, Srividya, K.B. Sundhara Kumar, ”Survey of Pre-processing Techniques for Mining Big Data”, International Conference on Computer, Communication and Signal Processing (ICCCSP), 2017

[9] A. Saleem, K. H. Asif, A. Ali, S. M. Awan, M. A. Alghamdi, ”Pre-processing Methods of Data Mining”, IEEE/ACM 7th International Conference on Utility and Cloud Computing, 2014..

[10] L. Capodiferro, L. Constantini, F. Mangiatordi, E. Pallotti, ”Data Preprocessing to Improve SVM Video Classification”, 10th International Workshop on Content-Based Multimedia Indexing (CBMI), 2012

[11] Abell‘an, J.n, J., Mantas, C. ,”Improving experimental studies about ensembles of classifiers for bankruptcy prediction and credit scoring. Expert Systems with Applica- tions” 41 , 38253830 , 2014.

[12] Chen, M.-C., Huang, S.-H. , ”Credit scoring and rejected instances reassigning through evolutionary computation techniques” Expert Systems with Applications 24 , 433441 , 2003

[13] Fatemeh nematikoutanaei ,hediehsajedi & mohammadkhanbabaeic ,”a hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring”,Journal of retailing and consumer services, volume 27november 2015,

[14]George H.John, RonKohavi & KarlPfleger, “Irrelevant Features and the Subset Selection Problem”,Machine Learning Proceedings 1994,Proceedings of the Eleventh International Conference, Rutgers University, New Brunswick, NJ, July 10–13, 1994.

[15]Edgar Acuna and Caroline Rodriguez, “The Treatment of Missing Values and its Effect on Classifier Accuracy”,from book Classification, Clustering, and Data Mining Applications: Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago, 15–18 July 2004.