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 highquality 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.

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

1. INTRODUCTION

As a tool for Knowledge Discovery in Databases (KDD), Data Mining has become very important in highly competitive business market for companies to extract some hidden information and patterns that can help them stay ahead of their competitors. It can help find unknown profitability, improve efficiency or help company's management make more correct decisions for the future. It is involved in different business domains, and so is in institutions and companies from Financial sector. 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 classification algorithms : Generalized Liner Model (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. The GLM generalizes linear

regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value .Decision Trees (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. Support Vector Machines(SVM) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.Naive Bayes algorithm is a simple but surprisingly powerful algorithm for predictive modeling.Naive Bayes is a classification algorithm for binary (two-class) and multi-class classification problems. The technique is easiest to understand when described using binary or categorical input values. The calculation of the probabilities for each hypothesis are simplified to make their calculation tractable . Extraction of hidden information and patterns requires implementation and appliance of advanced algorithms with careful analysis in order to choose technique that suits structure of given data sample the best. However, for the data mining model to be efficient and correct as much as possible, data pre-processing is of the crucial importance. It involves various tasks in order to prepare data so that data mining technique applied to it produces high-quality and accurate output patterns. Some of data pre-processing techniques are: data aggregation,feature selection and creation, data discretization and variable transformation.

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 [2],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 [2] 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 [3], 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 [3] 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 [4], 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[5], 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. [5] 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 [5], 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 signif-icant difference between these models. The result in [5] 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.

In [6], author investigate the extent to which features derived from bank statements provided by loan applicants, and which are not declared on an application form, can enhance a credit scoring model for a New Zealand lending company. Exploring the potential of such information to improve credit scoring models in this manner has not been studied previously. Author construct a baseline model based solely on the existing scoring features obtained from the loan application form, and a second baseline model based solely on the new bank statement-derived features. A combined feature model is then created by augmenting the application form features with the new bank statement derived features. The experimental results using ROC analysis show that a combined feature model performs better than both of the two baseline models, and show that a number of the bank statement-derived features have value in improving the credit scoring model. The target data set used for modelling was highly imbalanced, and Naive Bayes was found to be the best performing model, and outperformed a number of other classifiers commonly used in credit scoring, suggesting its potential for future use on highly imbalanced data sets.

In [7], The need for controlling and effectively managing credit risk has led financial institutions to excel in improving techniques designed for this purpose, resulting in the development of various quantitative models by financial institutions and consulting companies. Hence, the growing number of academic studies about credit scoring shows a variety of classification methods applied to discriminate good and bad borrowers. In[7] author aims to present a systematic literature review relating theory and application of binary classification techniques for credit scoring financial analysis. The general results show the use and importance of the main techniques for credit rating, as well as some of the scientific paradigm changes throughout the years.

In [8], Big Data analytics has become important as many administrations, or-ganizations, and companies both public and private have been collecting and analyzing huge amounts of domain-specific information, which can contain useful information about problems such as national intelligence, cyber security, fraud detection, marketing, and medical informatics. With more and more data being generated the ever dynamic size, scale, diversity, and complexity has made the requirement for newer architectures, techniques, algorithms, and analytics to manage it and extract value from the data collected. The progress and innovation is no longer hindered by the ability to collect data but, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion as well as a credible clean and noise free data sets. Author mainly makes an attempt to understand the different problems to solve in the processes of data preprocessing, to also familiarize with the problems related to cleaning data , know the problems to apply data cleaning and noise removal techniques for big data analytics and to mitigate the imperfect data, together with some techniques to solve them and also to identify the shortcomings in the existing methods of the reduction techniques in the necessary respective areas of application and also to identify the current big data pre-processing proposals effectiveness to various data sets.

In [9], Data Mining has become essential tool for discovery of hidden patterns and information in databases. However, for a Data Mining model to be meaningful and effective, data pre-processing is one of the key factors in successful model preparation. In [9] author have investigated how data pre-processing affects real dataset when applying Data Mining technique for the purpose of predicting default clients in a micro-financing institution. Therefore, several data pre-processing techniques have been described and applied to the dataset. Results are shown and compared for both of the cases with Generalized Linear Model and Decision Tree being the two Data Mining classification algorithms used for Credit Scoring model. It is concluded that Credit Scoring Model is much more accurate and efficient when it is executed on data that has been carefully prepared and pre-processed.

In [10], A pre-processing strategy to improve the performances of SVM in video clips classification is proposed. The segmentation of a video clip and the extraction of key frames, whose representation in terms of lowlevel features constitute the basic elements for the generation of the SVM data sets, are generally performed in an automatic way. This approach may produce several noise data, and it is therefore desirable to find a removal strategy. Noise key frames are usually detected when video includes color bars, test cards or other homogeneous frames. Duplicated key frames, generated when video is steady for a long while, also need to be removed. Author propose a data clustering method that performs an automatic pre-processing of SVM data sets, to minimize the presence of noise. Our experiments show an example of classification of historical sport video clips, demonstrating that the proposed pre-processing strategy improves the overall performances of SVM.

In [11], Previous studies about ensembles of classifiers for bankruptcy prediction and credit scoring have been presented. In these studies, different ensemble schemes for complex classifiers were applied, and the best results were obtained using the Random Subspace method. The Bagging scheme was one of the ensemble methods used in the comparison. However, it was not correctly used. It is very important to use this ensemble scheme on weak and unstable classifiers for producing diversity in the combination. In order to improve the comparison, Bagging scheme on several decision trees models is applied to bankruptcy prediction and credit scoring. Decision trees encourage diversity for the combination of classifiers. Finally, an experimental study shows that Bagging scheme on decision trees present the best results for bankruptcy prediction and credit scoring.

In [12], The credit industry is concerned with many problems of interest to the computation community. This study presents a work involving two interesting credit analysis problems and resolves them by applying two techniques, neural networks (NNs) and genetic algorithms (GAs), within the field of evolutionary computation. The first problem is constructing NN-based credit scoring model, which classifies applicants as accepted (good) or rejected (bad) credits. The second one is better understanding the rejected credits, and trying to reassign them to the preferable accepted class by using the GA-based inverse classification technique. Each of these problems influences on the decisions relating to the credit admission evaluation, which significantly affects risk and profitability of creditors. From the computational results, NNs have emerged as a computational tool that is well-matched to the problem of credit classification. Using the GA-based inverse classification, creditors can suggest the conditional acceptance, and further explain the conditions to rejected applicants. In addition, applicants can evaluate the option of minimum modifications to their attributes.

In[13]Authors have developed a hybrid data mining model of feature selection and ensemble learning classification algorithms on the basis of three stages. The first stage, deals with the data gathering and pre-processing. In the second stage, four FS algorithms were employed, including principal component analysis (PCA), genetic algorithm (GA), information gain ratio, and relief attribute evaluation function.After choosing the appropriate model for each selected feature, they are applied to the base and ensemble classification algorithms. In this stage, the authors choosed the best FS algorithm with its parameters for the modeling stage of the proposed model.At last,the results exhibited that in the second stage, PCA algorithm is the best FS algorithm. Authors proposed that the hybrid model is an operative and strong model for performing credit scoring.

In[14]Authors have addressed the problem of finding a subset of features that allows a supervised induction algorithm to induce small high-accuracy concepts. They examined the notions of relevance and irrelevance, and showed that the definitions used in the machine learning literature do not adequately partition the features into useful categories of relevance. They presented the definitions for irrelevance and for two degrees of relevance and those definitions improved the understanding of the behavior of previous subset selection algorithms, and help define the subset of features that should be sought. They concluded that the features selected should depend not only on the features and the target concept, but also on the induction algorithm. Authors described a method for feature subset selection using cross-validation that is applicable to any induction algorithm, and discused the experiments conducted with ID3 and C4.5 on artificial and real datasets.

In[15] Authors have experiented with twelve datasets to evaluate the effect on the misclassification error rate of four methods for dealing with missing values: the case deletion method, mean imputation, median imputation and KNN imputation procedure.The classifiers considered by them were the Linear discriminant analysis (LDA) and the KNN classifier. Authors concluded that in datasets with an small amount of instances containing missing values there is not much difference between case deletion and imputation methods for both type of classifiers. But this is not the case for datasets with a high percentage of instances with missing values.Overal results showed that KNN imputation seems to perform better than the other methods because it is most robust to bias when the percentage of missing values increases. In general doing imputation does not seems to hurt too much the accuracy of the classifier even sometimes with a high percentageof instances with missing values.Finally, authors recomended that we can deal with datasets having up to 20 % of missing values.

3. SYSTEM ARCHITECTURE

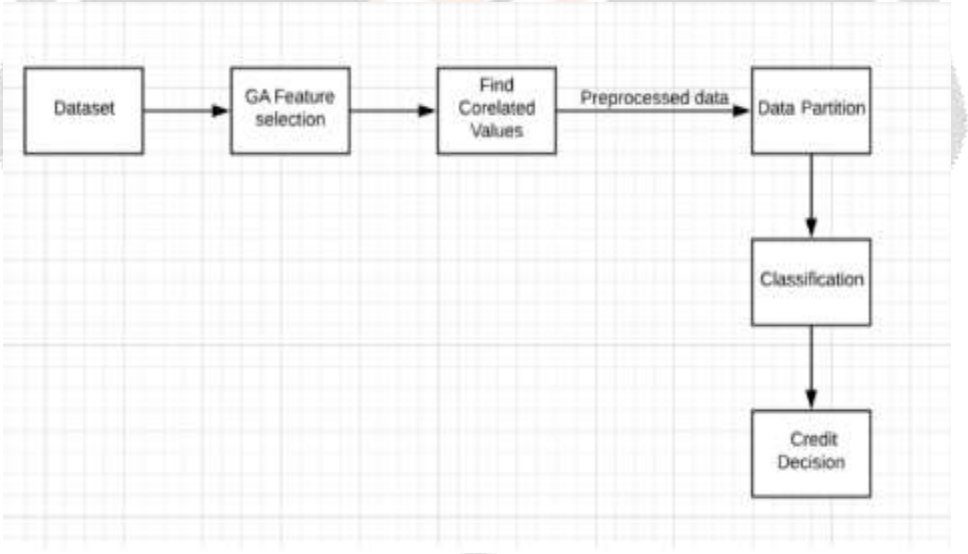

Figure: System Architecture

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

This review explains various technologies used for credit scoring model. Previous surveys have concentrated on the data which is used for building the credit scoring model is not preprocessed data. Predictive models can help in the process of accessing credit worthiness during the credit evaluation process . Decision tree algorithms have been used for the classification purpose. To improve the performance of imbalanced credit data various resampling techniques have been used.

Also various reviews shows that the data which is used for building the credit scoring model if it preprocessed then it increases the efficiency and improves the performance of credit scoring model. 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.