Importance of Data Pre-processing in Credit Scoring Models Based on Data Mining Approaches

Jasmina Nalić* and Amar Švraka **
* J.J. Strossmayer University of Osijek / Computer Science and Information Technology, Osijek, Croatia
** Sarajevo School of Science and Technology/Computer Science, Sarajevo, Bosnia and Herzegovina
jasmina.nalic@gmail.com/amar.svraka@stu.ssst.edu.ba

***Abstract* - 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 this paper, we 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.**

Keywords - Classification, Credit Scoring, Data Mining, Data Pre-processing, Decision Tree, Generalized Linear Model

I. 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. They recognized that 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 such as Generalized Liner Model (GLM), Decision Trees, Support Vector Machines and Naive Bayes algorithm [1].

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].

II. RELATED WORK

Regarding investigation of data pre-processing techniques, their impact and appliance, several papers have been published.

Chandrasekar, Qian, Shahriar & Bhattacharya 2017 [3] investigated how data pre-processing methods can be used to improve the prediction accuracy of decision tree. Authors described decision tree algorithms J48 and C4.5 as data mining techniques used for classification. They used data from a real world leukemia microarray experiment and Weka as software for Data Mining for their research. Next, they described and applied supervised discretization filter on J48 algorithm to construct a decision tree in Weka. Finally, results were compared with J48 algorithm without discretization. Paper concluded that J48 algorithm shows much better accuracy and performance when appropriate data pre-processing is applied to it.

Huang, Li, Keung, Yu & Chan 2017 [4] analysed three-stage data pre-processing for analogy-based software effort estimation. Firstly, concept of analogy-based software was explained and how it can be used. Emphasis was put on data pre-processing as one of the key factors to ensure validity of the selected features. Therefore, three data pre-processing techniques were explained in detail: missing data imputation, data normalization, and feature selection for analogy-based effort estimation. Results of the research showed that those techniques have had significant impact on the accuracy of effort estimation. Authors proposed that experiments on more datasets should be conducted to verify the findings, more advanced estimation methods should be involved as well as more advanced missing imputation approaches.

Garcia, Marques and Sanchez 2012 [5] showed how data pre-processing on imbalanced credit data is used for improving risk predictions. Authors described dataset used for their research as well as concept of imbalanced credit data where class of defaulters is under-represented in comparison to the class of non-defaulters. They investigated if data resampling technique could be used to improve accuracy of learners built from imbalanced credit

datasets. Two resampling methods were described (over-sampling and under-sampling) with both of them applied in the experiment described in paper. Results showed that learning with resampled dataset provides much better performance and accuracy than the learning with the original one. Paper concluded that even a small improvement in accuracy and performance in credit scoring model can lead to future savings and affect commercial implications in a very positive way.

A. *Paper Contribution*

This paper compares results of Data Mining algorithm applied to the data that was not pre-processed against data that went through several data pre-processing techniques. It describes these techniques in detail with focus on real dataset provided by a micro-financing institution from Bosnia and Herzegovina. Furthermore, it analyses and thoroughly compares how the outcomes of applied Data Mining algorithm are affected by data pre-processing and verifies importance of appliance of such techniques.

III. RESEARCH METHODOLOGY

For our research, we used Oracle Data Miner (ODM), software package for data mining by Oracle company. For the given problem of classifying results in one of the two predefined categories, the best approach was usage of Classification algorithm. ODM uses four different algorithms for the classification: Support Vector Machine (SVM), Naive Bayes (NB), Generalized Linear Model (GLM) and Decision Tree (DT). All four algorithms were applied in our experiment, but the best results were shown by decision tree and generalized linear model, so these two data mining algorithms will be considered later in this paper.

In this research, two credit scoring model were built, both based on DT and GLM algorithms. The first model was trained and tested on real dataset before pre-processing and the second one on dataset after pre-processing. After that, the outcomes of the experiment were measured, compared and evaluated outcomes of the experiment.

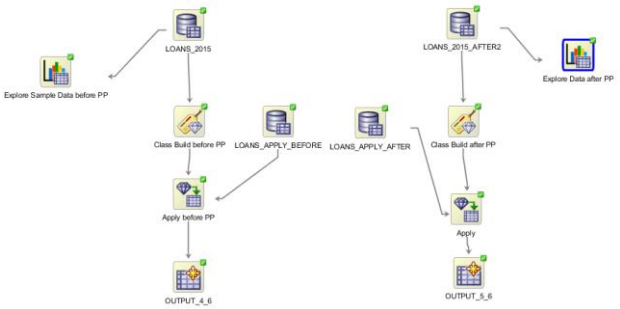

Figure 1. Credit Scoring models in ODM

Figure 1 depictures credit scoring models created in ODM. Credit scoring model applied on data before pre-processing is shown left in the picture and the right model is credit scoring model applied on the pre-processed dataset. After training of both models and evaluating the results, each of them were validated through applying on new dataset.

A. *Dataset Description*

In this paper a real dataset provided by a micro-financing institution from Bosnia and Herzegovina was used, which involved data about their clients, loans and repayment history. The original dataset consisted of 23615 records described with 33 attributes. These attributes are described in the following table:

TABLE I. DATASET DESCRIPTION

Name of Attribute	**Attribute Domain**	**Description**
UNIQUE_ID	Unique ID, Case Id	Unique id of loan request
SOC_STATUS	categorical	Social status of the client
DEMOG_STATUS	categorical	Demographic status of the client
QUALIFIACTION	categorical	Professional qualification of the client
GENDER	categorical	Client's gender (M/F)
TYPE_OF_CLIENT	Categorical	Type of client (F - individual, P - legal subject)
OWN_BUSINESS	categorical	Is client working in own business (Y/N)
OWNER_OF_BUSSINES	categorical	Is client owner of business (Y/N)
NATIONALITY	categorical	Client's nationality
URBAN_RURAL	categorical	Is client from urban or rural environment
MARTIAL_STATUS	categorical	Client's marital status
AGE	numerical	Client's age
DURATION_ON_ADDRESS	numerical	Duration of living on the same address in years
SOURCE_OF_INCOME	categorical	Client's source of income
TYPE_OF_CONTRACT	categorical	Client's type of contract with his/her employer (Unlimited time, Limited time, Seasonal etc.)
DURATION_OF_EMPLOYMENT	numerical	Duration of employment in years
LIVING_PLACE	categorical	Client's living place (Own, Rented, Parents, Others)
HH_NO_OF_INCOMES	numerical	Number of incomes in household
HH_NO_OF_NONADULTS	numerical	Number of non-adults in household
HH_NO_OF_ADULTS	numerical	Number of adults in household
HH_NO_OF_PETS	numerical	Number of pets in household
HH_NO_OF_DEPENDENT_MEMBERS	numerical	Number of dependent household members

AMOUNT_OF_LOAN	numerical	Amount of Loan
NUMBER_OF_INSTALMENT	numerical	Number of installments
TYPE_OF_LOAN	categorical	Purpose of taking loan (Agriculture, Trade, Production, etc.)
CLIENT_CYCLE	numerical	Ordinal number of client's loan request
COLLATERALS	categorical	Does client have collaterals (Y/N)
LOANS_IN_OTHER_BANKS	Categorical	Client has loan in other bank (Y/N)
CLIENTS_SEGMENT	categorical	Client's employment status (Employed, Unemployed, Family member is employed)
HH_INCOMES	numerical	Total amount of household income
HH_COSTS	numerical	Household costs
HH_AMOUNT_OF_INSTALMENT	numerical	Total amount of installments in household
DEFAULT_01	Target value	Is Client default (0 - Non-default / 1 - Default)

Original dataset consisted of loan requests that were submitted during one year period. Case Id of the dataset is unique id of loan request and the target attribute is classification attribute default_01. Clients are classified as bad or default if they have ever been late in paying loan obligations more than 30 days. Otherwise, they are classified as good or non-default. Since, our main target is recognizing default loans, positive class is the default one.

B. Data Mining Algorithms

Decision Tree is a supervised learning algorithm used to represent knowledge in intuitive and understandable form. It is based on recursion by selecting an attribute as a root node and making branches that represent an association between attributes and classes, as a result of attribute value test [6]. Furthermore, it uses historical data in form of training set to develop its decision rules in the form of binary trees and to enable classification of the cases [7]. This algorithm is able to generate clear structure that provides for easy interpretation of the rules which is its main advantage for credit scoring model.

ODM's generalized linear model uses logistic regression for classification. Logistic regression is a general linear model that models a binary outcome (0/1, good/bad, non-default/default etc.) on a certain number of predictors. It can be described with group of explanatory variables $X = \{X_i .. X_p\}$ and response variable of two categories, $Y = (y_1, y_2)$. This model can be represented by the following formula [7]:

$$\pi_i = \frac{\exp\{X_i\beta\}}{1-\exp\{X_i\beta\}}, \qquad (1)$$

where π_i is the probability of the i^{th} individual to belong to the certain category conditioned to X_i with β representing a vector that contains model's coefficients. Logistic regression is the most frequently used algorithm for credit scoring as it provides weighted linear sum of the attributes [8].

C. Data pre-processing

Various data pre-processing methods have been applied to the original dataset in order to produce improved version of it for our experiment. Firstly, we used data dimension reduction, as a method for reducing number of attributes for analysis by removing ones that are not relevant and important. High dimensionality of the dataset can make mined patterns more specific and less significant and at the same time make search space diverse. The consequence of such situation is dramatic increase of time consumption for the algorithm to find useful patterns [9]. After removing the attributes non-influential for the credit scoring, we decided to reduce dimension of the original data set by removing attributes type of client (all of the clients where of type F - individuals) and loans in other banks (all clients in the dataset had no loans in other banks or this information was incorrect). Also, the attribute amount of installment was removed due to redundancy, since it is highly related with attributes amount of loan and number of installments.

Next, we applied data aggregation as a technique for data summarization to solve a task of deriving the right level of data detail for Data Mining. It involves combining particular existing attributes into a new, single one to create more abstract dataset which will lead to a more compact dataset, better performance of the model and easier structure of the dataset for analysis [10]. We derived a new attribute *hh_incomes_real* by using the following formula: *hh_incomes_real = hh_incomes* - (*hh_costs +hh_amount_of_instalment*). The new attribute showed real amount of incomes in the household, and at the same time removed two unnecessary attributes from our data set. Previously described techniques directly affect data mining algorithm accuracy, since redundant or irrelevant attributes produce false results of algorithm applied.

When dealing with missing values, we identified that certain number of rows in the dataset (loans) have large number of attributes with null values such as: duration on address, duration of employment, source of income, type of contract, number of incomes in household, number of adults in household, number of non-adults in household, and number of pets in household. Since ODM offers two automatic methods for dealing with the problem of missing values: mean value and deleting rows, to avoid reducing number of rows for the training and thus having different size of dataset before and after pre-processing, we chose option mean value for missing values, while in pre-processed dataset with the

technique described previously we modified the original size of the dataset by replacing null values with meaningful data. We replaced null values of categorical attributes with the category "unknown", while numerical attributes with null values we replaced with the value of 0. This automatic option of ODM decreased our intent to improve importance of data pre-processing, but still, we were persistent in showing improvement by using other data pre-processing methods.

Next, we used equal interval binning, equal frequency binning and variable transformation as the methods for supervised discretization of our dataset. The main idea behind equal interval binning is to divide the domain of continuous attribute to a number of equal-length intervals (categories). The possible issue with this method is when category frequencies differ highly, and thus make the dataset skew towards certain categories. Therefore, on some attributes, we applied equal frequency binning also and compared results, as shown later in Results chapter. Variable transformation is a technique that transforms domain of an attribute to another for more efficient data usage and management. Furthermore, this method can make attribute values easier to compare [11]. The following table shows attributes from our dataset that were discretized:

TABLE II. DISCRETISED ATTRIBUTES

Name of Attribute	Attribute Domain	Description
QUALIFICATION	Categorical	Even though this attribute had registered value, many of these were duplicated or insufficient. We grouped then into four main groups: University degree, High School Education, Low education and Other. These are groups with approximate number of records belonging to the group.
DURATION_ON_ADDRESS	Numerical -> Categorical	We applied equal interval binning, but also equal frequency binning to categorize this attribute into four different groups.
SOURCE_OF_INCOME	Categorical	Same as with attribute qualification, we grouped these values into seven groups(Small private company, Big public company, Big private company etc.)
DURATION_OF_EMPLOYMENT	Numerical -> Categorical	Same as with attribute duration on address.
AMOUNT_OF_LOAN	Numerical -> Categorical	Same as with attribute duration on address.
NUMBER_OF_INSTALMENTS	Numerical -> Categorical	Same as with attribute duration on address.

IV. RESULTS

As mentioned earlier, the first credit scoring model was created based on a dataset without preprocessing where 60% of the dataset was used for training model and 40% for testing of it. Considering that many of the attributes with categorized, registry values were already discretized, even model trained on dataset without pre-processing has shown solid results. Predictive confidence for the decision tree algorithm was 45.6078% and for generalized linear model 39.7457%. Next figure shows model's results, including predictive confidence, average accuracy and overall accuracy.

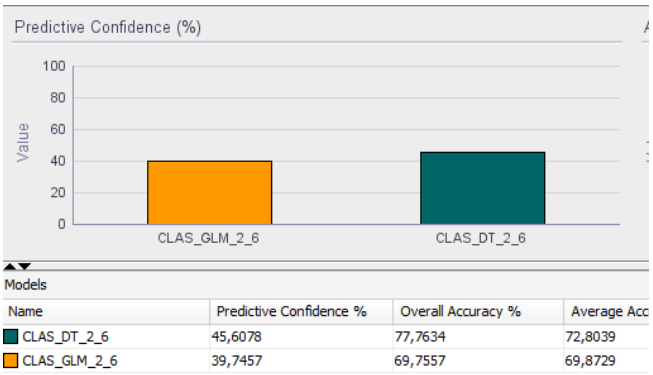

Figure 2. Results of classification credit scoring model before pre-processing

Next table represents performance matrix of DT credit scoring model built on original dataset before pre-processing.

TABLE III. PERFORMANCE MATRIX FOR DT MODEL BEFORE DATA PRE-PROCESSING

		PREDICTED			
		Non default	Default	Total	Correct (%)
ACTUAL	Non default	6046	1421	7467	80.9696
	Default	645	1179	1824	64.6382
	Total	6691	2600	9291	
	Correct (%)	90.3602	45.3462		

Out of 9291 records in the sample, model made 7225 right predictions, so the overall accuracy of the model before pre-processing is 77.7634%. Accordingly, error rate of the model before data pre-processing is 22.2366%.

TABLE IV. PERFORMANCE MATRIX FOR GLM MODEL BEFORE DATA PRE-PROCESSING

		PREDICTED			
		Non default	Default	Total	Correct (%)
ACTUAL	Non default	5203	2264	7467	70.2692
	Default	546	1278	1824	69.5175
	Total	5749	3542	9291	
	Correct (%)	90.5027	36.0813		

Performance matrix of GLM credit scoring model built on original dataset before pre-processing is shown in

previous table. Out of 9291 records in the sample, model made 6481 right predictions, so the overall accuracy of the model before pre-processing is 69.7557% and error rate is 30.2443%.

During pre-processing of the original dataset we applied pre-processing techniques described in previous chapter. Regarding discretization, two different techniques were applied: equal interval binning and binning with same frequency of values. The first technique showed better results for DT credit scoring model and the second one had better impact on results of GLM model. The following figures show results of applying equal interval and equal frequency binning on age attribute. The same techniques were applied on other discretized attributes.

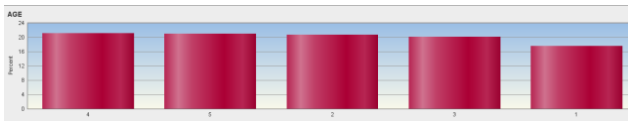

Figure 3. Histogram of attribute Age when equal frequency binning applied

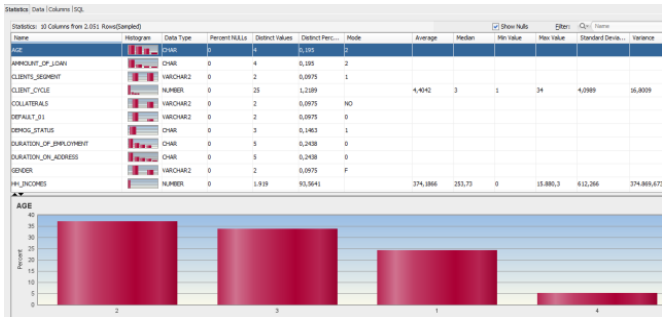

Figure 4. Histogram of attribute Age when equal interval binning applied

After applying described data pre-processing methods on original dataset, credit scoring model shown quite improvement for both of the trained algorithms. Next table compares actual data and predictions made with created DT and GLM credit scoring models before data pre-processing of dataset. Results of DT model are based on equal interval binning discretization and results of GLM model on equal frequency binning discretization.

TABLE V. COMPARISON OF PREDICTION BETWEEN DT AND GLM ALGORITHMS BEFORE DATA PRE-PROCESSING

	Measure	Before (%)	After (%)	Improvement (%)
DT algorithm	Predictive confidence	45.6078	62.5795	16.9717
	Overall accuracy	77.7634	91.3034	13.5400
	Average accuracy	72.8039	81.2892	8.4853
GLM algorithm	Predictive confidence	39.7457	43.2372	3.4915
	Overall accuracy	69.7557	72.2958	2.5401
	Average accuracy	69.8729	71.6189	1.7460

Thus, GLM model based on pre-processed dataset has shown 43.2372% of predictive confidence, 72.2958% overall accuracy and 71.6189% average accuracy, and DT model showed predictive confidence 62.5795%, overall accuracy 91.3034% and average accuracy 81.2892%. If compared with results of the credit scoring model based on the original set of data, we can see that data pre-processing improved results, but credit scoring model based on DT algorithm showed better improvement than the one based on GLM algorithm. It is excepted since discretized data enables DT to generate more precise rule sets. Performance matrix of this model is shown in the next table:

TABLE VI. PERFORMANCE MATRIX FOR DECISION TREE MODEL AFTER DATA PRE-PROCESSING

		PREDICTED			
		Non default	Default	Total	Correct (%)
ACTUAL	Non default	7301	166	7467	97.7769
	Default	642	1182	1824	64.8026
	Total	7943	1348	9291	
	Correct (%)	91.9174	87.6855		

As shown, out of 9291 approved loans, DT credit scoring model using pre-processed data correctly recognized 8484. That makes 1259 more correct prediction than DT model before pre-processing has predicted.

V. EVALUATION

Evaluation of results in this research is done by applying new real dataset on same, already built, credit scoring model. New dataset consisted of 21457 records and same set of attributes that were used for building the models. This dataset is applied on both credit scoring models, before and after pre-processing, but since it has shown greater improvement, only results given by DT algorithm are taken into consideration.

Next tables represent matrix of positive and negative classes for DT credit scoring model before and after data pre-processing.

TABLE VII. POSITIVE AND NEGATIVE CLASSES MATRIX FOR DECISION TREE MODEL BEFORE DATA PRE-PROCESSING

		PREDICTED		
		Non default	Default	Total
ACTUAL	Non default	13795	6472	20267
	Default	1001	189	1190
	Total	14796	6661	21457

TABLE VIII. POSITIVE AND NEGATIVE CLASSES MATRIX FOR DECISION TREE MODEL AFTER DATA PRE-PROCESSING

		PREDICTED		
		Non default	Default	Total
ACTUAL	Non default	18503	1764	20267
	Default	696	494	1190
	Total	19199	2258	21457

Model with applied data pre-processing techniques made 5013 correct predictions more than model without

application of these techniques. In relation with 21457 total predictions, it gives 23.3630% better predictions.

Out if 1190 loans that are really defaults, the first model recognized only 189 or 15.8823% loans as defaults and the second one recognized 41.5126% (494 loans). Therefore, the improvement of true positive rate is 25.6303%. As opposed to that, out of 20267 non-default loans, model built before data pre-processing, predicted 6472 as defaulters and the model built after pre-processing predicted only 1764 or 4708 mistakes less. According to that, false positive rate of first model is 31.9337% and of the second one is 8.7038% . This means improvement of 23.2299%.

When analyzing model predictions, the first model predicted 6661 loans as a default. Out of that number, only 189 loans were really defaulters, so the precision (positive) of this model is 2.8374%. The second model recognized 2258 loans as defaulters and 494 were really defaulters, so the positive precision of this model is 21.8778%.

Out of 21457 loans, credit scoring model before data pre-processing made 7473 wrong decisions and the one built after data pre-processing made 5013 wrong decisions less. So, the improvement of error rate is 23.3630%.

All mentioned measures of performance for both models are given in the table 9.

TABLE IX. MEASURES OF MODEL'S PERFORMANCES BEFORE AND AFTER PRE-PROCESSING

Measure	Before (%)	After (%)	Improvement (%)
Overall accuracy	65.1722	88.5352	23.3630
Average accuracy	41.9731	66.9403	24.9672
True positive rate	15.8823	41.5126	25.6303
False positive rate	31.9337	8.7038	23.2299
Precision(positive)	2.8374	21.8778	19.0404
Error rate	34.8278	11.4648	23.3630

Apart from improvement in accuracy, and in other showed measures, running time of applied algorithms has been reduced from 0.0833 minutes to 0.0467 minutes, which is also great improvement in algorithm execution time.

VI. CONCLUSION

This research showed impact of data pre-processing on results of credit scoring models based on data mining algorithms. Several data pre-processing techniques were applied on the real dataset and those are: dimension reduction, redundant attribute removing, data summarization (aggregation), dealing with missing values and supervised discretization (equal interval binning, equal frequency binning and variable transformation).

For the purpose of comparing results, two credit scoring models were built, one based on the data before data pre-processing and the other one based on data after pre-processing. Models were created using Oracle Data Miner application which uses different data mining algorithms for classification, so for both models decision tree and generalized linear algorithms were applied.

Even though the original dataset consisted of some already discretized data and the problem of dealing with missing values was automatically solved by Oracle Data Miner for both datasets, the improvement of the results for the model before and after applying pre-processing techniques is still notable. When validating both models improvement in overall, accuracy between model before and after data pre-processing is 23.3630% and improvement of positive precision is 19.0404%. All other measures of performances including running time also showed determined improvement, so we can conclude that detailed and comprehensive data preparation is of great importance for data mining approaches.

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] Du H., Data Mining Techniques and Applications, Cengage Learning EMEA, United Kingdom, 2010

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] F. Louzada, A. A. Guilherme, B. Fernandes, "Classification methods applied to credit scoring: Systematic review and overall comparison", February 2016. -8

[9] 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.

[10] 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.

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