Development of Credit Scoring Model to Determine the Creditworthiness of Borrowers

Amjad Ali[1], Muhammad Rafi[2]

[1,2]Shaheed Zulfikar Ali Bhutto Institute of Science and Technology (SZABIST) Karachi, Pakistan

[1]amjad1971@gmail.com

[2]rafi.muhammad@gmail.com

Abstract: The explosive growth of data in banking sector is common phenomena. It is due to early adaptation of information system by Banks. This vast volume of historical data related to financial position of individuals and organizations compel banks to evaluate credit worthiness of clients to offers new services. Credit scoring can be defined as a technique that facilitates lenders in deciding to grant or reject credit to consumers. A credit score is a product of advanced analytical models that catch a snapshot of the consumer credit history and translate it into a numeric number that signify the amount of risks that will be generated in a specific deal by the consumer. Automated Credit scoring mechanism has replaced onerous, error-prone labour-intensive manual reviews that were less transparent and lacks statistical-soundness in almost all financial organizations. The credit scoring functionality is a type of classification problem for the new customer. There are numerous data classification algorithm proposed and each one has its pros and cons. This independent study focuses on comparing three data classification algorithms namely: Naïve Bayes, Bayesian Network and Bagging, for credit scoring task. An extensive series of experiments are performed on three standard credit scoring datasets: (i) German credit dataset, (ii) Australian credit dataset and (iii) Pakistan credit dataset. One of the main contributions of this study is to introduced Pakistan credit dataset; it is collected from local credit repository, and transformed accordingly to be used in the study. The studies compare the experimental results of different selected algorithms for classification, their standard evaluation measures, performance on the three datasets, and conclude the major findings.

***Keywords*----Credit scoring, Data Mining, WEKA, Classification, Credit Dataset, Classification techniques.**

I. INTRODUCTION

Statistics and probability modeling in financial risk evaluation is now commonly used by financial institutions because of rapid growth in availability of consumer credit over the last few decades. Extensive quantitative analysis techniques are used by financial institutions for better decision making related to financial risks associated with consumer's loans. Financial risk evaluation is now one of the most significant functions of banks.

In the last few decades, adoption of various international standards such as Basel II, International Financial Reporting Standards, International Accounting Standards etc. deployment of risk assessment techniques through laws or regulations have become mandatory on financial institutions in various countries [1].

Adoption of Basel II standard by various central banks has forced banks to develop internal rating mechanism to measure credit and operational risk to determine the capital levels as fixed by their central banks. [2]. Financial institutions across the world have now developed and implemented various strategies for risk control and sensitivity analysis to ensure sustainability by reducing credit risk significantly [3].The compliance to various regulations issued by central banks under these international standards, financial institutions interest have increased manifold in novel and sophisticated data mining techniques.

The term credit score is used to classify the customer by using various statistical methods into estimable and decomposed classes. The prior class has high probability to repay the loan on time and later has the high probability to default on loan. Credit scoring systems facilitate financial institutions to accept or reject the credit request of a customer based on statistical techniques in place of human judgment and thus these systems can decrease the probability of defaults by prospective customers. Credit scoring involves collection of information on various attributes such as payment status and payment performance of a borrower over a particular period of time on a financial obligation. This information along with other attributes is used to calculate a credit score. This score provides the information about the likelihood that customer will default; however, it does not define when insolvency will occur [4]. Credit Scoring systems provide objective methods for decision making that are satisfactory alternative of using conventional method of human judgment [5].In financial industry, availability of information about consumer through different sources on one hand and advancement of data mining techniques on other hand such as linear discriminant analysis, logistic regression, expert system, genetic programming model, neural network and support vector machine has helped financial institutions to extensively deploy and use credit scoring systems. Despite extensive use of credit scoring for credit evaluation, credit scoring have many limitations such as it cannot account for the local economic conditions and cannot forecast the business cycles peaks and troughs [6]. These two factors play a key role in repayment of credit by consumers among many. Data patterns also changes with passage of time due to which credit scoring systems needs consistent updates to be relevant. In spite of these limitations, credit scoring systems are extensively used to determine the creditworthiness at the time of origination of loan [7].

Various classification techniques are now available for credit scoring and selection of proper technique is a challenging task because it can considerably improve the accuracy in credit scoring. However, selection of any such data mining technique cannot be considered and declared best and optimal technique because different classification techniques can also be integrated to enhance the effectiveness of credit scoring. Considering this challenge, selected three data mining techniques have been selected: Naïve Bayes, Bayesian Network, and Bagging for data analysis. The objective is to apply these data mining techniques on German and Australian credit datasets publically available on UCI machine learning repository [8] and Pakistani credit dataset taken from to local credit depository to determine which classification technique produce good results in terms of accuracy

II. RELATED WORK

In the past, several techniques have been used to construct the credit scoring models. It includes both statistical methods and machine learning techniques.

John Mylonakis tried to build the model on the basis of discriminant analysis for prospective credit card holders on the basis of information available on the existing credit card holders [9]. He indicated that banks are issuing credit cards to enhance their customer base without considering demographic characteristics that play key role in development of credit score. He concluded that discriminant analysis cannot only be used in this specific area of lending for developing credit scoring model.

Kočenda and Vojtek used logistic regression to build a credit scoring model to determine the most important financial and behavioral characteristics of default [10]. They identified that amount of loan, purpose of loan, marital status, level of education and number of years the customer have account with bank will be most relevant factors to determine the characteristics of client behavior in situations of economic changes and financial instability.

Data Mining is possible because of advances in computer sciences and machine learning techniques. These techniques have delivered new algorithms that can automatically sift deep into data at the individual record level to discover patterns, relationships, factors, clusters, associations, profiles, and predictions autonomously. Finding these patterns, relationships, factors, clusters, associations, profiles, and predictions from lager data is nearly impossible without data mining techniques.

Huang *et al* have used artificial intelligence technique: Support Vector Machine to construct the credit scoring model [11]. They concluded that credit scoring model built on Support Vector Machine provides the same level of classification accuracy that has been achieved by many researchers through back propagation neural network and genetic programming. Pacelli and Azzollini analyzed the ability of neural network to forecast the credit risk [12].They concluded that neural network will perform better when combined with traditional statistical methods to forecast credit risk. However, their investigation revealed that neural network can be used to construct the credit scoring model when dependent and independent variables display complex non-linear relationships. One of drawback of neural network as mentioned in their research paper is that neural networks based on supervised learning need extended training process to build the optimal network topology on one hand and on the other hand neural networks cannot be used to explain their decisions.

Zhang *et al* explored the vertical bagging decision tree model for credit scoring. They concluded that vertical bagging decision tree model is more robust and accurate in terms of classification accuracy when compared with the other models such as neural network and support vertical machine [13].

Individual credit scoring model have its own strengths and weaknesses where one model can perform better for one segment of data and another model perform better in another segment of data. Due to limitations of individual credit scoring system, hybrid systems have also been proposed where two models have been combined to procedure more accuracy rate.

Chen *et al* used clustering and support vector machine for building credit scoring model [14]. They used two stages in building model where in first stage (cluster stage), they grouped the samples into homogenous clusters, and then they identified the isolated samples for deletion and relabeled inconsistent samples. In second stage, they used support vector machine to construct the credit scoring model. Their research used more than two classes to classify the customers and their research confirmed that better classification accuracy can be achieved by choosing the proper cut off point.

Kumar and Rathee [15] have combined clustering and classification techniques for comparison of results with individual classification techniques. Their experiment revealed that combination of clustering and classification techniques produce better results than simple classification technique.

Abdou and Pointon [16] presented a comprehensive review of the 214 studies that have been carried out in development of credit scoring model. They presented the comparison of statistical techniques over the machine leaning techniques and concluded that machine learning techniques perform better to predict the customer as worthy or probably defaulters. The literature revealed that no optimal technique exist to develop a credit scoring model that will be best suited for all types of circumstances. The best technique depends on various factors such as data structure, details provided in loan application, choice of classification technique and segregation of classes depending upon the available features. They provided some interesting issues that have been ignored in development of credit scoring model and it includes the ranking of features in terms of importance in construction of credit scoring model, identification of behavioral perspective such as healthcare cost, rising education costs, mortgage payments etc. that play substantial role in default and identification of social and economic variables that will be suitable in changing economic conditions.

III. PROBLEM STATEMENT

Credit scoring is very important task to evaluate the credit applications of customers. This research will explore:-

1. What will be good method of credit scoring among the data mining methods: Naive Bayes, Bayesian Net and Bagging?
2. The decision will be based on the classification accuracy of each model.
3. Datasets used in this research are German credit dataset, Australian credit card approval dataset and Pakistani consumer credit dataset.

IV. DATASETS

Two datasets (German and Australian) are taken from UCI Machine Learning Repository and one dataset is taken from local credit depository mentioned in table 1. Main characteristics of datasets are:-

Table 1. Characteristics of credit scoring Datasets

Dataset	# Instances	# Classes (Good & Bad)	Nominal Attributes	Numerical Attributes	Good/Bad Credit %
German	1000	2	13	7	70/30
Australian	690	2	8	6	55.50/44.50
Pakistani	1706	2	6	3	75.85/24.15

V. EXPERIMENTAL METHODOLOGY

Experimental research has been used as a research methodology. Waikato Environment for Knowledge Analysis (WEKA) [17] has been used for performing classification on three data sets. For this research and model validation, data has been partitioned into training sample (70%) and test sample (30%). Before partitioning, data has been randomized using WEKA and then has been divided into training and test samples in ratio of 70/30 using percentage remove option of WEKA. Then, experiments were performed using classification techniques: Naïve Bayes, Bayesnet on both training and data samples, where local and global search under Hill Climbing, Repeated Hill Climbing and Simulated Annealing algorithms have been performed. The bagging technique was also used for experiments under which two algorithms J48 and NB Tree were used.

VI. EXPERIMENTAL RESULTS

In this section, results of various classification techniques applied on three datasets are presented. The most common measure to evaluate any classifier is Accuracy that can be expressed as under:-

$$\text{Accuracy} = \frac{TP+TN}{TP + TN+FP+FN} \qquad \text{(i)}$$

True positive (TP) are those credit records that has been accepted correctly and true negative (TN) are those records that have been correctly rejected. False positive (FP) are those credit rejected records that have been classified as accepted and false negative (FN) are those accepted credit records that have been classified as rejected.

Confusion matrix is also used to evaluate the classifier. In confusing matrix columns represent predictive class and rows represent the actual class.

Precision and recall are other functions that are used for true positives and false negatives:-

$$\text{Precision} = \frac{TP}{TP + FN} \qquad \text{(ii)}$$

$$\text{Recall} = \frac{TN}{FP+TN} \qquad \text{(iii)}$$

F measure combines precision and recall and it is harmonic mean of precision and recall

F-measure = 2 * ((Precision. Recall)/Precision+Recall)) (iv)

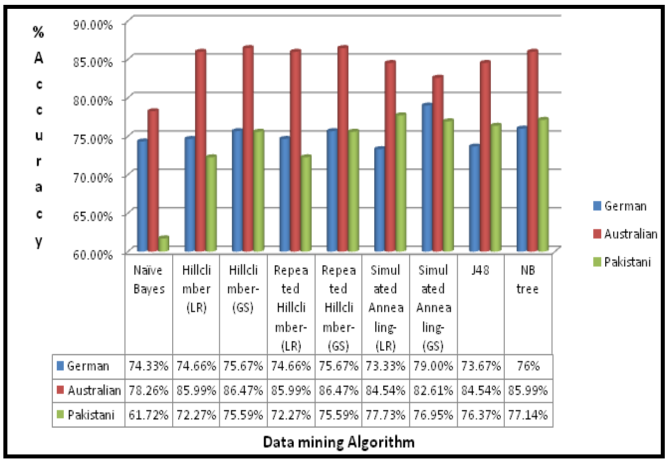
Table 2. Comparison of classification accuracy of different credit scoring techniques on test set

Data-set		Ger	Aus	Pak	Avg
Naïve Bayes		74.33%	78.26%	61.72%	71.44%
Bayes-net	Hillclimber - local Search	74.66%	85.99%	72.27%	77.64%
	Hillclimber -Global Search	75.67%	86.47%	75.59%	79.24%
	Repeaded Hillclimber -Local Search	74.66%	85.99%	72.27%	77.64%
	Repeaded Hillclimber -Global Search	75.67%	86.47%	75.59%	79.24%
	Simulated Annealing-Local Search	73.33%	84.54%	77.73%	78.53%
	Simulated Annealing-Global	79.00%	82.61%	76.95%	79.52%
Bagg-ing	J48	73.60%	84.54%	76.37%	78.17%
	NB Tree	76.00%	85.99%	77.14%	79.71%

IV. DISCUSSION

Summarized results of above experiments are mentioned below:-

According to table 2, the highest calcification accuracy of the German dataset is 79.00% that is achieved under the Simulated Annealing algorithm performed through Global Search where classification accuracy of Australian dataset is 86.47% achieved through Hillclimber algorithm performed through Global Search. Highest classification accuracy of Pakistani dataset is 77.73% achieved through Simulated Annealing algorithm performed through Local Search. All these algorithms belong to Bayesian Network and it can be concluded that Bayesian network have outperformed other classification techniques: Bagging and Naïve Bayes in terms of classification accuracy for three datasets. However, when precision (good credit), recall (good credit) and F score are considered, it can be seen that scores varies for each datasets.



	Naïve Bayes	Hillclimber (LR)	Hillclimber-(GS)	Repeated Hillclimber-(LR)	Repeated Hillclimber-(GS)	Simulated Annealing-(LR)	Simulated Annealing-(GS)	J48	NB tree
German	74.33%	74.66%	75.67%	74.66%	75.67%	73.33%	79.00%	73.67%	76%
Australian	78.26%	85.99%	86.47%	85.99%	86.47%	84.54%	82.61%	84.54%	85.99%
Pakistani	61.72%	72.27%	75.59%	72.27%	75.59%	77.73%	76.95%	76.37%	77.14%

Fig. (1). Correctly Classified Instances (%)-Credit scoring data sets

Table 3. Test set classification results on class attribute (Precision, recall and F scores) on credit scoring data sets

Techniques	German			Australian			Pakistani		
	Precision	Recall	F Score	Precision	Recall	F Score	Precision	Recall	F Score
Naïve Bayes	0.839	0.796	0.817	0.879	0.611	0.72	**0.864**	0.588	0.699
Hillclimber-LS	0.813	0.843	0.827	**0.884**	0.8	0.84	0.817	0.817	0.817
Hillclimber-GS	0.839	0.819	0.829	0.876	0.821	0.848	0.793	0.918	0.851
Repeated Hillclimber-LS	0.813	0.843	0.827	0.884	0.8	0.84	0.817	0.817	0.817
Repeated Hillclimber-GS	**0.839**	0.819	0.829	0.876	0.821	**0.848**	0.793	0.918	0.851
Simulated Annealing- LS	0.797	0.882	0.837	0.839	0.821	0.83	0.811	0.92	0.862
Simulated Annealing- GS	0.786	0.79	0.788	0.824	0.789	0.806	0.815	0.899	0.855
J48	0.807	0.833	0.82	0.806	**0.87**	0.838	0.797	**0.92**	0.855
NB Tree	0.795	**0.898**	**0.843**	0.859	0.832	0.845	0.816	0.902	**0.857**
Average	0.814	0.836	0.824	0.859	0.797	824	0.814	0.856	0.829

Table 4. Test set AUC on credit scoring data sets

Technique	German	Australian	Pakistani
Naïve Bayes	**.805**	.905	.72
Hillclimber-LS	.759	.909	.743
Hillclimber-GS	.794	**.922**	.755
Repeated Hillclimber-LS	.759	.909	.743
Repeated Hillclimber-GS	.794	.922	.755
Simulated Annealing- LS	.758	.913	**.791**
Simulated Annealing- GS	.799	.901	.765
J48	.764	.915	0.757
NB Tree	.795	.912	0.768
Average	0.781	0.912	0.771

Table 3 indicates that German dataset have highest scores in terms of precision under Bayesian network and recall and F score under Bagging, that are 0.839, 0.898 and 0.843. Australian dataset on other have highest precision and F score are under Bayesian network that are 0.884 and 0.848 and its highest recall score is achieved under J48 technique.

Table 5. Test set confusion matrices on credit scoring data sets

Technique	Desired Results	German		Australian		Pakistani	
		Output Results		Output Results		Output Results	
		Accepted	Rejected	Accepted	Rejected	Accepted	Rejected
Naïve Bayes	Accepted	172	44	58	37	228	160
	Rejected	33	51	8	104	36	88
Hillclimber-LS	Accepted	182	34	76	19	317	71
	Rejected	42	42	10	102	71	53
Hillclimber-GS	Accepted	177	39	78	17	356	32
	Rejected	34	50	11	101	93	31
Repeated Hillclimber-LS	Accepted	182	34	76	19	317	71
	Rejected	42	42	10	102	71	53
Repeated Hillclimber-GS	Accepted	177	39	78	17	356	32
	Rejected	34	50	11	101	93	31
Simulated Annealing- LS	Accepted	186	36	78	17	357	31
	Rejected	44	40	15	97	83	41
Simulated Annealing- GS	Accepted	187	29	75	20	349	39
	Rejected	34	50	16	96	79	45
J48	Accepted	180	36	**83**	**12**	**358**	**30**
	Rejected	43	41	**20**	**92**	**91**	**33**
NB Tree	Accepted	**194**	**22**	79	16	350	38
	Rejected	**50**	**34**	13	99	79	45

Similar behavior is observed for Pakistani dataset where highest precision score is achieved under Naïve Bayes that is 0.864 whereas recall and F score is achieved under Bagging technique that are 0.923 and 0.855.

Another classification measure is area under the ROC curve (AUC) that measures the discriminating ability of a binary classification model. The value of AUC indicate the probability of assigning positive cases being positive where high AUC value indicate the likelihood that an actual positive case have been assigned a higher probability of being positive than an actual negative case. mixed results on AUC for various datasets have been achieved. Table 4 indicate that AUC score for German dataset is 0.805 under Naïve Bayes, 0.922 for Australian dataset under Bayesian network and 0.791 for Pakistani dataset under Simulated Annealing under Local search. Confusion matrix is also used to visualize the performance the algorithm where columns represent the predictive class and rows represent actual class. Table 5 indicate the bagging technique have outperformed other techniques while predicting the good and bad classes.

V. CONCLUSION AND FUTURE WORK

In this study, three classification techniques have been used on three datasets to determine which classification technique is better in terms of accuracy. Credit scoring is now widely used by financial institutions for risk analysis for consumer loans. Therefore, development of credit scoring systems has gained serious attention of the researchers over the last few decades due to development in computing and machine learning techniques. The objective of this paper is to study different credit scoring techniques and their classification accuracy. Credit-scoring was performed on three datasets to evaluate classification accuracy of these techniques. Analytic results show that better classification accuracy rate was achieved through Bayesian network as compared to other techniques. This case study has many limitations that can be addressed in future work like other data mining techniques can be applied on three datasets to determine which technique is better for all three datasets, same set of attributes can be used for three datasets to determine the impact of attributes on classification accuracy.

VI. LIMITATIONS OF RESEARCH

Although credit scoring models have many benefits and are extensively used by financial institutions for classification of customer, these models have many limitations that need to be considered. One such limitation is that these models can be built on biased datasets where good customers represent the large portion of dataset and bad customer representation in dataset is relatively small. This model cannot be generalized and applied to actual large population. The second limitation is change of data patterns over the passage of time. Credit scoring model are built on the assumption that model have predictive capability to predict the future based on the past. This assumption is based on facts that characteristics used for classifying the customers as good or bad, credit can also be used for new applicant. However, there is high probability that characteristics will change over time and thus credit scorings models needs consistent update to be relevant. Another disadvantage of credit scoring model is that loan officers become too reliant on technology that they do not apply their personal judgment and experience in evaluation of those loan cases that needs prudent judgment. Further, credit scoring models have not been standardized and differ from market to market, needs extensive capital investment to deploy and train the loan officers. Credit scoring systems are criticized because they cannot incorporate economic characteristics like inflation and social characteristics (rising living cost) that also plays important role in repayment behavior of borrowers. Despite these limitations, credit scoring systems have many benefits for financial institution like (i) reduced cost of risk analysis (ii) automation of loan processing (iii) effective and efficient utilization of resources (iv) setting the competitive interest rate depending upon classification of customers (v) better risk management (vi) swift disposal of loan applications and increased objective analysis of loan applications

VII. ACKNOWLEDGEMENT

I would like to extend my appreciation to Mr. Muhammad Rafi for supervising this study and providing his technical knowledge, support and for meticulously reviewing this work. He provided extremely useful and critical suggestions in all matters of this study.

REFERENCES

[1] Credit Information Bureau. Reference: Available from: http://www.sbp.org.pk/ecib/

[2] W. R. Emmons, V. Lskavyan and T. J. Yeager, "Basel II Will Trickle Down to Community Bankers, Consumers" *The Regional Economist*, April 2005.

[3] D. Rösch "An Empirical Comparison of Default Risk Forecasts from Alternative Credit Rating Philosophies" *International Journal of Forecasting*, vol. 21, no. 1, pp: 37-51, 2005

[4] J. Morrison, "Introduction to Survival Analysis in Buisness", *The Journal of Business Forecasting*, 2003.

[5] J. D. Akhavein, W. S. Frame and L. J. White, and "The Diffusion of Financial Innovations: An Examination of the Adoption of Small Business Credit Scoring by Large Banking Organizations", *Journal of Business*, vol. 78, no. 2, March 2005

[6] R. B. Avery, P. S. Calem and G. B. Canner, "Consumer Credit Scoring: Do Situational Circumstances Matter" BIS Working Papers No. 146, January 2004.

[7] L. J. Mester, "What's the Point of Credit Scoring" *Buisness Review*, pp: 3-16, 1997.

[8] German and Australian credit datasets. Refernce: Available from: http://archive.ics.uci.edu/ml/datasets.html

[9] J. Mylonakis and G. Diacogiannis, "Evaluating the Likelihood of Using Linear Discriminant Analysis as A Commercial Bank Card OwnersCredit Scoring Model", *International Business Research,* vol. 3, no. 2, April 2010.

[10] E. Kočenda and M. Vojtek, "Default Predictors and Credit Scoring Models for Retail Banking", CESifo Working Paper Series No. 2862, December 2009.

[11] C-L. Huang, M-C. Chen and C-J. Wang, "Credit scorin with a data mining approach based on support vector machines", *Expert Systems with Applications*, vol. 33, no. 4, pp: 847–856, November 2007.

[12] V. Pacelli and M. Azzollini, "An Artificial Neural Network Approach for Credit Risk Management", *Journal of Intelligent Learning Systems and Applications*, vol. 3, no. 2, pp: 103-112, May 2011.

[13] D. Zhang, X. Zhou, S. C. H. Leung and J. Zheng, "Vertical bagging decision trees model for credit scoring", *Expert Systems with Applications*, vol. 37, no. 12, pp: 7838–7843, December 2010.

[14] W. Chen, G. Xiang, Y. Liu and K. Wang, "Credit risk Evaluation by hybrid data mining technique", *Systems Engineering Procedia*, vol. 3, pp: 194 -200, 2012.

[15] V. Kumar and N. Rathee, "Knowledge discovery from database Using an integration of clustering and classification", *International Journal of Advanced Computer Science and Applications*, vol. 2, no. 3, March 2011

[16] H. Abdou and J. Pointon, "Credit scoring, statistical techniques and evaluation criteria: a review of the literature", *Intelligent Systems in Accounting, Finance & Management*, vol. 18, no. 2-3, pp: 59-88, September 2011

[17] Weka 3: Data Mining Software in Java, Refeence: Available from: http://www.cs.waikato.ac.nz/ml/weka/