JCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

Predicting Credit Worthiness of Bank Customer with Machine Learning Over Cloud

A. Motwani [1*], P. Chaurasiya [2], G. Bajaj [3]

[1,2]Computer Science & Engineering, Sagar Institute of Science Technology & Research, RGPV, Bhopal, India
[3]Computer Science & Engineering, S.V. Polytechnic College, Bhopal, India

Corresponding Author: motwani.personal@gmail.com, Tel.: +91-8818965776

Available online at: www.ijcseonline.org

Accepted: 19/Jul/2018, Published: 31/Jul/2018

Abstract— Using Machine Learning (ML) and data analytics in the banking organizations is more than a trend and has become essential to keep up with the market competition and reduce credit risks. In recent years, Customer's Credit Worthiness is becoming more crucial for financial organizations. In past many credit risk models, that are actually statistical tools, are used to infer the future probabilities of customers to become default. At the same time with the massive increase in the volume, variety and velocity of data generated through various banking and business transactions pose a great computational and storage challenge for data analysis and intelligence tasks. To address the challenges for intelligence tasks Cloud Computing (CC) paradigm is evolved. The data and computation can be distributed to any CC environment with minimal effort nowadays. Also, CC paradigm turned out to be valuable alternatives to speed-up ML platforms. This paper aims to build and assess the performance of the 03 machine learning models, for prediction of credit card payment defaulter, over Microsoft Azure Machine Learning Platform. Finally a predictive analytics framework for classifying and predicting payment default by credit holder is proposed. For developing and testing the model a large, real and recent dataset of credit card, obtained from UCI repository, is used. The key focus of the work is on detection of Credit Worthiness which is defined as the 'probability of default' on the loan or credit from financial organizations like banks. The efficacy of model is demonstrated, on the basis of prediction accuracy and other metrics, against benchmark classifiers. Proposed work also demonstrates the use of Microsoft Azure cloud which is one of the foremost cloud environments for ML. The results attained by the proposed model are promising and the obtained results have potential to direct the future research work in domain.

Keywords—Artificial Intelligence, Predictive Modelling, Cognitive Computing, Computer Vision, Credit Risk, Credit Worthiness, Data Classification, Financial Organization, Machine Learning, Microsoft Azure, Cloud Computing

I. INTRODUCTION

Credit Worthiness (CW) is a valuation performed by lenders that determines the possibility a borrower may default on his debt obligations [1]. It is represented as a credit score by Financial Organizations. The creditworthiness of a company or individual is determined by using credit rating systems. A high credit score grants high CW. Payment history or credit history, health status and credit score depicts how a person meets debt obligations, which establishes credit worthiness or the financial character of a person. Payment history counts generally counts for 35% of a person's credit score [1].

Lending institutions also consider the amount of available assets and the amount of liabilities to determine the probability of a customer's default. In addition, it sees other factors such as age, health status, income, employment status, financial obligations, debt owed, accounts, length of payment history and the capability to repay debt. Banks also determines the interest rate, loan and other fees and fines, terms and conditions of a credit or loan on the basis of score. In case of unavailability of history of defaults some banks models for predicting credit risk with Moody's KMV [2]. In fact, banks have to take suitable actions to lessen credit risks to decrease costs as much as possible.

The need of cloud platforms for analysis and prediction from credit datasets can be seen as the shortcomings of present methods of credit scoring that runs over a single server and faces various problems including volume, speed, scalability and heterogeneity of data.

1. Massive increase in the volume of transaction data, generated through banks and other financial organizations worldwide, has been observed. The growing data Volumes and Velocity of generation is really a challenging and time consuming;

2. Classification and Regression tasks of ML can require a great deal of memory and processing power. So, speedy and scaled analysis is the key requirement for running ML tasks and cloud platforms are scalable to deal such requirements;
3. From the perspective of Data Analysis data from various sources are often unstructured and imbalanced. The Cloud platforms supports languages like Python, R etc. along with various statistical tools that enables development of new techniques to fix such datasets.
4. The Data Analysis and developing predictive models incurs too much cost in terms of infrastructure and computational tasks. Additional cost involves cost of learning and parameter optimization. CC paradigm eradicates the requirement of dedicated server at client end.

This model enables hiring of resources on demand and in cost-effective manner. Addressing the computational and storage needs for business intelligence tasks, Cloud Computing (CC) paradigm is evolved. Data and computation can be assigned to any CC environment with minimal effort.

A brief survey of: ML Models for Detection of Credit Worthiness and available Cloud Platforms for ML tasks are presented in section II. The proposed Cloud based Predictive Analytics Model over Microsoft Azure ML Platform for Detection of Credit Worthiness is presented in section III. Section IV includes introduction to Azure Machine Learning Studio, Dataset description and methodological steps for implementing the proposed work is discussed. The results that are evaluated and compared with 3 ML models, in terms of important performance metrics, are presented in section V. Section VI concludes the work and recommends the future work to be done regarding proposed predictive model / framework.

II. Related Work

A. Survey of Machine Learning Models for Detection of Credit Worthiness

Evaluation and Prediction of customer's Credit worthiness is a key for preventing losses for the banking sector. The survey is presented in Table 1.

Table 1. Survey of ML Models for Detection of Credit Worthiness

Authors	Contributions	Dataset used	Tool used
[9]	Experimented with 15 different ML methods. Proposed a model based on Linear Regression. Model used 3 Features.	Taiwan Bank Credit Card dataset [9]	Scikit-Learn [9] and MATLAB
[10]	Authors examined the use of incremental and batch classifiers. Feature selection is performed using ID3.	Malaysian Bank Credit card dataset	
[11]	Credit scoring and bankruptcy prediction is done using Ensemble [5] ML paradigm. Random Subspace (RS) ensemble method, with Neural Net classifier, performed better than other EMs.	Australian credit, German credit and Japanese credit dataset.	Different suitable toolkits
[12]	Performance comparison of ML techniques using three different decision tree classifiers: J48, Decision Tree and Random Tree, is done.	luxurious vehicle credit range' dataset	
[13]	Three Ensemble Methods namely Bagging [14, 15], AdaBoost [16] and Random Forest combined with three ML algorithm, to assess credit risk. Feature selection is	German credit dataset [9]	Weka toolkit [26]

	applied to select the important attributes.		
[18]	3 popular EMs, i.e., Bagging [14, 15], Boosting [16], and Stacking [17], based on 4 basic learners: Logistic Regression, Decision Tree (DT), ANN and SVM, are compared.	German credit and Australian credit dataset [9]	Weka toolkit [26]
[19]	Proposed a Model based on Bagging EM with REP Tree, for Credit risk prediction.	Taiwan Bank Credit Card dataset [9]	Weka toolkit [26]
[20]	Proposed a Model based on Bagging EM with REP Tree, for Credit risk prediction. Achieved accuracy of more than 81% with minimum number of features selected, i.e. 3.	Taiwan Bank Credit Card dataset [9]	Weka toolkit [26]
[21]	Proposed Neural Network based approach and compared with discriminate analysis.		

B. Survey of Cloud Platforms for Building Machine Learning Models

Amazon ML [22] provides visualization tools and wizards that guide users and practitioners through the process of building ML models without having to learn complex ML algorithms and technology. It is based on the highly proven and scalable ML technology. It allows obtaining predictions using simple APIs, once the model gets ready, without having to implement custom code or manage any infrastructure. Then, Amazon ML uses these models to process new data and generate predictions for user's application.

Google Cloud [23] Machine Learning Engine enables building of sophisticated, scalable ML models that cover a wide range of scenarios ranging from sophisticated regression models to image classification. It is fully managed, portable and works in integration with other Google products and Cloud Data platform. Once a trained model is obtained, model applies to obtain prediction for new instances. ML Engine offers two types of prediction: Online Prediction and Batch Prediction. These prediction models can be deploy over multiple frameworks. Cloud AutoML [24] at Google is a suite of ML products enables developers to train and build high quality models with limited ML expertise.

Azure ML Studio [25] is a simple and powerful browser-based, visual drag-and-drop based creative environment is briefly discussed in section IV.

III. PROPOSED WORK

A. Proposed Machine Learning Framework for Predicting Credit Worthiness over Azure Cloud

The model proposed for Credit Worthiness is presented in Figure 1. It employs the Neural Network Method with Filter based feature selection method for quicker and precise convergence of model. The feature selection method selects the attributes that are highly correlated to class. The utilization of filters at appropriate times for conversion to suitable forms and closer evaluation will definitely improve model building time. The methods pre-processing and filters contribute a lot in final evaluation of classifiers results also. The computation on numeric values of features and application of filters is an issue with stand-alone systems, but with scalability feature of cloud computing, it is not the issue. So, here the model is applied over all numeric values of features. Figure 1 showing screenshot of real experiment performed on Azure Machine Learning Workspace.

B. Model Phases

All four phases of proposed work are depicted in same figure. In First phase, data is uploaded to the cloud and brought into the workspace. After summarizing it is seen that the data contains lot of irregularities. So in this phase, data Pre-Processing with suitable filters that are necessary for Predictive modeling are applied. Here filter based feature selection is applied to get relevant features, i.e. the features having higher score with target class. Finally 04 relevant features are extracted.

In Second phase suitable split is applied for training and testing the model for good convergence of model. The best model for classifying the data is chosen in this phase by iteratively applying various models and scored the model with Test data on the basis of performance metrics. After the model becomes well tuned the classifier is finally scored on the basis of test split. Test dataset (50%) is supplied to this module. In this phase we applied 4 ML methods including proposed method. The model is built over following algorithms:

1. Bayes Point Machine
2. Logistic Regression
3. Decision Tree
4. Neural Network (NN) (Proposed)

Here, fully connected case of NN is used for proposed model. The fully connected network means that the hidden layer is fully connected to the input layer and output is fully connected to the hidden layer. The number of hidden layer here is one. Input refers to the number of input attributes and the number of nodes in the input layer is equal to the number of attributes in training data. The number of nodes in the hidden layer is set by the user. The default value is 100. The number of output nodes equals the number of classes. For a two-class neural network, this means that all inputs must map to one of two nodes in the output layer

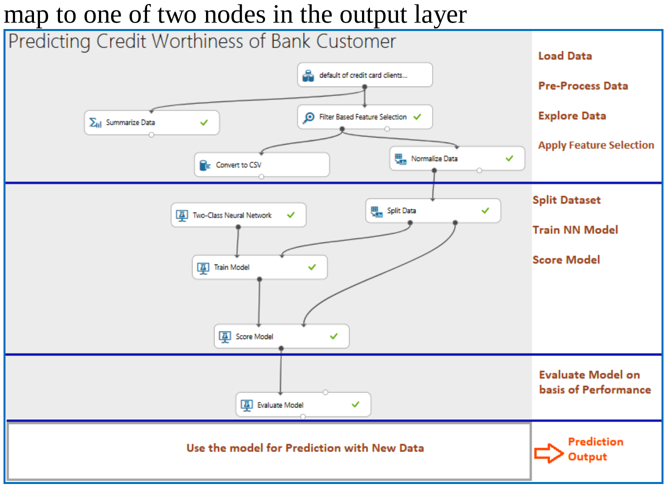

Figure 1. Proposed Predictive Model

In third phase, the predictive model is evaluated on basis of performance. The output of score model is given to 'Evaluate model' module. The classifier is evaluated on the basis of correct predictions that are performed. Utilize the "test" set predictions to calculate all the performance metrics (Measure Accuracy and other parameters).

The last phase the deployed model is used as a tool on new unclassified data. The model is tested with new data to get predictions. The overall correct response out of total test instances provided gives the accuracy of model. The evaluation is done on basis of few more metrics to check the model's viability.

IV. EXPERIMENTAL SETUP AND METHODOLOGY

A. *Microsoft Azure Machine Learning Studio (MAMLS)*

MAMLS provides ML Workspace with (a) ML studio, (b) ML Gallery and (c) ML Web Service Management. Azure ML studio is a graphical tool that is in use to organize and conduct the process of ML model building, testing and deployment of predictive analytics solutions. It includes: a collection of data pre-processing modules; a collection of best-in-class ML algorithms; An Azure ML API to deploy model as web application on Azure. ML Studio allows a user to import new datasets, pre-processing methods, ML algorithms and more onto its workspace. Figure 2, shows what those components are and where they fit into the ML process.

ML Studio lets a user drag and drop datasets, data pre-processing modules, ML algorithms and more onto its design surface. The user can connect these together graphically, and then execute the experiment. Once the model is built, a data scientist can run the experiment to evaluate the model created. When he has what he believes is the best possible model, he can use ML Studio to deploy this model to Microsoft Azure, where applications can use it. Rather than relying on ad hoc, largely manual mechanisms to do all of this, ML Studio provides a single tool for controlling the entire ML process. ML Studio publishes models as web services that can easily be consumed by custom apps or Business Intelligence tools such as Excel.

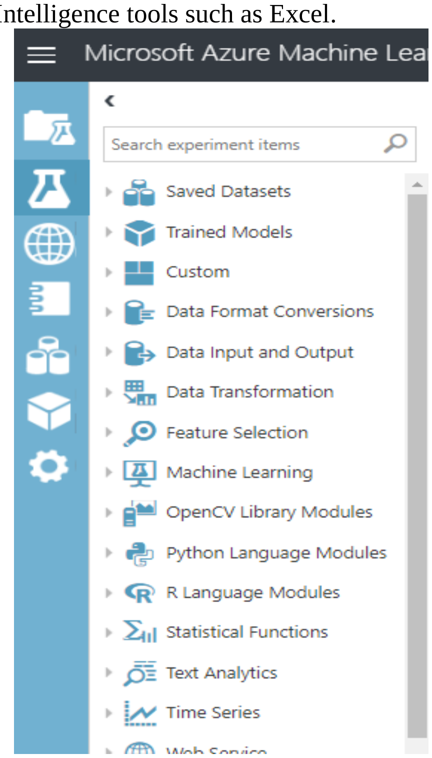

Figure 2. Azure Machine Learning Components

B. About Credit Dataset

For this research, the dataset available at UCI repository [27] is used. It contains details of payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. The dataset contains either the customers who paid or defaults. The total numbers of instances are 30,000. The No. of Instances (Yes) = 23364 (77.88%) and No. of Instances (No) = 6636 (22.12%).

C. Methodology of Implementing Proposed Model

The ML practitioner can sign up for Workspace using a Microsoft Account at the address given below:

https://studio.azureml.net/

Machine Learning Studio is where data science, predictive analytics, cloud resources, and our data meet.

The methodology of building Predictive Model (Classifier) is revealed here in Figure 3. Azure ML platform allows configuring several simulation parameters.

Figure 3. Methodology of Proposed Work

V. SIMULATION RESULTS AND ANALYSIS

A. Performance Metrics

The simulation runs are performed several times with different configurations to increase the performance i.e. with base classifiers, and then with proposed work. The results visualization can be done in various ways including confusion matrix. (Refer Figure 4 for results of benchmark classifier and Figure 5 for results of proposed Model. Then performance is evaluated on the basis of various performance metrics. The Metrics that is interesting to measure when studying classifiers / machine learning techniques are Accuracy, Recall, Prediction Rate, and False Alarm rate etc.

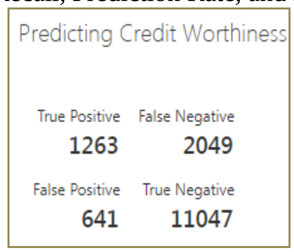

Figure 4. Confusion Matrix of Decision Tree Classifier

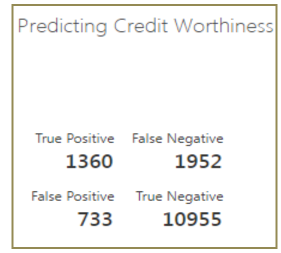

Figure 5. Confusion Matrix of Proposed NN Classifier

The parameters are presented below:

a) *Accuracy:* The accuracy of model is measured generally on basis of correctly classified instances. The comparison is depicted in Figure 6.

$$\text{Accuracy} = \frac{TP + TN}{No.of\ Instances} * 100$$

b) *True Positive:* It represents number of correctly identified instances from among the total number of correct instances. The comparison is depicted in Figure 7.

c) *Prediction rate:* Prediction rate refers to the percentage of correct predictions among all test data, and is defined as follows:

$$\text{Prediction Rate} = \frac{TP}{TP + TN} * 100$$

The comparison is depicted in Figure 8.

d) *Recall***:** It is also called Sensitivity. It is defined as number of positive cases that are correctly identified. The comparison is depicted in Figure 9.

$$\text{Recall} = \frac{TP}{TP + FN}$$

B. Comparison of Results

The results evaluated are mentioned in Table 2. The Graphs based on obtained results along with evaluation and analysis is presented below:

Table 2. Comparison of Results

Algorithm	Accuracy	True Positive	Recall	Prediction Rate
Bayes Point	80.4	568	0.171	0.047
Logistic Regression	80.89	682	0.206	0.056
Decision Tree	82.00	1263	0.381	0.103
Proposed Work	82.20	1360	0.411	0.110

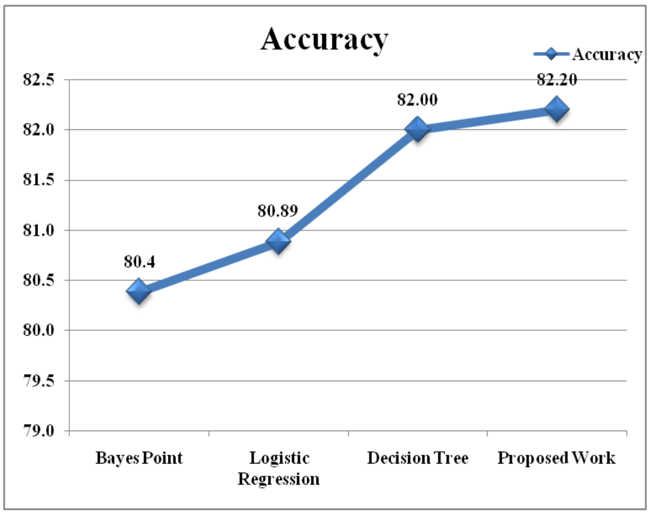

Figure 6. Comparison of Accuracy

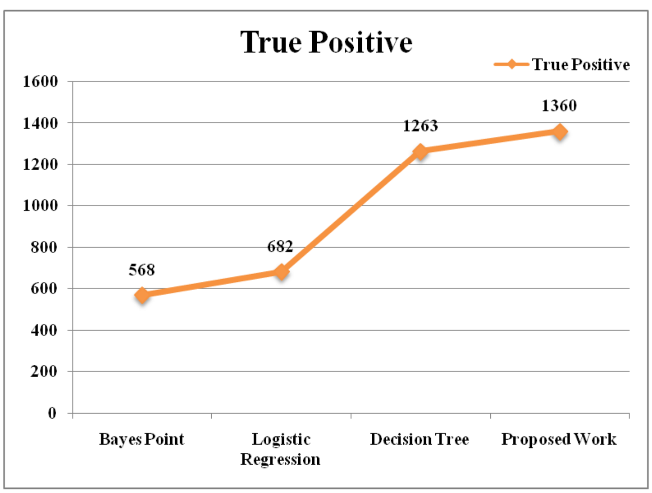

Figure 7. Comparison of True Positive

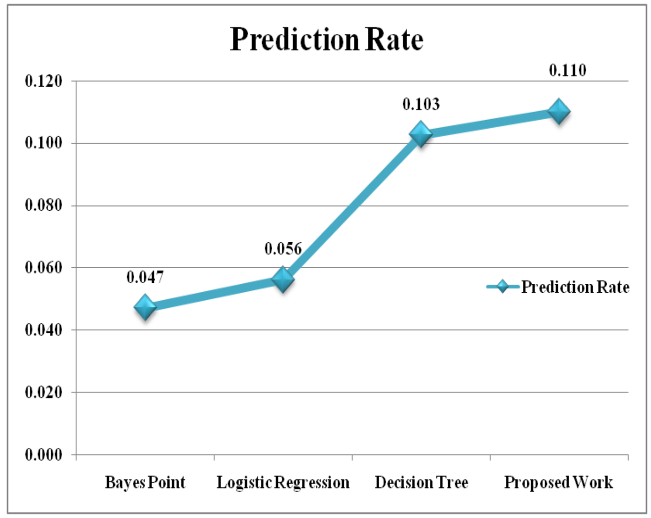

Figure 8. Comparison of Prediction Rate

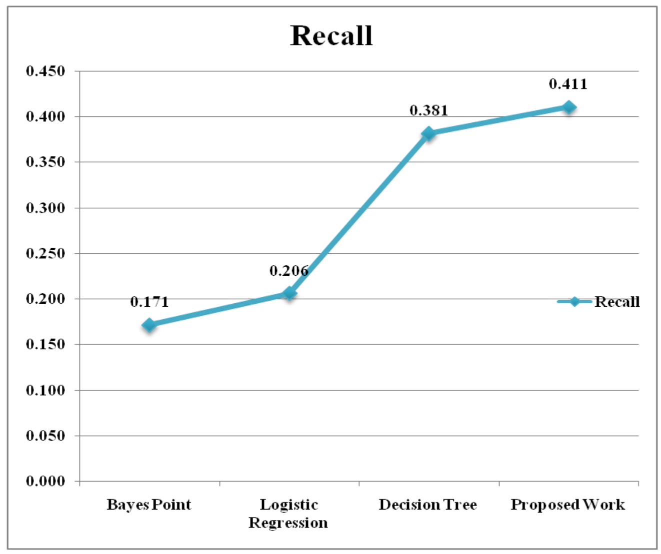

Figure 9. Comparison of Recall

The values of True Positive, Recall and Prediction rate are much higher than benchmark techniques.

VI. CONCLUSION AND FUTURE WORK

Banks and financial organizations are really facing the challenge of identifying risk factors, which should be considered while advancing the loans/credit to customers. So, Banks have to realize that ML potential can help them to focus their financial decisions, resource utilization and to make smarter decisions. The best model based on ML algorithm algorithms is chosen by data scientist to decide many aspects to generate more useful results.

Here in this work, Predictive modelling for detection of Credit Worthiness of Bank Customer is done. The model is based on NN model of ML. The model is completely built and tested over 'Microsoft Azure Cloud' platform. "Microsoft Azure Cloud" is a collection of integrated cloud services like:

analytics; database; storage; mobile computing; networking; and web. Building a Predictive model with ML for credit worthiness detection on existent cloud platform (Azure) and measuring its effectiveness will be the chief focus of this work. CC platform is used to avoid the computational limitations. The performance of proposed model is shown by comparing its efficacy against three other popular ML algorithms. The learners are compared on basis of various parameters and the proposed model found to be better in terms of prediction rate and other parameters. The findings of work have a lot of implications. In future, we intend to build up an automated risk assessment system over cloud for financial organizations that will incorporate key features to determine credit worthiness of customers.

REFERENCES

[1] https://www.investopedia.com/terms/c/credit-worthiness.asp
[2] https://www.moodys.com/sites/products/ProductAttachments/CreditMonitor_brochure.pdf
[3] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "From Data Mining to Knowledge Discovery in Databases",American Association for Artificial Intelligence. All rights reserved. 0738-4602-1996.
[4] Sebastiaan Tesink," Improving Intrusion Detection Systems through Machine Learning" Tilburg University March 2007.
[5] K. Tumer and N. C. Oza, "Decimated input ensembles for improved generalization," in Proceedings of the International Joint Conference on Neural Networks (IJCNN '99), pp. 3069–3074, Washington, DC, USA, July 1999.
[6] Belady C. "In the data center, power and cooling costs more than the it equipment it supports" 2007. URL: http://www.electronics-cooling.com/articles/2007/feb/a3/.
[7] National Institute of Standards and Technology (2011) NIST cloud computing reference architecture: Version 1. NIST Meeting Report
[8] P. Mell, T. Grance, "The NIST Definition of Cloud Computing, National Institute of Standards and Technology", ver. 15, 9 July 2010.
[9] Regina Esi Turkson, Edward Yeallakuor Baagyere, Gideon Evans Wenya, "A Machine Learning Approach for Predicting Bank Credit Worthiness", ISBN: 978-1-4673-9187-0, IEEE 2016
[10] Ling Kock Sheng, and Teh Ying Wah, "A comparative study of data mining techniques in predicting consumers' credit card risk in banks", African Journal of Business Management Vol. 5 (20), pp. 8307-8312, 16 September, 2011
[11] Loris Nanni , Alessandra Lumini, "An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring", Elsevier, Expert Systems with Applications 36 (2009) 3028–3033.
[12] U Bhuvaneswari, P. James Daniel Paul, Siddhant Sahu, "Financial Risk Modelling in Vehicle Credit Portfolio", 978-1-4799-4674-7/14/$31.00, IEEE 2014
[13] C.R.Durga devi, Dr.R.Manicka chezian, "A Relative Evaluation of the Performance of Ensemble Learning in Credit Scoring", IEEE International Conference on Advances in Computer Applications (ICACA), 978-1-5090-3770-4/16, 2016
[14] David Opitz and Richard Maclin, "Popular Ensemble Methods: An Empirical Study", Journal of artificial intelligence research 169-198, 1999.
[15] L. Breiman, "Bagging predictors," Machine Learning, vol. 24, no. 2, pp. 123–140, 1996.
[16] Freund, Y., & Schapire, R. (1996), "Experiments with a new boosting algorithm", in Proceedings of the thirteenth international conference on machine learning, Bari, Italy, pp. 148–156.
[17] Wolpert, D. H. (1992). Stacked generalization, Neural Networks, 5(2), 241–259.
[18] Gang Wang, Jinxing Hao, Jian Mab, Hongbing Jiang, "A comparative assessment of ensemble learning for credit scoring", Expert Systems with Applications 38 (2011) 223–230.
[19] K. Tumer and N. C. Oza, "Decimated input ensembles for improved generalization," in Proceedings of the International Joint Conference on Neural Networks (IJCNN '99), pp. 3069–3074, Washington, DC, USA, July 1999.
[20] Anand Motwani, Goldi Bajaj, Sushila Mohane, "Predictive Modelling for Credit Risk Detection using Ensemble Method", International Journal of Computer Sciences and Engineering, Vol.6, Issue.6, pp.863-867, 2018.
[21] Sihem Khemakhem ,and Younés Boujelbènea, "Credit risk prediction: A comparative study between discriminant analysis and the neural network approach", Accounting and Management Information Systems Vol. 14, No. 1, pp. 60-78, 2015
[22] https://aws.amazon.com/aml/
[23] https://cloud.google.com/products/machine-learning/
[24] https://cloud.google.com/ml-engine/
[25] https://azure.microsoft.com/en-in/services/machine-learning-studio/
[26] Weka, University of Waikato, Hamilton, New Zealand., http://www.cs.waikato.ac.nz/~ml/weka/index.html
[27] http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

Authors Profile

Prof. Anand Motwani pursuing his Ph.D. in Computer Science and Engineering. He is currently an Associate Professor and Head at Sagar Group of Institutions (SISTec-R), Bhopal, India. He is member of IEEE-'Cloud Computing', IEEE-'Internet of Things' Society and ORCID. He is a brilliant academician and researcher who brought with him the combination of both industry as well as rich academic experience. He has over a decade of experience with renowned educational institutions and industries and attended several national level conferences, workshops, seminars and FDPs. He has published an engineering book with Pearson Education and also published and presented several papers in quality conferences and journals. His domain of expertise includes: Software Engineering, Cloud Computing, Data Analytics, Machine Learning and Wireless Networking.