INTERNATIONAL RESEARCH JOURNAL OF ENGINEERING & APPLIED SCIENCES
ISSN: 2322-0821(O)
ISSN: 2394-9910(P)
VOLUME 8 ISSUE 4
Oct 2020 - Dec 2020
www.irjeas.org

CREDIT THREAT ESTIMATION BY MACHINE LEARNING TECHNIQUES OVER CLOUD PLATFORM

Upendra Kumar Kachhwaha[1], Anurag Shrivastava[1]

[1]M.Tech Scholar, Department Of Computer Science Engineering, NIRT, Bhopal, India
[2]Asst. Professor, Department Of Computer Science Engineering, NIRT, Bhopal, India

Abstract - Banks and financial organizations are really facing the challenge of identifying Threat factors, which should be considered while advancing the loans/credit to customers. An Ensemble ML algorithm is suitable for studying bank credit dataset. In future, we intend to build up a ML system risk automated system over cloud for financial organizations that will incorporate key features to determine credit value of customers.

Most methods for credit threat detection require previous data to build and validate models. Applying ML algorithms for credit threat determination and building prediction model is facing a major problem of data incompleteness. Most of the financial organizations do not share their information with other organizations, so determining credibility of customer is difficult. Another major issue faced by researchers in building model for Threat detection is the presence of noise in the data.

For this work, several ML techniques are explored and evaluated on real credit card datasets. Most ML methods have achieved an accuracy of less than 81 %. Finally, a Predictive model for Credit Threat Detection is proposed which is based on ensemble technique. The proposed model is evaluated on basis of various performance metrics and comparison is done with base classifier (learner). It gave around 82 % prediction accuracy.

Keywords – Cloud, Credit threats, Machine Learning

I. INTRODUCTION

The data landscape has changed over the years. What you can or should do with data has changed. Storage costs have dropped dramatically as data collection continues to grow. Some data arrive quickly and constantly require collection and observation. Other data arrives more slowly, but in very large blocks, often in the form of decades of historical data. There may be a problem with advanced analytics, or it might require machine learning. Credit value is represented as a credit score by Financial Organizations. A high credit score grants high credit value. 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. Credit worthiness also impacts eligibility for employment, insurance, business funding and professional licenses or certifications. Credit value is the evaluation examined by lenders that find the possibility a borrower may default on his debt obligations. [1] Supervising the liquidity risk and credit risk is one of the main issues of Bank Risk Management. "Liquidity Risk" is the risk in the lack of marketability of an investment, when the underlining asset cannot be bought or sold quickly enough for prevention or mitigation of a loss. "Credit Risk" is the main risk of holding a bond. [2]

Management models including 'Probability of Default', Expected Loss etc. Here, in this work we find 'Probability of Default' of credit card using Predictive model which is based on neural network technique. The aspects of credit threat management as shown in Figure 1.1.

Figure 1.1: Credit Threat Management – Aspects

II. LITERATURE REVIEW

The authors [1] explained that Recursive Feature Elimination with Cross-Validation and Principal component Analysis have been used for dimensionality reduction. Metrics such as F1 score, AUC score, prediction accuracy, precision and recall have been used to evaluate each model. Among all the models, the combination of a tuned Support Vector Machine (SVM) and Recursive Feature Elimination (RFE) with Cross-Validation have shown great promise in identifying loan defaulters. The support vector machines can outperform

other tree-based models or regression models if the setup of the experiment is similar to that of ours and recursive feature elimination with cross-validation can outperform models based on principal component analysis. For future improvements we would like to use more current data and from different sources for illustrating a better understanding of the trends present in this field.

Authors [12] compares the long list of Fin-Techs, one of the most attractive platforms is the Peer-to-Peer (P2P) ending which aims to bring the investors and borrowers hand in hand, leaving out the traditional intermediaries like banks. This paper investigates the machine learning techniques on big data platforms, analysing the credit scoring methods. It is concluded that on a HDFS (Hadoop Distributed File System) environment, Logistic Regression performs better than Decision Tree and Random Forest for credit scoring and classification considering performance metrics such as accuracy, precision and recall, and the overall run time of algorithms.

Among the three methods, Logistic Regression has the best accuracy, precision and recall, compared to Decision Tree and Random Forest. Considering the general belief that Logistic Regression and Random Forest are the most accepted and used methods for credit scoring, we saw that this is also true for HDFS. Both Logistic Regression and Random Forest have better results than Decision Tree. According to accuracy and precision of models, runs with more Data Nodes have performed better than others while the non-HDFS has performed almost as good as a three-node configuration.

Authors [6] The Threat assessment method consist identification and rank. The power market settlement Threats by Threat identification include data Threat, credit Threat, tax Threat and policy Threat, and the Threat ranking is carried out by using triangular fuzzy numbers to determine the influence degree of the four Threats on the settlement of power market

A method based on triangular fuzzy numbers for Threat factor ranking is used in this paper. This method can effectively compare the severity of settlement Threat factors in power market and analysis results can provide a basis for more accurately and effectively selecting the corresponding control measures in power market.

The case analysis in the paper shows that when most of the language fuzzification descriptions are selected during analysis, the values of the four power market settlement Threats calculated by the triangular fuzzy numbers are not much different, which indicates data Threat, credit Threat, tax Threat and policy Threats significantly affect all the security of power market settlements.

LI Changjian et al. [3] the purpose of this paper is to evaluate credit Threat for the rural credit cooperatives using artificial neural network model. We establish credit Threat assessment index system for rural credit cooperatives. Then, a kind of credit Threat assessment model based on particle swarm optimized neural network is put forward. Using neural network technology to identify the credit Threat can achieve very high accuracy rate and overcome the credit of many uncertain factors. The model can provide scientific reference to the rural credit cooperatives credit policy and credit Threat management.

R.S.Ramya et al. [5] Information gain measure identifies the entropy value of each specific feature. The amount of information gain or entropy is used to decide whether the feature is selected or deleted. Gain ratio applies normalization technique to information gain using spilt information value. The correlation based feature selection uses heuristic search strategies to estimate how the features are correlated with the class attribute and how they are important of each other. The feature selection techniques such as information gain, gain ratio, chi square correlation were applied to the German credit dataset available in UCI Machine Learning Repository. These feature selection techniques selected the features that will be useful for classification of clients and the ones that are irrelevant and redundant are omitted. The performance, robustness and usefulness of data mining algorithms are improved when relatively few and relevant features are involved in the process.

III. METHODOLOGY

The methodology of proposed work is explained with the help of Figure 3.1 showing screenshot of real experiment performed on Azure Machine Learning Workspace. Azure ML platform allows configuring several simulation parameters.

The methodology of building Predictive Model (Classifier) is revealed here in Figure 3.1.

International Research Journal of Engineering & Applied Sciences, IRJEAS
www.irjeas.org, ISSN(O): 2322-0821,Volume 8 Issue 4, Oct-Dec 2020, Page 11-16

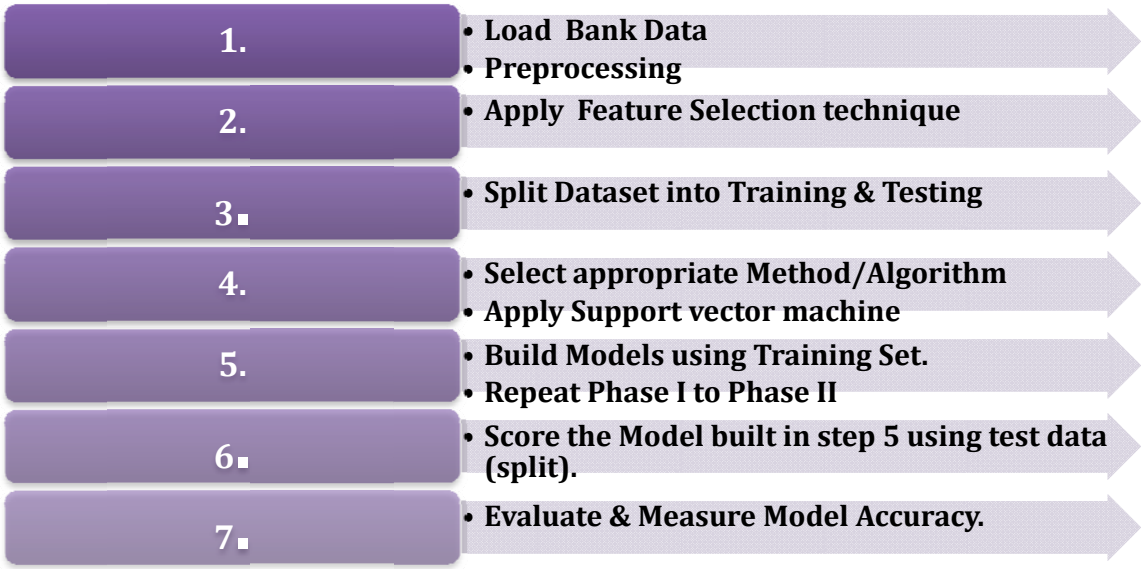

Figure 3.1: Methodology of Proposed Work

In the proposed model, the Cloud Computing platform is used to avoid the computation limitations. Some filters are applied for better prediction accuracy, faster evaluation. Application of proposed framework provides better data classification, better predictive accuracy than some benchmark classifiers. The implementation of predictive classifier (model) is done over Microsoft Azure ML studio.

In Second phase suitable split is applied for training and testing the modelfor 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. In this phase we applied three ML methods including proposed method. The phases of model are depicted in Figure 3.1. The model is built over following algorithms:

1. Bayes Point Machine

2. Logistic Regression

3. Deep support vector machine

IV. RESULTS

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. It includes: a collection of data pre-processing modules; a collection of ML algorithms; An Azure ML API to deploy model as application on Azure. ML Studio allows a user to import new datasets, pre-processing methods, ML algorithms and more onto its workspace.

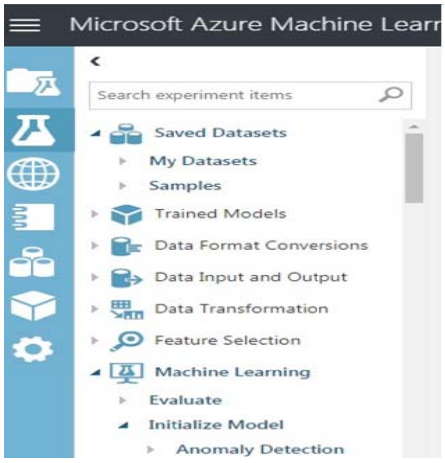

Figure 4.1: Azure Machine Learning Components

As the Figure 4.1and 4.2 suggest, Machine Learning Studio lets a user drag and drop datasets, data cleaning, pre-processing modules, machine learning algorithms and more onto its workspace. The user can connect these together, and then execute the experiment. Once the model is builta user can run the experiment to evaluate the model created. User can use ML Studio to deploy this model to Microsoft Azure, where applications can use it. ML Studio provides a single tool for controlling the entire machine learning process.

International Research Journal of Engineering & Applied Sciences, IRJEAS
www.irjeas.org, ISSN(O): 2322-0821,Volume 8 Issue 4, Oct-Dec 2020, Page 11-16

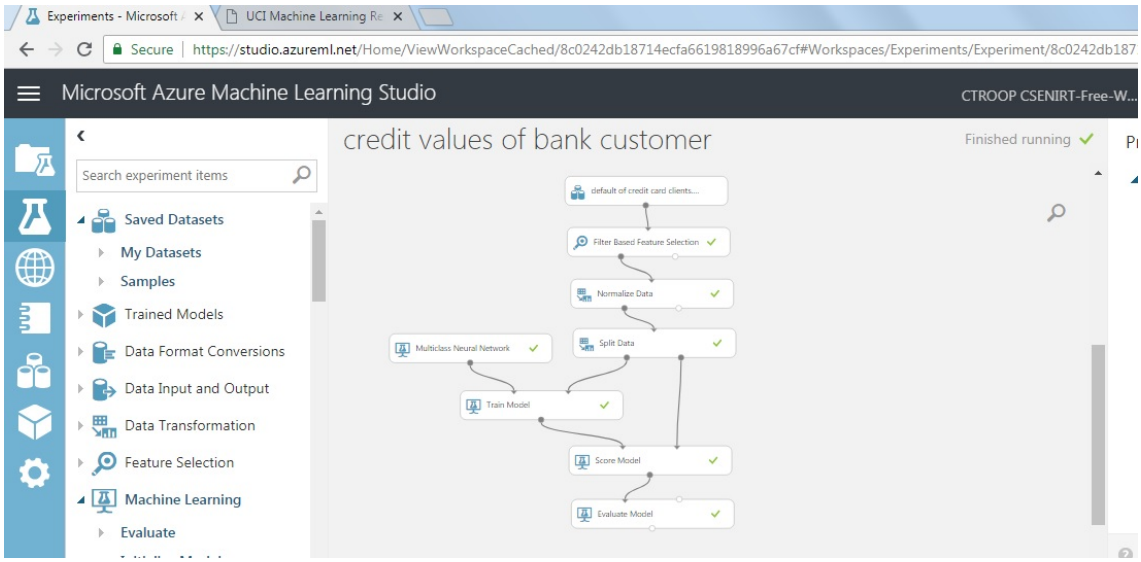

Figure 4.2: Models on Machine Learning Studio (Microsoft Azure)

V. SIMULATION RESULTS AND ANALYSIS

In the past, many works in this field are proposed. The frameworks or classifiers proposed for these tasks are based on well-known DM techniques and Machine Learning algorithms and giving good accuracy rates. Although the accuracy is increased by 1% in this work and other metrics like correlation coefficient and MAE is improved. Also, the model building time is also less for our work as compared to base model.

The parameters along with result evaluation and analysis are presented below:

Accuracy: The accuracy of model is measured generally on basis of correctly classified instances. The comparison is depicted in Figure 5.1.

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

True Positive: It represents number of correctly identified instances from among the total number of correct instances.

Recall: It is also called Sensitivity. It is defined as number of positive cases that are correctly identified.

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

The results mentioned here are as per the simulation scenario shown in the section 4.6. The results are shown in Table 5.1 for graphical representational of results. For evaluation the results of Base Learners and Proposed model are compared (refer Table 5.1). The Accuracy and Model Building time is also improved in comparison to base learners.

Table 4.3: Comparison of Results

Comparison Of Result							
Algorithms	**Accuracy**	**True Positive**	**FN**	**FP**	**TN**	**Precison**	**Recall**
Two Class Bayes	80.80%	659	2653	232	11456	0.74	0.199
Two Class Logistic Regression	81.20%	760	2552	271	11417	0.737	0.229
Praposed Work	82.20%	1291	2021	645	11043	0.667	0.39

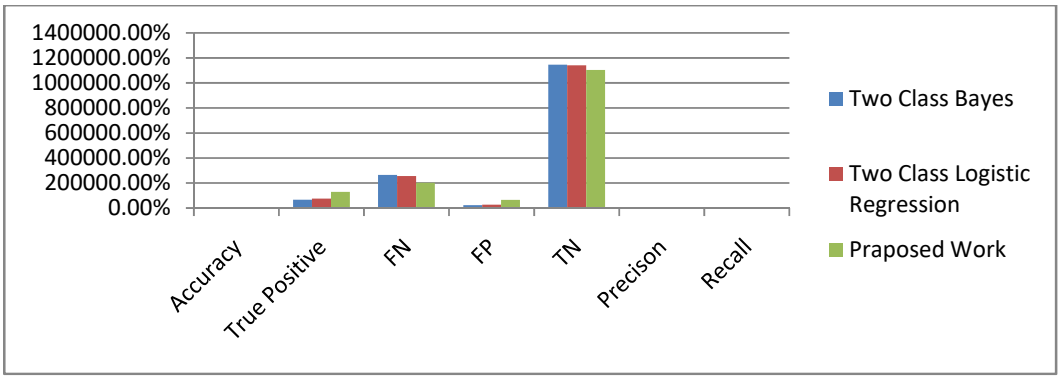

Figure 5.1: Comparison of Result

REFERENCES

[1] Syed Zamil Hasan Shoumo et al., "Application of Machine Learning in Credit Threat Assessment: A Prelude to Smart Banking" 2019 IEEE Region 10 Conference (TENCON 2019), 6/19/2019 IEEE, ISSN 978-1-7281-1895, 2022- 2028.

[2] Yavuz Selim Hindistan et al., "Alternative Credit Scoring and Classification Employing Machine Learning Techniques on a Big Data Platform" (UBMK'19) 4rd International Conference on Computer Science and Engineering – ISSN 978-1-7281-3974, 731-734.

[3] Dianning Wu et al., "Analysis and Evaluation of Settlement Threat in Power Market based on Triangular Fuzzy Number"2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM), 978-1-7281-4691-1/19, 13-17.

[4] LI Changjian, HU Peng, "Credit Threat Assessment for Rural Credit Cooperatives based on Improved Neural Network", 2017 International Conference on Smart Grid and Electrical Automation, ISSN 978-1-5386-2813-3/17, DOI 10.1109, 227-230.

[5] R.S.Ramya, Prof. S.Kumaresan, "Analysis Of Feature Selection Techniques In Credit Threat Assessment",2015 International Conference on Advanced Computing and Communication Systems (ICACCS -2015), Jan.05 – 07,2015, ISSN 978-1-4799-6438-3/15.

[6] H. Zhou *et al.*, "Big Data Mining Approach of PSO-Based BP Neural Network for Financial Threat Management With IoT" VOLUME 7, 2019, DOI 10-1109-2948949,154035-154043.

[7] Li Jinjuan , "Research on Enterprise Credit Threat Assessment Method Based on Improved Genetic Algorithm", 2017 9th International Conference on Measuring Technology and Mechatronics Automation, 2157-1481/17,213-218.

[8] Hui Sun, Mingyuan Guo, "Credit Threat Assessment Model of Small and Medium-Sized Enterprise Based on Logistic Regression" 2015 IEEE IEEM, 978-1-4673-8066-9/15, 1714-1718.

[9] Yun Lin, Yuan Zhang, "Credit Threat Assessment based on Neural Network"2012 8th International Conference on Natural Computation (ICNC 2012), 978-1-4577-2133-5/10, 402-405.

[10] Yun Lin, Yuan Zhang, "Credit Threat Assessment based on Neural Network", 2012 8th International Conference on Natural Computation (ICNC 2012), 978-1-4577-2133-5/10, 402-404.

[11] Wei Sun, Qiu-Shi Du, Bo Cui, "The Model Of Credit Threat Assessment In Power Industry Base On Rs-Svm" Proceedings of the Ninth International Conference on Machine Learning and Cybernetics, Qingdao, 11-14 July 2010, 978-1-4244-6527-9/10, 2021-2024.

[12] Xincun Wang, Jinpeng Huang, "Credit Threat Assessment Based on Fuzzy Comprehensive Assessment", sixth International Conference on Fuzzy Systems and Knowledge Discovery 2009, 978-0-7695-3735-1/09, 603-607.

[13] Bhekisipho Twala, "Multiple classifier application to credit threat assessment", Expert Systems with Applications 37 (2010) 3326–3336.

[14] Polikar, R. (2006). Ensemble based systems in decision making. IEEE Circuits and Systems Magazine, 6(3), 21–45.

[15] Zhou, Z. H. (2009). Ensemble. In L. Liu & T. Özsu (Eds.), Encyclopedia of database systems. Berlin: Springer.

[16] 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.

[17] 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.

[18] Sebastiaan Tesink," Improving Intrusion Detection Systems through Machine Learning" Tilburg University March 2007.
[19] Zhang, "Z., Research of credit threat of commercial bank's personal loan based on CHAID decision tree", Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011.
[20] T. G. Dietterich, "Machine-learning research: four current directions," AI Magazine, vol. 18, no. 4, pp. 97–136, 1997