International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 11 | Nov 2021 www.irjet.net p-ISSN: 2395-0072

Machine Learning for Microfinance Institutions – A Review

Komal Bakshi[1]

[1]*B. Tech Graduate, Computer Science and Engineering, Vellore Institute of Technology, Vellore, India*

Abstract - *Microfinance is a category of financial service that caters to small businesses, individuals with low-income or rural population that lacks access to traditional models of banking services. Machine learning (ML) algorithms have seen a wide-range of applications across the non-microfinance banking environments. However, adoption of machine learning in microfinance is currently not at par with the traditional banking sectors resulting in financial exclusion. The of lack recorded credit histories and the relevant data points present a different set of challenges for application of machine learning in microfinance environment. Tree-based algorithms such as decision trees, random forest and approaches such as fuzzy logic help in tackling the information asymmetry in this sector. This paper conducts a survey on research done on machine learning algorithms for microfinance institutions across various countries.*

***Key Words*: Machine Learning, Micro-Credit, Microfinance, Decision Trees, Fuzzy Logic, Artificial Neural Networks**

1.INTRODUCTION

Most of the financial institutions use historical data of customers, previous borrowing habits and so on for credit scoring. Farmers, women, low-income individuals tend to not have documented credit histories, making it tough for them to access to financial products offered by banking institutions. Microfinance Institutions tackle this divide by enabling the availability of financial services to these subsets of population. Artificial Intelligence (AI) and Machine Learning (ML) based technologies can be instrumental in helping out Microfinance Institutions (MFIs) for assessing credit scores. Machine Learning Systems can help in developing a comprehensive profile of the borrower's current level on income, ability to repay and employment opportunities. Aiding MFIs with ML based technologies can help lender's make more informed decisions, covering a more comprehensive set of attributes and data.

The microfinance space in India has seen a positive disruption due to the digital revolution and success of Pradhan Mantri Jan Dhan Yojna (PMJDY) scheme launched by Government of India in 2014. The scheme has led to opening of 417 million new bank accounts and has resulted in 80% increase in India's banking penetration. Despite this evolution, a report by KPMG in 2021 has indicated that only half of Microfinance Institutions leverage AI/ML for credit risk assessment.

The main aim of this paper is to explore various available literature studies on application of machine learning algorithms in a microfinance context. The research in this survey is spread across various countries such as Ghana, Peru and Morocco. The findings and recommendations of existing research work can help in identifying appropriate ML algorithms for credit assessment specifically for rural borrowers. This study can also help fintech startups, banking and non-banking financial institutions in India to develop more financially inclusive products.

2. LITERATURE SURVEY

Jia Wu and et al. (2010) [1] conducted a comparative study on data mining methodologies for developing a loan risk assessment system for sub-prime lenders within the microfinance space. The datasets used are from East Lancashire Moneyline (ELM), a non-profit, sub-prime lending company in United Kingdom and the second one being a German Credit data set. The authors have used the popular WEKA tool for implementing data mining algorithms – Decision Tree, Clustering, Naïve Bayes Classifier, Family Resemblance Exemplar Based Model (has foundations in Bayesian Networks). The accuracy of these models on ELM dataset were very low with K-means clustering giving an accuracy of 38.76% with a standard deviation of 9.62%. This could be because the ELM dataset is small with only 109 loan cases. The German Credit Dataset showed promising results with Naïve Bayes model giving the highest accuracy of 75.09% with a standard deviation of 2.19%. The authors have indicated that decision trees with an accuracy of 71.44 % might be a better model because it is better understood by staff who are not from technical background and hence are able to analyze the sets of rules given as an output by the decision tree.

Alper Ozpinar and et al. (2016) [2] In this paper the authors have proposed Credit Risk Evaluation as a service by using a novel cloud-based service-oriented architecture. The proposed solution is intended for all kinds of financial institutions including microfinance institutions. The proposed service will store customer data in private cloud to maintain privacy and trade secrets. The service works on an

underlying Artificial Neural Network (ANN) decision engine which will evaluate customers credit risk. The input parameters for the Artificial Neural Network (ANN) will be received from the institution. The system also provides database storage of existing customer information in a secure environment and has suggested using the historical data with new customer data for continuous improvement of the underlying ANN engine. The preliminary results from sample data gave a performance of 80% accuracy for 'Low Risk Accept' class, 95% accuracy for 'Medium Risk Analyse' class and 89% accuracy for 'High Risk Reject' class.

Ghita Bennouna and et al. (2018) [3] explored fuzzy logic approach for setting up a credit risk model for Microfinance Institutions in Morocco. The authors have highlighted that information asymmetry in microfinance sector has led to failure of traditional credit scoring systems in Microfinance institutions. Fuzzy Logic approach will help in addressing the information asymmetry since it is well-equipped to mimic human decision-making behaviour much accurately. Three major customer variables are used as an input to the fuzzy model – descriptive variable, behavioural variable and variable characterizing loans contracted by customer. Rule base was generated for each variable on the basis of the weight assigned to the variables. The model was applied on data obtained from Microfinance Institutions in Morocco to describe customers profile and classify their behaviour (High Risky, Medium Risky, Low Risky) taking in consideration statistical data as well as objective data based on opinions of portfolio manager who is in direct contact with the customers. The fuzzy model classified 6% of customers as High Risky, 75% as Medium Risky and 19% as Low Risky for the given data. The authors have indicated that the result for the given model can be variable since the input classification can vary over time, hence the filters will have to be updated as and when required.

Jennifer Ifft and et al. (2018) [4] explored nine common machine learning approaches under the umbrella of General Linear Models, Bayesian Models and Ensemble models to predict the demand for new credit. They also compared the efficacy of machine learning models with standard econometric approaches. The dataset used for the study was procured from 2014 Agricultural Resource Management Survey, which is the primary source of United States farm business financial performance. The results clearly indicated that the machine learning models performed significantly better than typical econometric studies with Gaussian naïve bayes approach having the highest average recall score of 80%. Although the Gaussian naïve approach was best at predicting the target potential customers for micro-credit, it had a lower precision as compared to logistic regression. The authors have also highlighted that they have considered a cost-based model approach for their evaluation and the results of the study can vary on the basis of how the model performance is being evaluated. Machine learning models had a higher recall score when learning from featured engineering dataset but have higher tendencies for false positives. Furthermore, the authors have concluded that an expanded set of features could also lead to overfitting hence the machine learning models performance can be assessed on a case-by-case basis depending on the dataset, the business requirements and evaluation metrics.

Sofie De Cnudde and et al. (2019) [5] investigated the power of Meta (previously known as Facebook) data to automate the process of credit scoring in a microfinance space. The dataset is from Philippines and have taken three sets of features for their models – socio-demographic data (name, age, religion, birthdate, education level), interest data which considers people with similar interest, backgrounds and activities and finally social network data which considers the default history of first-degree connections of the customer. The authors have used existing research methodologies to represent the interest data as a bipartite graphs and the social network data as a unigraph. The bipartite graphs and unigraphs are parsed using a linear Support Vector Model. The authors have used two ensemble models- one uses baseline Support Vector Model with twenty-nine socio-demographic variables for each loan applicant and other uses Support Vector Model on unweighted unigraph. The results from the two ensemble models are applied over interest data and social network data for classification. The highest Area Under Curve is 0.9219 for interest-based data. The authors have indicated at further research on using Meta (previously known as Facebook) data for applications other than automating assessment of creditworthiness.

Emmanuel Awoin and et al. (2020) [6] employed Machine Learning approach to develop a financial distress model for rural banks in Ghana. The goal for the model is to effectively predict the financial status of the rural bank. The authors have used random forest for feature selection, which helped in identifying 13 relevant predictors out 16 for the model. The classification model uses three Decision

Tree algorithms – C5.0, Classification and Regression Tree (CART) and C4.5. The prediction accuracy was recorded with using all predictors and then using only the top 13 predictors. The C5.0 model gave a 100% accuracy for both sets of predictors when implemented on the test data, followed by Classification and Regression Technique (CART) with 87.88% and 81.31% accuracy with all predictors and 13 predictors respectively. For future studies, the authors have indicated on using a dataset with larger amount of data, using similar models on commercial banks and lastly using deep learning for prediction of rural banks performance.

Henry Ivan Condori-Alejo and et al. (2020) [7] proposed a machine learning model for accurately identifying credit risk for microfinance institutions based in rural sector in Peru. The authors have used rural variable specification to identify top 34 empirical variables for credit risk analysis in a rural setting. Artificial Neural Network (ANN) gave an accuracy of 93.27% followed by Decision Tree with an accuracy of 88.80%. The study aimed at aiding microfinance institutions for default detection has improved on the traditional methodology accuracy by 16.91%, with the Artificial Neural Network model's accuracy of 93.27%. The paper recommends performing further study of machine learning models on other rural entities for further validation of variables.

Apostolos Ampountolas and et al. (2021) [8] performed novel machine learning based approach on assessing credit risk in a micro-credit risk environment using real life data from Africa. They performed data pre-processing to identify accurate parameters (customer's age, gender, marital status) for prediction in micro-credit context, reduce variance by taking log values of loan amounts and avoid overfitting by removing highly correlated variables. Data imbalance was tackled using SMOTENC algorithm. They have used classification algorithms to classify customers into three risk classes – "good", "average", or "poor". The authors concluded that tree-based algorithms are powerful in predicting default risk in micro-credit space, since such datasets have many categorical features. The top three best performing algorithms – Random Forest, XGBoost and AdaBoost are all ensemble classifiers and tree-based algorithms as well. All ensemble classifiers reported an overall accuracy of at least 80% on the validation set with Adaboost giving best performance of 81.2071% prediction accuracy. The authors have indicated at further areas of research focusing on including temporal aspects of credit risk for micro-credit risk and default detection.

TABLE-1: LITERATURE SURVEY ON MACHINE LEARNING FOR MICROFINANCE INSTITUTIONS

Sr No	Paper -Topic	Source -Year	Algorithm recommended-Benefits	Dataset Description	Accuracy
1	A Comparison of Data Mining Methods in Microfinance	IEEE -2010	Decision Tree – helps non-technical people within banking sector analyze the set of rules given as an output	German Credit Dataset - includes 1000 cases, where each example has 20 attributes. It has two classes which are 'Good' and 'Bad'. There are 700 good cases and 300 bad cases	71.44%
2	Credit Risk Evaluation as a Service (CREaaS) based on ANN and Machine Learning	International Journal on Recent and Innovation Trends in Computing and Communication -2016	Artificial Neural Networks – inputs can be received from banking institution	Custom Dataset	80% for 'Low Risk Accept' class, 95% accuracy for 'Medium Risk Analyse' class, 89% accuracy for 'High

					Risk Reject'
3	Fuzzy logic approach applied to credit scoring for microfinance in Morocco	ScienceDirect - 2018	Fuzzy Logic – mimics human-decision making more accurately. This will help in maintaining human touch in the microfinance space.	Data obtained from Microfinance Institutions in Morocco to describe customers profile and classify their behaviour (High Risky, Medium Risky, Low Risky) taking in consideration statistical data as well as objective data based on opinions of portfolio manager who is in direct contact with the customers	-
4	Can machine learning improve prediction – an application with farm survey data	International Food and Agribusiness Management Review - 2018	Gaussian naïve bayes – better than typical econometric approach for credit risk evaluation	Agricultural Resource Management Survey	Recall Score – 80%
5	What does your Facebook profile reveal about your creditworthiness? Using alternative data for microfinance	Journal of the Operational Research Society - 2019	Social Network Analysis + Support Vector Model	Meta (previously known as Facebook) data from Philippines and considers three sets of features for the ML models – 1. Socio-demographic data (name, age, religion, birthdate, education level) 2. Interest data which considers people with similar interest, backgrounds and activities 3. Social network data which considers the default history of first-degree connections of the customer	92.19 % for Interest Data
6	Predicting the Performance of Rural Banks in Ghana Using Machine Learning Approach	Research article in Advances in Fuzzy Systems - 2020	Decision Tree	Data collected from rural banks in Ghana	C5.0 model – 100%
7	Rural Micro Credit Assessment using Machine Learning in a Peruvian microfinance institution	ScienceDirect - 2020	Artificial Neural Networks	Data collected from microfinance institutions in Peru	93.27%
8	A Machine Learning Approach for Micro-Credit Scoring	Special Issue-Interplay between Financial and	Adaboost – tree-based algorithms are powerful since datasets in the	Real life data from Africa	81.2071%

		Actuarial Mathematics - 2021	microfinance context have a lot of categorical features		

"– "– Not Mentioned

3. CONCLUSION

Microfinance Institutions can achieve higher credit penetration by integrating machine learning algorithms for credit scoring. This survey gives an insight into already existing research on the benefits of different machine learning algorithms across various microfinance settings. Some studies have shown that tree-based algorithms such as decision tree algorithm and adaboost are found to be powerful in such settings where the dataset tends to have many categorical variables. It has also been highlighted that the existing microfinance space deals with information asymmetry and a lot of human-touch (since decisions are taken by lenders who have first hand knowledge of rural environments), these elements can be incorporated in machine learning systems by implementing fuzzy logic which is more intuitive in nature. Furthermore, studies have indicated that analysis of social media data can also help in building predictive models for rural settings where there has been a penetration of mobile applications and online banking applications. Lastly, some research methodologies have also explored deep learning models such as artificial neural networks (ANN) for micro-credit score prediction.

REFERENCES

[1] J. Wu, S. Vadera, K. Dayson, D. Burridge and I. Clough, "A comparison of data mining methods in microfinance," 2010 2nd IEEE International Conference on Information and Financial Engineering, 2010, pp. 499-502, doi: 10.1109/ICIFE.2010.5609408.

[2] A. O. A. B., "Credit Risk Evaluation as a Service (CREaaS) based on ANN and Machine Learning", *IJRITCC*, vol. 4, no. 4, pp. 459–465, Apr. 2016.

[3] Ghita Bennouna et al. "Fuzzy logic approach applied to credit scoring for microfinance in Morocco" 2018, Procedia Computer Science, Volume 127, 2018, Pages 274-283. https://doi.org/10.1016/j.procs.2018.01.123

[4] Jennifer Ifft et al. "Can machine learning improve prediction – an application with farm survey data" ,2018 International Food and Agribusiness Management Review: 21 (8)- Pages: 1083 – 1098, https://doi.org/10.22434/IFAMR2017.0098

[5] Sofie De Cnudde, Julie Moeyersoms, Marija Stankova, Ellen Tobback, Vinayak Javaly & David Martens (2019) What does your Facebook profile reveal about your creditworthiness? Using alternative data for microfinance, Journal of the Operational Research Society, 70:3, 353-363, DOI: 10.1080/01605682.2018.1434402

[6] Emmanuel Awoin, Peter Appiahene, Frank Gyasi, Abdulai Sabtiwu, "Predicting the Performance of Rural Banks in Ghana Using Machine Learning Approach", *Advances in Fuzzy Systems*, vol. 2020, Article ID 8028019, 7 pages, 2020. https://doi.org/10.1155/2020/8028019

[7] Henry Ivan Condori-Alejo, Miguel Romilio Aceituno-Rojo, Guina Sotomayor Alzamora,Rural Micro Credit Assessment using Machine Learning in a Peruvian microfinance institution,Procedia Computer Science,Volume 187,2021,Pages 408-413,ISSN 1877-0509,https://doi.org/10.1016/j.procs.2021.04.117.

[8] Ampountolas, A.; Nyarko Nde, T.; Date, P.; Constantinescu, C. A Machine Learning Approach for Micro-Credit Scoring. *Risks* 2021, *9*, 50. https://doi.org/10.3390/risks9030050