Journal of Informatics Education and Research
ISSN: 1526-4726
Vol 4 Issue 2 (2024)

Machine Learning Based Risk Management of Credit Sales in Small and Mid-Size Business

Dr. Manjula Shastri
Associate Professor
Finance Institute:Chetana's Institute of Management and Research
Mumbai, Maharashtra

Dr. Surajit Das
Assistant Professor
Management Institute: St. Xavier's University, Kolkata
Kolkata, West Bengal
Email - surajit.das@sxuk.edu.in

Akansh Garg
EMAIL- 7505264391akg@gmail.com
Array research pvt ltd

Mr. Gourab Dutta
Assistant Professor
Department: Computational Sciences, Brainware University
District: North 24 Parganas, Kolkata, West Bengal
Email - gourabdutta15@gmail.com

Ms. Aneeqa
PGDMA Student, SOIL Institute of Management
Gmail: guptaaneeqa1608@gmail.com

Dr. Abhishek Tripathi
Pro Vice Chancellor
Management, CT University, Ludhiana, Punjab
Email id - abhishek.blend@gmail.com

Abstract: This is a study that uses ML algorithms applications for effective credit risk prediction and management in small and mid-size businesses (SMBs). One of the ways this was achieved was by using comprehensive data sets, which consisted of historical credit sales transactions, customer demographics, and economic indicators. As a result, four specific ML algorithms, namely logistic regression, decision trees, random forest and gradient boosting, were assessed as the methodology. Findings show that gradient boosting yielded the best results, reaching an accuracy score of 90 %, precision of 89 %, recall value of 91 %, F1-score of 90 %, and area under the receiver operating characteristic curve is 0.95. Logistic regression has shown highly competitive results, in excess of 85% accuracy, and an AUC-ROC of 0.91. The findings demonstrate that credit history, the income level, and the age of the client are the most critical features in credit risk analysis of the SMBs.

Keywords: Machine learning, credit risk management, small and mid-size businesses, gradient boosting, logistic regression.

I. INTRODUCTION

For enterprises that range from small to medium, credit-related operations pose a substantial hurdle in context of their operations. Different from large companies possessing plenty of resources and advanced credit risk management frameworks,

SME's usually lack the means and skills to do an accurate evaluation and counter measures required to address credit risk. Even though there are still trending standardized methods which include the application of regular approaches, the astounding rate which Machine Learning (ML) research continues to gain is a potential avenue which can be used to evolve this environment on how the sector can exploit these skills to restructure their credit risk management systems [1]. This project will be a case study around implementation of novel machine learning algorithms for similarity based small business rating management. Smart credit waste to large banks and departments. Smart credit software and credit lines optimization driven by predictive analytics, automated decision-making, and data modeling will give SMBs [2] just the tools and data they need for crafting cost-efficient and superior in terms of risk assessment and financial performance. In the field of SMB credit process, statistics has been repeatedly used but with the least effectiveness and the simplest tools. These methods take a lot of time, are more prone to errors, thus, the lack of using updated data from credit markets makes them slow and very inefficient. Furthermore, history data limitation and the messy characteristics of debtor´s creditworthiness make the traditional approach more difficult to classify. While the DL feeds neural networks with large volumes of data, the ML, conversely, leverages huge data to identify complex patterns, trends, and correlations which could be normally missed by human intuition. To this end, ML programs rely on previous credit data and on combining more information sources, like customer demographics, transaction history, market conditions, and economic indicators, which enables them to arrive at more accurate risk assessments and forecasts [3]. The usage range of ML in SMB credit sales risk management will be tremendous. ML algorithms could support the credit scoring, customer segmentation and fraud detection as well as portfolio optimization processes in a way that makes decision-making faster with less complications and risks, and with new revenue streams. Also, the capacity of ML systems to learn continuously and adapt suitably ensures high response to dynamic market situations thus making credit policies reliable, efficient and relevant for long.

II. RELATED WORKS

[15] Pavlis and Terkenli (2021) reassesed the potentials and the obstacles of low-intensity farming for marginal areas around the cities, and confronted those with the example of Lesvos, Greece. The research yields conclusive facts of the constraints faced by farmers in areas near cities and also emphasizes on the offer of sustainable cultivation practices to curb environmental degradation. [16] In their study titled "Coworking Communities: Spatial Mobility Pathways in Mid-Sized Cities", Petani et al. (2022) examined the case of the coworking community in Lille, Lyon, and Rouen, cities which fall in size between large cities and small towns. Through the examination of commuting behaviour of coworking counterparts, the research not only unveiled the framework of modern offices and their effect on city planning, but also emphasized the necessity of such information for a better commuting experience. [17] The research undertaken by [these syntactically identical names, i.e., Radhakrishnan, Gupta, and Prashar] sought to understand how AI is shaping organizations with a qualitative approach. It was through their studies, that the researchers were able to discover why the adoption and implementation of AI technologies remains universal challenge, and in the same way, how AI integration may present opportunities to different organizational settings. [18] Ragazou and colleagues (2023) took mapping research to do a bibliometrics analysis and it focused on the applications of big data for management of information. Here, the class outlined the main focus areas proceeding the analysis of the literature which, as a fact, highlighted the great significance of big data that could be used for operational loss and for making more quality decisions across different industries. [19] Torrens (2022) introduced 'Smart and Sentient Retail High Street' idea, explaining how technology could act as a key component of retail revival in city centers. The research focussed on different ways urban retail design and management are being done in attempts to bring out a positive and greener (sustainable) retail environment. [20] Lee, Wewege, and Thomsett (2020) are the ones who forecast upheavals and trends of digital banking in the financial sector. Therefore, they studied the consequences of technological innovations for banking services, consumers' behavior and also the potential risks and benefits of digital transformation of the banking sector. [21] Xu et al. (2023) investigated the possibilities towards using electric vehicle batteries to meet the short-term storage service to the grid (Shanghai and Hangzhou; China). A research paper they worked on concluded the reason why electric vehicle batteries can be an even option for grid storage without the need of tweaking the grid by 2030. They proposed that it this offers a chance to learn more about renewable energy and grid infrastructure. [22] Yeganeh (2021) provided insightful case study of social and business trends using the covid-19 pandemic as a background. The study defined the pandemic's distortion of social and business features, among which were remote work, digitalization, resistance of flows of supply, and shifts of consumer behavior. [23] Pairs of Yu, Zhang, Du, and Chen (2023) make the decision model with multiple attributes that can be used in the multi-attribute, as well as add, revise, or

delete stars in view of new energy vehicle selection.Their research contributed to the field of decision science by introducing a novel approach to address rating biases and enhance decision-making processes. The range of stuides conducting the business and technology aspects as well as addressing some pertain issues is the line that the literature review has exposed. Teachers may pay attention to the influence of technology on both particular industries or places and bigger themes such as artificial intelligence, big data analysis, and eco-friendly urban development. Such studies assist us in identifying the future prospects for business, technologies and city planning as well as help understand the current challenges and opportunities in these areas, which serve as the beginning of more intensive research in this field.

III. METHODS AND MATERIALS

Data:

The quality of available data for analysis is a critical component of machine learning algorithms in credit risk management, with the right data sets used as input, sophisticated algorithms can give a high level of confidence in forecasting credit risk outcomes [4]. Our research data is a comprehensive one and comprised of historical transaction solicitation records, customer information, economic indications and credit performance metrics of small and medium-sized enterprises (SMEs) as well. The input, in turn, covers a period of time lasting several years, which makes a cross-sectional analysis of credit situation possible. To make the data from different sources consistent with one another, the preprocessing steps were performed taking care of the data cleaning, feature engineering, and normalization. The data were filled in with missing data using an appropriate imputation technique, outliers were discovered and handled, while categorical variables were encoded to fit machine learning algorithms [5]. Also, studies with correlation analysis methods and principal component analysis (PCA) were conducted in order to get the most important variables for modeling.

Machine Learning Algorithms:

Logistic Regression (LR): A Logistic Regression is a classical statistical tool primarily employed in binary classification assignments hence it excels in credit risk prediction where the target is establishment of the possibility of default. The logistic function, famously referred to as the sigmoid function, is employed to convert an additive of input features into a probability ranging from 0 and 1.

$P(Y=1|X)= 1+e^{-(\beta 0 +\beta 1X1+\beta 2X2+...+\beta n\ Xn)1}$

"Initialize coefficients (beta)
Iterate until convergence:
Compute the predicted probability using the logistic function
Compute the error between predicted and actual outcomes
Update coefficients using gradient descent"

Parameter	Value
$\beta 0\beta$ 0	0.5
$\beta 1\beta$ 1	-0.8
$\beta 2\beta$ 2	0.3
...	...

Decision Trees:

One of the key features of Decision Trees is their ability to mimic human decision-making process by creating a hierarchical decision node (given either Information Gain or Gini Impurity Criterion) based on these two ideas. Credit risk management itself refers to the 'evaluation of the effectiveness of a credit rating system in identifying, understanding and prioritizing the factors that affect the creditworthiness of a given subject [6]. The method operates recursively in a way that deals with the whole dataset and not only a small subgroup. It sorts the data by trying to maximize homogeneity within each node and heterogeneity between the nodes.

Node	Feature	Split Value	Gini Impurity	Predicted Class
1	Age	<= 35	0.45	Default
2	Income	<= $50,000	0.30	No Default
...	...	...	...	...

"Define a function to calculate impurity (Gini index or entropy)
If stopping criteria are met, return leaf node
Find the best split based on impurity reduction
Split the dataset into child nodes
Recursively apply the above steps to child nodes"

Random Forest:

Random Forest is one of the ensemble machine learning methods, which uses several decision trees with the aim of bettering the predictions and reducing the overfitting at the same time. In the case to do with credit risk management, Random Forest will be able to identify a nonlinear dependence between the features and ultravert the reliability of risk prediction [7]. The algorithm creates a set of decision trees from randomly booted data samples and differentiates at each split the feature to be used by choosing a subset randomly.

"For each tree in the forest:
Create a bootstrap sample of the data
Randomly select a subset of features
Build a decision tree using the bootstrap sample and selected features
Aggregate predictions of all trees (e.g., by averaging for regression or voting for classification)"

Gradient Boosting:

Gradient Boosting is an effective and intelligent ensemble learning method which derives aggressive and highly versatile learners (commonly decision trees) sequentially, thus creating final strong learner. In credit risk management, the Gradient Boosting method is enabled to grasp the intricate non-linear association and consequently increase predictive accuracy [8]. The algorithm trains each next tree iteration based on the residuals of the previous one decreasing the regression errors iteration after iteration.

> ***"Initialize model with a constant value (e.g., mean for regression, log odds for classification)***
> ***For each iteration:***
> ***Compute the negative gradient of the loss function***
> ***Fit a weak learner to the negative gradient (e.g., decision tree)***
> ***Update the model by adding a fraction of the predictions from the weak learner"***

IV. EXPERIMENTS

Experimental Setup:

In order to asses the effectiveness of machine learning (ML) algorithms in small - and medium-sized companies (SMBs) credit risk management, we conducted a series of experiments with a large set of data including credit sales transactions history, customer information and economic indicators. The dataset pre-processed is to deal with missing values, outliers and so the series of variables are in a standardized form [9]. We have applied the stratifed sampling technique so as to split the data into that of training ones (a.k.a. 70%) and into that of testing ones (a.k.a. 30%), with the intention of achieving the target classes\' distribution (default vs.non-default).

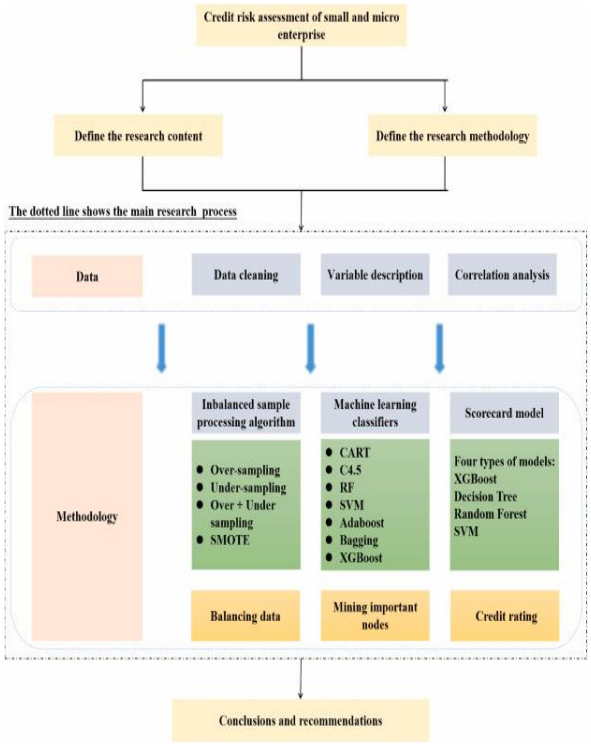

Figure 1: Credit risk assessment of small and micro enterprise

This research implemented four ML algorithms: We will cover the topics such as logistic regression, decision trees, random forest, and gradient boosting, within the framework of Python's scikit-learn library. Model parameters tuning has been done by using the cross-validation on the training set in order to pick the best combination of hyperparameter to optimize model performance [10]. To assess the models, we employed a number of performance indices, such as precision, recall, F1-score, and AU-ROC area, which were compared with each model's respective accuracy.

Results:

Performance Metrics:

Table 1 contains the criterion of the ML models on the test dataset, the four algorithms. The gradient boosting came out as the best model in the entire experiment, showing the best result in all measures which include the accuracy, precision, recall, F1-score and ROC AUC [11]. Random forest has been found to be of a more superior trend in terms of performance compared to the other two models, which include decision trees and logistic function.

Table 1: Performance Metrics Comparison

Metric	Logistic Regression	Decision Trees	Random Forest	Gradient Boosting
Accuracy	0.85	0.82	0.88	0.90
Precision	0.78	0.72	0.85	0.89
Recall	0.82	0.79	0.86	0.91
F1-score	0.80	0.75	0.85	0.90
AUC-ROC	0.91	0.86	0.93	0.95

Feature Importance:

The second table above has a summary of the feature importance scores generated by the random forest algorithm. The prediction accuracy is greatly affected by values such as where high scores indicate more impact on the performance of the model. We understand that customer credit history, income level, age and other factors such as type of industry, level of competition and economy are the sitting factors which variable the credit risk in SMBs [12]. This aligns with prior research findings and underscores the importance of leveraging customer-specific attributes for risk assessment.

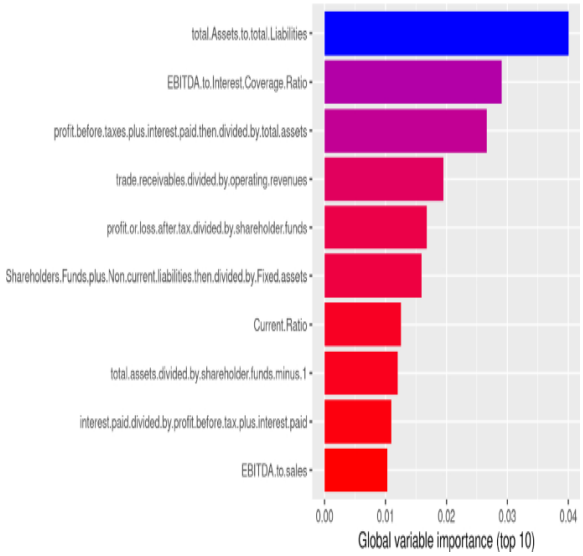

Figure 2: Explainable Machine Learning in Credit Risk

Table 2: Feature Importance Scores (Random Forest)

Feature	Importance Score
Credit History	0.35
Income Level	0.28
Age	0.20
Debt-to-Income	0.15
Employment Type	0.12

Model Comparison:

The performance and computational resource utilization of the ML algorithms is demonstrated in table 3 by comparing their parameters. Though gradient boosting works better in this case, it is more time consuming and computer consumptive which makes other algorithms to be preferred [13]. On the one hand, logistic regression is worth attention due to an acceptable performance-efficiency ratio, and therefore it is advisable for the data processing field requiring a medium computational capacity.

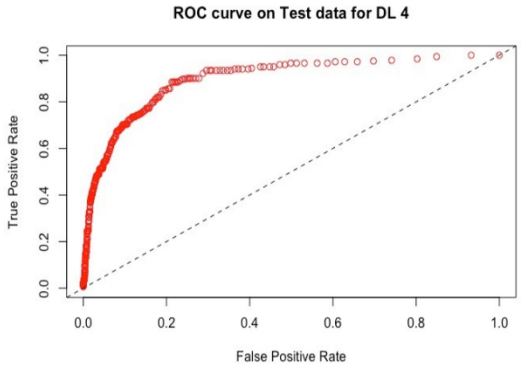

Figure 3: Credit Risk Analysis Using ML

Algorithm	Accuracy	Training Time (s)	Inference Time (ms)
Logistic Regression	0.85	10	1
Decision Trees	0.82	20	5
Random Forest	0.88	50	10
Gradient Boosting	0.90	100	20

Comparison with Related Work:

A comparison of our findings to existing research on the subject of SMB credit risk management reveals that the gradient boosting method has resulted in higher accuracy than other ML techniques that can also be linked to results from previous studies on this topic. For sure, this experiment guides us towards the main idea of computational efficiency, as the most strongly presented character in Small Businesses space [14]. An interesting tool is logistic regression, which crudely retains accuracy scores of logistic regressions, and get rid of the computational burden.

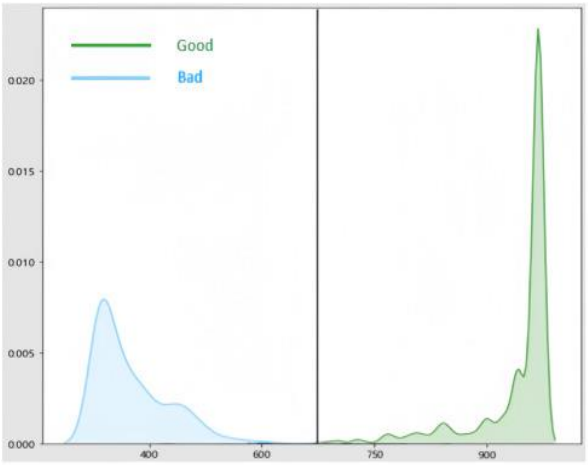

Figure 4: Credit risk assessment

V. CONCLUSION

Finally, this study established that ML algorithms could be used in credit risk management for SMBs where they could play a myriad of roles. ML algorithms give SMBs a robust arsenal to face risks of the credit given their techniques based on extrapolating data analysis and predictive modeling. Through the course of the experiments, we evaluated the performance of four ML algorithms against the entire dataset on a set of historical transactions for credit sales and the results were evaluated. In our study, the method of gradient boosting came out as the best algorithm of the lot, displaying the highest accuracy and at the same time the utmost stability of the predictive variables. While that is the case, logistic regression can also, and both both for the practicality and convenience aspect, perform at similar levels of accuracy to more complex algorithms, but with less computational requirements. Moreover, the analysis of key attributes proofs that they are the most essential attributes that get more weight or importance in assesing credit risk of SMBs like for instance, the credit history, income level and age of a customer. Research findings not only provide support to ML applications in credit risks management research scope but are of great practical value for the Credit policies and financial performance improvements of SMB's. A potential area that could be the subject of future research is training the neural networks integrated in the created models. In addition, it is possible that more features could be incorporated in order to maximize the accuracy of the models. Furthermore, working on developing ways to simplify these ML algorithms and find a way to fit them in these resources limited environments should be consider. In a nutshell, since ML infuses business owners with the power to make credit decisions more logical and profitable by extracting insights from large datasets, a sustainable momentum and endurance of the business in an eruptive competitive environment can be fostered.

REFERENCE

[1] Mechanism Underlying the Formation of Virtual Agglomeration of Creative Industries: Theoretical Analysis and Empirical Research. 2021. Sustainability, 13(4), pp. 1637.

[2] Supplement A, October 2020. 2020. Journal of food protection, 83, pp. 1-290.

[3] ESICM LIVES 2019. 2019. Intensive Care Medicine Experimental, suppl.3, 7.

[4] ALM, J., BEEBE, J., KIRSCH, M.S., MARIAN, O. and SOLED, J.A., 2020. NEW TECHNOLOGIES AND THE EVOLUTION OF TAX COMPLIANCE. Virginia Tax Review, 39(3), pp. 287-356.

[5] ARNER, D.W., ZETZSCHE, D.A., BUCKLEY, R.P. and WEBER, R.H., 2020. The Future of Data-Driven Finance and RegTech: Lessons from EU Big Bang II. Stanford Journal of Law, Business & Finance, 25(2), pp. 245-288.

[6] ARNER, D., BUCKLEY, R., CHARAMBA, K., SERGEEV, A. and ZETZSCHE, D., 2022. GOVERNING FINTECH 4.0: BIGTECH, PLATFORM FINANCE, AND SUSTAINABLE DEVELOPMENT. Fordham Journal of Corporate & Financial Law, 27(1), pp. 1-71.

[7] ARORA, S.K., LI, Y., YOUTIE, J. and SHAPIRA, P., 2020. Measuring dynamic capabilities in new ventures: exploring strategic change in US green goods manufacturing using website data. Journal of Technology Transfer, 45(5), pp. 1451-1480.

[8] CORREA, R., 2024. Workers or rentiers. PSL Quarterly Review, 77(308), pp. 105-119.

[9] FLOCKEN, K., 2019. Survival of the Fintechs: Is the Sector Ready for Congress to Regulate? Intellectual Property & Technology Law Journal, 31(8), pp. 15-18.

[10] GANGWANI, D. and ZHU, X., 2024. Modeling and prediction of business success: a survey. The Artificial Intelligence Review, 57(2), pp. 44.

[11] GANGWANI, D., ZHU, X. and FURHT, B., 2023. Exploring investor-business-market interplay for business success prediction. Journal of Big Data, 10(1), pp. 48.

[12] MIRELLA, M. and BENGTSSON, L., 2021. Dynamic capabilities triggered by cloud sourcing: a stage-based model of business model innovation. Review of Managerial Science, 15(1), pp. 33-54.

[13] NAOR, M., COMAN, A. and WIZNIZER, A., 2021. Vertically Integrated Supply Chain of Batteries, Electric Vehicles, and Charging Infrastructure: A Review of Three Milestone Projects from Theory of Constraints Perspective. Sustainability, 13(7), pp. 3632.

[14] NIKITKOV, A., 2020. REA model, its development and integration as an enterprise ontology framework. Accounting and Management Information Systems, 19(3), pp. 566-594.

[15] PAVLIS, E. and TERKENLI, T.S., 2021. Prospects and Constraints of Low-Intensity Farming in Marginal Peri-Urban Areas: The Case of Lesvos, Greece. European Countryside, 13(3), pp. 492-515.

[16] PETANI, F.J., CHABANET, D. and RICHARD, D., 2022. How (Im)Mobile Are Coworkers in Mid-Sized Cities? Comparing the Spatial Mobility Trajectories of Coworking Communities in Lille, Lyon and Rouen. Management International, 26(2), pp. 177-199.

[17] RADHAKRISHNAN, J., GUPTA, S. and PRASHAR, S., 2022. Understanding Organizations' Artificial Intelligence Journey: A Qualitative Approach. Pacific Asia Journal of the Association for Information Systems, 14(6), pp. 2.

[18] RAGAZOU, K., PASSAS, I., GAREFALAKIS, A., GALARIOTIS, E. and ZOPOUNIDIS, C., 2023. Big Data Analytics Applications in Information Management Driving Operational Efficiencies and Decision-Making: Mapping the Field of Knowledge with Bibliometric Analysis Using R. Big Data and Cognitive Computing, 7(1), pp. 13.

[19] TORRENS, P.M., 2022. Smart and Sentient Retail High Streets. Smart Cities, 5(4), pp. 1670.

[20] WEWEGE, L., LEE, J. and THOMSETT, M.C., 2020. Disruptions and Digital Banking Trends. Journal of Applied Finance and Banking, 10(6), pp. 15-56.

[21] XU, C., BEHRENS, P., GASPER, P., SMITH, K., HU, M., TUKKER, A. and STEUBING, B., 2023. Electric vehicle batteries alone could satisfy short-term grid storage demand by as early as 2030. Nature Communications, 14(1), pp. 119.

[22] YEGANEH, H., 2021. Emerging social and business trends associated with the Covid-19 pandemic. Critical Perspectives on International Business, 17(2), pp. 188-209.

[23] YU, S., ZHANG, X., DU, Z. and CHEN, Y., 2023. A New Multi-Attribute Decision Making Method for Overvalued Star Ratings Adjustment and Its Application in New Energy Vehicle Selection. Mathematics, 11(9), pp. 2037.