Comparative Analysis of Machine Learning Models for Credit Scoring: A Case Study on the South German Credit Dataset

Vikas Mendhe[1], Aamol Gote[2], Roshan Mahant[3]
Senior Consultant, Austin, Texas, USA[1,3]
Solution Architect, Plainsboro, New Jersey, USA[2]
LinkedIn: https://www.linkedin.com/in/vikas-mendhe-69260012[1]
LinkedIn: https://www.linkedin.com/in/aamolgote[2]
LinkedIn: https://www.linkedin.com/in/roshan-mahant/[3]
vikas.mendhe@gmail.com[1], aamolgotewrites@gmail.com roshanmahant@gmail.com[3]

Abstract: *In this research, proposed a novel hybrid machine learning approach that combines the strengths of Random Forest, Multi-Layer Perceptron (MLP), and LightGBM algorithms for classification tasks. This research work focuses on evaluating the performance of this hybrid model using the South German Credit dataset obtained from Kaggle, comprising bank client data, client last contact information, and labels. With 45,211 records and 16 attributes, this dataset provides a suitable environment for assessing the effectiveness of our proposed approach. Employ various evaluation metrics including accuracy, sensitivity and specificity and Receiver Operating Characteristic (ROC) to comprehensively analyze the model's performance. Through experiments, aim to demonstrate the efficacy of the hybrid approach in accurately classifying instances and providing insights into its potential applications in real-world scenarios*

Keywords: Machine Learning, Random Forest, Multi-Layer Perceptron (MLP), LightGBM

REFERENCES

[1]. Assef F, Steiner MT, Neto PJS, de Barros Franco DG (2019) Classification algorithms in financial application: credit risk analysis on legal entities. IEEE Lat Am Trans 17(10):1733–1740

[2]. Ben-David A (1995) Monotonicity maintenance in information-theoretic machine learning algorithms. Mach Learn 19(1):29–43

[3]. Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140

[4]. Breiman L (2001) Random forests. Mach Learn 45(1):5–32

[5]. Central Bank of Brazil (2007) Annual report. https://www.bcb.gov.br/pec/ boletim/banual2007/rel2007p.pdf

[6]. Central Bank of Brazil (2020) Consumer personal loan. https://www.bcb. gov.br/estatisticas/reporttxjuros/

[7]. Cornée S (2019) The relevance of soft information for predicting small business credit default: Evidence from a social bank. J Small Bus Manag 57(3):699–719

[8]. Crone SF, Finlay S (2012) Instance sampling in credit scoring: an empirical study of sample size and balancing. Int J Forecast 28(1):224–238

[9]. Damrongsakmethee T, Neagoe V (2019) C4.5 decision tree enhanced with AdaBoost versus multilayer perceptron for credit scoring modeling. In: Silhavy R, Silhavy P, Prokopova Z (eds) Computational statistics and mathematical modeling methods in intelligent systems. CoMeSySo 2019. Advances in intelligent systems and computing, vol 1047. Springer, Cham, pp 216–226

[10]. Dastile X, Celik T, Potsane M (2020) Statistical and machine learning models in credit scoring: a systematic literature survey. Appl Soft Comput 106263

[11]. Davis R, Edelman D, Gammerman A (1992) Machine-learning algorithms for credit-card applications. IMA J Manag Math 4(1):43–51

[12]. Feng X, Xiao Z, Zhong B, Dong Y, Qiu J (2019) Dynamic weighted ensemble classification for credit scoring using Markov Chain. Appl Intell 49(2):555–568

[13]. Galindo J, Tamayo P (2000) Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications. Comput Econ 15(1/2):107–143

[14]. Kamalloo E, Saniee Abadeh M (2014) Credit risk prediction using fuzzy immune learning. Adv Fuzzy Syst 2014:1–11

[15]. Khandani AE, Kim AJ, Lo AW (2010) Consumer credit-risk models via machine-learning algorithms. J Bank Finance 34(11):2767–2787

[16]. Kozodoi N, Lessmann S, Papakonstantinou K, Gatsoulis Y, Baesens B (2019) A multi-objective approach for profit-driven feature selection in credit scoring. Decis Support Syst 120:106–117

[17]. Lantz B (2013) Machine learning with R. Packt Publishing Ltd, Birmingham

[18]. Lei K, Xie Y, Zhong S, Dai J, Yang M, Shen Y (2019) Generative adversarial fusion network for class imbalance credit scoring. Neural Comput Appl pp 1–12

[19]. Li W, Ding S, Chen Y, Wang H, Yang S (2019) Transfer learning-based default prediction model for consumer credit in China. J Supercomput 75(2):862–884

[20]. Luo C (2020) A comprehensive decision support approach for credit scor- ing. Ind Manag Data Syst 120(2):280–290

[21]. Morales EA, Ramos BM, Aguirre JA, Sanchez DM (2019) Credit risk analysis model in microfinance institutions in Peru through the use of Bayesian networks. In: 2019 Congreso Internacional de Innovación y Tendencias en Ingenieria (CONIITI), IEEE, pp 1–4

[22]. Moula FE, Guotai C, Abedin MZ (2017) Credit default prediction mode- ling: an application of support vector machine. Risk Manag 19(2):158–187

[23]. Niklis D, Doumpos M, Zopounidis C (2014) Combining market and accounting-based models for credit scoring using a classification scheme based on support vector machines. Appl Math Comput 234:69–81

[24]. Oreski S, Oreski G (2014) Genetic algorithm-based heuristic for feature selection in credit risk assessment. Expert Syst Appl 41(4):2052–2064

[25]. Anupam Khan, Soumya K. Ghosh (2023) Machine assistance for credit approval? Random wheel can recommend and explainVolume 215, 1 April 2023, 119231

[26]. https://in.mathworks.com/help/matlab/ref/varargin.html

[27]. Shen KY, Sakai H, Tzeng GH (2019) Comparing two novel hybrid MRDM approaches to consumer credit scoring under uncertainty and fuzzy judgments. Int J Fuzzy Syst 21(1):194–212

[28]. Vieira J, Barboza F, Sobreiro VA, Kimura H (2019) Machine learning models for credit analysis improvements: predicting low-income families' default. Appl Soft Comput83(105):640

[29]. Pławiak P, Abdar M, Pławiak J, Makarenkov V, Acharya UR (2020) DGHNL: a new deep genetic hierarchical network of learners for prediction of credit scoring. Inf Sci 516:401–418