Journal of Intelligent Decision and Computational Modelling

www.jidcm.reapress.com

J. Intell. Decis. Comput. Model. Vol. 2, No. 2 (2026) 101–111.

Paper Type: Original Article

Intelligent Analysis of Users Data for Investment in Digital Financial Platforms Using Machine Learning Algorithms

Vajiheh Torkian[1], Shahrooz Bamdad[1,*], Amir Homayoun Sarfaraz[1]

[1] Department of Industrial Engineering, ST.C., Islamic Azad University, Tehran, Iran; vajiheh.torkian@iau.ac.ir; sh_bamdad@azad.ac.ir; dr.ahsarfaraz@yahoo.com.

Citation:

Received: 18 December 2025 *Revised: 27 February 2026* *Accepted: 09 April 2026*	Torkian, V., Bamdad, Sh., & Sarfaraz, A. H. (2026). Intelligent analysis of users data for investment in digital financial platforms using machine learning algorithms. *Journal of intelligent decision and computational modelling, 2*(2), 101-111.

Abstract

Today, due to the high risk of investing in emerging markets, the application of Machine Learning (ML) to design systems tailored to specific financial environments has rapidly expanded, providing new opportunities for analyzing financial data and improving decision-making. Peer-to-Peer (P2P) lending, a form of crowdfunding, has attracted considerable attention from corporate investors, institutional investors, and credit rating agencies by providing an alternative source of financing for individuals, businesses, and entrepreneurs. In this study, to address the class imbalance problem in P2P lending credit scoring, undersampling, oversampling, and synthetic minority oversampling techniques were employed in conjunction with six widely used ML classifiers including Multilayer Perceptron (MLP), Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), Logistic Regression (LR), and Light Gradient Boosting Machine (LGBM) based on a large dataset from the Lending Club website. The performance of the models was evaluated using four metrics: precision, accuracy, recall, and the F1 score. The experimental results show that artificial minority oversampling with MLP method has the best performance due to its higher value in the three criteria compared to other methods, as well as previous studies, and achieved an Area Under the Curve (AUC) value of 0.925 in the Receiver Operating Characteristic (ROC) curve, indicating high class discrimination. The findings of this study provide a comprehensive, data-driven framework that can guide credit risk assessment and support decision-making in real-world financial risk management.

Keywords: Peer-to-peer lending, Online platform, Financial risk, Machine learning, Classification, Imbalanced data.

1 | Introduction

Peer-to-Peer (P2P) lending is a form of platform-based online crowdfunding that enables borrowers to obtain credit from a large number of individual lenders. Unlike other forms of crowdfunding, which may be driven by altruistic motives, P2P lending is primarily motivated by the expectation of financial returns on the part of

Corresponding Author: sh_bamdad@azad.ac.ir

https://doi.org/10.48314/jidcm.v2i2.86

lenders [1]. This form of lending disrupts traditional financial intermediation by providing an alternative credit system that enhances financial inclusion and expands access to investment opportunities [2]. It is facilitated through online platforms where borrowers submit loan applications and lenders make informed investment decisions based on the available information. A fundamental challenge in P2P lending is information asymmetry, whereby loan applicants possess more, and often higher-quality, information than lenders. To mitigate this issue, lending platforms employ various strategies to complement the conventional information provided in loan applications [3].

These lending platforms have become an important channel for microlending and private financing, but the lack of standardized risk assessment mechanisms has led to inconsistencies in borrower credit assessment and uneven default patterns across platforms [4]. Artificial Intelligence (AI) has emerged as a transformative technology across a wide range of sectors. Many countries are increasingly adopting AI-based systems to improve public service delivery and better serve their citizens [5], [6]. The intersection of AI and credit assessment has demonstrated transformative innovation in financial intermediation, and recent academic studies highlight the capacity of AI to go beyond traditional credit scoring metrics by combining diverse and multidimensional borrower data, including behavioral patterns, transaction history, and social network information, through advanced Machine Learning (ML) models that create dynamic credit profiles to better capture borrower risk [7].

From the perspective of application scale, ML can be divided into two categories: supervised learning and unsupervised learning. Supervised learning covers a structured knowledge base containing a wide range of information and comprehensive knowledge from multiple disciplines and topics, while unsupervised learning focuses on a structured knowledge base in a specific field or topic [8]. When doing financial portfolio management, we usually start with two relationships, one involving the correlation of assets and the other the market trend [9].

This study aims to address the problem of novice investors lacking the knowledge and experience to make investment decisions, which often leads to suboptimal choices, and provides an effective way to invest, as well as learn about financial markets, using ML algorithms. The remaining structure of this study is organized as follows: Section 2 reviews past research and Section 3 presents the theoretical foundations of the study. Section 4 discusses the database and various data preprocessing methods, and Section 5 analyzes the performance of the proposed model. Section 6 concludes with the results and future suggestions.

2 | Review of Recent Works

Ban et al. [10] investigated the limitations of traditional optimization methods and examined whether ML-based models could outperform conventional approaches. Their findings indicated that certain ML techniques can improve upon traditional optimization methods by reducing prediction errors and enhancing overall performance. Chen et al. [11] demonstrated that neural networks are capable of processing large volumes of data, thereby improving predictive accuracy.

Gu et al. [12] explained in a study that ML methods have the ability to process multiple predictors and traditional methods cannot capture linear dependencies. They discovered that ML models can capture nonlinearities. In order to reduce the default risk for P2P companies, Gao and Balian [13] used a gradient boosting-based model to analyze a large number of sample data from a representative P2P industry platform. Their results showed that the proposed model achieved an accuracy of 91.36% in predicting borrower default risk, demonstrating its effectiveness for credit risk assessment in P2P lending.

In the research conducted by Hu et al. [14], an online optimization framework was used to obtain investors' preferences for further financial product recommendations. The proposed method allowed updating investor preferences for newly entered datasets and coping with the situation where there are large amounts of missing values in investors' records.

Shuryhin and Zinawatna [15] proposed a system for generating personalized financial recommendations by leveraging transaction history, predicted expenses, and anomaly detection to provide tailored recommendations. Sharma and Chiu [16] evaluated the performance of several ML algorithms for P2P lending, including RF, Logistic Regression (LR), extreme gradient boosting, Multilayer Perceptron (MLP), and k-nearest neighbors, with the aim of improving borrower default risk prediction. Their results showed that the RF algorithm achieved the highest performance, attaining an accuracy of 91%.

3 | Theoretical Foundations of Research

3.1 | Artificial Intelligence

AI encompasses various research and technology areas such as big data, natural language processing, ML, data mining, and deep learning [17]. One of the most widely used techniques in data mining is the creation of systems that use classifiers [18] to manage large amounts of information and can be used to make assumptions about the names of the classified classes, class labels, and classifications of newly obtained data [19].

3.2 | Machine Learning

ML is a subset of AI that allows computers to analyze and interpret data without explicit programming. As input data increases, the algorithm learns and improves its performance and accuracy [20]. The following are the ML algorithms used for classification in this study.

3.2.1 | Multilayer perceptron

The MLP classifier is a feed-forward artificial neural network composed of an input layer, one or more hidden layers, and an output layer. The hidden layers, located between the input and output layers, enable the network to learn complex nonlinear relationships from the data. The computational complexity of an MLP depends largely on its architecture, particularly the number of hidden layers and neurons. As the number of layers' increases, the computational cost and training time generally increase accordingly.

According to *Eq. (1)*, the bias is represented by (b) and the weight by (w), and each neuron accepts inputs such as x_1, x_2, x_3, …, x_n. Also, the input is multiplied by the weight and the output is produced, which may be represented by (ϕ) based on the activation function [21].

$$y = \varphi\left(\sum_{i=1}^{n} \left(\left(w_i . x_i \right) + b \right) \right). \quad \textbf{(1)}$$

3.2.2 | Decision tree

A Decision Tree (DT) is an algorithm for classification and regression that consists of multiple branches, leaf nodes, and root nodes. This algorithm creates a tree-like structure by classifying samples and using a recursive partitioning algorithm, with class labels represented by a leaf node and branches representing test results. When a DT is trained on a training dataset, the nodes above are identified as important variables in that dataset, and feature selection is performed automatically. This algorithm is insensitive to missing values and outliers, and missing data cannot prevent the data from being split to build a DT [22].

3.2.3 | Random forest

In the RF algorithm, a set of predictors is constructed with several DTs that are expanded in randomly selected data subspaces. In this method, the output of all existing DTs is combined and the final prediction is produced. The final prediction of the algorithm is obtained either by sampling the results of each tree or by selecting one of the predictions that is most repeated among the trees [23]. According to *Eqs. (2)* and *(3)*, the Gini index or the entropy index can be used to construct nodes in the trees, which indicates the best combination of data with the maximum distance between its branches [24].

$$\text{Gini} = 1 - \sum_{i=1}^{c} (p_i)^2, \quad (2)$$

$$\text{Entropy} = \sum_{i=1}^{c} -p_i * \log_2 p_i, \quad (3)$$

In the above equation, pi is the relative frequency (or proportion) of the observed class and c represents the number of classes in the data set. The final result of the RF can be expressed by *Eq. (4)*.

$$RF = \arg\max_{j\in\{1,2,\ldots,c\}} \sum_{i=1}^{i} DT_{i,j}. \quad (4)$$

In this equation, the argmax function identifies the class with the highest score (i.e., the majority vote). The index (i) denotes the DT, ranging from the first tree to the (i)-th tree, while (j) represents the class index in the output feature space.

3.2.4 | Support vector machines

SVM is a supervised algorithm used in classification and regression problems, specifically designed for text classification problems where only a handful of labeled examples are needed, and is gaining traction in several biological fields. In this algorithm, the dataset trains the SVM to support new data, and it is useful because it allows us to consider more complex relationships between our data points while still using a classification algorithm [25].

3.2.5 | Logistic regression

One of the simplest and most widely used ML classification models is LR, which is a computationally efficient method that calculates the probability of a sample belonging to class 0 or 1 [26]. In the LR method, a linear combination of the input features is first constructed. The LR model is then formulated as shown in *Eq. (5)*, where the regression coefficient, (β), represents the logit [27].

$$\text{logit}(Y) = \ln\frac{\pi}{\pi - 1} = \alpha + \beta x. \quad (5)$$

3.2.6 | Light gradient boosting machine

Light Gradient Boosting Machine (LGBM) is a distributed ML framework based on gradient boosting that employs tree-based learning algorithms. It was originally developed by Microsoft. The LGBM algorithm adopts a leaf-wise tree growth strategy, in which the tree is expanded from the leaf that yields the greatest reduction in the loss function. This approach typically achieves a larger decrease in loss than level-wise (depth-wise) tree growth strategies used by many conventional boosting algorithms, resulting in improved predictive accuracy [28].

4 | Research Data and Preprocessing

This study is based on a real-world dataset of loans obtained from the Lending Club platform over a six-month period, with each loan having a repayment term of 36 months. During the data preprocessing stage, records containing missing, incomplete, or outlier values were removed, and categorical variables were encoded into numerical values. For example, fully paid loans were labeled as 0, whereas charged off loans were labeled as 1.

The research variables include loan status, loan amount requested by the borrower, loan interest rate paid by the borrower, annual installments, grade, subcategory, borrower's employment tenure, homeownership status, annual income, loan purpose, number of open lines of credit by the borrower, total number of lines of credit in the borrower's credit file, number of credit applications filed by the borrower in the past 6 months, number of open lines of credit during the borrower's credit life, amount of credit used by the borrower, total outstanding credit, debt-to-income ratio (high and low), number of months since the borrower's last outstanding debt, and total amount paid. The histogram of the research variables is shown in *Fig. 1*.

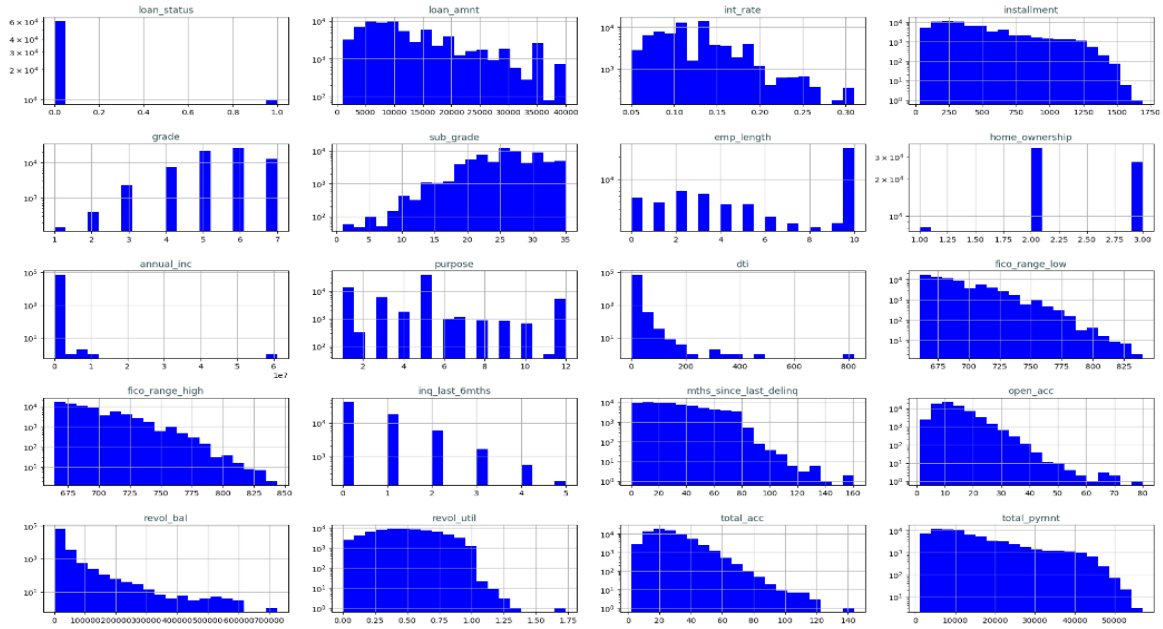

Fig. 1. Histogram of research variables.

These histograms provide a comprehensive overview of the distributions of the variables included in the proposed model, illustrating their central tendencies, dispersion, skewness, and typical value ranges. According to *Eq. (6)* to calculate the correlation between two features from the correlation relationship, the correlation between two features is represented by Corr, the correlation coefficient by E, the covariance between two features by COV, the mathematical expected value by E, and the mean of a feature in this context by μ.

$$\mathrm{Corr}(x,y)=\frac{\mathrm{Corr}(x,y)}{p_x p_y}=\frac{E\left[\left(x-\mu_x\right)\times\left(y-\mu_y\right)\right]}{p_x p_y}. \tag{6}$$

In *Fig. 2*, the heat map used to express the degree of correlation between different variables using the correlation method clearly identifies which financial and credit variables are highly correlated, highlighting the inverse relationship between credit rating and interest rate.

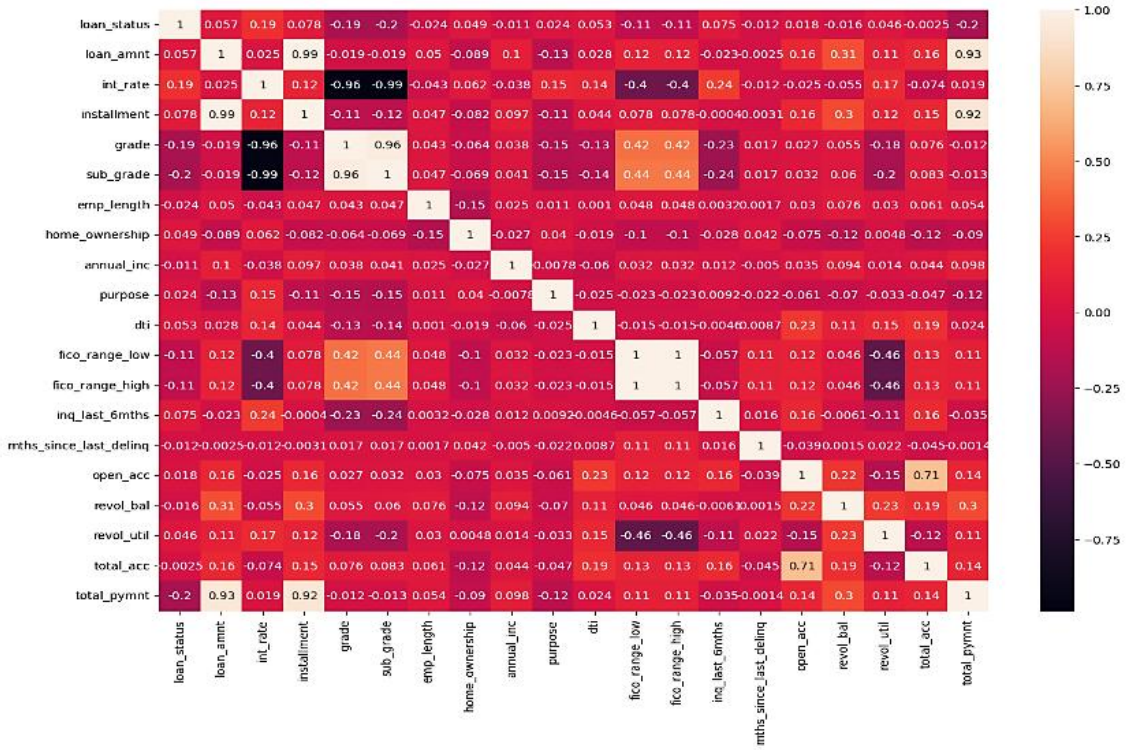

Fig. 2. Heatmap of correlation between features.

The heat map indicates that most pairs of variables exhibit low absolute correlation coefficients (i.e., values close to zero), suggesting weak or negligible linear relationships among the majority of the variables in the dataset. In the selected dataset, the number of fully paid versus failed loans is as shown in *Table 1*.

Table 1. Number of fully paid and failed loans.

Loan class	Loan type	Quantity	Percentage
0	Fully paid	60715	85.9985
1	Charged off	9885	14.0014

According to the information in the table, the number of paid loans is much higher than the number of failed loans, which causes class imbalance in the data set and balancing must be done. *Fig. 3* shows a boxplot of the distribution of repaid and defaulted loans. This plot allows for comparison of descriptive indicators such as median, interquartile range, data dispersion, and outliers between the two groups, and provides a better understanding of the difference in the distribution of variables. As can be seen in the figure, the median loan amount in the failed loan group is higher than that in the repaid loans. Also, the dispersion of the data in this group is higher and outliers are observed in both groups, indicating the relative difference in the distribution of loan amounts between the two repayment statuses.

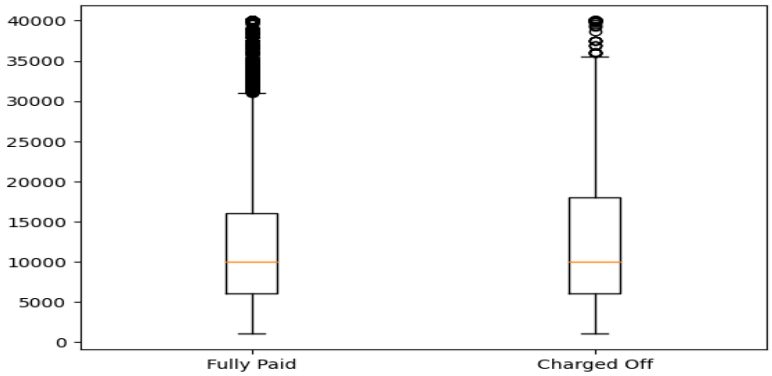

Fig. 3. Box plot of loan distribution by loan type.

As the scale of the dataset increases, class imbalance often becomes more pronounced, making accurate classification more challenging. To improve the performance of the analysis and predictive modeling, it is essential to address class imbalance through appropriate balancing techniques. Oversampling is a widely used data augmentation approach for balancing imbalanced datasets. It was proposed by Chawla et al. to overcome the overfitting problems associated with traditional random oversampling methods [29].

4.1 | Balancing

The processed data should be homogeneous, have a standard structure, and be coherent. Adherence to these principles allows for proper data integration and enables the data to be used in a manner consistent with the analysis, modeling, and interpretation processes. In this study, standardization according to *Eq. (7)* was used to normalize the data distribution or eliminate differences between data scales. According to this equation, the range of the data is between [1,0]. X_{std} represents the standard score for X_i, μ is the mean of the index, and δ is the standard deviation of the index for the data, and the goal of standardization is to obtain values with a mean of zero and a variance or standard deviation of 1.

$$x_{std} = \frac{x_i - \mu}{\delta}. \quad (7)$$

4.2 | Implementing the Classification Model

After performing the preprocessing steps, the dataset is divided into two subsets, training and testing, with proportions of 70 and 30 percent. The training set is used as the main input to ML algorithms and is used to extract hidden patterns, adjust model weights, and optimize its parameters. In contrast, the test set, which does not participate in the model training process, is used to evaluate the model's ability to predict new samples, measure generalizability, and prevent overfitting. This partitioning method allows for a fair assessment of model performance and a more accurate comparison of the results obtained from different classification algorithms.

Tables 2-4 present the performance of six baseline classifiers MLP, RF, DT, SVM, LR, and LGBM under three class balancing techniques: Random Under-Sampling (RUS), Random Over Sampling (ROS), and the Synthetic Minority Over Sampling Technique (SMOTE). The models are evaluated using four performance metrics: precision, accuracy, recall, and F1-score.

Table 2. Results of comparing different models with the RUS method.

Model	Precision	Accuracy	Recall	F1_score
MLP	0.2106	0.9544	0.6890	0.3154
RF	0.2107	0.9539	0.6535	0.3186
DT	0.1730	0.9182	0.5674	0.2652
SVM	0.2035	0.9546	0.6142	0.2136
LR	0.2117	0.9599	0.5450	0.3049
LGBM	0.2060	0.9621	0.6474	0.3126

Table 3. Results of comparing different models with the ROS method.

Model	Precision	Accuracy	Recall	F1_score
MLP	0.2164	0.9720	0.8042	0.3172
RF	0.2782	0.9628	0.0703	0.1122
DT	0.2000	0.9666	0.2078	0.2038
SVM	0.2058	0.9625	0.6175	0.2085
LR	0.2092	0.9615	0.5598	0.3046
LGBM	0.2138	0.9701	0.6072	0.3163

Table 4. Results of comparing different models with the SMOTE method.

Model	Precision	Accuracy	Recall	F1_score
MLP	0.2256	0.9849	0.9192	0.2492
RF	0.1314	0.9731	0.0514	0.0879
DT	0.1868	0.9532	0.2302	0.2062
SVM	0.2235	0.9548	0.8974	0.2080
LR	0.2109	0.9640	0.5888	0.3106
LGBM	0.1845	0.9536	0.2297	0.2046

According to the results presented in the tables, applying the SMOTE sampling technique during the training of the MLP model resulted in the best overall performance among the classification models, achieving the highest scores on three of the four evaluation metrics. Therefore, considering its superior statistical performance and predictive capability, the SMOTE-based MLP model was selected as the optimal model for the loan repayment prediction task. The results of the feature importance analysis using the MLP model are presented in *Fig. 4*.

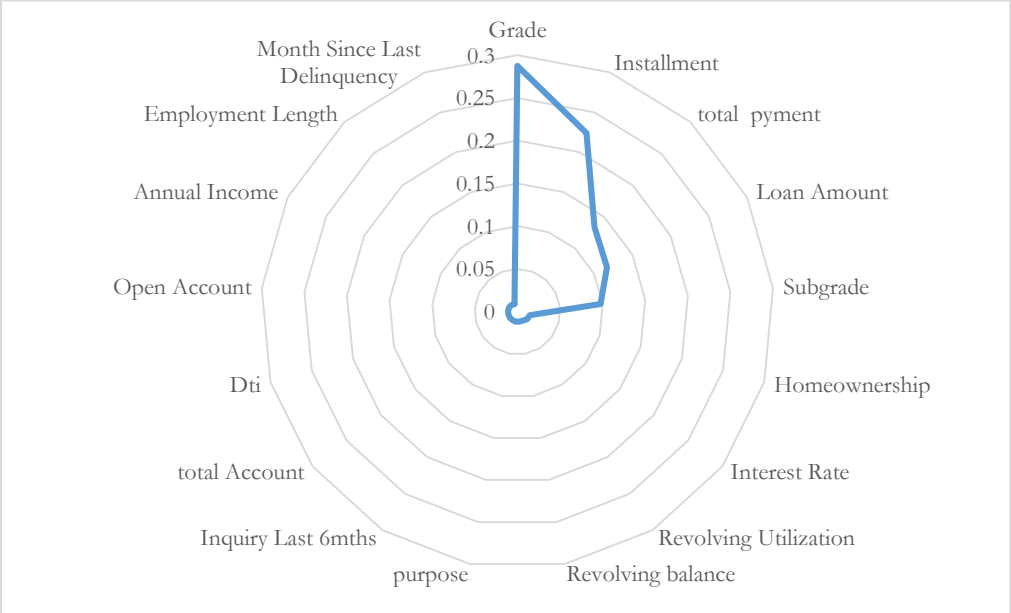

Fig. 4. Feature importance using the SMOTE technique and the MLP model.

As shown in figure, the input variables did not contribute equally to the performance of the classification model. Among the features examined, credit grade, number of installments, total amount repaid, loan amount, and credit sub-grade had the greatest influence on the model's predictive performance. These findings suggest that emphasizing these key features can reduce model complexity, improve computational efficiency, and enhance the model's generalizability in predicting loan repayment status.

5 | Performance Analysis with Receiver Operating Characteristic Curve

The Receiver Operating Characteristic (ROC) curve is a widely used metric for evaluating the performance of binary classification models. It measures a classifier's ability to distinguish between two classes by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) across different classification thresholds. The curve plots the TPR on the y-axis versus the FPR on the x-axis at different classification thresholds. They consider all potential thresholds and show the initial slope of the ROC curve, which shows how quickly performance degrades as the prediction dataset gets smaller.

This measure ranges from zero to one, such that an Area Under the Curve (AUC) value of 1 indicates perfect discrimination power and ideal performance of the model in classification, while a value of 0.5 indicates the

absence of significant discrimination ability and performance equivalent to random classification. The ROC curve and the results of evaluating the performance of the selected model based on this criterion are presented in *Fig. 5*.

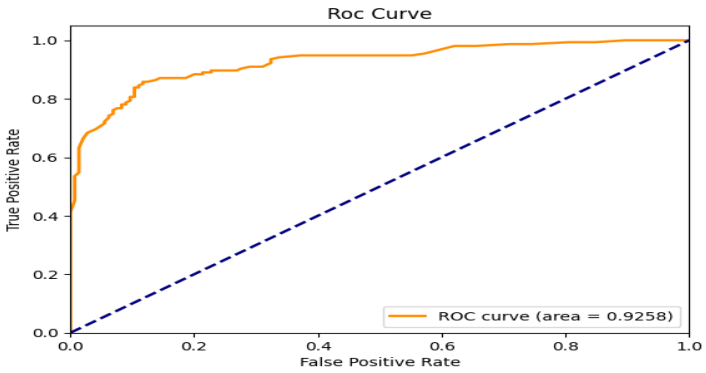

Fig. 5. ROC curve for evaluating model performance.

The diagonal line in this figure (TPR = FPR) shows the expected performance of a random classifier, and a classifier with a curve closer to the upper left corner of the graph is usually preferred. The dotted line in the figure corresponds to a classifier with performance comparable to a purely random classifier. A higher area under the ROC curve (AUC) indicates a better ability of the classifier to distinguish between the "fully paid" and "charged off" loan classes. As shown in *Fig. 5*, the proposed model achieved an AUC of 0.925, demonstrating well discriminative performance.

6 | Results and Future Suggestions

In today's world, due to the digitization of businesses, emerging lending platforms have become very popular with investors looking for new opportunities for growth and profitability. P2P lending platforms are a global lending marketplace for connecting borrowers and lenders using web applications that are accessible to everyone, regardless of their location or financial status. When it comes to AI and ML, the problem of unbalanced data that leads to inefficient classification must be solved in relation to the information recorded in this platform because it affects people's decision-making. ML, as one of the transformative technologies, plays an important role in data analysis and the development of intelligent systems. The use of ML algorithms can increase the accuracy of predictions and improve the decision-making process.

The aim of the present study is to predict and identify the factors affecting P2P lending using the Lending Club platform dataset. For this purpose, supervised learning models including MLP, RF, DT, SVM, LR and LGBM using RUS, ROS and SMOTE sampling techniques have been compared with each other in order to improve the performance of the models in the face of unbalanced data and the Python programming language. The results of the experiments show that unbalanced data is an important issue and the variables considered in this study using the MLP method and SMOTE sampling have greater predictive power than previous studies.

Future research could focus on mitigating the effects of class imbalance through the use of hybrid sampling techniques. In addition, the accuracy, robustness, and generalizability of predictive models may be further enhanced by employing more advanced deep learning architectures or hybrid models that integrate deep learning with traditional ML techniques. The adoption of these approaches, together with hyperparameter optimization and effective feature selection, has the potential to improve the accuracy of loan repayment and default prediction. Ultimately, these enhancements can contribute to more reliable credit scoring and more effective credit risk management systems.

Authors' Contributions

All aspects of the research and manuscript preparation were carried out by the author. The author has read and approved the final version of the manuscript.

Funding

This research has not received funding from the government, commercial, or non-profit sectors.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no competing interests.

Consent for Publication

The author confirms consent for the publication of this work

Ethics Approval and Consent to Participate

This article does not contain any studies with human participants performed by the author.

References

[1] Vallee, B., & Zeng, Y. (2019). Marketplace lending: A new banking paradigm? *The review of financial studies*, *32*(5), 1939–1982. https://doi.org/10.1093/rfs/hhy100

[2] Huang, C. H., & Nguyen, V. T. (2026). Revisiting the shifting landscape of P2P lending: A systematic review based on the affordance actualization perspective. *Electronic commerce research*, *26*(3), 2973–2999. https://doi.org/10.1007/s10660-026-10110-x

[3] Cummins, M., Lynn, T., an Bhaird, C., & Rosati, P. (2018). Addressing information asymmetries in online peer-to-peer lending. In *Disrupting finance: Fintech and strategy in the 21st century* (pp. 15–31). Springer. https://doi.org/10.1007/978-3-030-02330-0_2

[4] Silva, R. R., Garcia, L. A. A., Buso, A. L. Z., de Paula, F. F. S., Marques, D. S., & da Silva Santos, Á. (2022). Novas listas e novas ferramentas tecnológicas sobre medicamentos potencialmente inapropriados para idosos: Uma revisão integrativa. *Revista família, ciclos de vida e saúde no contexto social*, *10*(2), 340-369. https://doi.org/10.18554/refacs.v10i2.5970

[5] Wu, M. E., Syu, J. H., Lin, J. C. W., & Ho, J. M. (2021). Portfolio management system in equity market neutral using reinforcement learning. *Applied Intelligence*, *51*(11), 8119-8131. https://doi.org/10.1007/s10489-021-02262-0

[6] Wang, Y. F., Chen, Y. C., & Chien, S. Y. (2025). Citizens' intention to follow recommendations from a government-supported AI-enabled system. *Public policy and administration*, *40*(2), 372-400. https://doi.org/10.1177/09520767231176126

[7] Hafez, I. Y., Hafez, A. Y., Saleh, A., Abd El-Mageed, A. A., & Abohany, A. A. (2025). A systematic review of AI-enhanced techniques in credit card fraud detection. *Journal of big data*, *12*(1), 6. https://doi.org/10.1186/s40537-024-01048-8

[8] Kim, W., Kanezaki, A., & Tanaka, M. (2020). Unsupervised learning of image segmentation based on differentiable feature clustering. *IEEE transactions on image processing*, *29*, 8055–8068. https://doi.org/10.1109/TIP.2020.3011269

[9] Zakhidov, G. (2024). Economic indicators: Tools for analyzing market trends and predicting future performance. *International multidisciplinary journal of universal scientific prospectives*, *2*(3), 23–29. https://www.researchgate.net/publication/380519295

[10] Ban, G. Y., El Karoui, N., & Lim, A. E. B. (2018). Machine learning and portfolio optimization. *Management science*, *64*(3), 1136–1154. https://doi.org/10.1287/mnsc.2016.2644

[11] Chen, L., Pelger, M., & Zhu, J. (2024). Deep learning in asset pricing. *Management Science*, *70*(2), 714–750. https://doi.org/10.1287/mnsc.2023.4695

[12] Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The review of financial studies*, *33*(5), 2223–2273. https://doi.org/10.1093/rfs/hhaa009

[13] Gao, B., & Balyan, V. (2022). Construction of a financial default risk prediction model based on the LightGBM algorithm. *Journal of intelligent systems*, *31*(1), 767–779. https://doi.org/10.1515/jisys-2022-0036

[14] Hu, X., Chen, Y., Ren, L., & Xu, Z. (2023). Investor preference analysis: An online optimization approach with missing information. *Information sciences*, *633*, 27–40. https://doi.org/10.1016/j.ins.2023.03.066

[15] Shuryhin, K. A., & Zinovatna, S. L. (2024). Recommendation system for financial decision-making using artificial intelligence. *Applied aspects of information technology*, *7*(4), 348–358. https://doi.org/10.15276/aait.07.2024.24

[16] Sharma, A. K., & Chiu, K. C. (2026). Data-driven prediction of borrower default in P2P lending using feature-optimized ML models. *OPSEARCH*, 1–24. https://doi.org/10.1007/s12597-026-01166-2

[17] Najem, R., Amr, M. F., Bahnasse, A., & Talea, M. (2022). Artificial intelligence for digital finance, axes and techniques. *Procedia computer science*, *203*, 633–638. https://doi.org/10.1016/j.procs.2022.07.092

[18] Kumar, R., & Verma, R. (2012). Classification algorithms for data mining: A survey. *International journal of innovations in engineering and technology (IJIET)*, *1*(2), 7–14. https://d1wqtxts1xzle7.cloudfront.net/32141572/data-libre.pdf

[19] Nikam, S. S. (2015). A comparative study of classification techniques in data mining algorithms. *Oriental journal of computer science and technology*, *8*(1), 13–19. https://www.computerscijournal.org/pdf/vol8no1/vol_8_no_01_13-19.pdf

[20] Haleem, A., Javaid, M., Qadri, M. A., Singh, R. P., & Suman, R. (2022). Artificial intelligence (AI) applications for marketing: A literature-based study. *International journal of intelligent networks*, *3*, 119–132. https://doi.org/10.1016/j.ijin.2022.08.005

[21] Odeh, A. J., Keshta, I., & Abdelfattah, E. (2020). Efficient detection of phishing websites using multilayer perceptron. *International journal of interactive mobile technologies (IJIM), 14*(11), 22–31. https://doi.org/10.3991/ijim.v14i11.13903

[22] Madane, N., & Nanda, S. (2019). Loan prediction analysis using decision tree. *Journal of the gujarat research society*, *21*(14), 214–221. https://doi.org/10.1088/1757-899X/1022/1/012042

[23] Addo, P. M., Guegan, D., & Hassani, B. (2018). Credit risk analysis using machine and deep learning models. *Risks*, *6*(2), 38. https://doi.org/10.3390/risks6020038

[24] Flach, P. (2012). *Machine learning: The art and science of algorithms that make sense of data*. Cambridge University Press. https://www.amazon.com/Machine-Learning-Science-Algorithms-Sense/dp/1107422221

[25] Noble, W. S. (2006). What is a support vector machine? *Nature biotechnology*, *24*(12), 1565–1567. https://doi.org/10.1038/nbt1206-1565

[26] Serrano-Cinca, C., & Gutiérrez-Nieto, B. (2016). The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. *Decision support systems*, *89*, 113-122. https://doi.org/10.1016/j.dss.2016.06.014

[27] Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The journal of educational research*, *96*(1), 3-14. https://doi.org/10.1080/00220670209598786

[28] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*. https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[29] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of artificial intelligence research, 16*, 321–357. https://doi.org/10.1613/jair.953