ORIGINAL ARTICLE

An innovative approach for predicting default risk in peer-to-peer lending using stacking ensemble models with explainable machine learning

Markus Atef · Menna Ibrahim Gabr · Wafaa Seoud · Shimaa Ouf

Received: 4 December 2025 / Accepted: 11 May 2026

Abstract
Peer-to-peer (P2P) lending has increased significantly during the past few years on a global scale. However, there are several challenges associated with P2P lending's rapid rise. The major challenges are imbalanced datasets, which make machine learning difficult, an excessive number of features, and low-performing classification algorithms. Furthermore, machine learning models face another complex challenge referred to as the black-box problem. To address these challenges, an innovative approach was developed by first applying Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance in the Bondora dataset, followed by the implementation of multiple feature selection techniques: Chi-Square (filter), Sequential Backward Selection (SBS) (wrapper), and embedded methods such as Random Forest (RF), Gradient Boosting Machine (GBM), Light Gradient Boosting Machine (LightGBM), and Categorical Boosting (CatBoost). A range of classifiers, linear (Logistic Regression (LR)), non-linear (Support Vector Machine (SVM), Naive Bayes (NB), and tree-based models (Decision Tree (DT), RF, Adaptive Boosting (AdaBoost), CatBoost), were then used to predict loan defaults. The top-performing models were integrated into various stacking ensembles using GBM, Extreme Gradient Boosting (XGBoost), and LightGBM as meta-learners to enhance predictive accuracy. The results declared that LightGBM exhibited an outstanding performance with accuracy, F-score, and Area Under the Curve (AUC) values of 0.981, 0.980, and 0.994, respectively, showing better performance than that reported in the literature. Explainable models were employed to interpret predictions and enhance user trust. Specifically, the LightGBM stacking model was combined with the Local Interpretable Model-agnostic Explanations (LIME) framework to provide interpretable insights into its prediction results.

Keywords P2P lending · Loan default risk · Feature selection · Prediction algorithms · Stacking models · Explainable machine learning models

1 Introduction

Online peer-to-peer (P2P) lending is a process where individuals borrow and lend money directly to each other, bypassing traditional financial institutions like banks. This typically occurs through online platforms that link borrowers with investors (lenders). P2P has grown rapidly in popularity with the growth of internet finance.

Published online: 16 June 2026

However, the extraordinarily high rate of defaulted loans has resulted in significant loss to P2P platform operators and investors due to the absence of a perfect credit risk prediction approach. The main challenges with default risk prediction in P2P lending system are imbalanced datasets, which make machine learning difficult, an excessive number of features, and low-performing classification algorithms [1–3]. Thus, a novel approach for the management of credit risk prediction and effectively direct investors toward future investments in P2P lending is needed.

Although several studies [1–8] have explored the application of machine learning in P2P lending platforms, none have comprehensively integrated dataset balancing techniques, feature selection methods, individual machine learning models, and stacking ensemble approaches with explainable machine learning to predict loan defaults. Furthermore, the absence of a systematic performance comparison across these methods highlights the need for research that holistically addresses these challenges and advances the state of the art in P2P lending default prediction. Enhancing the performance of the classification algorithms for perfect credit risk prediction can be obtained by stacking models, where several classification algorithms are combined with meta learner to create a final model having accurate prediction [1, 2].

The term Explainable Artificial Intelligence (XAI) describes a set of methods and techniques that help human users comprehend the decisions and predictions made by artificial intelligence (AI) systems [9–12].

A careful inspection of the open literature indicates that P2P lending risk default predictions using dataset balancing, feature selection methods, machine learning models and stacking models with meta-learning have received very little attention [1–8].

The studied dataset by Kun et al. [5] was lending dataset from of the fourth quarter of 2018 on the Lending Club platform in the United States. To extract the key features, the Spearman-Boruta algorithm and information value (IV) algorithm were combined. Then, in predicting loan defaults, the stacking algorithm was used to integrate four basic classifiers: Artificial Neural Network (ANN), RF, AdaBoost, and XGBoost. The outcomes demonstrate that the stacking outperforms the single classifier in terms of total performance. The stacking model can successfully lower the error rate of incorrectly classifying defaulting borrowers as non-defaulting borrowers. Trivedi [8] investigated the combinations of feature selection methods and machine learning classifiers. He concludes that the optimum performance may be achieved while using Chi-Square for feature selection in conjunction with RF for data categorization. Siham et al. [7] picked features with the help of RF and GBDT and then classified the Australian credit dataset with the assistance of RF and AdaBoost. By using the selected FS methods with RF and AdaBoost classifiers, the accuracy of the classification increases. To predict the credit risk of borrowers on P2P lending, Munsarif et al. [6] suggested stacking with feature selection based on embedded techniques. The stacking model was constructed from a stack of meta-learners used in feature selection. The authors have stated that by using the stacking model and the right feature selection, the credit risk prediction for P2P lending could be improved. By balancing the data, applying selection features, and combining stacking model with the meta-learner, Much et al. [1] presented research aimed at increasing default risk prediction's accuracy. The authors used two datasets: the P2P lending club loan dataset and online P2P lending dataset. Three base-learner algorithms, namely KNN, SVM, and RF, were combined with the XGBoost meta-learner algorithm to build a model of stacking learning. The evaluation findings demonstrate that for both datasets, stacking XGBoost is the model that performs the best. Yin et al. [2] attempted to predict credit default probabilities for P2P lending by adopting machine learning techniques using 126,090 P2P loan transactions from RenRenDai, one of the largest online P2P websites in China. The authors suggested a stacking model for assessing the risk of loan default for P2P lending platforms. K-means clustering was used to exclude features that are not relevant after feature selection using the MRMR approach. Finally, the stacking model was developed to yield reliable predictions in the feature subset. Experimental findings indicate that stacking models produces great performance in terms of prediction accuracy as well as precision and recall. The stacking model has a lower error rate and gives more accurate predictions of

the probability of credit default than single classifiers[1]. The results also support the effectiveness of the suggested stacking model using the area under the ROC curve.

Yang et al. [13] conducted a study using data from Lending Club with the goal of improving default loan prediction in P2P lending. Stepwise and variance threshold approaches were both used in the feature selection. The class imbalance was addressed by using random undersampling. Several machine learning models, including LR, DT, KNN, RF, Gradient Boosting, Light GBM and XGBoost, were employed. Additionally, stacking and bagging techniques were used. The accuracy and AUC of the stacking model (XGB) were 0.85 and 0.91, respectively.

Using loan data from Lending Club, Akinjole et al. [14] performed a comparative analysis utilising RF, DT, SVMs, XGBoost, Adaboost, and MLP to predict loan defaults. SMOTE + ENN was used to address the class imbalance problem. Recursive Feature Elimination with Cross-Validation (RFECV) was employed to select features. To further improve performance, the stacking method was used. With an accuracy of 93.7%, precision of 95.6%, and recall of 95.5%, the suggested stacking model demonstrated the potential of stacking approaches to enhance loan default predictions.

To the best of our knowledge, no prior study has conducted a systematic performance comparison that integrates dataset balancing, diverse feature selection techniques, individual machine learning models, and stacking ensemble models incorporating meta-learning for P2P lending default prediction. Moreover, explainable machine learning has yet to be applied in this context using stacking ensemble models. This study addresses these gaps by introducing a novel approach that combines all these components to significantly improve prediction performance. Additionally, we propose a systematic framework to evaluate the interpretability of the LIME method within P2P loan default prediction. By comparing features identified through traditional feature selection methods with those highlighted by the best-performing stacking model enhanced with LIME, our study offers deeper insights into model behavior and decision-making. This integrated approach contributes to advancing both the interpretability of machine learning in financial contexts and the development of transparent, accountable AI systems.

2 Research framework

The research framework for this study is structured through several key stages, beginning with data collection and culminating in explainable ensemble modelling, Fig. 1. The dataset was obtained from Bondora, a leading peer-to-peer lending platform, encompassing loan characteristics and borrower transaction histories from January 2020 to January 2024. Loans with final status “Defaulted” or “Charged off” were labelled as default, while “Fully Paid” loans were labelled as non-default. Data cleaning was performed using explicit criteria, where duplicated records, entries with missing values in key features, logically inconsistent entries and loans with incomplete outcomes were removed. This process excluded 23,489 records (11.2% of the original dataset), resulting in a refined dataset of 186,111 records and 16 features.

These features were selected based on their proven significance in the literature for predicting loan defaults in P2P lending [1–14].

SMOTE was used to address class imbalance by generating synthetic minority-class samples. Unlike traditional oversampling, it preserves the dataset while enriching the feature space for a more balanced distribution. SMOTE-ENN extends this by removing noisy samples but is computationally intensive and may discard important boundary points. Random oversampling risks overfitting, while ADASYN can introduce noise by overemphasizing hard-to-learn or outlier samples. Undersampling methods like NearMiss or Tomek Links may discard useful majority-class data, and GANs, though capable of high-quality synthetic data, require complex training [1].

[1] While the model outputs are expressed as probabilities of default, they are primarily intended for ranking and classification rather than precise probability calibration. Calibration is evaluated as a complementary measure of reliability.

 https://doi.org/10.1007/s00521-026-12226-5

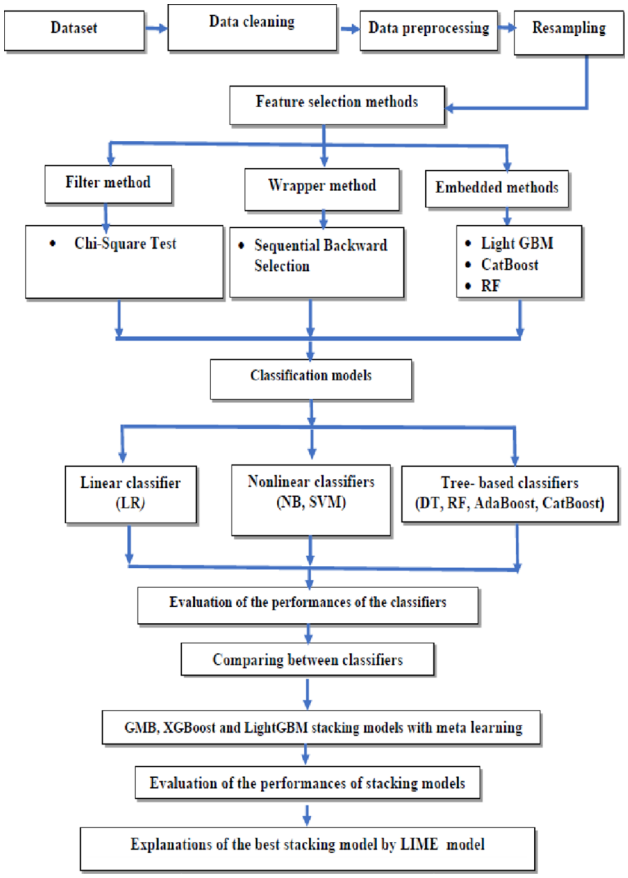

Fig. 1 Framework of the research

Stratified sampling was used to split the dataset into training and test sets. SMOTE was applied to the training data, and all feature selection and model training steps were performed on the balanced training set.

To identify the most predictive features, a hybrid feature selection strategy was employed. This included filter-based Chi-square tests, wrapper-based Sequential Backward Selection (SBS), and embedded methods such as LightGBM, Random Forest, and CatBoost. By integrating multiple approaches, the study ensured a robust and efficient feature subset that balances predictive power and computational cost.

The study explored a variety of classification algorithms grouped into linear (LR), non-linear (SVM and NB), and tree-based models (DT, RF, AdaBoost, CatBoost). This diversity enabled a comprehensive assessment of different modelling techniques for loan default prediction.

Threshold-dependent metrics (accuracy, F1-score, sensitivity, specificity) were computed using a default decision threshold of 0.5.

Building upon individual classifiers, stacking ensemble models using gradient boosting methods (GBM, XGBoost, and LightGBM) were developed. These models combined base learners' predictions into a meta-learner, enhancing overall accuracy by leveraging complementary strengths from diverse classifiers through a meta-learning framework.

The hyperparameters of all models were tuned using random search with 10-fold cross-validation. Each model was assigned a literature-based search space, and the best settings were selected according to mean cross-validation performance. The final models were then retrained with the optimal parameters on the full training data before being evaluated on the test set.

To assess the reliability of predicted default probabilities, calibration was evaluated alongside discrimination metrics. The Brier score was computed as the mean squared error between predicted probabilities and observed outcomes, providing a strictly proper scoring rule. The Expected Calibration Error (ECE) was estimated by

dividing predictions into equal-width bins and calculating the weighted average gap between predicted and empirical probabilities. Calibration curves (reliability diagrams) were also used to visualize deviations from perfect calibration.

Finally, to enhance model interpretability, LIME technique was applied to the best stacking model. LIME provides consistent and locally faithful explanations of complex, black-box models, effectively handling feature interactions and increasing transparency in the prediction process, which is crucial for practical deployment in financial decision-making [15–19]. Global feature importance is obtained by aggregating local explanations using the mean absolute feature weights across instances. We additionally perform a stability check by repeating the explanation process multiple times and comparing the resulting feature rankings.

3 Results and discussion

3.1 Feature selection techniques

In this study, a diverse range of feature selection techniques, encompassing filter (Chi-Square test), wrapper (Sequential backward selection (SBS)), and embedded (LightGBM, RF and CatBoost) methods were employed to identify the most influential features [20–24].

The Chi-Square technique feature selection findings are illustrated in Fig. 2. The figure displays the importance score percentage of the features, shown in descending order of their rankings. The decrease in the proportion of the importance scores of the features indicates the level of confidence in the technique. From Fig. 2, it is evident that the important score percentage of the features declines rapidly at the beginning. Subsequently, the differences in important score percentage of the features cannot be deemed noteworthy. The Chi-Square technique ranks education as the most prominent feature in the dataset. The following crucial features are homeownership type, employment duration current employer, interest, gender, verification type, applied amount, age and loan duration. Moreover, the score for feature importance declines rapidly, indicating that only a small number of significant features contribute to the default.

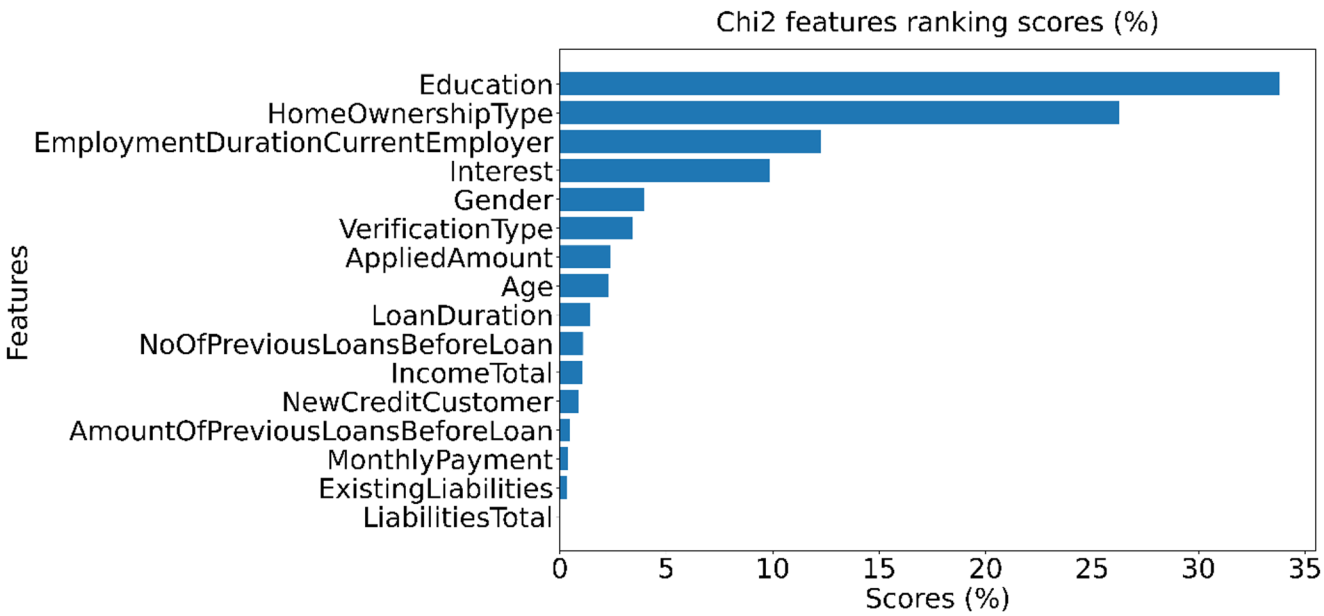

Fig. 2 Feature importance scores of Chi-square method

The sequential backward selection method is an iterative procedure that operates in the opposite direction of the forward selection method. This strategy initiates the process by considering all the features and subsequently eliminates the one that holds the least significance. This iterative process of elimination continues until the removal of features from the model no longer leads to an enhancement in its overall performance [20]. The top ten features suggested by SBS algorithm are age, interest, loan duration, monthly payment, income total, existing liabilities, amount of previous loans before loan, new credit customer, gender and home ownership type.

Figure 3 displays the outcomes of the LightGBM method. The two most important features are interest rate and income total, followed by age, amount of previous loan before the loan, applied amount, liabilities total, monthly payment and loan duration.

Figure 4 displays the most important features detected by RF approach, where interest rate and income total are the top ranked features, followed by age, liabilities total, monthly payment, applied amount and the amount of previous loan before the loan.

Figure 5 presents the results obtained using the CatBoost algorithm. As shown in Fig. 5, the top-ranked features are interest rate and total income, followed by applied amount. Other notable features include age, previous loan amount, total liabilities and loan duration.

The analysis of feature selection methods reveals a high degree of consistency across different techniques, as most methods identified a similar subset of features, albeit with slight variations in ranking. This consistency suggests that the predictive signal in the dataset is strongly concentrated in a core group of features rather than being method dependent.

In particular, loan-related characteristics, such as applied amount, loan duration, and interest rate, were consistently selected, highlighting their dominant role in determining default risk. The universal selection of interest rate across all feature selection techniques indicates that it is the most robust predictor, reflecting its direct relationship with borrower risk and lending conditions. In addition, several borrower-specific attributes, including age, income, employment duration, and credit history (e.g., number of previous loans), were repeatedly identified, suggesting that both financial capacity and past borrowing behaviour are critical factors in risk assessment. These findings provide an important insight: effective credit risk prediction in P2P lending relies on a combination of loan characteristics and borrower financial stability indicators, rather than on any single category of features. Moreover, the stability of feature selection across multiple methods enhances the reliability and interpretability of

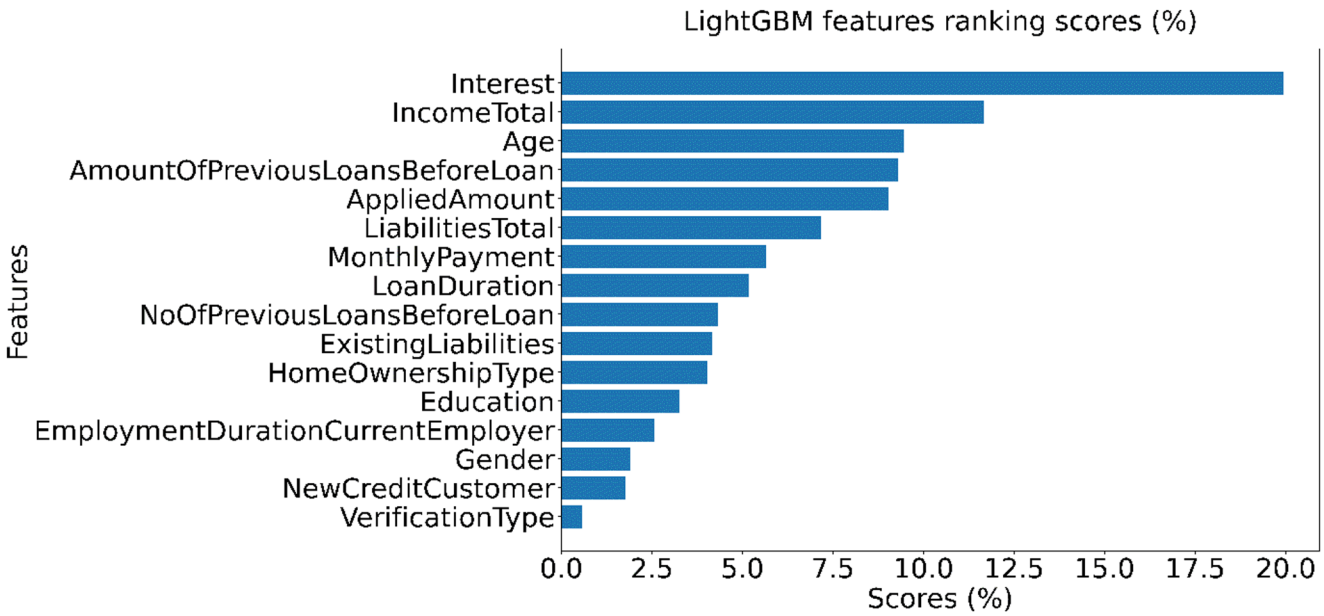

Fig. 3 Feature importance scores of LightGBM approach

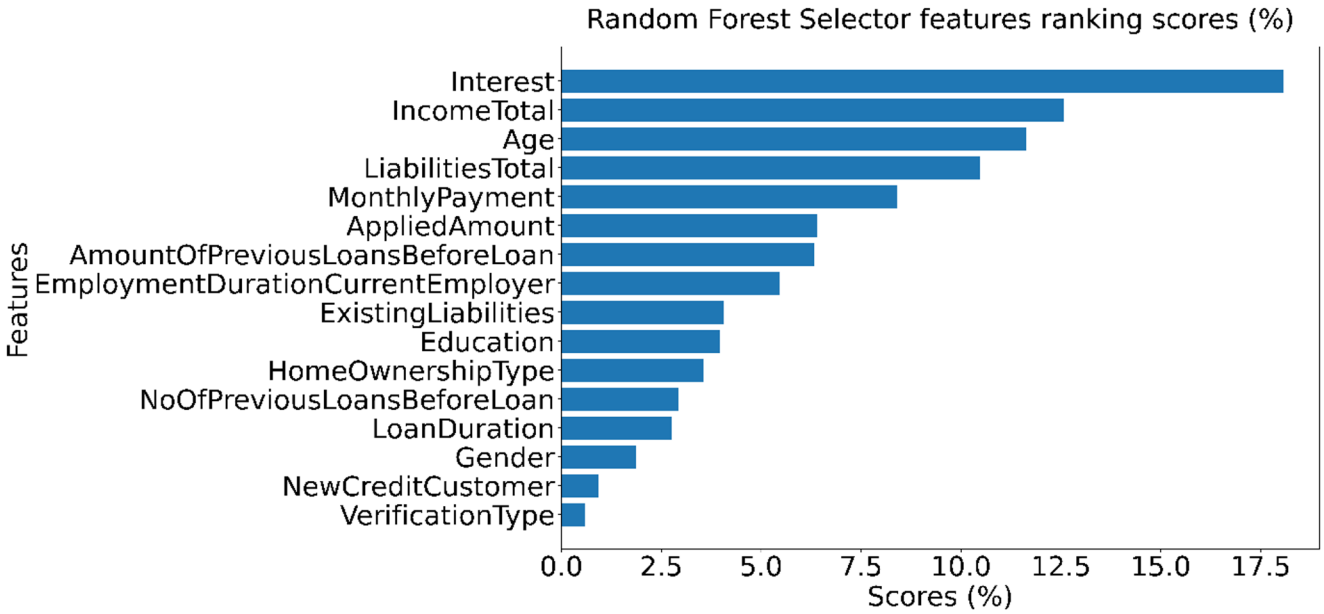

Fig. 4 Feature importance scores of random forest technique

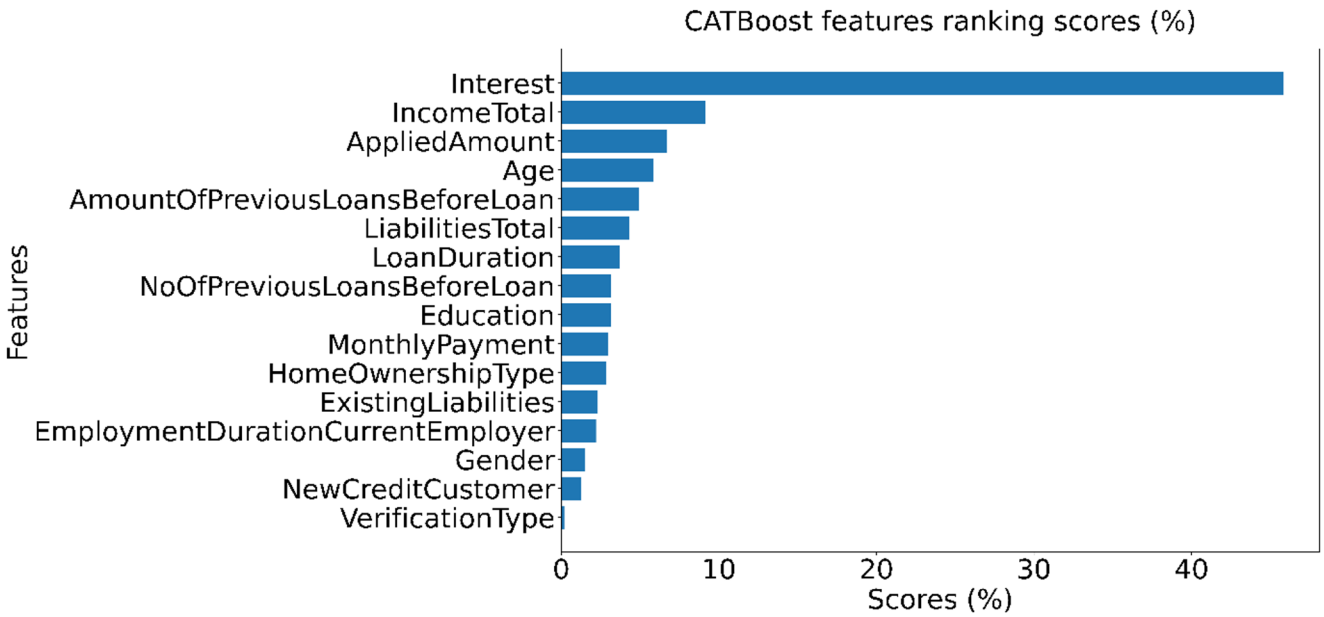

Fig. 5 Feature importance scores of CatBoost approach

the model. Based on this analysis, a subset of ten consistently selected features was used for model development, ensuring both strong predictive performance and reduced model complexity.

3.2 Results of single models

Linear model (LR), non-linear (SVM and NB), tree-based (DT, RF, AdaBoost and CatBoost) models were employed to predict the loan default [25–36].

 https://doi.org/10.1007/s00521-026-12226-5

Table 1 Prediction performance metrics of the LR, NB and SVM algorithms

Model	LR		NB		SVM	
Dataset	Imbalanced	SMOTE resampling	Imbalanced	SMOTE resampling	Imbalanced	SMOTE resampling
F-score	0.207	0.636	0.257	0.548	0.034	0.64
Precision	0.579	0.674	0.466	0.696	0.581	0.68
Sensitivity	0.126	0.601	0.178	0.451	0.017	0.609
FPR	0.022	0.291	0.049	0.195	0.003	0.287
Specificity	0.978	0.709	0.951	0.8055	0.997	0.713
FNR	0.874	0.399	0.822	0.549	0.983	0.390
ROC-AUC	0.71	0.713	0.679	0.682	0.653	0.723
Accuracy	0.812 ± 0.003	0.655 ± 0.003	0.800 ± 0.209	0.628 ± 0.006	0.806 ± 0.003	0.661 ± 0.003
KS	0.307	0.313	0.268	0.271	0.273	0.324
H-measure	0.057	0.115	0.057	0.089	0.007	0.124

Table 2 Prediction performance metrics of DT, AdaBoost CatBoost and RF algorithms

Model	DT		AdaBoost		CatBoost		RF	
Dataset	Imbalanced	SMOTE resampling	Imbalanced	SMOTE resampling	Imbalanced	SMOTE resampling	Imbalanced	SMOTE resampling
F-score	0.354	0.907	0.276	0.931	0.312	0.751	0.351	0.946
Precision	0.340	0.844	0.637	0.886	0.596	0.726	0.619	0.912
Sensitivity	0.369	0.982	0.176	0.980	0.211	0.779	0.245	0.984
FPR	0.174	0.182	0.024	0.126	0.035	0.294	0.036	0.0951
Specificity	0.826	0.818	0.976	0.874	0.965	0.706	0.964	0.905
FNR	0.631	0.018	0.824	0.0196	0.789	0.221	0.755	0.017
ROC-AUC	0.598	0.900	0.789	0.988	0.788	0.82	0.788	0.993
Accuracy	0.737 ± 0.003	0.900 ± 0.002	0.820 ± 0.003	0.927 ± 0.021	0.819 ± 0.003	0.742 ± 0.002	0.824 ± 0.003	0.944 ± 0.001
KS	0.196	0.800	0.429	0.923	0.425	0.485	0.432	0.938
H-measure	0.060	0.692	0.092	0.772	0.103	0.272	0.127	0.824

Tables 1 and 2 summarize algorithm performance on the imbalanced dataset and after SMOTE resampling across multiple evaluation metrics. The results highlight the impact of resampling on classification performance, particularly in improving sensitivity to the minority class (i.e., default prediction). For LR model, resampling significantly enhances class-sensitive metrics. While the accuracy decreases from 0.812 (imbalanced) to 0.655 (balanced), sensitivity improves markedly from 0.126 to 0.601, and precision increases from 0.579 to 0.674. The F-score, which combines precision and sensitivity, shows a substantial improvement from 0.207 to 0.636. Although specificity and false negative rate (FNR) are higher in the imbalanced dataset, the SMOTE resampling demonstrates superior Kolmogorov–Smirnov (KS) and H-measure values. Interestingly, the AUC remains constant, indicating that SMOTE resampling enhances discrimination between classes without affecting the model's overall ranking capability. The NB model shows a similar trend, with accuracy declining from 0.800 to 0.628 after resampling, but precision improving significantly from 0.466 to 0.696. Other metrics remain relatively stable, and the AUC is unaffected by the application of SMOTE resampling, suggesting limited sensitivity of the NB classifier to resampling beyond precision.

For SVM, resampling leads to a drop in accuracy from 0.806 to 0.661. Other performance metrics are improved after SMOTE resampling, except for specificity and the false negative rate (FNR). Except specificity and FNR, the performance of the DT algorithm is enhanced by SMOTE resampling. For RF algorithm, SMOTE resampling yields substantial gains. Accuracy increases from 0.824 to 0.944, and the F-score improves dramatically from 0.351 to 0.946. These improvements highlight the model's strong ability to capture patterns after SMOTE resampling, resulting in more reliable predictions across all key metrics. Similarly, AdaBoost demonstrates excellent performance after resampling. Accuracy rises from 0.820 to 0.927, while the F-score leaps from 0.276 to 0.931,

indicating that the model becomes significantly more effective at identifying defaulting borrowers after applying SMOTE resampling. Finally, the CatBoost model experiences a modest drop in accuracy after resampling (from 0.819 to 0.742). Nonetheless, precision, F-score, KS, FPR, H-measure and AUC are improved by SMOTE resampling, suggesting enhanced overall model effectiveness. The imbalanced dataset retains higher specificity and FNR, indicating a stronger emphasis on correctly identifying non-defaulters.

3.3 Discussion and findings of the single models

From the above findings, it can be observed that SMOTE resampling leads to a decrease in the accuracy of the LR, NB, SVM, and CatBoost algorithms, as shown in Tables 1 and 2. One possible reason is the structure of the dataset, which consists of four categorical features (40% of the total) and six numerical features. Categorical features may limit these models' ability to detect patterns in the data [26, 37]. Moreover, SMOTE resampling may inadvertently introduce bias and noise by generating synthetic outliers [26]. It also assumes feature independence and equal importance, which does not hold in this dataset, especially given the presence of low inter-feature correlations. This can result in unrealistic synthetic samples and overlapping between the minority and majority classes, thereby introducing further noise.

Additionally, the performance decline may be attributed to model sensitivity to changes in class distribution caused by SMOTE resampling. Some classifiers are more susceptible to such changes than others. For instance, LR, NB, and SVM rely on specific data distribution assumptions, such as linear separability for LR and SVM, and feature independence for NB, which may be disrupted by synthetic oversampling [38–41]. Despite the drop in accuracy, other performance metrics such as the F1 score (which is more responsive to minority class prediction), KS statistic, and H-measure show improvement, indicating better identification of minority class instances while preserving majority class performance.

In contrast, tree-based algorithms benefit significantly from SMOTE resampling. In imbalanced datasets, these algorithms tend to be biased toward the majority class, which hampers minority class classification. SMOTE resampling mitigates this issue by balancing class distribution and enabling better learning from the minority class [38]. By generating diverse and representative synthetic samples, SMOTE resampling enhances the generalization capability of tree-based models, which are sensitive to data diversity [42, 43]. This facilitates the development of more discriminative decision boundaries and improves overall classification performance. Furthermore, SMOTE resampling reduces overfitting on the majority class by limiting its dominance, allowing the model to better generalize to unseen instances [26, 41].

These findings demonstrate that SMOTE is not universally beneficial but interacts differently with model families, improving performance in flexible, non-parametric models while potentially degrading performance in assumption-driven classifiers. This highlights the importance of aligning resampling strategies with model characteristics in imbalanced credit risk prediction tasks.

3.4 Stacking with meta learner

In the present work, RF, AdaBoost and CatBoost algorithms exhibited the best evaluation performance for the imbalanced dataset. Thus, they were selected as the base learners for the stacking approach in this case. After applying SMOTE resampling, DT, RF, and AdaBoost exhibited the highest performance; therefore, they were selected as the base learners for the stacking approach. In addition to performance, the selection also considered the diversity and generalization behavior of the models. For the imbalanced dataset, CatBoost offered stable learning and effectively handled complex feature interactions, complementing the behaviors of RF and AdaBoost. After applying SMOTE, Decision Trees (DT) achieved higher predictive accuracy, increasing from 0.737 to 0.900, whereas the accuracy of CatBoost decreased from 0.819 to 0.742. Consequently, DT was chosen as a base learner for the stacking ensemble. The inclusion of DT after balancing contributed to simpler and more distinct decision boundaries, increasing diversity among the base learners and improving the generalization ability

of the stacking ensemble. Therefore, the chosen combinations ensured both strong individual performance and diversity, which are essential for effective stacking.

In a stacking architecture, the meta-learner operates on the outputs of multiple models, which often exhibit non-linear dependencies and interactions. Traditional linear approaches are limited in this context, as they assume additive and independent contributions of inputs. In contrast, gradient boosting models leverage a sequential tree-based structure to learn non-linear patterns and higher-order interactions more effectively. In addition, boosting algorithms incorporate regularization mechanisms that enhance generalization and reduce the risk of overfitting. Their capability to handle heterogeneous and correlated inputs further makes them well-suited for aggregating predictions from diverse base learners. Moreover, as reported in the literature, LightGBM, XGBoost, and GBM provide stable and high predictive performance [44–46]. Accordingly, LightGBM, XGBoost, and GBM are adopted as meta-learners due to their combined theoretical advantages and demonstrated effectiveness in modelling complex ensemble relationships.

Stacking was implemented using out-of-fold (OOF) predictions from stratified 10-fold cross-validation with fixed splits and random seeds across all base learners. Base models were trained on 9 folds and used to predict the held-out fold, with all preprocessing fitted only on training folds. The resulting OOF predictions formed the meta-level dataset used to train LightGBM, XGBoost, and Gradient Boosting.

3.5 Results of stacking with meta learner

Table 3 shows the GBM stacking performance metrics with RF, Adaboost, CatBoost algorithms as base models for imbalanced dataset in addition to DT, RF and Adaboost as base models after applying SMOTE resampling. For the imbalanced dataset, the GBM stacking model has a marginal influence on the performance, when compared with RF, Adaboost, CatBoost base models. However, when DT, RF, and AdaBoost are used as base models after SMOTE resampling, the GBM stacking model shows better performance than when base models are trained on the imbalanced dataset.

The performance of the XGBoost stacking model with RF, AdaBoost, and CatBoost as base models for the imbalanced dataset, and with DT, RF, and AdaBoost as base models after SMOTE resampling, is presented in Table 3. For the imbalanced dataset, the XGBoost stacking does not significantly affects the prediction performance, when compared with RF, Adaboost and CatBoost base models. However, the XGBoost stacking displays excellent performance, when compared with those of XGBoost stacking with RF, Adaboost, CatBoost as base models for the unbalanced dataset or base models (DT, RF and Adaboost) after applying SMOTE, Table 2.

The performance of the LightGBM stacking model using RF, Adaboost, and CatBoost algorithms as base models for the imbalanced dataset and DT, RF, and Adaboost as base models after SMOTE resampling are shown in Table 3. LightGBM stacking has no significant effect on performance measures for the unbalanced dataset

Table 3 Prediction performance metrics of GBM, XGBoost and LightGBM stacking algorithms

Model	GBM		XGBoost		LightGBM	
Dataset	Imbalanced	SMOTE resampling	Imbalanced	SMOTE resampling	Imbalanced	SMOTE resampling
F-score	0.387	0.980	0.395	0.980	0.391	0.980
Precision	0.638	0.989	0.621	0.991	0.644	0.991
Sensitivity	0.277	0.970	0.289	0.970	0.281	0.970
FPR	0.037	0.010	0.042	0.009	0.037	0.008
Specificity	0.963	0.990	0.959	0.991	0.963	0.992
FNR	0.723	0.030	0.712	0.030	0.719	0.030
ROC-AUC	0.817	0.993	0.819	0.993	0.819	0.994
PR-AUC	**0.634**	**0.669**	**0.549**	**0.632**	**0.563**	**0.633**
Accuracy	0.832	0.980	0.831	0.980	0.833	0.981
KS	0.471	0.960	0.479	0.961	0.478	0.961
H-measure	0.153	0.937	0.155	0.937	0.157	0.938

when compared to the RF, Adaboost, and CatBoost base models. After applying SMOTE resampling to the base learners, the LightGBM stacking model outperforms both the LightGBM stacking with base models trained on the imbalanced dataset (RF, AdaBoost, CatBoost) and the individual base models (DT, RF, and AdaBoost) after SMOTE resampling, as shown in Table 2.

Across all models, SMOTE leads to consistent and meaningful improvements in PR-AUC: GBM (0.634→0.669), XGBoost (0.549 → 0.632), and LightGBM (0.563 → 0.633). These gains demonstrate enable more effective identification of default cases and improving minority class discrimination. A comparison across stacking models shows that, although GBM achieves the highest PR-AUC (0.669), the differences between models after SMOTE are relatively small (within ~0.03–0.04). In contrast, under imbalanced conditions, the gap between models is more pronounced, with XGBoost exhibiting notably lower PR-AUC (0.549) compared to GBM (0.634). This indicates that SMOTE not only improves overall performance but also reduces variability across different stacking architectures, leading to more stable and consistent results.

These findings suggest that the primary driver of performance improvement is the integration of SMOTE within the stacking framework, rather than the choice of boosting algorithm itself. Overall, the results confirm that SMOTE plays a critical role in enhancing PR-AUC and harmonizing model performance, resulting in robust and generalizable improvements in default prediction.

LightGBM is selected as the best model due to its balanced and consistent performance across all evaluation metrics, achieving high F1-score, strong sensitivity, and very high precision. It also records the lowest false positive rate and highest specificity, indicating more reliable identification of non-default cases. In addition, LightGBM offers greater computational efficiency and scalability, making it well-suited for real-world credit risk applications. Overall, its combination of accuracy, stability, and practical efficiency makes it the most robust and deployable model.

3.6 Discussion and findings of stacking method with meta learner

Overall, stacking is advantageous as it leverages the strengths of multiple models while mitigating their individual limitations, thereby enhancing predictive accuracy. However, the results indicate that the GBM, LightGBM, and XGBoost stacking models, when combined with RF, AdaBoost, and CatBoost as base learners, do not perform effectively on the imbalanced dataset. This reduced performance can be attributed to the inherent challenges posed by imbalanced data, where learning algorithms tend to exhibit bias toward the majority class [47–49].

The current results show that GBM, LightGBM, and XGBoost stacking models, using DT, RF, and AdaBoost after SMOTE resampling as base learners, achieve excellent performance. This can be attributed to the complementary nature of the tree-based base models. Due to their structural differences, DT, RF, and AdaBoost capture diverse patterns in the balanced data, and their combined predictions in the stacking framework offer a more holistic view of the dataset. Moreover, their ability to model nonlinear relationships between features and the target variable allows them to capture complex interactions, thereby enhancing the overall predictive performance of the stacking models.

DT, RF, and AdaBoost are well-suited as base models in stacking due to their robustness to outliers, computational efficiency, and built-in regularization techniques such as pruning and shrinkage. These properties help prevent overfitting, particularly in the volatile P2P lending environment, where data patterns can shift due to economic or demographic changes. In this context, stacking offers enhanced generalization by combining multiple models, reducing reliance on any single predictor, and improving performance across diverse and noisy datasets [50]. Given the mixed nature of the current dataset, comprising both numerical and categorical features, stacking enables different base models to contribute their respective strengths, improving overall predictive capability. Furthermore, stacking can adapt to changing borrower behaviour or economic conditions by weighting more predictive models, thereby enhancing risk assessment and decision-making. This flexibility is crucial in the P2P domain, where prediction errors can lead to significant financial consequences. By diversifying model risk and

refining prediction accuracy, stacking supports more effective interest rate pricing, better risk management, and increased profitability for lenders, while offering more favourable terms for borrowers.

These findings demonstrate that effective stacking in imbalanced credit risk prediction requires a coordinated design of resampling strategy and base learner diversity. The results emphasize that carefully selected, complementary base learners combined with appropriate data balancing yield superior performance. This provides a practical guideline for designing robust stacking frameworks in P2P lending applications, where accurate risk estimation is critical for financial decision-making [50].

3.7 Probability reliability and calibration analysis

3.7.1 Quantitative calibration assessment using Brier score and Expected Calibration Error (ECE)

Table 4 presents a quantitative evaluation of probabilistic calibration using the Brier score, which measures the mean squared difference between predicted probabilities and actual outcomes. Lower Brier scores indicate better calibration, reflecting predictions that are both accurate and well-aligned with true event frequencies. Unlike discrimination metrics, the Brier score provides a strictly proper scoring rule that jointly captures calibration and refinement, making it particularly suitable for risk-sensitive applications such as credit default prediction.

The results reveal several important insights. First, the proposed stacking-based frameworks, particularly LightGBM and XGBoost stacking, exhibit substantial improvements in Brier score following SMOTE resampling. For instance, the Brier score of XGBoost stacking decreases markedly from 0.115 to 0.094, while LightGBM stacking improves from 0.104 to 0.094. This indicates that data balancing not only enhances classification performance but also significantly improves the reliability of predicted probabilities. Such improvements suggest that the models become less biased toward the majority class and better capture the true likelihood of default events.

Second, the convergence of Brier scores for LightGBM and XGBoost stacking after SMOTE (both achieving 0.094) is particularly noteworthy. This consistency indicates that the proposed stacking architecture, when combined with class balancing, produces stable and well-calibrated probability estimates across different gradient boosting paradigms. This stability is a key contribution, as prior studies often report improvements in discrimination without systematically demonstrating calibration consistency across ensemble variants.

In contrast, GBM stacking shows a slight degradation in calibration after SMOTE (from 0.091 to 0.095), suggesting that not all ensemble configurations benefit equally from resampling. This highlights an important methodological insight: the interaction between resampling techniques and ensemble learning is model-dependent, and improvements in class balance do not universally translate into better probabilistic calibration.

To complement the Brier score, calibration performance is further assessed using the Expected Calibration Error (ECE), which quantifies the discrepancy between predicted probabilities and observed outcomes across probability bins. Lower ECE values indicate better calibration. As shown in Table 4, resampling leads to a consistent reduction in ECE for all ensemble models, confirming improved probability reliability after addressing class imbalance. LightGBM stacking and XGBoost stacking exhibit the most substantial improvements, with ECE decreasing from 0.050 to 0.020 and from 0.052 to 0.026, respectively. These reductions indicate improved alignment between predicted and empirical probabilities, confirming that resampling effects are model-dependent and more pronounced in models sensitive to class imbalance.

Table 4 Comparative evaluation of model calibration before and after SMOTE using Brier Score and ECE

Metric	Dataset	LightGBM stacking	XGBoost stacking	GBM stacking
Brier Score	Imbalanced	0.104	0.115	0.091
	SMOTE resampling	0.094	0.094	0.095
ECE	Imbalanced	0.050	0.052	0.018
	SMOTE resampling	0.020	0.026	0.016

Reductions in both Brier score and ECE confirm that the proposed framework improves calibration in terms of overall probabilistic accuracy and bin-wise reliability. Their agreement provides robust and complementary evidence of enhanced probability estimation for credit risk decision-making.

3.7.2 Visual calibration analysis via reliability diagrams

To further validate the reliability and practical applicability of the proposed framework, a probability calibration analysis was conducted for the stacking ensemble models under both imbalanced and SMOTE-balanced settings, as illustrated in Fig. 6. The results reveal several important findings. First, under the imbalanced setting, the Light-GBM stacking model exhibits noticeable deviation from the diagonal line, particularly in the mid-probability range. This indicates that the model tends to underestimate or overestimate default risk when trained on skewed data, leading to unreliable probability outputs. In contrast, after applying SMOTE, the calibration performance

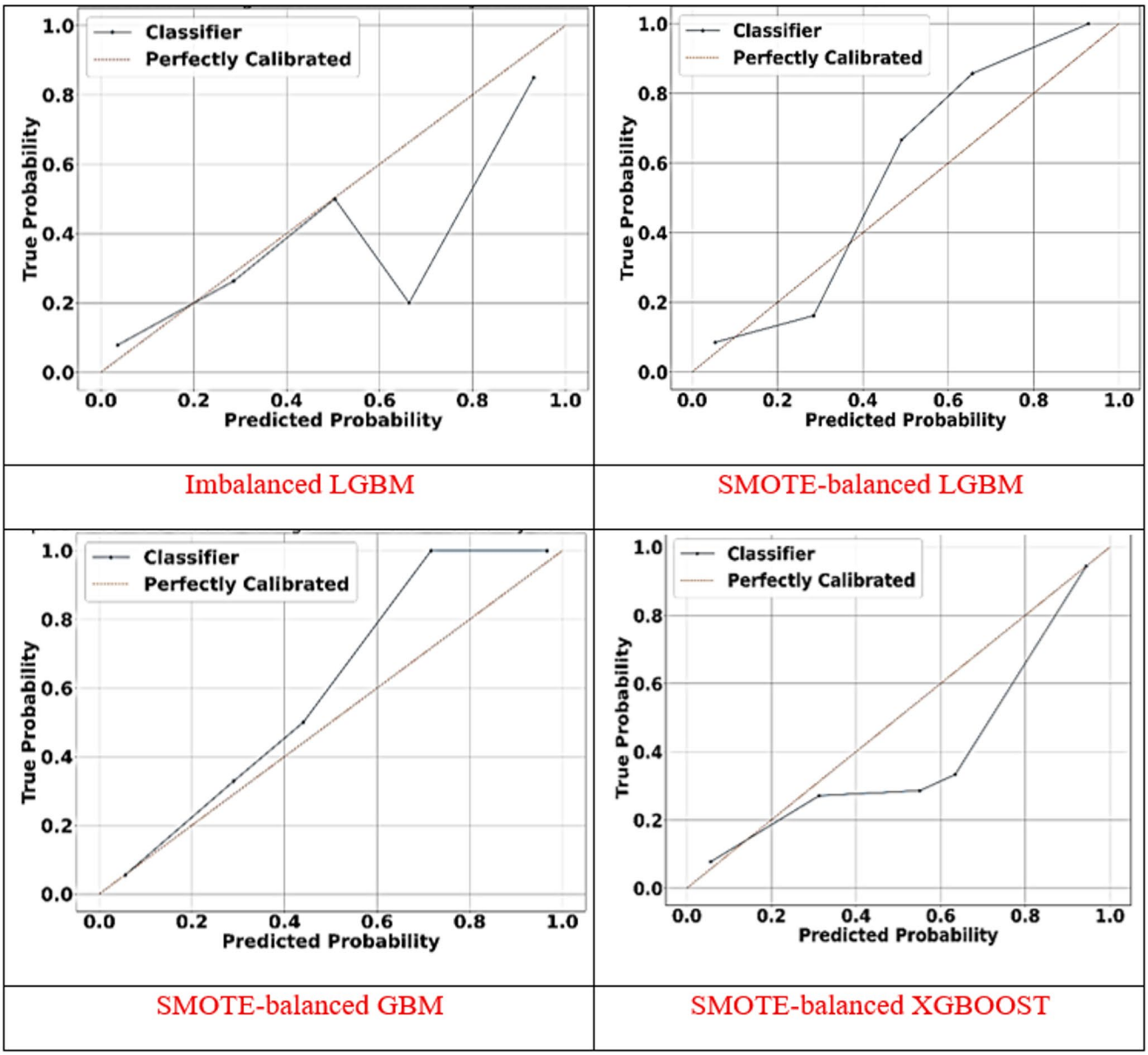

Fig. 6 Calibration curves of stacking models under imbalanced and smote-balanced settings

improves significantly. The LightGBM stacking model shows a much closer alignment with the diagonal, demonstrating that balancing the dataset enhances the reliability of predicted probabilities. This improvement confirms that SMOTE not only affects classification performance but also plays a critical role in improving probability estimation.

A comparative analysis across stacking models further highlights that the LightGBM-based stacking approach provides the most stable and well-calibrated predictions. While XGBoost and GBM stacking models also benefit from SMOTE, their calibration curves exhibit step-like patterns and deviations in certain probability ranges, suggesting less consistent probability estimation. This indicates that LightGBM is better suited as a meta-learner for producing reliable probability outputs in P2P credit risk prediction.

This analysis provides strong empirical evidence on the interplay between class imbalance handling, ensemble learning, and probability estimation using Brier score, Expected Calibration Error (ECE), and calibration curves. The results show that the proposed framework improves both predictive accuracy and probability reliability, as reflected by consistent reductions in Brier score and ECE and improved calibration curve alignment. By quantifying the effects of SMOTE and stacking across complementary metrics, the study tackles key challenges related to model reliability and reproducibility. Notably, the findings establish that integrating SMOTE with stacking ensembles systematically enhances probability calibration, rather than merely classification performance, thereby addressing a critical gap in P2P lending research where probability reliability is often overlooked. The combined evaluation framework provides compelling evidence that the proposed model yields well-calibrated, trustworthy risk scores, highlighting the importance of calibration-aware evaluation and positioning the hybrid pipeline (SMOTE + stacking) as an effective approach for jointly improving predictive performance and probabilistic reliability.

3.8 Component-wise ablation study and quantitative contribution analysis

To assess the contribution of individual components, a comprehensive ablation study was conducted to quantify the impact of SMOTE, feature selection (FS), and stacking on model performance. The objective of this analysis is not only to validate the effectiveness of the proposed framework, but also to explicitly identify which components drive the observed improvements. Table 5 presents the results obtained by systematically removing each component while keeping the remaining pipeline unchanged. The results demonstrate that each component contributes significantly to overall performance, although their effects differ across models.

The impact of SMOTE is particularly evident in boosting-based stacking models. For instance, in the LightGBM stacking model, PR-AUC decreases from 0.633 to 0.563 when SMOTE is removed, accompanied by a substantial drop in F1-score from 0.980 to 0.391. A similar trend is observed for the XGBoost stacking model (0.632

Table 5 Ablation analysis of SMOTE, feature selection (FS), and stacking: quantifying their impact on model performance

Model variant	SMOTE	FS	Stacking	PR-AUC	ROC-AUC	F1
LGBM stacking	✓	✓	✓	0.633	0.994	0.980
LGBM stacking	✗	✓	✓	0.563	0.819	0.391
LGBM stacking	✓	✗	✓	0.273	0.81	0.311
XGBoost stacking	✓	✓	✓	0.632	0.993	0.980
XGBoost stacking	✗	✓	✓	0.549	0.63	0.395
XGBoost stacking	✓	✗	✓	0.334	0.81	0.456
GBM stacking	✓	✓	✓	0.669	0.993	0.980
GBM stacking	✗	✓	✓	0.634	0.817	0.387
GBM stacking	✓	✗	✓	0.292	0.83	0.337
RF (Highest performance base learner)	✓	✓	✗	0.618	0.993	0.946
RF (Highest performance base learner)	✗	✓	✗	0.601	0.788	0.351
RF (Highest performance base learner)	✓	✗	✗	0.274	0.86	0.297

to 0.549) and, to a lesser extent, for the GBM stacking model (0.669 to 0.634). This indicates that class imbalance handling is essential for improving minority class detection and maintaining balanced predictive performance.

The removal of feature selection leads to the most pronounced degradation in performance across all models. In the LightGBM stacking model, PR-AUC drops sharply from 0.633 to 0.273, while F1-score decreases from 0.980 to 0.311. Similar reductions are observed for XGBoost (0.632 to 0.334) and GBM (0.669 to 0.292) stacking models. This confirms that the proposed feature selection strategy plays a critical role in eliminating irrelevant and noisy features, thereby significantly improving both ranking performance and generalization.

Furthermore, the comparison between the stacking models and the strongest individual base learner (Random Forest) shows that stacking models demonstrate more consistent performance across different evaluation metrics. This suggests that stacking effectively integrates complementary patterns learned by different base models, leading to more stable and balanced predictions.

An important observation is that each component contributes in a distinct manner. SMOTE primarily improves performance metrics such as ROC-AUC and F1-score, reflecting enhanced minority class detection. In contrast, feature selection has the largest impact on PR-AUC, indicating improved ranking and discrimination capability. Stacking contributes by stabilizing performance and reducing variability through the integration of diverse model behaviors.

The ablation study provides clear evidence that performance improvements are not the result of a simple combination of existing techniques. Instead, they arise from the complementary and statistically meaningful contributions of each component within the proposed framework. By explicitly quantifying these effects, the study offers new empirical insight into how resampling, feature selection, and ensemble learning interact in P2P credit risk prediction. These findings enhance both the interpretability and reproducibility of the proposed approach, while strengthening its methodological contribution beyond a purely integrative design.

3.9 Comparing current findings with the existing literature

Table 6 presents a comparison between the results of the present work and those reported in the literature using other P2P datasets. The framework and approach developed in this study lead to remarkable prediction

Table 6 Comparison between the results of the present work and the best results reported in literature

Dataset	Feature selection methods	Machine learning models	Resampling	Accuracy	F- score	AUC	References
Lending Club	Pearson Correlation and Recursive Feature Selection	XGBoost	SMOTE	0.83	0.78	-------	[3]
Australia	Recursive feature elimination	LightGBM	Random undersampling	-------	-------	0.933	[4]
Lending Club	GBDT, RF, AdaBoost, XGBoost, LGBM and DT	RF stacking (GBDT, RF, AdaBoost, XGBoost, LGBM, and DT)	No	0.925	-------	-------	[6]
Lending club loan	LGBM	XGBoost stacking (KNN, SVM, R F)	SMOTE	0.910	0.920	-------	[1]
RenRenDai	MRMR	DT stacking (RF, NN, DTa, KNN, SV)	No	0.930	0. 90	0.910	[2]
Lending Club	RFECV	Stacking RF, DT, SVM, XGBoost, Adaboost, MLP)	SMOTE + ENNs	0.937	0.955	0.980	[14]
Lending Club	Variance threshold, Stepwise	XGB stacking	Random undersampling	0.850	0.720	0.911	[13]
Bondora	**Chi-2, SBS, RF, LightGBM, CatBoost**	**GBM stacking (DT, RF, Adaboost)**	**SMOTE**	**0.980**	**0.980**	**0.993**	**Present work**
Bondora	**Chi-2, SBS, RF, LightGBM, CatBoost**	**LGBM stacking (DT, RF, Adaboost)**	**SMOTE**	**0.981**	**0.980**	**0.994**	**Present work**
Bondora	**Chi-2, SBS, RF, LightGBM, CatBoost**	**XGBoost stacking (DT, RF, Adaboost)**	**SMOTE**	**0.980**	**0.980**	**0.993**	**Present work**

performance results. The models employed in this study (LightGBM, GBM, and XGBoost) exhibit outstanding performance across multiple evaluation metrics, surpassing the results reported in previous studies found in the literature. In particular, LightGBM stands out with outstanding performance, achieving an accuracy of 0.981, an F-score of 0.980, and an AUC of 0.994. These values significantly exceed the performance metrics reported in earlier works.

3.10 Explainable model

Table 3 demonstrates that LGBM performs exceptionally well in this study in terms of accuracy, AUC, and F-score. However, since LGBM is a black-box model, it is difficult to understand how various features influence the prediction outcomes. To address this, the current study employs the LIME explainable machine learning model [9, 51, 52] to provide an interpretability analysis of the loan default predictions made by the LGBM model.

3.10.1 LIME model

Defaulted sample Figures 7 and 8 jointly provide an in-depth LIME-based explanation of the prediction for a sample with a high defaulting probability of 0.85. Figure 7 displays the LIME output for the local prediction. The figure highlights which feature contributed to the final classification decision of "default (YES)" and separates them based on their influence direction. Features that contribute towards the prediction of "YES" are shown in orange, while those pushing the prediction towards "NO" (non-default) are shown in blue. It can be observed that features such as interest, income total, gender, and education=primary education heavily influenced the decision toward defaulting. Figure 7 complements this visualization by quantifying the contribution of each feature using the "impact degree" scale. The interest rate being above 29.34% and Income Total less than or equal to 1100.00 are the two most influential factors increasing the probability of default, with the highest impact on the prediction. These two features alone capture the borrower's financial burden and limited repayment capacity. Conversely, Home Ownership Type=Owner and Loan Duration≤60 months slightly contribute toward lowering the default probability, as shown by the negative (red) impact bars. However, their mitigating effect is minimal compared to

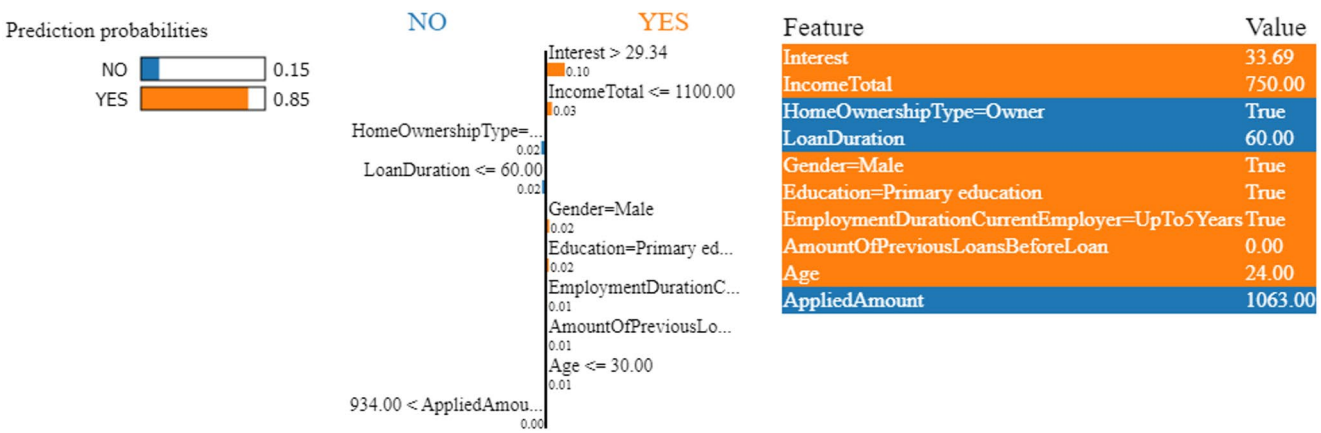

Fig. 7 Results from the LIME analysis of sample 419th with a high default rate

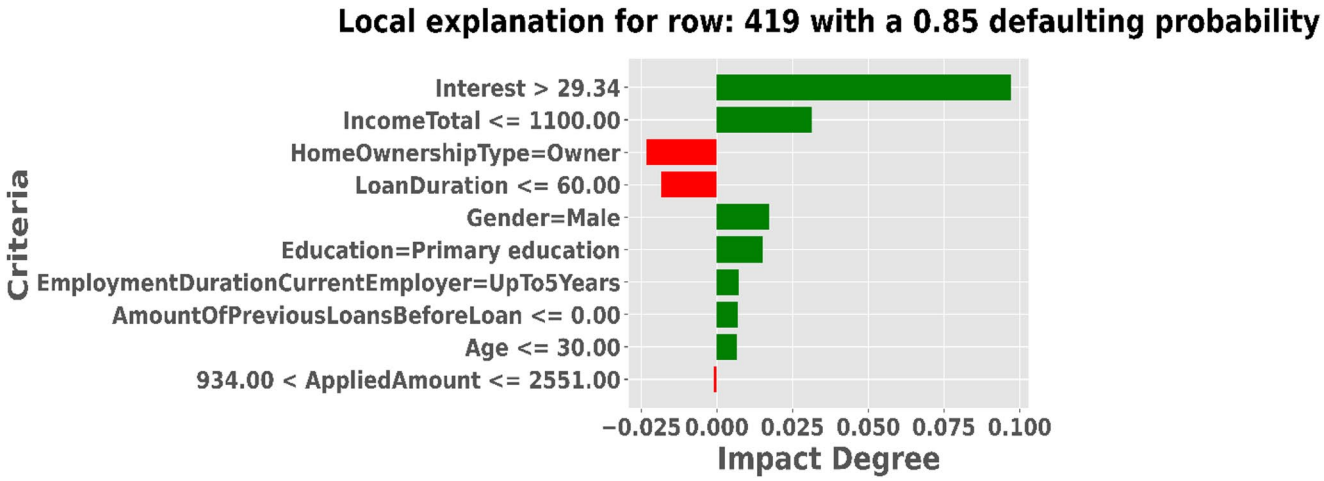

Fig. 8 Local explanation for default of sample 419th

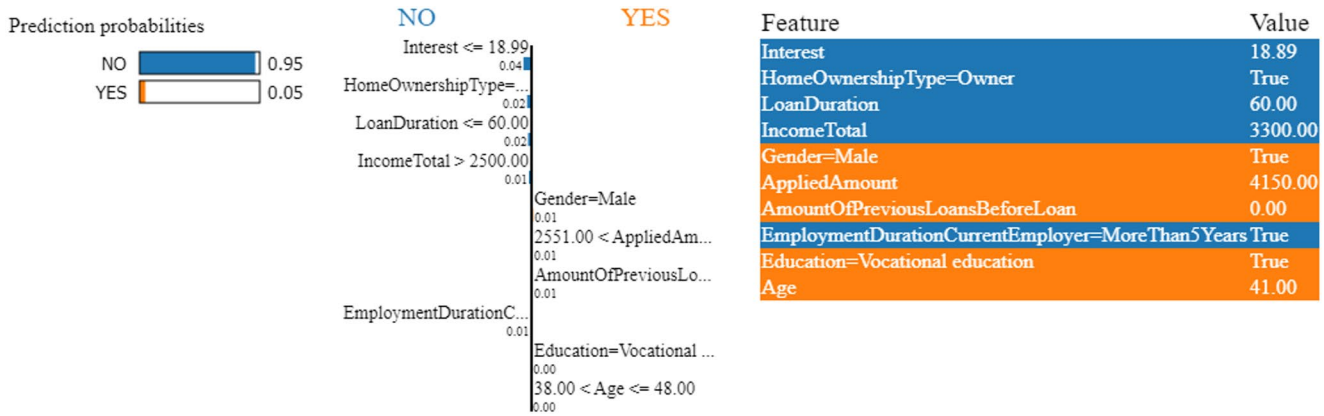

Fig. 9 Results from the LIME analysis of sample 20th with a high non-default rate

the dominant positive predictors. These red bars demonstrate that while some features suggest creditworthiness, their influence is insufficient to outweigh the overall risk profile.

The LIME explanation clearly indicates that the model's prediction is well-grounded and interpretable: the decision to classify this borrower as a defaulter is driven by financially sound reasoning; high interest burden, low income and lower education level.

In conclusion, Figs. 7 and 8 effectively illustrate how local interpretability can provide transparency into individual classification decisions. The combination of visual explanation and feature-wise impact quantification confirms that the model's high-confidence prediction for the sample is both explainable and justifiable. This detailed LIME-based explanation satisfies the need for transparent AI in high-stakes domains such as credit risk evaluation.

Non-defaulted Sample Figure 9 illustrates the LIME output for a non-defaulted sample with a predicted non-default probability of 0.95. The most influential features contributing to this prediction are visualized in Fig. 10. As seen in Fig. 10, the features "interest ≤ 18.99", "home ownership type = owner", and "loan duration ≤ 60.00" are the top contributors to the non-default prediction, as they appear with high negative impact scores (red bars). These features reduce the likelihood of default significantly. Other supporting features include "income total > 2500", and "employment duration > 5 years", indicating financial stability and low risk behaviour. This

 https://doi.org/10.1007/s00521-026-12226-5

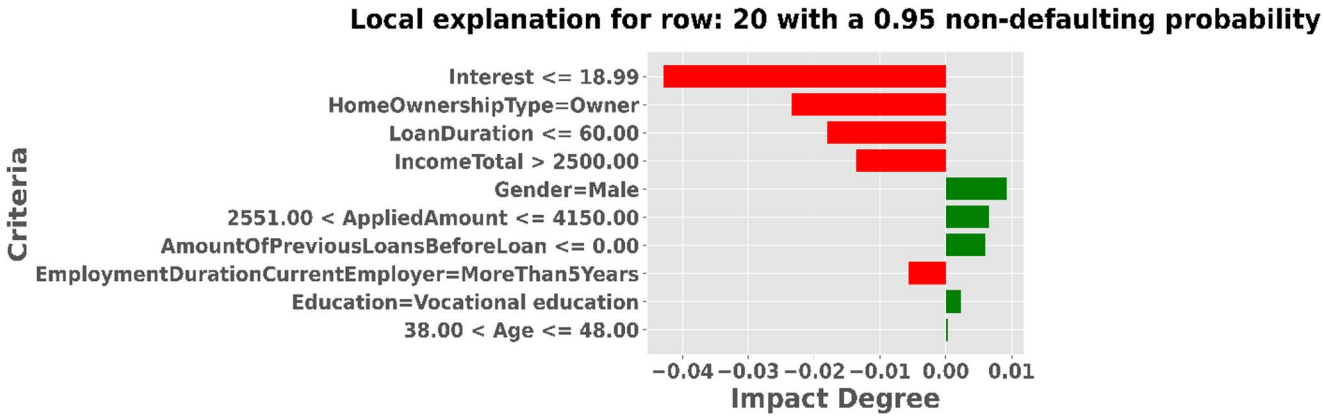

Fig. 10 Local explanation for non-default of sample 20th

Fig. 11 Degree of impact for defaulted samples with probability of 0.7–0.95

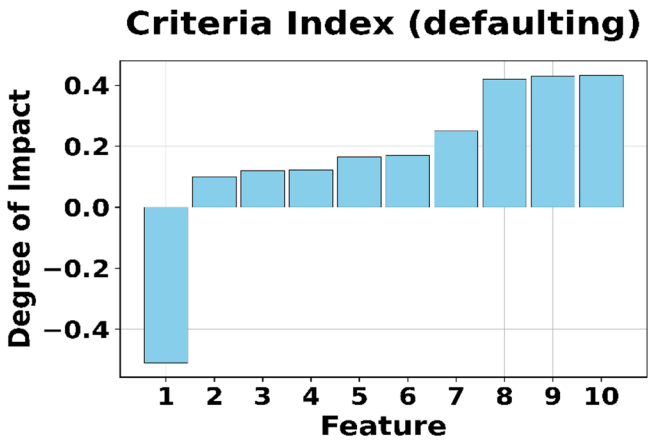

Table 7 Features and their importance of defaulted samples with probability of 0.7–0.95

Index	Feature	Impact degree
1	20.46 < Interest <= 29.34	− 0.510
2	Employment Duration Current Employer=UpTo1Year	0.100
3	Home Ownership Type=Living with parents	0.119
4	30.00 < Age <= 38.00	0.122
5	Loan Duration > 60.00	0.165
6	0.00 < Amount of Previous Loans Before Loan <= 1594.00	0.171
7	Gender=Female	0.250
8	Education=Secondary education	0.419
9	1100.00 < Income Total <= 1700.00	0.429
10	2551.00 < Applied Amount <= 4150.00	0.433

analysis reveals that low-interest loans, home ownership, long-term employment, and manageable applied amounts collectively play a critical role in reducing default risk in P2P lending platforms. The threshold-based conditions in this local surrogate model highlight the nuanced decision-making process of the classifier. These figures together illustrate the transparency of the model's decision-making process, emphasizing the interpretability and reliability of the credit risk model for individual predictions.

Interpretation of aggregating defaulted samples: Figure 11; Table 7 collectively illustrate the feature impact analysis for borrowers classified as highly likely to default, with predicted default probabilities ranging from

0.7 to 0.95. Figure 11 presents the criteria index bar chart, showing the degree of impact, each feature has on the model's decision for this group of high-risk borrowers. The features are indexed from 1 to 10, with the corresponding details provided in Table 7. The x-axis represents the feature indices, and the y-axis shows their impact degrees, indicating the magnitude and direction (positive or negative) of influence on the default prediction. According to Table 7, the most negatively contributing feature is interest rate between 20.46 and 29.34, with an impact degree of −0.510, suggesting that within this specific interest range, the risk of default might be comparatively lower—likely due to stronger loan performance history or compensating borrower characteristics. Conversely, the remaining nine features exert positive influence on the default prediction, contributing to higher default probabilities. Among these, the most significant contributors are: applied amount between 2551.00 and 4150.00 (+0.433), income total between 1100.00 and 1700.00 (+0.429), secondary education level (+0.419), female gender (+0.250). Other notable features include short employment duration (UpTo1Year), younger age group (30–38 years), living with parents, longer loan duration (>60 months), and limited prior borrowing experience, all with moderate positive impacts. This analysis underscores how combinations of socioeconomic factors, limited financial history, and moderate loan amounts contribute to the model's assessment of high default risk. The results affirm the model's ability to detect nuanced patterns that align with real-world credit behaviour, offering crucial transparency for lenders and analysts in high-stakes financial decision-making.

Interpretation of aggregating non-defaulted samples Figure 12 aggregates LIME results from non-defaulted samples with a predicted non-default probability of 0.7–0.95. The corresponding impact degree for the top 10

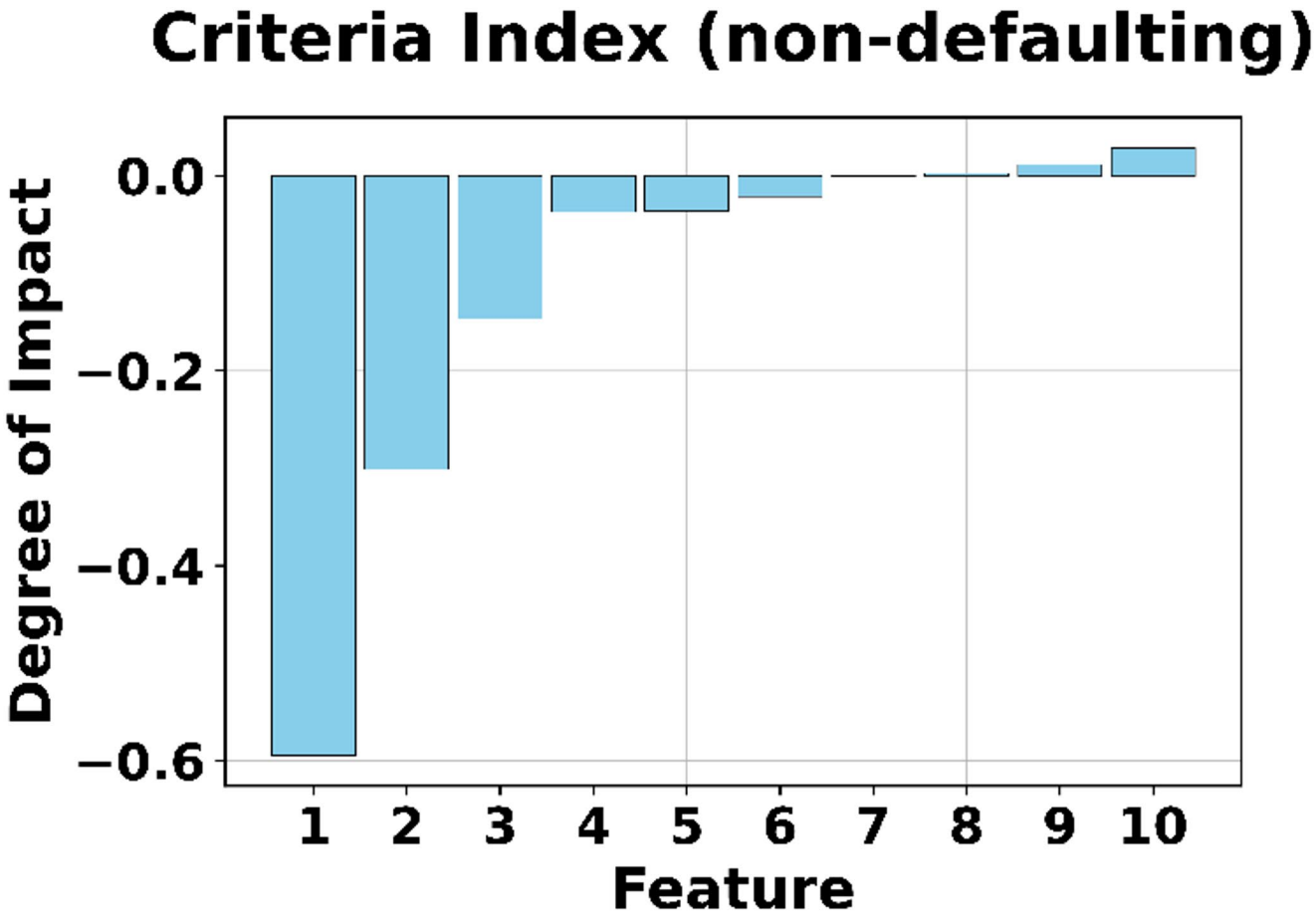

Fig. 12 Degree of impact for non-defaulted samples with probability of 0.7–0.95

Table 8 Features and their importance of non-default samples with probability of 0.7–0.95

Index	Feature	Impact degree
1	Interest<=18.99	− 0.595
2	Loan Duration<=60.00	− 0.301
3	1700.00<Income Total<=2500.00	− 0.147
4	934.00<Applied Amount<=2551.00	− 0.037
5	Gender=Female	− 0.036
6	Home Ownership Type=Mortgage	− 0.021
7	Employment Duration Current Employer=UpTo1Year	− 0.000
8	30.00<Age<=38.00	0.002
9	Education=Secondary education	0.011
10	0.00<Amount Of Previous Loans Before Loan<=1594.00	0.029

features is visualized in Fig. 12 and detailed in Table 8. The most impactful feature is "interest≤18.99", showing a strong effect on default probability with an impact score of -0.595, followed by "Loan duration≤60" and "income total between 1700–2500". These variables consistently support the model's confidence in identifying low-risk borrowers. Interestingly, "gender=female", "home ownership=mortgage", and "short employment duration (≤1 year)" show minor contributions, suggesting subtle influence when considered collectively. On the contrary, features like "education=secondary" and "amount of previous loans>0" show marginal contributions, possibly reflecting a slight elevation in risk, although not enough to shift the classification. This aggregate explanation demonstrates the model's alignment with financial reasoning, confirming that lower interest rates, shorter loan durations, and moderate-income levels are decisive indicators of repayment ability. Such insights are critical for stakeholders to trust and interpret model decisions in real-world applications.

Figures 11 and 12; Tables 7 and 8 support the interpretability analysis. LIME is applied to the final stacking ensemble, and the resulting local explanations are aggregated and systematically analyzed to reveal consistent structural patterns across the dataset, enabling the identification of stable directional effects and context-dependent interactions, where feature effects vary with the values of other features. The results indicate that key financial variables, particularly interest rate, loan amount, and loan duration, consistently act as primary drivers of default risk. Higher interest rates and larger loan amounts increase repayment burden and default likelihood, while income plays a stabilizing role by mitigating adverse loan conditions. However, these effects are not globally uniform; the model captures heterogeneous decision boundaries in which identical feature values may lead to different outcomes depending on their interaction context. For example, the adverse effect of high interest rates is significantly amplified under low-income conditions, while longer loan durations introduce a temporal risk dimension by increasing exposure to financial instability. Education further operates as a latent socio-economic factor, indirectly influencing repayment capacity through its relationship with income stability and financial behaviour.

These findings demonstrate that default risk emerges from non-linear, interaction-driven relationships that cannot be fully captured by traditional global feature importance methods. Importantly, LIME is used as an analytical layer to identify recurring feature patterns and interaction effects in the predictions generated by the stacking ensemble.

To provide a more complete understanding of global model behaviour, the stacking framework is interpreted as a two-level learning system in which base learners capture diverse non-linear relationships, and the LightGBM meta-learner integrates out-of-fold predictions by adaptively weighting their contributions. From this perspective, the ensemble can be understood as a dynamic decision system rather than a black box: strong financial signals such as interest rate and loan amount dominate predictions, while in weaker signal scenarios the model increases reliance on demographic and contextual features.

By linking aggregated local explanations to the stacking model, the proposed framework moves beyond a straightforward combination of existing techniques and enables the identification of stable feature effects,

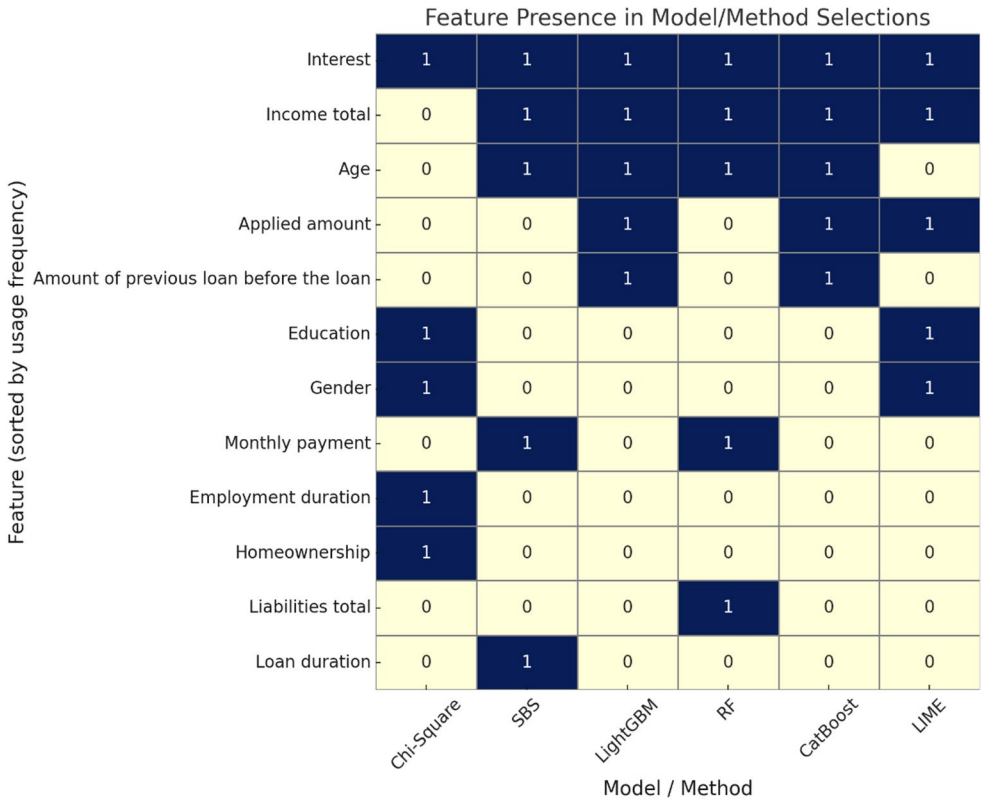

Fig. 13 Visualization indicating the features identified as important by each method

interaction mechanisms, and adaptive decision strategies. Consequently, the stacking model is transformed from a purely predictive tool into an interpretable knowledge discovery framework, providing actionable insights into how borrower characteristics jointly influence default risk while enhancing transparency, consistency, and applicability in real-world financial decision-making.

3.10.2 Exploring interpretability in explainable AI (xAI)

Figure 13 highlights the key features deemed essential by the feature selection methods employed in this study (Chi-Square, SBS, LightGBM, RF, CatBoost), as well as those identified through the integration of the LGBM stacking model (comprising DT, RF, and AdaBoost) with the explainable AI technique (Aggregated LIME).

LIME vs. Chi-square: An analysis of LIME (aggregated) versus chi-square demonstrates notable differences in the types of features each method identifies as influential in predicting P2P loan defaults. The Chi-square test, which assesses the statistical independence between categorical variables, identifies education, homeownership, employment duration, interest, and gender as significant features. This reflects that Chi-Square focus on uncovering relationships and dependencies between categorical variables. Education and gender are particularly notable because they often correlate with socioeconomic status and decision-making factors. Similarly, homeownership and employment duration are chosen due to their relevance in understanding financial stability and long-term commitments, which are key indicators of risk or ability to repay loans. On the other hand, the aggregated LIME model emphasizes interest, income total, education, and gender, with a stronger focus on income total as it directly correlates with financial capacity and repayment ability. Although both models include education and gender, LIME places greater importance on financial attributes, as these are more directly linked to prediction outcomes in the aggregated setting. Chi-Square, however, captures broader categorical associations, such as homeownership and employment duration, to highlight the relationships that may influence loan decisions at a macro level. This comparison underlines the difference in approach: Chi-Square is more focused on identifying

statistical associations within categorical data, while LIME prioritizes financial and demographic features to provide interpretable insights into the model's predictions.

LIME vs. SBS When comparing the results of LIME (aggregated) and SBS the differences in feature selection and model interpretation become evident. The LIME model focuses on interest, income total, education, and gender, with an emphasis on income total. This reflects LIME's objective of offering a clearer and more understandable explanation of predictions, where financial factors like income total are key to understanding the ability of an individual to repay a loan. Education and gender are also selected by LIME as important features, reflecting their influence on financial decision-making and outcomes. LIME (aggregates insights across predictions, allowing for a broader understanding of feature importance across the dataset. In contrast, SBS, a feature selection technique that removes features one at a time to minimize model performance degradation, highlights interest, loan duration, monthly payment, income total, and age as the most important variables. SBS removes features that do not contribute significantly to predictive accuracy, leading to a model that emphasizes variables like income total and age, which are directly tied to financial stability and the ability to repay loans. While both models agree on the importance of income total, SBS includes loan duration and monthly payment, which are key to assessing loan terms and financial obligations, whereas LIME does not prioritize these factors as heavily. This comparison illustrates how LIME (aggregated) offers more interpretable insights into the model's behavior by focusing on demographic and financial features, while SBS provides a more streamlined, predictive feature set by removing less relevant variables, with a stronger emphasis on financial metrics such as loan duration and monthly payment.

LIME vs.. LightGBM *Comparing LIME with LightGBM reveals distinct perspectives on feature importance in the context of P2P loan default prediction. LightGBM identifies income total, applied amount, and age as the most influential variables, with the amount of previous loan before the loan also contributing notably to the model's overall predictive performance. These findings demonstrate the model's focus on financial and demographic factors. On the other hand, the aggregated LIME model highlights a different set of key features, with interest, income total, education, and gender being the most impactful. While both models agree on the importance of financial factors like income total, LIME places greater emphasis on demographic attributes such as education and gender, which are less prominent in the LightGBM model. This discrepancy illustrates how the two models prioritize different sets of features, with LightGBM focusing more on financial data and previous loan history, while LIME shows the importance of demographic characteristics. This comparison highlights the varying perspectives in feature importance between a complex ensemble model and an interpretable explanation method.*

LIME vs. RF When contrasting LIME and RF, significant differences in feature selection and model behavior emerge. The aggregated LIME model highlights interest, income total, education, and gender, with a particular emphasis on income total, reflecting its focus on providing interpretable insights into how these features influence loan decisions across the dataset. Education and gender are prioritized as demographic factors that can shape financial decision-making, while interest and income total are essential for understanding an individual's financial capacity. On the other hand, the RF model emphasizes interest, income total, age, liabilities total, and monthly payment. RF, being an ensemble learning method, focuses on more complex interactions between financial variables, with liabilities total and monthly payment representing critical indicators of loan repayment capacity and financial stability. While both models agree on the importance of interest and income total, RF places greater importance on financial terms, such as liabilities total and monthly payment, which are not as comprehensively weighted by LIME. The aggregated LIME model, by contrast, provides a more interpretable, demographic-focused perspective, emphasizing education and gender, which are less prominent in RF. This comparison reveals

that while LIME emphasizes transparency and demographic factors for interpretability, RF focuses on capturing complex financial relationships, offering greater predictive power with less emphasis on interpretability.

LIME vs. CatBoost In comparing LIME and CatBoost, notable differences in feature selection and model behavior emerge. The aggregated LIME model highlights interest, income total, education, and gender, with particular emphasis on income total, reflecting its focus on providing interpretable insights into how financial and demographic factors influence loan decisions across the dataset. Education and Gender are prioritized as key demographic factors that impact financial decision-making, while interest and income total are crucial for evaluating financial capacity. CatBoost is a gradient boosting algorithm optimized for handling categorical data, identifies interest, income total, applied amount, age, and amount of previous loan before the loan as the most important features. The ability of CatBoost to capture complex interactions between categorical and continuous variables leads to a focus on transactional features like applied amount and amount of previous loan before the loan, which are essential for predicting loan approval and repayment potential. While both models emphasize interest and income total, CatBoost gives more weight to financial aspects such as applied amount and amount of previous loan before the loan, while LIME (aggregated) focuses more on demographic factors like education and gender. This comparison highlights how LIME prioritizes transparency and demographic factors, while CatBoost is better suited for capturing complex data interactions for more precise predictions.

This study contributes to the growing field of explainable artificial intelligence (XAI) by integrating both traditional feature selection techniques and model-agnostic interpretability tools such as LIME to understand the factors influencing P2P loan default. While many previous studies rely solely on global feature importance derived from classifiers, the inclusion of LIME (aggregated) offers a novel perspective on model behavior, uncovering context-specific drivers of risk that are often obscured in aggregate analyses. By comparing LIME (aggregated) with classical feature selection methods (e.g., Chi-Square, SBS) and ensemble-based algorithms (e.g., LightGBM, RF, CatBoost), this work demonstrates how explainable AI techniques can complement conventional machine learning approaches to provide deeper, more transparent insights into financial decision-making. The integration of these perspectives not only improves the interpretability of credit risk models but also promotes fairer and more accountable AI systems. As such, the methodology and findings presented in this study offer a meaningful advancement in the application of XAI to the financial technology sector.

Overall, the findings of this study are consistent with recent cross-domain research demonstrating that ensemble-based and imbalance-aware learning strategies enhance predictive robustness and stability in complex decision-making problems [53]. Prior studies have shown that combining data-level balancing with advanced learning frameworks can significantly improve model reliability [54]. Moreover, learning-based decision systems have been successfully applied to high-stakes optimization problems, emphasizing the importance of robustness and generalization [55]. At the same time, research on increasingly complex machine learning models highlights the trade-off between predictive performance and interpretability, underscoring the necessity of explainable mechanisms when deploying machine learning models in critical applications [56, 57].

4 Research implications

4.1 Theoretical implications

The current research is expected to make a valuable contribution to the existing body of scientific literature in the rapidly growing P2P lending business, offering new insight into predicting the default risk of P2P lending. Furthermore, the present research has the potential to expand the theoretical understanding of ensemble learning in the scientific community, particularly stacking models, in default risk prediction. Moreover, the incorporation of

explainable machine learning techniques into stacking models presents a new theoretical framework for explaining complex, black-box models, making a significant impact to the body of scientific literature. By proposing a new approach that combines a stacking model with explainable AI, this work could advance knowledge within the scientific community on how complex financial problems, such as loan default prediction, can be addressed using cutting-edge machine learning techniques.

4.2 Practical implications

P2P lending platforms will directly benefit from the current effort. It can assist platforms in making more informed lending choices by increasing the accuracy of default risk predictions. This can lower lenders' financial risk and improve the P2P lending market's overall sustainability. Moreover, more accurate default predictions can reduce default rates by allowing lenders to better assess borrowers' creditworthiness, resulting in lower interest rates and better loan terms for borrowers who are judged less risky. Furthermore, P2P platforms benefit from accurate default risk prediction since it enables lenders to allocate capital more efficiently and make better decisions. This lowers the possibility of default-related financial loss and draws additional lenders to the platform. Before granting loans, P2P platforms can detect risky borrowers with the aid of accurate default risk predictions. Thus, lenders can improve loan portfolio management and save money on collections by decreasing defaults. As a result, profitability increases, and overall operating expenses decreases.

The study has practical implications for enhancing transparency since it incorporates explainable AI technique (LIME). In the context of financial decision-making and regulatory compliance, it is crucial for investors and regulators to comprehend the rationale behind a model's predictions. The P2P platforms can increase user trust (both borrowers and lenders) and lower ambiguity surrounding automated decision-making by providing an explanation of why a specific borrower is likely to default. This is particularly crucial for stakeholders who might be doubtful of unclear machine learning models.

The present work constructs a new credit risk management strategy using feature selection methods and machine learning approaches to direct investors in P2P lending toward successful investment opportunities. Our results are especially valuable for the control of the P2P enterprises through assistance platforms to choose better borrowers to be included on their websites.

5 Recommendations to the professionals

Based on the LIME results from the feature importance, several key attributes emerged as locally significant in influencing predictions of P2P loan defaults. These findings offer valuable insights from a business perspective, helping organizations better understand borrower behavior, fine-tune risk assessment processes, and design more effective credit policies. The following features were identified as most impactful: interest, applied amount, income total, education, and gender.

From these attributes, several business-relevant rules and strategies can be inferred:

- Interest rate: Higher interest rates were frequently associated with default cases. Businesses should treat these as red flags and consider offering risk-adjusted pricing or interest rate reductions for marginal applicants to mitigate default risks.
- Applied loan amount: Larger requested amounts often indicate increased risk. Implementing stricter evaluation criteria for high-value loans or offering tiered loan products could help manage exposure.
- Income total: Higher borrower income is a strong indicator of lower default risk. Organizations should prioritize applicants with stable and substantial incomes, possibly streamlining approval for such profiles through automation.

- Education level: Certain education levels were linked to reduced default risk, likely due to improved job stability and earning potential. Educational attainment can serve as a soft feature to enhance creditworthiness assessment, particularly when financial history is limited.
- Gender: Although LIME flagged gender as an influential factor in specific cases, it should not be used directly in credit decisions to avoid ethical and legal concerns. This finding highlights the broader risk of inadvertently embedding discriminatory patterns into machine learning-based credit scoring models. The transparency provided by LIME is valuable here, as it allows practitioners and regulators to inspect how sensitive or proxy variables influence individual predictions, making potential bias visible rather than hidden inside the model. In this context, gender-related signals may point to underlying structural disparities (e.g., income stability, employment continuity, access to financial resources) that require attention, rather than being used as direct predictors. Therefore, explainability tools help ensure that improving predictive performance does not come at the expense of fairness and regulatory compliance.

6 Future work

For future work, it may be possible to combine stacking models with other explainable machine learning models to improve the trade-offs between interpretability and predictive accuracy. Developing approaches that can continuously adjust to new information (online learning or incremental learning) might increase their effectiveness rapidly in fast-changing environments. The generalizability of the suggested approach could be assessed by testing them on various P2P systems. By exploring the future directions mentioned, the research could lead to the development of more robust, adaptable and fair systems for predicting default risk in P2P lending platforms. Moreover, Future work will focus on incorporating out-of-time validation to evaluate the temporal stability of the proposed model and assess its robustness under evolving economic conditions, thereby enhancing its applicability for real-world financial decision-making. The absence of such temporal validation constitutes a notable limitation of the present study, which will be explicitly addressed in future research.

7 Conclusions

This study explored the credit risk within the Bondora dataset, Europe's leading P2P lending platform, by introducing a comprehensive framework for credit risk prediction that enhances investment decision-making in the P2P lending space. To address class imbalance, the SMOTE technique was employed, significantly improving model performance. Through the implementation of feature selection strategies, including filter-based (Chi-square), wrapper-based (SBS), and embedded methods (GBM, LGBM, AdaBoost, CatBoost), the most influential predictors of loan default were identified.

A range of classification models, linear (Logistic Regression), non-linear (SVM, Naïve Bayes), and tree-based (Decision Tree, Random Forest, AdaBoost, CatBoost), were evaluated, with top-performing classifiers used as base learners in stacking ensembles such as GBM, XGBoost, and LightGBM. Among these, LightGBM demonstrated superior performance, achieving an accuracy of 0.981, F-score of 0.980, and AUC of 0.994, outperforming benchmarks reported in existing literature.

To enhance interpretability, the best-performing model (LightGBM) was integrated with the LIME explainable AI framework. The use of LIME (aggregated) revealed several useful and interesting patterns in predicting P2P loan defaults. Unlike traditional global feature selection methods, which find features with consistent predictive power across the entire dataset, LIME focuses on how features influence predictions across groups and the overall dataset. This approach uncovered detailed, context-specific relationships between borrower characteristics and default risk, providing a more understandable view of how the model behaves at an aggregate level. Demographic

factors such as Education and Gender, which are often less important in global models, showed significant influence in specific cases, especially when financial indicators were unclear or weak. One important pattern involved the relationship between Applied amount and Income total; higher loan requests relative to income were linked to a higher risk of default, highlighting the need for personalized debt-to-income assessments beyond traditional financial ratios. Interest rate consistently stood out in both global and aggregated interpretations, with LIME reinforcing its importance at the group level, showing that borrowers are sensitive to loan pricing. Furthermore, LIME uncovered hidden risk factors, particularly for applicants with limited financial data but strong demographic signals, enabling more careful manual assessments or hybrid decision-making. Lastly, social features like Education were found to distinguish risk levels among borrowers with similar financial profiles, showing that softer features can improve model interpretability and sensitivity when used appropriately.

Author contributions Conceptualization, Markus Atef and Shimaa Ouf; methodology, Markus Atef and Shimaa Ouf; software, Markus Atef; investigation, Markus Atef, Shimaa Ouf, Wafaa Seoud and Menna Gabr; writing-original draft preparation, Markus Atef; writing-review and editing, Shimaa Ouf, Wafaa Seoud and Menna Gabr; supervision, Shimaa Ouf, Wafaa Seoud and Menna Gabr.

Funding No funding was received for the research reported upon in the paper.

Data availability The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests The authors declare that there are no conflicts of interest.

References

1. Much AM, Nikmah TL, Pertiwi DA, Jumanto S, Iswanto YD (2023) New model combination meta-learner to improve accuracy prediction P2P lending with stacking ensemble learning. Intell Syst Appl 18:200204. https://doi.org/10.1016/j.iswa.2023.200204
2. Yin W, Kirkulak-Uludag B, Zhu D, Zhou Z (2023) Stacking ensemble method for personal credit risk assessment in peer-to-peer lending. Appl Soft Comput 142:110302. https://doi.org/10.1016/j.asoc.2023.110302
3. Nguyen L, Ahsan M, Haider J (2024) Reimagining Peer-to-Peer Lending Sustainability: Unveiling Predictive Insights with Innovative Machine Learning Approaches for Loan Default Anticipation. FinTech 3(1):184–215. https://doi.org/10.3390/fintech3010012
4. Souadda LI, Halitim AR, Benilles B, Oliveira JM, Ramos P (2025) Optimizing Credit Risk Prediction for Peer-to-Peer Lending Using Machine Learning. Forecasting 7(3):35. https://doi.org/10.3390/forecast7030035
5. Kun Z, Weibin F, Jianlin W (2020) Default identification of P2P lending based on stacking ensemble learning. Proc Int Conf Econ Manag Model Eng (ICEMME) 2020:992–1006. https://doi.org/10.1109/ICEMME51517.2020.00203
6. Munsarif M, Muhammad S, Safuan S (2022) Peer to peer lending risk analysis based on embedded technique and stacking ensemble learning. Bull Electr Eng Inf 11:3483–3499. https://doi.org/10.11591/eei.v11i6.3927
7. Siham A, Sara S, Abdellah A (2021) Feature selection based on machine learning for credit scoring: an evaluation of filter and embedded methods. Proc Int Conf Innov Intell Syst Appl (INISTA) 1–6. https://doi.org/10.1109/INISTA52262.2021.9548410
8. Trivedi SK (2020) A study on credit scoring modeling with different feature selection and machine learning approaches. Technol Soc 63:101413. https://doi.org/10.1016/j.techsoc.2020.101413
9. Ribeiro MT, Singh S, Guestrin C (2016) Why should I trust you? Explaining the predictions of any classifier. Proc ACM SIGKDD Int Conf Knowl Discov Data Min 22:1135–1144. https://doi.org/10.1145/2939672.2939778
10. Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 30:4765–4774. https://doi.org/10.48550/arXiv.1705.07874
11. Sundararajan M, Taly A, Yan Q (2017) Axiomatic attribution for deep networks. Proc Int Conf Mach Learn 70:3319–3328. https://doi.org/10.48550/arXiv.1703.01365
12. Wachter S, Mittelstadt B, Floridi L (2017) Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv J Law Technol 31:841–887. https://doi.org/10.48550/arXiv.1711.00399

13. Yang R (2024) Machine learning-based loan default prediction in peer-to-peer lending. Highlights Sci Eng Technol 94:310–318. https://doi.org/10.53555/kuey.v30i5.5637
14. Akinjole A, Shobayo O, Popoola J, Okoyeigbo O, Ogunleye B (2024) Ensemble-based machine learning algorithm for loan default risk prediction. Mathematics 12:3423. https://doi.org/10.3390/math12213423
15. Knab P, Marton S, Schlegel U, Bartelt C (2025) Which LIME should I trust? Concepts, challenges, and solutions. https://doi.org/10.48550/arXiv.2503.24365
16. Salih AM, Raisi-Estabragh Z, Galazzo IB, Radeva P, Petersen SE, Lekadir K, Menegaz G (2025) A perspective on explainable artificial intelligence methods: SHAP and LIME. Adv Intell Syst 7:2400304. https://doi.org/10.1002/aisy.202400304
17. Li H, Wu W (2024) Loan default predictability with explainable machine learning. Financ Res Lett 60:104867
18. Zhu X, Chu Q, Song X, Hu P, Peng L (2023) Explainable prediction of loan default based on machine learning models. Data Sci Manag 6:123–133. https://doi.org/10.1016/j.dsm.2023.04.003
19. Vishwarupe V, Joshi PM, Mathias N, Maheshwari S, Mhaisalkar S, Pawar V (2022) Explainable AI and interpretable machine learning: A case study in perspective. Procedia Comput Sci 204:869–876. https://doi.org/10.1016/j.procs.2022.08.105
20. Chandrashekar G, Sahin F (2014) A survey on feature selection methods. Comput Electr Eng 40:16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024
21. McHugh M (2013) The Chi-square test of independence. Biochem Med 23:143–149. https://doi.org/10.11613/BM.2013.018
22. Li ZS, Yao X, Liu ZG, Zhang JC (2021) Feature selection algorithm based on LightGBM. J Northeast Univ (Nat Sci) 42:1688–1695. https://doi.org/10.12068/j.issn.1005-3026.2021.12.003
23. Strobl C, Boulesteix AL, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8:25. https://doi.org/10.1186/1471-2105-8-25
24. Dorogush AV, Ershov V, Gulin A (2018) CatBoost: Gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363. https://doi.org/10.48550/arXiv.1810.11363
25. Nasrabadi NM (2007) Pattern recognition and machine learning. J Electron Imaging 16(4):049901. https://doi.org/10.1117/1.2819119
26. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res 16:321–357. https://doi.org/10.1613/jair.953
27. Noble WS (2006) What is a support vector machine? Nat Biotechnol 24:1565–1567. https://doi.org/10.1038/nbt1206-1565
28. Provost F, Fawcett T (2013) Data science for business: What you need to know about data mining and data-analytic thinking. O'Reilly Media, Sebastopol
29. Mohri M, Rostamizadeh A, Talwalkar A (2012) Foundations of machine learning. MIT Press, Cambridge
30. Song YY, Lu Y (2015) Decision tree methods: Applications for classification and prediction. Shanghai Arch Psychiatry 27:130–135. https://doi.org/10.11919/j.issn.1002-0829.215044
31. Salman HA, Kalakech A, Steiti A (2024) Random Forest algorithm overview. Babylonian J Machine Learn 2024:69–79. https://doi.org/10.58496/BJML/2024/007
32. Wang H, Chen K, Zhu W, Song Z (2015) A process model on P2P lending. Financ Innov 1:1–8. https://doi.org/10.1186/s40854-015-0002-9
33. Wen X, Shao L, Xue Y, Fang W (2015) A rapid learning algorithm for vehicle classification. Inf Sci 295:395–406. https://doi.org/10.1016/j.ins.2014.10.040
34. Finlay S (2011) Multiple classifier architectures and their application to credit risk assessment. Eur J Oper Res 210:368–378. https://doi.org/10.1016/j.ejor.2010.09.029
35. Ma L, Zhao X, Zhou Z, Liu Y (2018) A new aspect on P2P online lending default prediction using meta-level phone usage data in China. Decis Support Syst 111:60–71. https://doi.org/10.1016/j.dss.2018.05.001
36. Xia Y, He L, Li Y, Liu N, Ding Y (2019) Predicting loan default in peer-to-peer lending using narrative data. J Forecast 39:260–280. https://doi.org/10.1002/for.2625
37. Ghorbani R, Ghousi R (2020) Comparing different resampling methods in predicting students' performance using machine learning techniques. IEEE Access 8:67899–67911. https://doi.org/10.1109/ACCESS.2020.2986809
38. Douzas G, Bacao F (2018) Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. Inf Sci 465:1–20. https://doi.org/10.1016/j.ins.2018.06.056
39. Kovács G (2019) An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. Appl Soft Comput 83:105662. https://doi.org/10.1016/j.asoc.2019.105662
40. Teslenko D, Sorokina A, Khovrat A, Huliiev N, Kyriy V (2023) Comparison of dataset oversampling algorithms and their applicability to the categorization problem. Innov Technol Sci Solut Ind 2:161–171. https://doi.org/10.30837/ITSSI.2023.24.161
41. Wongvorachan T, He S, Bulut O (2023) A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. Inf 14:54. https://doi.org/10.3390/info14010054

42. Fernández A, García S, Herrera F, Chawla N (2018) SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. J Artif Intell Res 61:863–905. https://doi.org/10.1613/jair.1.11192
43. Bergstra J, Bengio Y (2012) Random search for hyper-parameter optimization. J Mach Learn Res 13:281–305
44. Daoud EA (2019) Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset. Int J Comput Inf Eng 13:6–10. https://doi.org/10.5281/zenodo.3607805
45. Xia Y, Liu C, Li Y, Liu N (2017) A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Syst Appl 78:225–241. https://doi.org/10.1016/j.eswa.2017.02.017
46. Lu T, Zhang Y, Li B (2019) The value of alternative data in credit risk prediction: Evidence from a large field experiment. ICIS 2019 Proceedings, 10, pp. 1–16. https://aisel.aisnet.org/icis2019/data_science/data_science/10
47. He H, Garcia EA (2009) Learning from imbalanced data. IEEE Trans Knowl Data Eng 21:1263–1284. https://doi.org/10.1109/TKDE.2008.239
48. Khan A, Chaudhari O, Chandra R (2024) A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. Expert Syst Appl 244:122778. https://doi.org/10.1016/j.eswa.2023.122778
49. Chen W, Yang K, Yu Z et al (2024) A survey on imbalanced learning: latest research, applications and future directions. Artif Intell Rev 57:137. https://doi.org/10.1007/s10462-024-10759-6
50. Wolpert DH (1992) Stacked generalization. Neural Netw 5:241–259. https://doi.org/10.1016/S0893-6080(05)80023-1
51. Hasan Md (2024) Understanding model predictions: A comparative analysis of SHAP and LIME on various ML algorithms. J Sci Technol Res 5:17–26. https://doi.org/10.59738/jstr.v5i1.23(17-26).eaqr5800
52. Gao W, Ju M, Yang T (2023) Severe weather and peer-to-peer farmers' loan default predictions: Evidence from machine learning analysis. Financ Res Lett 58:104287. https://doi.org/10.1016/j.frl.2023.104287
53. Fachrie M, Musdholifah A, Pulungan R (2025) Effectiveness of data resampling and ensemble learning in multiclass imbalance learning. Artif Intell Rev 58:368. https://doi.org/10.1007/s10462-025-11357-w
54. Peng T, Ma C, Zhang Z, He R, Nazir MS, Zhang C (2026) An integrative approach to enhance photovoltaic power forecasting via TimeGAN-augmented data balancing and DES-improved autoformer model. Comput Electr Eng 130:110858. https://doi.org/10.1016/j.compeleceng.2025.110858
55. Nazir MS, Nazir HMR, Ullah H, Ji J, Zhang C (2026) Optimized energy system using a novel learning approach for low-carbon economic dispatch. Eng Res Express 8(1):015308. https://doi.org/10.1088/2631-8695/ae305c
56. Zhang X, Zhang C, He R, Ma C, Yao J, Nazir MS, Peng T (2025) A pyramidal attention-based transformer model based on improved differential innovation search algorithm and feature extraction for solar radiation prediction considering relevant factors. Renew Energy 253:123666. https://doi.org/10.1016/j.renene.2025.123666
57. Uddin N, Mahmud P, Ahammed M et al (2026) Explainable ensemble learning model for cardiovascular disease prediction with feature optimization and data balancing. Discov Comput 29:44. https://doi.org/10.1007/s10791-026-09930-0

Authors and Affiliations

Markus Atef[1,2] · Menna Ibrahim Gabr[2] · Wafaa Seoud[3] · Shimaa Ouf[2]

✉ Markus Atef
markatef@msa.edu.eg

[1] Faculty of Management Sciences, October University for Modern Sciences and Arts (MSA), Giza, Egypt

[2] Department of Information Systems, Faculty of Commerce and Business Administration, Helwan University, Cairo, Egypt

[3] Department of Business Administration, Faculty of Commerce and Business Administration, Helwan University, Cairo, Egypt