Expert Systems With Applications 165 (2021) 113986

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A benchmark of machine learning approaches for credit score prediction

Vincenzo Moscato, Antonio Picariello, Giancarlo Sperlí *

Department of Electrical Engineering and Information Technology (DIETI), University of Naples "Federico II", Via Claudio 21, Naples, Italy

ARTICLE INFO

Keywords:
Credit score prediction
Benchmark
Supervised learning
Machine learning
Explainable artificial intelligence

ABSTRACT

Credit risk assessment plays a key role for correctly supporting financial institutes in defining their bank policies and commercial strategies. Over the last decade, the emerging of social lending platforms has disrupted traditional services for credit risk assessment. Through these platforms, lenders and borrowers can easily interact among them without any involvement of financial institutes. In particular, they support borrowers in the fundraising process, enabling the participation of any number and size of lenders. However, the lack of lenders' experience and missing or uncertain information about borrower's credit history can increase risks in social lending platforms, requiring an accurate credit risk scoring. To overcome such issues, the credit risk assessment problem of financial operations is usually modeled as a binary problem on the basis of debt's repayment and proper machine learning techniques can be consequently exploited. In this paper, we propose a benchmarking study of some of the most used credit risk scoring models to predict if a loan will be repaid in a P2P platform. We deal with a class imbalance problem and leverage several classifiers among the most used in the literature, which are based on different sampling techniques. A real social lending platform (*Lending Club*) data-set, composed by 877,956 samples, has been used to perform the experimental analysis considering different evaluation metrics (i.e. AUC, Sensitivity, Specificity), also comparing the obtained outcomes with respect to the state-of-the-art approaches. Finally, the three best approaches have also been evaluated in terms of their explainability by means of different *eXplainable Artificial Intelligence* (XAI) tools.

1. Introduction

Nowadays, digital financial services represent one of the main *Big Data* sources. In fact, in the last two years, the global payments' revenue has grown of 12%, reaching the value of 1.9 trillion dollars in 2018 (McKinsey, 2010) by processing 14 trillion financial transactions per day.

The wide use of financial services has focused the attention of researchers on the credit risk management for developing models aiming, on one hand, to reduce financial risks and, on the other hand, to increase the related profits. According to Basel Committee on Banking Supervision (BCBS), banking risks can be classified in: (i) credit, (ii) market and (iii) operational risks. As described in Buehler et al. (2008), about 60% of the banks' threat is represented by credit risks, mainly due to the emerging of *Social Lending Platform*, also known as Peer-to-Peer (P2P) lending.

Through these platforms, lenders and borrowers can interact with each other without involving financial institutes. They support borrowers in the fundraising process, enabling the participation of every number and size of lenders. The report of the Online Lending Industry in 2017 has described how the amount of P2P platforms' transactions in China increased by 35.90% with respect to 2016's ones, while it had a growth rate of 50% since late 2017. Lender's risks in returning from their investment concern credit risk analysis, mainly due to borrowers' failure to pay back the loans, whose computation process is the credit risk assessment, the so-called *credit scoring*.

The credit risk assessment's problem of financial operations, including those supported by social lending platforms, is usually modeled as a binary classification problem (Hens & Tiwari, 2012; Verbraken et al., 2014), on the basis of debt's repayment. The loan payment status is dichotomous and fully paid loans are denoted as "0", while default loans are represented as "1". In 2016, about 30% of the unsecured installment loan sector was represented by social lending platforms (TransUnion, 2017).

Indeed, P2P lending platforms arise different challenges with respect to traditional methods due to the high dimension, sparse and imbalanced data (Sameer et al., 2019; Soui et al., 2019). Furthermore, in P2P lending platforms, the default risk increases with respect the traditional methods because a lender could be not able to effectively evaluate the risk level of borrowers (Guo et al., 2016). In particular, the main challenge concerns the evaluation of loan applicants' creditworthiness due to the lack of borrowers' credit history whose result could not improve by adding more features (Orsenigo & Vercellis, 2013).

* Corresponding author.
E-mail addresses: vincenzo.moscato@unina.it (V. Moscato), antonio.picariello@unina.it (A. Picariello), giancarlo.sperli@unina.it (G. Sperlí).

https://doi.org/10.1016/j.eswa.2020.113986
Received 25 May 2020; Received in revised form 7 September 2020; Accepted 7 September 2020
Available online 9 September 2020

P2P platforms produce, therefore, a large amount of unlabeled data that require online analysis for supporting lenders' real-time decisions (Kim & Cho, 2017). According to Namvar et al. (2018), credit scoring predictive models can be classified into two categories: statistical approaches and artificial intelligence methods.

Different statistical approaches have been proposed although they do not properly cover non-linear effects among different variables.

In this paper, we propose a benchmarking study of some credit risk scoring models based on the most used machine learning techniques in the literature to predict if a loan will be repaid in a P2P platform. We deal with a class imbalance problem by leveraging several classifiers based on different sampling techniques. An experimental analysis was conducted by considering different evaluation metrics (i.e. AUC, Sensitivity and Specificity) and the most promising state-of-the-art approaches, also in terms of the explainability of the adopted models using several *eXplainable Artificial Intelligence* (XAI) tools.

To the best of our knowledge, this is the first benchmark study aiming to evaluate at the same time classifiers' accuracy performances (in reducing false positive) and their explainability; in our opinion, it is very important to provide understandable results by human experts with a full transparency of decisions especially for the credit risk assessment problem.

The paper is organized as follows. Section 2 analyzes state-of-art approaches about credit risk prediction. Section 3 describes the proposed benchmark methodology for credit risk prediction. The evaluation made on a real P2P platform has been discussed in Section 4, while Section 5 summarizes the obtained results and suggests several areas for future work.

2. Related work

Credit risk assessment supports financial institutes in defining bank policies and commercial strategies. According to Wu et al. (2014) financial risk assessment is characterized by the following three properties: interconnection, dependence, and complexity. After the 2008 financial crisis, credit risk scoring has increasingly grown in importance becoming a critical means in credit risk management. In particular, it aims to support practitioners in the decision making process about loan's assignment to an applicant on the basis of different parameters.

Several approaches (Hayashi, 2016; Soui et al., 2019) rely on rules generation for evaluating credit risks. In particular, Hayashi (2016) generated a set of rules by three versions of *Re-RX* algorithm to evaluate credit risk from a Pareto optimal perspective. However, this approach is difficult to apply on large amounts of data due to different issues in the rules generation process.

In a similar way, in Soui et al. (2019) a credit risk evaluation model based on multi-optimization strategy produced a set of classification rules aiming, on one hand, to minimize the complexity of the generated solution, and, on the other hand, to maximize weights representing rules importance.

In the last years traditional services have been disrupted by online virtual places, called *social lending platforms*, in which lenders and borrowers interact among them without involving financial institutes. However, lack of lenders' experience and missing or uncertain information about borrower's credit history can increase risks in social lending platforms requiring an accurate credit risk scoring. These platforms provide several features that have been investigated by Emekter et al. (2015) and Malekipirbazari and Aksakalli (2015) for respectively evaluating their relevance using logistic regression and predicting borrower status through random forest based classification. In turn, Li et al. (2016) analyzed the Chinese P2P company's risk distribution by using a Cox Hazard Model with the aim to define policy implications for P2P sector.

Nevertheless, it is really hard to design model for credit risk prediction due to the high number of missing values, high-dimension and class-imbalanced data.

Concerning data preparation and feature selection issues, an analysis about different levels of attribute noise on several ensemble models has been performed in Twala (2010) showing how the impact of noise is based on the type of classification. Furthermore, Wang et al. (2014) introduced a feature selection algorithm into boosting, whilst (Koutanaei et al., 2015) exploited feature selection algorithms as a first stage to remove noisy attributes that are used on AdaBoost, BAGGING, random forest, and stacking. In addition, García et al. (2019) investigated the possible connection between the performances of classifier ensembles and the positive samples characterizing the complexity of 14 financial databases characterized by different kinds of positive samples. In García et al. (2012), the same authors analyzed how application of filtering algorithms can increase the accuracy of instance-based classifiers in the context of credit risk assessment.

A multi-criteria optimization based on one-norm regularization has been, then, proposed in Zhang et al. (2019) for credit risk assessment. Kim and Cho (2019) designed an interesting method that combines label propagation transductive support vector machine (TSVM) with Dempster–Shafer theory for social lending default prediction. In addition, using a social lending platform data, a semi-supervised approach based on SVM has been designed by Li et al. (2017) for reject inference in credit scoring.

Sun et al. (2018) developed a Decision Tree ensemble model based on the synthetic minority over-sampling technique (SMOTE) and the Bagging ensemble learning algorithm for credit risk evaluation about 138 Chinese listed companies with negative net profit. Another ensemble method has been proposed by Feng et al. (2018), whose classifiers are selected on the basis of their classification performance, for credit scoring. Furthermore, Marqués et al. (2012) evaluated prediction performances of five different ensemble methods with the aim to suggest the appropriate classifiers for each ensemble approach in the context of credit scoring. An ensemble strategy based on Choquet fuzzy integral has also been proposed by Namvar and Naderpour (2018) to improve credit score prediction in social lending platforms. Moreover, Xia et al. (2018) propose an ensemble credit model that integrates the bagging algorithm with the stacking method for credit score prediction. In Li et al. (2020) an heterogeneous ensemble learning is proposed combining different classifiers using a linear weight ensemble for social lending default prediction.

Finally, an ensemble of classifiers based on a distance-to-model and adaptive multi-view clustering (DM-ACME) learning method has been designed in Song et al. (2020) for predicting default risk in P2P.

In the last years several efforts have been made to produce intelligent agents that are able to explain their decisions (Townsend et al., 2019). According to Adadi and Berrada (2018), the interpretability concerns the understanding of a given model according to two different granular levels (Adadi & Berrada, 2018): (i) *Global* (analyzing the entire model behavior) and (ii) *Local* (focusing on a single prediction). Furthermore, interpretability techniques can also be classified according to their application in model *agnostic*, that are independent from the model, or *specific*, which are designed for specific problems. Different applications use black box prediction, being a data-mining and machine-learning obscure model, which requires to be explained (for more details see Guidotti, Monreale, Ruggieri, Turini et al., 2018). The desiderata about an explanation model can been summarized in the following three peculiarities (Doshi-Velez & Kim, 2017; Freitas, 2014): (i) interpretability, that can be understandable by human; (ii) accuracy, the model capability to predict unseen instances; (iii) Fidelity, because it is able to accurately imitate a black box predictor. Furthermore, explanation models for black box prediction can be summarized into two main classes (see Molnar (2020) for an extensive survey): *rule-based* (Grover et al., 2019; Guidotti, Monreale, Ruggieri, Pedreschi et al., 2018) or *features-based* (Lundberg & Lee, 2017; Ribeiro et al., 2016, 2018).

In this paper, we propose a benchmark of several machine learning classifiers from the discussed literature, using different sampling strategies with the aim to deal with loan assignment prediction on the basis of

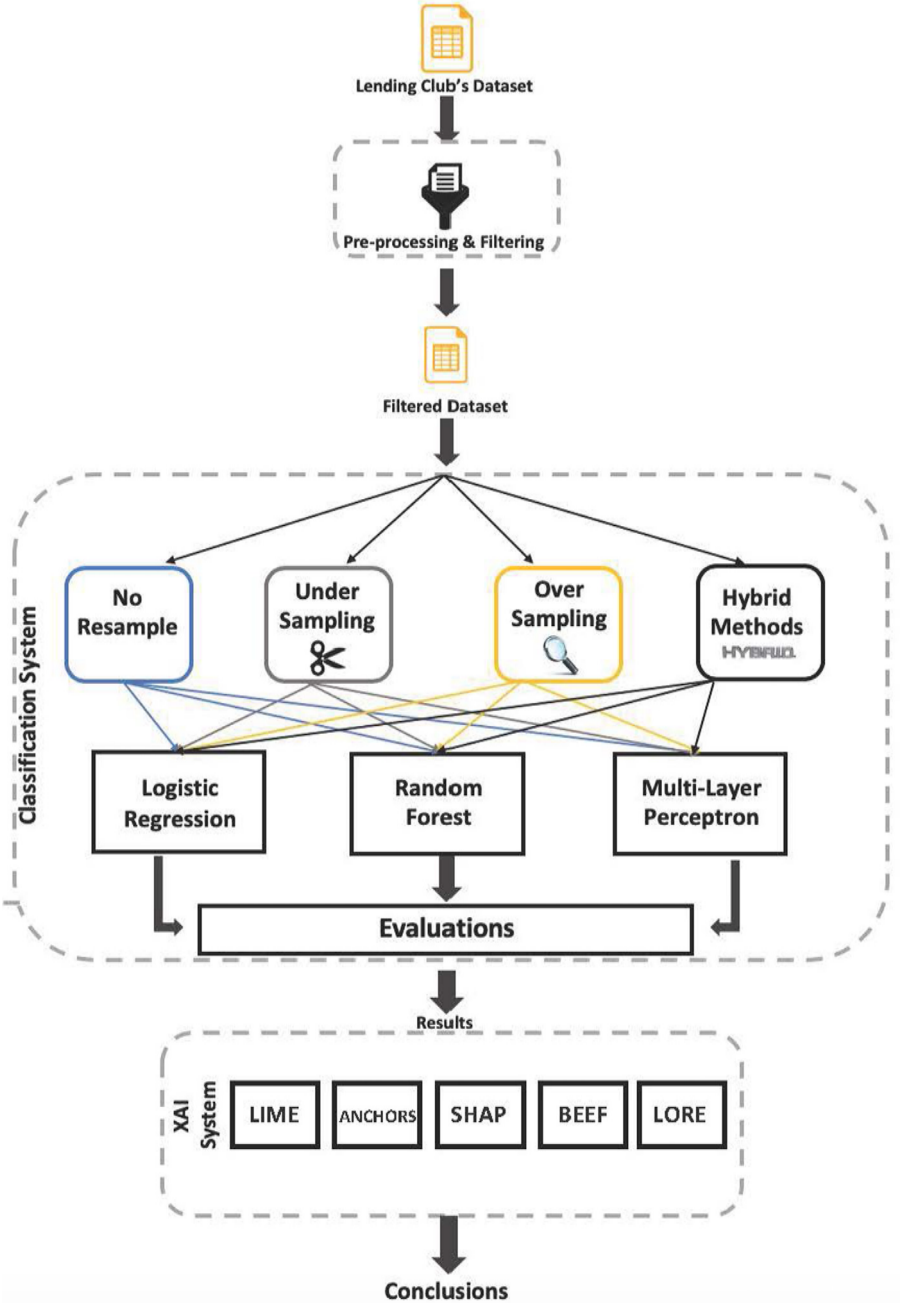

Fig. 1. Benchmark testbed.

data coming from social lending platforms. Our analysis aims to reduce false positives with respect to Namvar et al. (2018) and Marqués et al. (2012) because the misclassification costs are typically much higher than those associated to good loans class (as also shown in García et al. (2019). Furthermore, we performed a different pre-processing strategy with respect to Namvar et al. (2018). In particular: (i) first, we have carried out a statistical analysis to identify not relevant features that were excluded from our analysis according to a percentage of missing values; in turn, Namvar et al. (2018) removed all the missing values from dataset and used a data leakage strategy to identify features to discharge; (ii) then, we have leveraged a further analysis based on features' standard deviation and correlation for dropping out additional not relevant features; (iii) eventually, we have substituted the remaining missing values of relevant features with the related median, whilst in Namvar et al. (2018) the authors only removed them. In addition, we also focus on the explainability of the adopted models through the adaption of different *Explainable Artificial Intelligence* (XAI) tools with respect to Namvar et al. (2018) and Marqués et al. (2012).

3. The proposed benchmark methodology

Credit risk evaluation is a multidimensional and imbalanced problem, mainly based on a large volume of historical data such as: job status, credit history, personal account status and so on. Nevertheless, using all features, on one hand, increases the coverage while, on the other hand, could decrease the accuracy, requiring, therefore, a feature selection approach for properly handling high dimension data.

The proposed benchmark aims to deal with credit risk prediction problem for supporting investors in the evaluation of potential borrowers within social leading platforms. In these platforms a registered member fills out a comprehensive application about its financial history

and loan's reason to ask a lend without involving social financial intermediaries. Lenders can earn higher returns compared to savings and investment product offered by banks while borrowers ask for money at lower interest rates.

The benchmark testbed has been provided in Fig. 1, in which it is possible to note three main modules: *ingestion*, *classification* and *explanation*.

The ingestion module has the aim to crawl data from social lending platforms and to enhance data reliability by cleaning information and performing feature selection on the basis of the chosen classifier. In particular, data are firstly cleaned by removing features with a relevant number of missing or null values and zero variance attributes from the data-set. Once data are cleaned, several transformations are made on the data-set such as include converting categorical features to numeric ones, changing date attributes into numerical values. Furthermore, correlation analysis is computed with respect to the loan status to provide a better understanding of the data and their attributes' trend. The second component performs the credit prediction with respect to a given user that is affected by the imbalance problem, typical of the social lending platform, due to the high number of loans rejected compared to those requested.

In the classification stage we have chosen three of the most used classifiers for credit score prediction (Malekipirbazari & Aksakalli, 2015; Namvar & Naderpour, 2018; Sun et al., 2018), but any other classification techniques can be easily added to our framework. We have chosen these classifiers because, as shown in Marqués et al. (2012) and Namvar et al. (2018), they constitute the best solutions for the credit score prediction.

To deal with the imbalance problem, we used three different sampling strategies: random under-sampling, random over-sampling and smoothing. In particular, over-sampling creates new samples in the minority class, randomly duplicating the minority samples to balance the distribution of data. Synthetic minority oversampling technique (SMOTE) is another oversampling technique that uses k-nearest neighbors to produce new instances based on the distance between the minority data and some randomly selected nearest neighbors. In turn, under-sampling discards samples from majority class, randomly eliminating examples from the majority class to balance the class distribution.

Finally, the third module concerns the comparison of different XAI techniques for explaining the obtained results. More in details, the module is applied to the model classification to:

- Explain results of each prediction to underline how decisions are made.
- Explain to gain information about financial domain, to which Lending Club belongs.
- Explain to improve models in terms of performance and computational costs.

In particular, we compared five different XAI tools: *LIME*, *Anchors*, *SHAP*, *BEEF* and *LORE*.

LIME (Ribeiro et al., 2016) can be classified as a Post-Hoc and Model-Agnostic method that provides a Local explanation about the prediction realized. In turn, *Anchors* (Ribeiro et al., 2018) can be defined as a method with the same LIME's characteristics: it is a Post-Hoc, Model Agnostic that provides Local explanation by using rules that sufficiently "anchors" the predicted locally. Furthermore, *SHapley Additive exPlanations* (SHAP) (Lundberg & Lee, 2017) is a method to explain individual predictions based on the game theoretically optimal Shapley Values whose aims is to investigate how each feature affects the prediction.

Grover et al. (2019) developed *Balanced English Explanations of Forecasts* (BEEF) that uses global information, retrieved by clustering algorithm over the entire dataset, in order to generate a local explanation (in both supporting and counterfactual rules). Finally, Guidotti, Monreale, Ruggieri, Pedreschi et al. (2018) proposed *Local Rule-Based Explanations* (LORE) that firstly learns a local interpretable predictor and successively it derives the explanation as a decision rule.

Table 1
Data-set characterization.

Label	Dataset
Current	395.901
Fully paid	354,994
Charged off	107,384
Late (31–120 days)	12,550
In-grace period	4703
Late (16–30 days)	2393
Default	31
Total	**877,956**

Table 2
Pre-processing: Features excluded from our analysis because they have a percentage of missing values greater than 55%.

Features name	% of missing values
il_util	86.9
mths_since_recent_inq	90.7
emp_title	93.2
emp_length	93.3
num_tl_120dpd_2m	95.1
title	96.4
mths_since_rcnt_il	97.4
mo_sin_old_il_acct	97.4
bc_util	98.8

Table 3
Classifier/Resampling methods combination.

Classifier	Resampling methodology		Performance measure
Logistic regression Random forest Multi-layer perceptron	Under-sampling	RUS IHT	G-MEAN AUC Sensitivity (TPR) Specificity (TNR) False-positive rate Accuracy (ACC)
	Over-sampling	ROS SMOTE ADASYN	
	Hybrid method	SMOTE-TOKEN SMOTE-EN	

4. Experimental evaluation

The aim of our evaluation is to compare several classification engines' performances using several sampling strategies according to different evaluation metrics (see for more details Section 4.1).

For the proposed analysis we have selected the data-set provided by *Lending Club*,[1] a real P2P lending platform, focusing on financial data about loans grant in 2016 and 2017 composed by 877,956 samples and 151 features. According to Malekipirbazari and Aksakalli (2015) and Namvar et al. (2018), we have considered *loan_status* as target class for our problem, whose values are shown in Table 1.

Since we considered the binary problem if a loan will be paid back, we only take into account the labels "FullyPaid" or "Charged off", producing an unbalanced data-set in which 0.77% samples are fully paid whilst 0.23% are charged off.

In particular, we performed a 10-cross validation in which the data-set has been divided into a training and test set at ratio $75:25$ for each step. Furthermore, the performance of each classifier has been stored at each step for computing its mean and its standard deviation.

Finally, we have also compared the obtained results with respect to the ones computed in Namvar et al. (2018) and Song et al. (2020) using the same evaluation metrics.

The benchmark has been performed on Google Colab,[2] equipped with one single core hyper threaded Xeon Processors @2.3Ghz, 12 GB of RAM and a Tesla K80 having 2496 CUDA cores and 12 GB GDDR5 VRAM, using Python 3.6 with scikit-learn 0.23.1.[3]

[1] https://www.lendingclub.com/.
[2] https://colab.research.google.com/.
[3] https://scikit-learn.org/stable/index.html.

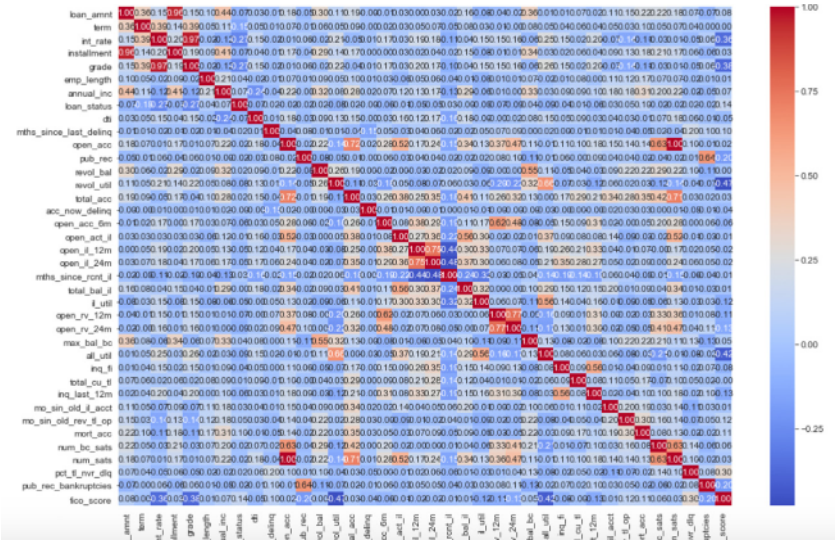

Fig. 2. Matrix correlation.

Table 4
Classification results (Under-Sampling approach).

Classifier	AUC	TPR	TNR	FP-Rate	G-Mean	ACC
LR - RUS	0.710 ± 0.014	0.658 ± 0.012	0.640 ± 0.019	0.356 ± 0.018	0.650 ± 0.015	0.650 ± 0.013
RF - RUS	**0.717 ± 0.008**	**0.630 ± 0.010**	**0.680 ± 0.011**	**0.320 ± 0.012**	**0.656 ± 0.011**	**0.640 ± 0.010**
MLP - RUS	0.680 ± 0.031	0.589 ± 0.028	0.650 ± 0.026	0.350 ± 0.033	0.590 ± 0.027	0.730 ± 0.028
LR - IHT	0.700 ± 0.018	0.350 ± 0.016	0.877 ± 0.015	0.120 ± 0.019	0.557 ± 0.018	0.470 ± 0.021
RF - IHT	0.694 ± 0.012	0.400 ± 0.013	0.580 ± 0.022	0.146 ± 0.019	0.580 ± 0.022	0.500 ± 0.021
MLP - IHT	0.696 ± 0.029	0.380 ± 0.031	0.850 ± 0.028	0.130 ± 0.029	0.570 ± 0.034	0.485 ± 0.028

4.1. Evaluation metrics

In credit risk prediction, several measures have been defined to evaluate the efficacy of the related models. According to Abellán and Castellano (2017), *accuracy* (ACC) measure may not be accurate because it does not consider that false positives are more important than false negatives. More suitable measures can be used as well as *Sensitivity* (TPR) and *Specificity* (TNR) that evaluate the accuracy of positive and negative samples respectively. Another suitable measure is the *G-mean* that can be used to evaluate the balance between classification performance in both minority and majority classes.

Furthermore, *Precision* and *FP-Rate* describe how good a model predicts positive and negative classes respectively. Finally, the *Area Under Curve* (AUC) measure determines the area under ROC curve, that summarizes the trade-off between the true positive rate and false positive rate for a predictive model exploiting different probability thresholds.

The equations for the above defined measures are shown below.

$$Sensitivity = \frac{TruePositive}{TruePositive + FalseNegative} \quad (1)$$

$$Specificity = \frac{TrueNegative}{TrueNegative + FalsePositive} \quad (2)$$

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \quad (3)$$

$$FP - Rate = \frac{FalsePositive}{FalsePositive + TrueNegative} \quad (4)$$

$$G - MEAN = \sqrt{Sensitivity * Specificity} \quad (5)$$

4.2. Data preparation

The aim of this section is to enhance data reliability using data cleaning and feature selection techniques. The former has been firstly performed by removing features with a percentage of missing values (shown in Table 2) greater than 55%: *il_util*, *mths_ since_recent_inq*, *emp_title*, *emp_length*, *num_tl_120dpd_2m*, *title*, *mths_ since_rcnt_il*, *mo_sin_old_il_acct*, *bc_util*.

We also remove *id* and *hardship_flag* features because they have one value in the data-set. Furthermore, we remove *emp_title* and *emp_length* due to the high standard deviation and we dropped out *sub_grade* because it has very similar value with respect to grade.

Fig. 2 shows correlation between features whose analysis allows to drop out the additional following ones: *mo_sin_old_rev_tl_op*, *installment*, *total_acc*, *pub_rec_bankruptcies*, *int_rate*, *num_bc_sats*, *num_sats*, *open_il_24m*, *open_il _12m*, *open_rv_24m*, *open_rv_12m*.

Finally, we have substituted missing values with feature's median and we transformed nominal features in binary data.

4.3. Experimental results

We evaluated the performance of classifiers according to different sampling strategies as shown in Table 3.

In particular, we have chosen three classifiers (Random Forest, Logistic Regression and Multilayer perceptron) using different of sampling strategies, whose best pairs are compared. In particular, we used the RandomForestClassifier with n_estimators = 100 and max_depth = 4, the LogisticRegression with penalty = "l2" and solver = "lbfgs" and the MLPClassifier with hidden_layer_sizes = 100, solver = "adam" and alpha = 0.0001

From a G-Means perspective, all classifiers in combination with Random Under Sampling (RUS) significantly outperform other under-sampling combinations, whose best result corresponds to the *RF-RUS* (G-Means 0.656) as shown in Table 4.

Table 5 describes the classifiers' performances using over-sampling strategies.

As it is easy to note, lowest values of G-Mean and Specificity are obtained using Random Forest while it performs the highest FP-Rate

Table 5
Classification results (Over-Sampling approach).

Classifier	AUC	TPR	TNR	FP-Rate	G-Mean	ACC
LR - ROS	**0.710 ± 0.009**	**0.659 ± 0.011**	**0.642 ± 0.012**	**0.360 ± 0.009**	**0.6503 ± 0.014**	**0.650 ± 0.0012**
RF - ROS	0.718 ± 0.009	0.855 ± 0.014	0.392 ± 0.012	0.600 ± 0.011	0.5800 ± 0.010	0.746 ± 0.015
MLP - ROS	0.685 ± 0.032	0.595 ± 0.029	0.658 ± 0.034	0.300 ± 0.031	0.6100 ± 0.030	0.675 ± 0.027
LR - SMOTE	0.709 ± 0.004	0.660 ± 0.005	0.639 ± 0.004	0.360 ± 0.003	0.6502 ± 0.004	0.656 ± 0.002
RF - SMOTE	0.710 ± 0.004	0.976 ± 0.004	0.100 ± 0.004	0.900 ± 0.004	0.314 ± 0.004	0.770 ± 0.004
MLP - SMOTE	0.700 ± 0.015	0.600 ± 0.013	0.686 ± 0.016	0.280 ± 0.018	0.637 ± 0.018	0.530 ± 0.019
LR - ADASYN	0.709 ± 0.011	0.679 ± 0.011	0.620 ± 0.011	0.380 ± 0.011	0.649 ± 0.011	0.667 ± 0.011
RF - ADASYN	0.715 ± 0.009	0.977 ± 0.012	0.100 ± 0.019	0.900 ± 0.014	0.314 ± 0.013	0.774 ± 0.014
MLP - ADASYN	0.680 ± 0.021	0.583 ± 0.024	0.678 ± 0.029	0.384 ± 0.022	0.615 ± 0.024	0.660 ± 0.021

Table 6
Classification results (Hybrid approach).

Classifier	AUC	TPR	TNR	FP-Rate	G-Mean	ACC
LR - SmoteToken	**0.710 ± 0.011**	**0.660 ± 0.009**	**0.640 ± 0.014**	**0.360 ± 0.011**	**0.650 ± 0.012**	**0.656 ± 0.010**
RF - SmoteToken	0.714 ± 0.020	0.978 ± 0.018	0.100 ± 0.019	0.900 ± 0.021	0.310 ± 0.022	0.770 ± 0.019
MLP - SmoteToken	0.690 ± 0.030	0.536 ± 0.032	0.740 ± 0.027	0.347 ± 0.029	0.626 ± 0.031	0.605 ± 0.032
LR - SmoteEnn	0.700 ± 0.035	0.435 ± 0.032	0.820 ± 0.039	0.180 ± 0.032	0.600 ± 0.032	0.507 ± 0.031
RF - Enn	0.710 ± 0.029	0.900 ± 0.026	0.290 ± 0.031	0.710 ± 0.032	0.510 ± 0.029	0.760 ± 0.028
MLP - SmoteEnn	0.690 ± 0.041	0.376 ± 0.038	0.850 ± 0.037	0.100 ± 0.037	0.558 ± 0.039	0.527 ± 0.039

Table 7
Our best Classification results.

Classifier	AUC	TPR	TNR	FP-Rate	G-Mean	ACC
RF - RUS	**0.717**	**0.630**	**0.680**	**0.320**	**0.6560**	**0.640**
LR - ROS	0.710	0.659	0.642	0.360	0.6503	0.650
LR - SmoteToken	0.710	0.660	0.640	0.360	0.6500	0.656
Logistic regression	0.685	0.983	0.069	0.960	0.2600	0.770
Random forest	0.720	0.983	0.084	0.920	0.2870	0.773
MLP	0.704	0.990	0.040	0.945	0.2060	0.771

Table 8
Best result in Namvar et al. (2018).

Classifier	AUC	TPR	TNR	FP-Rate	G-Mean	Accuracy
RF - RUS	**0.6960**	**0.717**	**0.582**	**0.420**	**0.65**	**0.6920**
Linear discrimination analysis - SMOTE	0.7000	0.630	0.650	0.350	0.643	0.6400
LR - SmoteToken	0.7000	0.638	0.648	0.352	0.643	0.6400
Logistic regression	0.7030	0.988	0.048	0.950	0.218	0.8173
Random forest	0.6960	0.996	0.015	0.980	0.12	0.8176

Table 9
Result in Song et al. (2020).

	Method	AUC	TPR	TNR	G-Mean	Accuracy
	Song et al. (2020)	0.6697	0.4607	0.7678	0.6009	0.7231
Over-sampling	GBDT	0.6207	0.6168	0.6246	0.6207	0.6235
	Random forest	0.5795	0.3107	0.8423	0.5134	0.7701
	AdaBoost	0.5224	0.1925	0.8523	0.4050	0.7562
	Decision tree	0.5231	0.1934	0.8527	0.4060	0.7568
	Logistic regression	0.5600	0.5558	0.5642	0.5597	0.5630
	Multilayer perceptron	0.4892	0.1572	0.8211	0.3593	0.7245
Under-sampling	GBDT	0.6140	0.6292	0.5989	0.6138	0.6033
	Random forest	0.6207	0.6623	0.5791	0.6193	0.5912
	AdaBoost	0.5408	0.5577	0.5238	0.5404	0.5288
	Decision tree	0.5421	0.5558	0.5283	0.5418	0.5323
	Logistic regression	0.5615	0.5437	0.5794	0.5609	0.5742
	Multilayer perceptron	0.4892	0.1572	0.8211	0.3593	0.7245

values with respect to the other combinations. Therefore, Random Forest is not a suitable classifier with all over-sampling approach. In turn, Logistic Regression obtains the highest values of G-Mean by using ROS (G-Means 0.6503) and SMOTE (G-Means 0.6502) and it has also similar AUC values (0.71 and 0.709 using ROS and SMOTE respectively).

In conclusion, among all over-sampling methods LR-ROS emerged as the best approaches.

Table 6 describes performance measures about hybrid sampling methods. whose the best result is represented by LR-Smote-Token (G-Means value 0.65).

In Table 7 we compared the best combination obtained by using different sampling strategies, including also the results of the three classifiers without sampling strategies.

The non-sampling classifiers have the highest values of accuracy due to they are biased towards the majority class, whilst they performed worst in terms of G-Mean, specificity, and false positive rates. In term of G-Mean, MLP obtains the lowest value (0.206) while RF-RUS has the highest value (0.656). However Logistic Regression, Random Forest and MLP have Sensitivity values higher with respect to sampling approaches. This difference between non-sampling and sampling proves the effectiveness of the latter techniques on the prediction's performance. RF-RUS emerged as the best method for predicting a borrower's status in a social lending marketplace.

4.4. Comparison with state-of-the-art results

Finally, we compared the obtained results with respect to the best ones in Namvar et al. (2018) and Song et al. (2020).

Our best combination (RF-RUS) (Table 7) shows lowest accuracy with respect to the best ones in Namvar et al. (2018), shown in Table 8, whilst our AUC value (0.717) and Specificity (0.68) are higher with respect to the best ones in Namvar et al. (2018). It is worth to note that it is important to reduce the number of false positive in credit score prediction because the misclassification costs are typically much higher than those associated to good loans class (as also shown in García et al., 2019).

In turn, Table 9 shows higher values of specificity with respect to our results although its sensitivity value is much lower than ours.

4.5. Explanation results

The last evaluation has concerned the performance comparison of several XAI tools (*LIME*, *Anchors*, *SHAP*, *BEEF* and *LORE*) in terms of *Precision* measure – what fraction of the predictions were correct – on the our three best classifiers' combinations (Random Forest & Random Under-Sampling, Logistic Regression & Random Over-Sampling and Logistic Regression & Smote-Token) according to the experimental protocol described in Ribeiro et al. (2016).

First of all, to simulate trust on an individual prediction, we randomly chose a group of possible features (25% of the total) that must be consider "untrustworthy", assuming that a user, that can recognize

Table 10
Comparison between Anchors, Lime, SHAP, BEEF and LORE in terms of Precision measure.

	Random -Forest Random under-sampling (Precision value)	Logistic regression Random over-sampling (Precision value)	Logistic regression Smote -Token (Precision value)
Anchors	0.907	0.547	0.747
Lime	0.872	0.918	0.676
SHAP	0.891	0.924	0.752
BEEF	0.881	0.741	0.725
LORE	0.913	0.878	0.781

them, does not want to trust on these features. It is important to note that these features are chosen more than one time because we select a different bunch of unreliable features at each round. For each combination of the chosen features, an oracle has been developed by labeling test set predictions from a black-box classifier as "untrustworthy" if the prediction changes when untrustworthy features are removed from the instance, and trustworthy otherwise.

Furthermore, we assume that users consider predictions untrustworthy from the XAI tools if the prediction of the black-box classifier changes when all untrustworthy features that appear in the explanations are removed (the simulated human "discounts" the effect of untrustworthy features). Finally, for each test set prediction, we can evaluate whether the simulated user trusts it, using each explanation method, and compare it to the trustworthiness oracle. Several explanations have been generated by using different sets of instances that are computed by several random sampling (10 runs) from data-set.

As we can see in Table 10, *LORE* shows better results because it combines local predictions with the use of counterfacts to generate the explanation that supports the user's understanding the changes in the instance's features that lead to a different outcome. Furthermore, *LIME* achieves good coverage values for the three configurations since it models the prediction as a weighted sum making it easy to understand how the prediction is generated. In turn, *SHAP*, which is also based on feature importance, obtains statistically more reliable outcomes than LIME through the use of shap values, whose computational complexity, typically approximated by different heuristics, could affect the efficiency of the explanation. Finally, *BEEF* and *Anchors*, relying on axis-aligned hyper-rectangle and specific rules (called *Anchors*), could be limited in expressive power (as we can see for Logistic Regression).

5. Conclusion

Risk prediction score is one of the main challenges in finance sector for supporting people in their investments. Nevertheless, different challenges are faced by P2P lending platform with respect to the traditional ones due to the high dimension and imbalanced data.

The aim of the proposed approach is to design a benchmark for machine learning approaches for credit risk prediction for social lending platforms, also able to manage unbalanced data-sets. The evaluation made on real world social lending platforms shows the feasibility of some of the analyzed approaches w.r.t. their explainability.

Future works will be devoted to improve the evaluation considering also other real P2P lending platforms and to investigate other more recent approaches as well as deep learning or ensemble strategies that in some cases could show better performances.

CRediT authorship contribution statement

Vincenzo Moscato: Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Drafting the manuscript, Revising the manuscript critically for important intellectual content, Approval of the version of the manuscript to be published. **Antonio Picariello:** Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Drafting the manuscript, Revising the manuscript critically for important intellectual content, Approval of the version of the manuscript to be published. **Giancarlo Sperlí:** Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Drafting the manuscript, Revising the manuscript critically for important intellectual content, Approval of the version of the manuscript to be published.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abellán, J., & Castellano, J. G. (2017). A comparative study on base classifiers in ensemble methods for credit scoring. *Expert Systems with Applications*, *73*, 1–10.

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160.

Buehler, K., Freeman, A., & Hulme, R. (2008). The new arsenal of risk management. *Harvard Business Review*, *86*(9), 93–100.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Emekter, R., Tu, Y., Jirasakuldech, B., & Lu, M. (2015). Evaluating credit risk and loan performance in online peer-to-peer (p2p) lending. *Applied Economics*, *47*(1), 54–70.

Feng, X., Xiao, Z., Zhong, B., Qiu, J., & Dong, Y. (2018). Dynamic ensemble classification for credit scoring using soft probability. *Applied Soft Computing*, *65*, 139–151.

Freitas, A. A. (2014). Comprehensible classification models: a position paper. *ACM SIGKDD Explorations Newsletter*, *15*(1), 1–10.

García, V., Marqués, A., & Sánchez, J. (2012). On the use of data filtering techniques for credit risk prediction with instance-based models. *Expert Systems with Applications*, *39*(18), 13267–13276.

García, V., Marqués, A. I., & Sánchez, J. S. (2019). Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction. *Information Fusion*, *47*, 88–101.

Grover, S., Pulice, C., Simari, G. I., & Subrahmanian, V. S. (2019). Beef: Balanced english explanations of forecasts. *IEEE Transactions on Computational Social Systems*, *6*(2), 350–364.

Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., & Giannotti, F. (2018). Local rule-based explanations of black box decision systems. arXiv preprint arXiv:1805.10820.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, *51*(5), 1–42.

Guo, Y., Zhou, W., Luo, C., Liu, C., & Xiong, H. (2016). Instance-based credit risk assessment for investment decisions in p2p lending. *European Journal of Operational Research*, *249*(2), 417–426.

Hayashi, Y. (2016). Application of a rule extraction algorithm family based on the re-rx algorithm to financial credit risk assessment from a Pareto optimal perspective. *Operations Research Perspectives*, *3*, 32–42.

Hens, A. B., & Tiwari, M. K. (2012). Computational time reduction for credit scoring: An integrated approach based on support vector machine and stratified sampling method. *Expert Systems with Applications*, *39*(8), 6774–6781.

Kim, A., & Cho, S.-B. (2017). Dempster-shafer fusion of semi-supervised learning methods for predicting defaults in social lending. In *International conference on neural information processing* (pp. 854–862). Springer.

Kim, A., & Cho, S.-B. (2019). An ensemble semi-supervised learning method for predicting defaults in social lending. *Engineering Applications of Artificial Intelligence*, *81*, 193–199.

Koutanaei, F. N., Sajedi, H., & Khanbabaei, M. (2015). A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring. *Journal of Retailing and Consumer Services*, *27*, 11–23.

Li, W., Ding, S., Wang, H., Chen, Y., & Yang, S. (2020). Heterogeneous ensemble learning with feature engineering for default prediction in peer-to-peer lending in china. *World Wide Web*, *23*(1), 23–45.

Li, J., Hsu, S., Chen, Z., & Chen, Y. (2016). Risks of p2p lending platforms in china: Modeling failure using a cox hazard model. *The Chinese Economy*, *49*(3), 161–172.

Li, Z., Tian, Y., Li, K., Zhou, F., & Yang, W. (2017). Reject inference in credit scoring using semi-supervised support vector machines. *Expert Systems with Applications*, *74*, 105–114.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems, Vol. 30* (pp. 4765–4774). Curran Associates, Inc..

Malekipirbazari, M., & Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert Systems with Applications*, *42*(10), 4621–4631.

Marqués, A., García, V., & Sánchez, J. (2012). Exploring the behaviour of base classifiers in credit scoring ensembles. *Expert Systems with Applications*, *39*(11), 10244–10250.

McKinsey (2010). Global payments report 2019. https://www.mckinsey.com/~/media/mckinsey/industries/financial%20services/our%20insights/tracking%20the%20sources%20of%20robust%20payments%20growth%20mckinsey%20global%20payments%20map/global-payments-report-2019-amid-sustained-growth-vf.ashx (Online; accessed 20 March 2020).

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Namvar, A., & Naderpour, M. (2018). Handling uncertainty in social lending credit risk prediction with a choquet fuzzy integral model. In *2018 IEEE international conference on fuzzy systems (FUZZ-IEEE)* (pp. 1–8). IEEE.

Namvar, A., Siami, M., Rabhi, F., & Naderpour, M. (2018). Credit risk prediction in an imbalanced social lending environment. *International Journal of Computational Intelligence Systems*, *11*(1), 925–935.

Orsenigo, C., & Vercellis, C. (2013). Linear versus nonlinear dimensionality reduction for banks' credit rating prediction. *Knowledge-Based Systems*, *47*, 14–22.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *Thirty-second AAAI conference on artificial intelligence*.

Sameer, F., Bakar, M., Zaidan, A., & Zaidan, B. (2019). A new algorithm of modified binary particle swarm optimization based on the gustafson-kessel for credit risk assessment. *Neural Computing and Applications*, *31*(2), 337–346.

Song, Y., Wang, Y., Ye, X., Wang, D., Yin, Y., & Wang, Y. (2020). Multi-view ensemble learning based on distance-to-model and adaptive clustering for imbalanced credit risk assessment in p2p lending. *Information Sciences*, *525*, 182–204.

Soui, M., Gasmi, I., Smiti, S., & Ghédira, K. (2019). Rule-based credit risk assessment model using multi-objective evolutionary algorithms. *Expert Systems with Applications*, *126*, 144–157.

Sun, J., Lang, J., Fujita, H., & Li, H. (2018). Imbalanced enterprise credit evaluation with dte-sbd: Decision tree ensemble based on smote and bagging with differentiated sampling rates. *Information Sciences*, *425*, 76–91.

Townsend, J., Chaton, T., & Monteiro, J. M. (2019). Extracting relational explanations from deep neural networks: A survey from a neural-symbolic perspective. *IEEE Transactions on Neural Networks and Learning Systems*.

TransUnion (2017). Fintechs taking larger share of personal loan market while increasing portfolio risk-return performance. https://newsroom.transunion.com/fintechs-taking-larger-share-of-personal-loan-market-while-increasing-portfolio-risk-return-performance/ (Online; accessed 20 March 2020).

Twala, B. (2010). Multiple classifier application to credit risk assessment. *Expert Systems with Applications*, *37*(4), 3326–3336.

Verbraken, T., Bravo, C., Weber, R., & Baesens, B. (2014). Development and application of consumer credit scoring models using profit-based classification measures. *European Journal of Operational Research*, *238*(2), 505–513.

Wang, G., Ma, J., & Yang, S. (2014). An improved boosting based on feature selection for corporate bankruptcy prediction. *Expert Systems with Applications*, *41*(5), 2353–2361.

Wu, D. D., Chen, S.-H., & Olson, D. L. (2014). Business intelligence in risk management: Some recent progresses. *Information Sciences*, *256*, 1–7, Business Intelligence in Risk Management.

Xia, Y., Liu, C., Da, B., & Xie, F. (2018). A novel heterogeneous ensemble credit scoring model based on bstacking approach. *Expert Systems with Applications*, *93*, 182–199.

Zhang, Z., He, J., Gao, G., & Tian, Y. (2019). Sparse multi-criteria optimization classifier for credit risk evaluation. *Soft Computing*, *23*(9), 3053–3066.