A benchmark of credit score prediction using Machine Learning

Vincenzo Moscato[1], Giancarlo Sperlì[1,*]

[1]Department of Electrical Engineering and Information Technology (DIETI), University of Naples ”Federico II”, Via Claudio 21, 80125, Naples, Italy

Abstract

One of the main relevant financial services is the credit risk assessment, whose aim is to support financial institutes in defining their policies and strategies. In the last years, traditional credit risk services have been disrupted by the arise of Social Lending Platforms. This paper reports an experimental analysis relying on the use of different machine learning models to deal with credit risk in social lending platform. For this reason, we use a real world dataset, composed by 877,956 samples, to compare our results w.r.t. state-of-the-art baselines and benchmarks, also evaluating the explanaibility of the proposed three best models using different well-known XAI tools. Hence, the proposed study aims to design both effectiveness and explainable credit risk models.

Keywords

Credit Score Prediction, Benchmark, Machine Learning, Explainable Artificial Intelligence

1. Introduction

In the last years, the pervasive use of Artificial Intelligence (AI) models has brought effectiveness improvements in several application domains, including the financial sector. Nowadays, several financial services have benefited from the introduction of artificial intelligence-based models by defining a new generation of financial technology (FinTech)-based systems, which have enabled the definition of a range of services such as lending, payment, risk and regulatory management [1, 2]. Hence one of the main challenge is the large of data produced by digital financial services; in fact, the financial transaction processed per day hanno raggiunto il valore di 14 trillioni, generando un incremento delle revenue del global payments del 12% negli ultimi due anni raggiungendo un valore pari a 1.9 trilions of dollars in 2018 [3].

In particular, researchers and practitioners have been increasingly interested in defining AI-based methodologies with the aim to jointly increase their revenues and minimize associated risks, leading to new opportunities and challenges, as discussed in [4]. The Basel Committee on Banking Supervision (BCBS) has classified banking risks into three categories, namely credit, market, and operational risks. According to [5], credit risks account for approximately 60% of banks’ risks., which is mainly due to the arise of Social Lending Platforms.

These platforms enable communications among lenders and borrowers without any transaction costs, that are typically for traditional financial institute. These platforms facilitate the fundraising process for borrowers by allowing lenders of all sizes to participate. In fact, a study shows how the Social Lending platform transactions in China had grown by 35.90% and 50% w.r.t. the previous years and since late 2017, respectively. Nevertheless, lenders are exposed to risks when investing in P2P lending, particularly in the form of credit risk, which is assessed through the process of *credit scoring*. This risk arises primarily from the possibility that borrowers may be unable to repay their loans.

Typically, credit risk assessment for financial operations, including Social Lending transactions, is defined as a binary classification problem [6, 7], where the focus is on whether debts are repaid or not. The loan payment status is classified as either fully paid (represented as ”0”) or defaulted (represented as ”1”). Nevertheless, according to a report made by TransUnion in 2016, social lending platforms made up approximately 30% of the unsecured installment loan sector [8].

Hence, Social lending platforms pose unique challenges w.r.t. traditional methods, dealing with high-dimensionality, sparsity, and imbalance data ([9, 10]). Furthermore, the risk of defaults in P2P lending platforms is generally higher than in traditional methods due to the issues of lenders in accurately assessing borrowers’ risk levels ([11]). Hence, the primary challenge concerns how it is possible to evaluate creditworthiness of loan applicants, since borrowers often lack a sufficient credit history, and simply adding more features may not necessarily improve the accuracy of the assessment [12, 13].

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

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy

*Corresponding author.
vincenzo.moscato@unina.it (V. Moscato); name.surname@organization.it (G. Sperlì)
0000-0002-0754-7696 (V. Moscato); 0000-0003-4033-3777 (G. Sperlì)

CEUR Workshop Proceedings (CEUR-WS.org)

among different variables.

This paper represents an extended abstract of our previous study [14], where we designed a benchmark of machine learning models for credit scoring prediction, whose results have been compared w.r.t. the state of the art ones. In particular, the credit scoring task has been designed as a binary problem corresponding to the decision whether a loan or no on Social Lending platforms. The results have been investigated using several sampling strategies for dealing with the unbalanced issues in these datasets and different measures, also using eXplainable Artificial Intelligence (XAI) tools for explaining the prediction of the analyzed machine learning models.

2. Methodology

The proposed benchmark is designed to deal with the credit risk prediction task with the aim tosupport investors in evaluating potential borrowers on social lending platforms. In particular, members, registered on these platforms, complete a detailed application regarding their financial history and the reason for seeking a loan, without the involvement of financial intermediaries. Lenders can earn higher returns than what is typically offered through banks' savings and investment products, while borrowers can access funds at lower interest rates.

Figure 1 shows the three main components in the benchmark testbed: *ingestion*, *classification* and *explanation*.

The ingestion module is responsible for crawling data from social lending platforms, also performing data cleaning and feature selection operations on the basis of the chosen classifier. Firstly, data is cleaned by removing features having a significant number of missing or null values, as well as zero variance attributes. Successively, several transformations are performed on the dataset, such as converting categorical features into numeric ones and changing date attributes into numerical values. Additionally, a correlation analysis is conducted with respect to the loan status to gain a better understanding of the data and their attribute trends.

The second component is responsible for credit prediction for a given user, which is impacted by the imbalance problem, typical issue in Social Lending platforms. This imbalance problem arises due to the high number of rejected loans compared to those that are requested.

For the classification stage, three of most efficacy models in credit score prediction have been selected we selected three of the most commonly used classifiers for credit score prediction [15, 16, 17, 18, 19].

Furthermore, we train the chosen machine learning models on the basis of different sampling strategies to address data imbalance issues: random under-sampling and over-sampling, that respectively and smoothing.

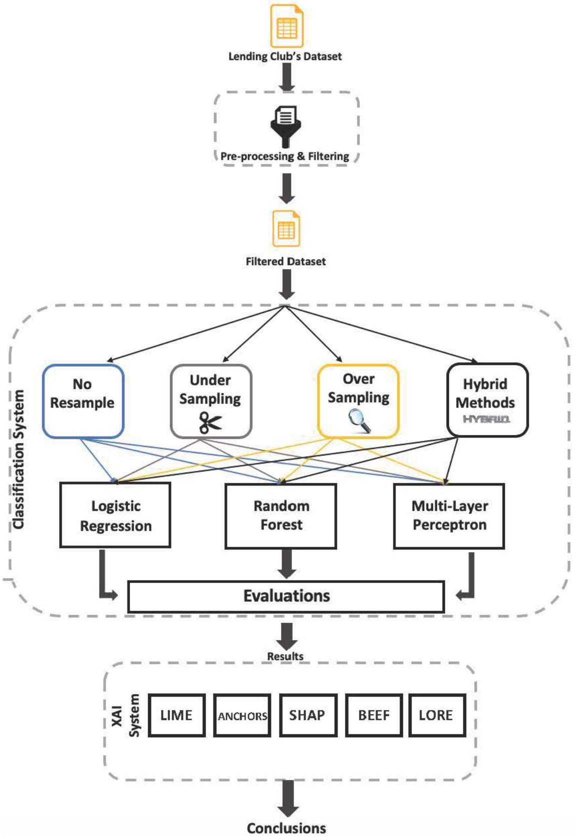

Figure 1: Benchmark Testbed

The third module deals with comparing different XAI techniques to explain the results obtained with the aim to explaine prediction outcome for highlighing how decisions are made. In particular, we compared five different XAI tools: *LIME* [20], *Anchors* [21], *SHapley Additive exPlanations* (SHAP) [22], *Balanced English Explanations of Forecasts* (BEEF) [23] and *Local Rule-Based Explanations* (LORE) [24].

3. Experimental Evaluation

In this section, we describe the analysis made for evaluating the effectiveness of different classification models on the basis of several sampling strategies and evaluation metrics.

In particular, we have used a dataset from a real-world Social Lending platform, named Lending Club[1], including 877, 956 samples and 151 features, with the target class for our problem being the loan status. As suggested by

[1] https://www.lendingclub.com/

previous research ([15, 19]), we have used the values of the loan status, which are presented in Table 1.

Loan Status	Samples number
Current	395.901
Fully Paid	354,994
Charged Off	107,384
Late (31-120 days)	12,550
In-grace period	4,703
Late (16-30 days)	2,393
Default	31
Total	**877,956**

Table 1
Data-set characterization

We only included ”FullyPaid” or ”Charged off” labels due to we are intereting in predicting wheter a loan would be paid back or not. Under this assumption, we generate an imbalanced dataset, in which 77% and 23% of samples are fully paid and charged off, respectively. Furthermore, we perform a 10-fold cross-validation, in which we split the dataset according to 75:25 ratio for each fold, computing mean and standard deviation for each classifier during the training process.

The best results have been compared w.r.t. the ones in [19, 25] on the basis of several metrics (Precision, FP-Rate, Area Under Curve (AUC), accuracy (ACC), Sensitivity (TPR), Specificity (TNR), and G-mean).

The analysis has been made on a Platform-as-a-Service (PaaS) Google Colab[2], providing 12 GB of RAM and a Tesla K80 with 2496 CUDA core and a software stack composed by Python 3.6 with scikit-learn 0.23.1[3].

4. Results

In this section, our highest results shown in Table 2 w.r.t. the best ones in ([19]) and ([25]).

It is easy to note that our RF-RUS configuration, shown in Table 2 achieves lower accuracy measure w.r.t. the best outcome in [19] in Table 3 while AUC (0.717) and Specificity (0.68) values are higher than the best results in ([19]). Furthermore, our aim is to reduce the number of false positive because the misclassification cost are more higher than assigning good loans [26]. On the other hand, Table 4 shows higher specificity values compared to our results while achieving lower sensitivity value than ours.

Furthermore, we investigate the epxlanation of the individual predictions by randomly selecting a group of possible features (25% of the total) that were considered ”untrustworthy”, being unrecognized by users. An oracle has been designed for each combination of the chosen features to label test set by classifying as ”untrustworthy” if the prediction changed when untrustworthy features were removed from the instance (simulating human discounts), and trustworthy otherwise. In conclusion, we evaluate test set prediction through different explanation methods, whose results are compared with the trustworthiness oracles (see Table 5) and performing 10 random sampling from the dataset.

It is worth to note in Table 5 that *LORE* achieves highest outcomes w.r.t. the other ones by combining local predictions and counterfacts explanation for providing user-friendly explanation in understanding which features affect changes in predictions. In turn, *LIME* achieves higher coverage because it describe each prediction as a weighted sum while *SHAP* provides more reliable outcomes through the use of SHAP values, whose expensive computational complexity can be addressed by using several heuristics. In conclusion, *BEEF* and *Anchors* suffer of limited expressive power, being based on rules.

5. Conclusion

Predicting credit risk is a relevant challenge in the finance industry, particularly in Social lending platforms where high dimensionality and imbalanced data present unique challenges. This study proposes a benchmark for evaluating the effectiveness of machine learning techniques for credit risk prediction in real-world social lending platforms, with a focus on managing imbalanced data sets and ensuring explainability.

Future work will focused on considering additional Social Lending platforms, also designing novel techniques such as deep learning and ensemble strategies that may offer improved performance (see [27]) although they are less explainable.

References

[1] V. Murinde, E. Rizopoulos, M. Zachariadis, The impact of the fintech revolution on the future of banking: Opportunities and risks, International Review of Financial Analysis 81 (2022) 102103. doi:https://doi.org/10.1016/j.irfa.2022.102103.

[2] S. Luo, Y. Sun, F. Yang, G. Zhou, Does fintech innovation promote enterprise transformation? evidence from china, Technology in Society 68 (2022) 101821. doi:https://doi.org/10.1016/j.techsoc.2021.101821.

[3] McKinsey, 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-sustaine

[2] https://colab.research.google.com/

[3] https://scikit-learn.org/stable/index.html

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 2
The best classification results we achieved.

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 3
Highest results achieved in ([19])

d-growth-vf.ashx, 2010. Online; accessed 20 March 2020.

[4] L. Cao, Ai in finance: Challenges, techniques, and opportunities, ACM Comput. Surv. 55 (2022). URL: https://doi.org/10.1145/3502289. doi:10.1145/3502289.

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

[6] A. B. Hens, M. K. Tiwari, Computational time reduction for credit scoring: An integrated approach based on support vector machine and stratified sam-

	Method	AUC	TPR	TNR	G-Mean	Accuracy
	[25]	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

Table 4
Result in ([25])

	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

Table 5
Comparison between Anchors, Lime, SHAP, BEEF and LORE in terms of Precision measure.

pling method, Expert Systems with Applications 39 (2012) 6774–6781. doi:https://doi.org/10.1016/j.eswa.2011.12.057.

[7] T. Verbraken, C. Bravo, R. Weber, B. Baesens, Development and application of consumer credit scoring models using profit-based classification measures, European Journal of Operational Research 238 (2014) 505–513. doi:https://doi.org/10.1016/j.ejor.2014.04.001.

[8] TransUnion, 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/, 2017. Online; accessed 20 March 2020.

[9] M. Soui, I. Gasmi, S. Smiti, K. Ghédira, Rule-based credit risk assessment model using multi-objective evolutionary algorithms, Expert Systems with Applications 126 (2019) 144–157. doi:https://doi.org/10.1016/j.eswa.2019.01.078.

[10] F. Sameer, M. A. Bakar, A. Zaidan, B. Zaidan, A new algorithm of modified binary particle swarm optimization based on the gustafson-kessel for credit risk assessment, Neural Computing and Applications 31 (2019) 337–346. doi:https://doi.org/10.1007/s00521-017-3018-4.

[11] Y. Guo, W. Zhou, C. Luo, C. Liu, H. Xiong, Instance-based credit risk assessment for investment decisions in p2p lending, European Journal of Operational Research 249 (2016) 417–426. doi:https://doi.org/10.1016/j.ejor.2015.05.050.

[12] C. Orsenigo, C. Vercellis, Linear versus nonlinear dimensionality reduction for banks' credit rating prediction, Knowledge-Based Systems 47 (2013) 14–22. doi:https://doi.org/10.1016/j.knosys.2013.03.001.

[13] N. Kozodoi, J. Jacob, S. Lessmann, Fairness in credit scoring: Assessment, implementation and profit implications, European Journal of Operational Research 297 (2022) 1083–1094. doi:https://doi.org/10.1016/j.ejor.2021.06.023.

[14] V. Moscato, A. Picariello, G. Sperlí, A benchmark of machine learning approaches for credit score prediction, Expert Systems with Applications 165 (2021) 113986. doi:https://doi.org/10.1016/j.eswa.2020.113986.

[15] M. Malekipirbazari, V. Aksakalli, Risk assessment in social lending via random forests, Expert Systems with Applications 42 (2015) 4621–4631. doi:https://doi.org/10.1016/j.eswa.2015.02.001.

[16] A. Namvar, M. Naderpour, Handling uncertainty in social lending credit risk prediction with a choquet fuzzy integral model, in: 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2018, pp. 1–8. doi:10.1109/FUZZ-IEEE.2018.8491600.

[17] J. Sun, J. Lang, H. Fujita, H. Li, Imbalanced enterprise credit evaluation with dte-sbd: Decision tree ensemble based on smote and bagging with differentiated sampling rates, Information Sciences 425 (2018) 76–91. doi:https://doi.org/10.1016/j.ins.2017.10.017.

[18] A. Marqués, V. García, J. Sánchez, Exploring the behaviour of base classifiers in credit scoring ensembles, Expert Systems with Applications 39 (2012) 10244–10250. doi:https://doi.org/10.1016/j.eswa.2012.02.092.

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

[20] M. T. Ribeiro, S. Singh, C. Guestrin, "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, 2016, pp. 1135–1144.

[21] M. T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[22] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions (2017) 4765–4774.

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

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

[25] Y. Song, Y. Wang, X. Ye, D. Wang, Y. Yin, Y. Wang, Multi-view ensemble learning based on distance-to-model and adaptive clustering for imbalanced credit risk assessment in p2p lending, Information Sciences 525 (2020) 182–204. doi:https://doi.org/10.1016/j.ins.2020.03.027.

[26] V. García, A. I. Marqués, J. S. Sánchez, Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction, Information Fusion 47 (2019) 88–101. doi:https://doi.org/10.1016/j.inffus.2018.07.004.

[27] M. Ala'raj, M. F. Abbod, M. Majdalawieh, L. Jum'a, A deep learning model for behavioural credit scoring in banks, Neural Computing and Applications (2022) 1–28. doi:https://doi.org/10.1007/s00521-021-06695-z.