Credit Score Prediction Relying on Machine Learning

Flora Amato[1], Antonino Ferraro[1], Antonio Galli[1], Francesco Moscato[3], Vincenzo Moscato[1,2] and Giancarlo Sperlí[1,2]

[1]Department of Electrical Engineering and Information Technology (DIETI), Naples, Italy
[2]CINI - ITEM National Lab, Complesso Universitario Monte S.Angelo, Naples, Italy
[3]DIEM, University of Salerno, Fisciano, Italy

Abstract

Financial institutions use a variety of methodologies to define their commercial and strategic policies, and a significant role is played by credit risk assessment. In recent years, different credit risk assessment services arose, providing *Social Lending* platforms to connect lenders and borrowers in a direct way without assisting of financial institutions. Despite the pros of these platforms in supporting fundraising process, there are different stems from multiple factors including lack of experience of lenders, missing or uncertain information about the borrower's credit history. In order to handle these problems, credit risk assessments of financial transactions are usually modeled as a binary problem based on debt repayment, going to apply Machine Learning (ML) techniques. The paper represents an extended abstract of a recent work, where some of the authors performed a benchmarking among the most used credit risk assessment ML models in the field of predicting whether a loan will be repaid in a P2P platform. The experimental analysis is based on a real dataset of Social Lending (Lending Club), going to evaluate several evaluation metrics including AUC, sensitivity, specificity and explainability of the models.

Keywords

Credit Score Prediction, Machine Learning, eXplainable Artificial Intelligence,

1. Introduction

The recent development of digital financial services has led researchers to pay attention to the management of credit risk, proposing useful models to reduce such a risk but also to obtain profits from the investment. Banking risks can arise from different factors including: operational risks, market, credit, and the last one represents 60% of problems for banks [1].

The main cause of credit risk is the spread of Social Lending (SL) platforms, known as Peer-to-Peer (P2P) lending. These platforms allow lenders and borrowers to be interconnected without involving financial institutions; they support borrowers in the fundraising process and allow lending entities to participate. One challenge that needs to be addressed in this context is the credit risk analysis, due to possible non-repayment of loans by borrowers, where risk assessment is calculated through credit scoring.

The credit risk assessment of financial transactions on SL platforms is performed through a binary classification problem, based on debt repayment [2, 3].

SEBD 2022: The 30th Italian Symposium on Advanced Database Systems, June 19-22, 2022, Tirrenia (PI), Italy
flora.amato@unina.it (F. Amato); antonino.ferraro@unina.it (A. Ferraro); antonio.galli@unina.it (A. Galli); fmoscato@unisa.it (F. Moscato); vincenzo.moscato@unina.it (V. Moscato); giancarlo.sperli@unina.it (G. Sperlí)

Additionally, it is important to note that P2P platforms produce large amounts of unlabeled data so additional analysis is required to support real-time decisions [4]. An additional critical issue with these platforms is the risk of default, which is higher than standard methods, this is due to the fact that a lender may not always be able to effectively assess the risk level of borrowers [5], thus the main issue is due to a lack of credit history of borrowers.

Predictive models of credit scoring can be classified into two broad categories [6]: statistical approaches and artificial intelligence methods. Regarding statistical approaches, they have been proposed, but suffer from coverage problems inherent in nonlinear effects among the variables involved. Credit risk assessment is characterized by the following properties: dependence, complexity, and interconnectedness [7], thus credit scoring estimation is very complex as it is dependent on different parameters. Several methodologies have been proposed that rely on rule generation to evaluate credit risks [8, 9], however, these approaches may be limited in the process of generating rules on large amounts of data. Another problem is the lack of lender experience or the uncertainty of borrower history information, these factors greatly increase credit risk. Some platforms incorporate borrower status prediction models, particularly using logistic regression [10] and Random Forest-based classification [11].

However, the development of credit risk prediction models is difficult due to different factors, including high data size and imbalance and high number of missing values. For these reasons, additional approaches have then been proposed, such as Support Vector Machine (SVM) based semi-supervised approach [12], while [13] has introduced an ensemble Decision Tree model for credit risk assessment on 138 Chinese companies with loss-making corporate earnings. Another ensemble method was developed by Feng et al. (2018)[14], in which classifiers are selected based on performance related to credit scoring. While [15] designed a hybrid model that relies on transductive support vector machine (TSVM) and Dempster-Shafer theory to predict social loan defaults. Finally, [16] has described a combination of different classifiers using linear weight ensemble to predict SL default, instead Song et al. (2020)[17] an ensemble of classifiers based on distance-model learning method and adaptive multi-view clustering (DM-ACME).

In this paper, which represents an extended abstract of our previous work [18], we propose a benchmarking for credit risk scoring using the most advanced machine learning (ML) techniques used in the literature , to understand whether a loan will be repaid on a P2P platform. The performance was evaluated using different scoring metrics such as Sensitivity, AUC, Specificity. In addition, eXplainable Artificial Intelligence (XAI) approaches were used to obtain a high degree of explainability of the models. The goal is to evaluate both in terms of accuracy performance of the classifiers but also to provide results understandable by domain experts, ensuring transparency in decisions, this is particularly required for credit risk assessment.

2. Proposed benchmark architecture

The proposed benchmark architecture in Fig.1 stems from the need to be able to offer support to the risk prediction problem, thus an investor can evaluate potential borrowers within social lending platforms. The main challenge to address is that credit risk assessment is a multidimensional and unbalanced issue because it is based on huge amounts of historical data, including, credit history (obtained by filling out a comprehensive application), bank account

status, employment status, etc.

In addition, using all these features increases coverage but decreases accuracy, thus it is essential to apply a feature selection approach. In particular, the proposed architecture that is based on three macro-modules:

1. Ingestion,
2. Classification,
3. Explanation.

The ingestion phase aims at crawling the data from the social lending platforms, cleaning and filtering the obtained data and performing feature selection based on the chosen classifier. In details, the data are cleaned by removing features with many missing or null values and attributes with zero variance from the dataset. After cleaning, several transformations are applied, such as converting categorical features to numeric and changing date attributes to numeric values. The second macro-block performs credit prediction, here we have to deal with a problem of imbalance because usually a user of P2P platforms have a high number of rejected loans compared to those requested. The classifiers chosen in our architecture are: Logistic-regression, Random Forest and Multi-Layer Perceptron , being the most suitable ones for credit prediction [19, 6] and the most used in this context [11, 20, 13]. To handle the unbalance problem, the following techniques are used: random subsampling, random oversampling, and smoothing. Specifically, oversampling merely creates new minority class samples, the Synthetic Minority Oversampling Technique (SMOTE) is based on oversampling using k-nearest neighbors. While subsampling eliminates the majority class samples randomly. Finally, the third macro-block is concerned with explaining the results of each prediction, i.e., the decisions made by the classifiers to obtain information about the financial domain being analyzed. In particular, five XAI tools are used: LIME, Anchors, SHAP, BEEF and LORE. LIME [21] is a Post-Hoc and Agnostic method that provides a local explanation on the prediction, Anchors [22] is also of the same type, a Post- Hoc, Model Agnostic method that provides a local explanation but using rules that sufficiently "anchor" the predictor locally. SHapley Additive exPlanations (SHAP) [23] is a method for explain individual predictions based on the game theoretically optimal Shapley Values in order to analyze how each feature influences the prediction. Balanced English Explanations of Forecasts (BEEF) [24] exploits global information, retrieved by the clustering algorithm on the entire dataset, in order to generate a local explanation. Finally, Local Rule-Based Explanations (LORE), proposed by Guidotti et al. (2018)[25] is first based on learning an interpretable local predictor and then deriving the explanation as a decision rule.

3. Experimental evaluation

The purpose of our experimentation is to compare different classification models, evaluating them according to some metrics (for more details see Section 3.1). The dataset is provided by Lending Club[1], a P2P lending platform, in particular we focused on loans disbursed between 2016 and 2017, it consists of 877,956 samples and 151 features, where the most important ones are loan_amount and term.

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

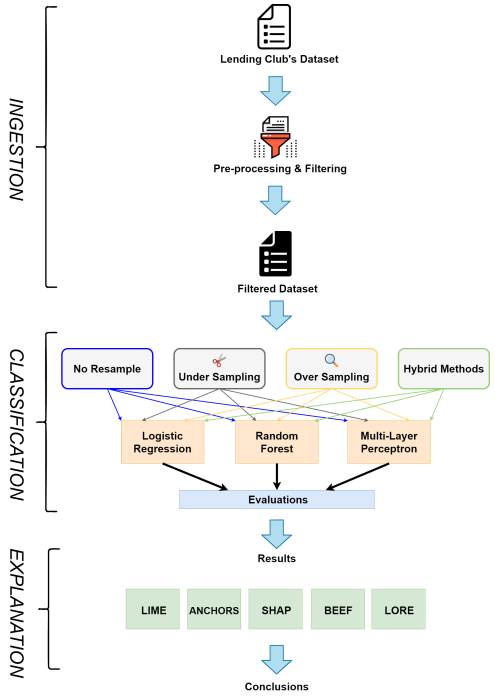

Figure 1: Architecture

According to [11] and [6], we considered $loan\_status$ as the target class for our problem.

Only the labels "FullyPaid" or "Charged off" were considered, since we classified the problem as binary, whether the loan will be repaid or not, this leads to unbalanced data, respectively 0.77% of the samples are fully paid , the remaining 0.23% are unpaid. A 10-cross validation was performed in which the dataset was divided into a training set and a test set with a ratio of 75/25.

Finally, the results obtained were compared against those presented in Namvar et al. (2018)[6] and Song et al. (2020)[17]. The benchmark was run on Google Colab[2], with Xeon single core hyper threaded processor @2.3Ghz, 12 GB RAM, NVIDIA Tesla K80 with 2496 CUDA cores and 12 GB GDDR5 VRAM, using Python 3.6 with scikit-learn 0.23.1.3[3]

3.1. Evaluation metrics

The following metrics were used to evaluate and compare the effectiveness of the considered models: *Sensitivity (TPR), Specificity (TNR), G-mean, Precision, FP-Rate, Area Under Curve (AUC). Accuracy* (ACC) was not used as an evaluation metric because it does not consider that false positives are more important than false negatives, thus it results in an inaccurate evaluation. Instead, TPR and TNR are suitable because they assess the accuracy of positive and negative samples, respectively. While *G-mean* is an appropriate metric for assessing the balance of

[2]https://colab.research.google.com/
[3]https://scikit-learn.org/

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 1
Our best Classification results.

classification performances for both majority and minority classes.

In turn, *Precision* and *FP-Rate* are useful for understanding how well the model predicts positive and negative classes. Finally, AUC determines the area under the ROC curve, thus it is used to assess the trade-off between the rate of true positives and true negatives in the evaluated model.

3.2. Feature Engineering

The aim of this section is to explain the criterion of improving the data through their cleaning and feature selection. Specifically, all features with missing values greater than 55%, and also those with high standard deviation were removed. Finally, the missing values were replaced with the median of the features, furthermore the nominal features were converted to binary data (more details are reported in [18]).

3.3. Experimental results

The classifiers used are Random Forest (RF), Logistic Regresion (LR) and Multi Layer Perceptron and have been evaluated according to different sampling strategies: Under-sampling (RUS, IHT), Over-sampling (ROS, SMOTE, ADASYN), Hybrid-Method (SMOTE-TOKEN, SMOTE-EN). In Table 1 we report the best combination between the classifiers and the sampling strategies, comparing them also against the performance obtained by the classifiers without any strategy, this last comparison highlights the effectiveness of the latter techniques on the prediction performance. The experiment decrees that RF-RUS turns out to be the best method for predicting a borrower's status in a social lending market.

3.3.1. Comparison with state-of-art results

We compared our results against the best results of Namvar et al. (2018)[6] and Song et al. (2020)[17], it can be seen that our best combination (RF-RUS) (see in Table 1) has the lowest accuracy while our AUC value and Specificity are higher than the best of [6]. This is important in our context because reducing false positives avoids the serious economic damage of misclassification, i.e., the loss of a user's loan. In addition, Table 2 shows higher values of Specificity than our results even though its sensitivity value is much lower than our model.

	Method	AUC	TPR	TNR	G-Mean	Accuracy
	RF -RUS	0.717	0.630	0.680	0.6560	0.640
	Song et al.[17]	0.6697	0.4607	0.7678	0.6009	0.7231
Namvar et al.[6]	Linear discrimination analysis - SMOTE	0.7000	0.630	0.650	0.643	0.6400
	LR - SmoteToken	0.7000	0.638	0.648	0.643	0.6400
	Logistic regression	0.7030	0.988	0.048	0.218	0.8173
	Random forest	0.6960	0.996	0.015	0.12	0.8176
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 2
Results.

3.3.2. Explanation results

In this last part of the evaluation, we compare the performance of several XAI tools: LIME, Anchors, SHAP, BEEF, and LORE. In particular, the metrics are based on the Accuracy measure, according to the protocol described in Ribeiro et al. (2016)[21], evaluated on the three best classifiers: Random Forest & Random Subsampling, Logistic Regression & Random Oversampling, and Logistic Regression & Smote-Token. Several explanations were generated, using different sets of instances computed with different random sampling (10 runs) from the dataset. Analyzing the results in Table 3, LORE is the best because it combines local predictions with the use of counterfactuals for explanation generation, while LIME achieves good results for all three classifiers, this is because the prediction is modeled as a weighted sum and this makes it easy to interpret the prediction generation. SHAP, on the other hand, based on the importance of features, offers statistically more significant results than LIME, this is given by the use of shap values, whose computational complexity, even if dampened by different heuristics, can affect the efficiency of the explanation. Finally, regarding BEEF and Anchors, they can be limited in the expressiveness of the explanation, as noted for Logistic Regression, since they are based on axis-aligned hyper-rectangle and specific rules (called anchors).

4. Conclusion

Determining the risk prediction score is one of the biggest challenges in finance. The aim of the proposed approach is to support people in their investments, proposing a reference model based

	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 3
Comparison between Anchors, Lime, SHAP, BEEF and LORE in terms of Precision measure.

on Machine Learning approaches for the prediction of credit risk in social lending platforms, going to manage what are the major criticalities in P2P platforms: the high dimension of data to be analyzed and unbalanced data. The evaluation done on a real dataset demonstrates the goodness of the proposed approach, as well as the fact of being able to provide an explanation for the prediction obtained, which is very significant in the financial field to be able to motivate a positive or negative judgment to provide a loan. Developments of future work may be to consider different P2P lending platforms and use additional classification approaches such as Deep Learning or ensemble learning techniques in order to achieve better performance.

References

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

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

[3] 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.

[4] A. Kim, S.-B. Cho, Dempster-shafer fusion of semi-supervised learning methods for predicting defaults in social lending, in: International Conference on Neural Information Processing, Springer, 2017, pp. 854–862.

[5] 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.

[6] A. Namvar, M. Siami, F. Rabhi, M. Naderpour, Credit risk prediction in an imbalanced social lending environment, arXiv preprint arXiv:1805.00801 (2018).

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

[8] Y. Hayashi, 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 (2016) 32–42.

[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.
[10] R. Emekter, Y. Tu, B. Jirasakuldech, M. Lu, Evaluating credit risk and loan performance in online peer-to-peer (p2p) lending, Applied Economics 47 (2015) 54–70.
[11] M. Malekipirbazari, V. Aksakalli, Risk assessment in social lending via random forests, Expert Systems with Applications 42 (2015) 4621–4631.
[12] Z. Li, Y. Tian, K. Li, F. Zhou, W. Yang, Reject inference in credit scoring using semi-supervised support vector machines, Expert Systems with Applications 74 (2017) 105–114.
[13] 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.
[14] X. Feng, Z. Xiao, B. Zhong, J. Qiu, Y. Dong, Dynamic ensemble classification for credit scoring using soft probability, Applied Soft Computing 65 (2018) 139–151.
[15] A. Kim, S.-B. Cho, An ensemble semi-supervised learning method for predicting defaults in social lending, Engineering Applications of Artificial Intelligence 81 (2019) 193–199.
[16] W. Li, S. Ding, H. Wang, Y. Chen, S. Yang, Heterogeneous ensemble learning with feature engineering for default prediction in peer-to-peer lending in china, World Wide Web 23 (2020) 23–45.
[17] 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.
[18] V. Moscato, A. Picariello, G. Sperlí, A benchmark of machine learning approaches for credit score prediction, Expert Systems with Applications 165 (2021) 113986.
[19] V. García, A. Marqués, J. S. Sánchez, On the use of data filtering techniques for credit risk prediction with instance-based models, Expert Systems with Applications 39 (2012) 13267–13276.
[20] 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), IEEE, 2018, pp. 1–8.
[21] 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.
[22] M. T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations, in: Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018.
[23] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).
[24] S. Grover, C. Pulice, G. I. Simari, V. Subrahmanian, Beef: Balanced english explanations of forecasts, IEEE Transactions on Computational Social Systems 6 (2019) 350–364.
[25] 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).