Class Imbalance Problem in Credit Risk Prediction

Subjects: Computer Science, Artificial Intelligence

Contributor: Zixue Zhao, Tianxiang Cui, Shusheng Ding, Jiawei Li, Anthony Graham Bellotti

Credit, as defined by financial institutions such as banks and lending companies, represents a vital loan certificate issued to individuals or businesses. This certification mechanism plays a pivotal role in ensuring the smooth functioning of the financial sector, contingent upon comprehensive evaluations of creditworthiness.

Keywords: credit risk prediction ; resampling ; class imbalance

1. Introduction

Credit, as defined by financial institutions such as banks and lending companies [1], represents a vital loan certificate issued to individuals or businesses. This certification mechanism plays a pivotal role in ensuring the smooth functioning of the financial sector, contingent upon comprehensive evaluations of creditworthiness. The evaluation process inherently gives rise to concerns regarding credit risk, encompassing the potential default risk associated with borrowers. Assessing credit risk entails the utilization of credit scoring, a method aimed at distinguishing between “good” and “bad” customers [2]. This process is often referred to as credit risk prediction in numerous studies [3][4][5][6][7]. Presently, the predominant approaches to classifying credit risk involve traditional statistical models and machine learning models, typically addressing binary or multiple classification problems.

Credit data often exhibit a high number of negative samples and a scarcity of positive samples (default samples), a phenomenon known as the class imbalance (CI) problem [8]. Failure to address this issue may result in significant classifier bias [9], diminished accuracy and recall [10], and weak predictive capabilities, ultimately leading to financial institutions experiencing losses due to customer defaults [11]. For instance, in a dataset comprising 1000 observations labeled as normal customers and only 10 labeled as default customers, a classifier could achieve 99% accuracy without correctly identifying any defaults. Clearly, such a classifier lacks the robustness required. To mitigate the CI problem, various balancing techniques are employed, either at the dataset level or algorithmically. Dataset-level approaches include random oversampling (ROS), random undersampling (RUS), and the synthetic minority oversampling technique (SMOTE) [12], while algorithmic methods mainly involve cost-sensitive algorithms. Additionally, ensemble algorithms [13] and deep learning techniques, such as generative adversarial networks (GANs) [14], are gradually gaining traction for addressing CI issues.

Indeed, there is no one-size-fits-all solution to the CI problem that universally applies to all credit risk prediction models [15][16][17]. On the one hand, the efficacy of approaches is constrained by various dataset characteristics such as size, feature dimensions, user profiles, and imbalance ratio (IR). Notably, higher IR and feature dimensions often correlate with poorer classification performance [18]. On the other hand, existing balancing techniques exhibit their own limitations. For instance, the widely used oversampling technique, SMOTE, has faced criticism for its failure to consider data distribution comprehensively. It solely generates new minority samples along the path from the nearest minority class to the boundary. Conversely, some undersampling methods are deemed outdated as they discard a substantial number of majority class samples, potentially leading to inadequately trained models due to small datasets. Additionally, cost-sensitive learning hinges on class weight adjustment, which lacks interpretability and scalability [11].

2. Class Imbalance Problem in Credit Risk Prediction

In the domain of credit risk prediction, accurately identifying potential defaulting users holds paramount importance [19]. Banks meticulously gather user characteristics and devise scoring systems to scrutinize customers and allocate loan amounts judiciously. Upon identifying potential risks, they may either reduce the loan quota or decline lending altogether. This dynamic is evident in the data, where positive samples (minority class) bear greater significance than negative samples (majority class). This poses a dilemma, as the classifier requires substantial information about the minority class to effectively identify positive samples, yet it inevitably tends to be more influenced by the majority class [20]. Consequently, oversampling and cost-sensitive algorithms have been favored in addressing credit risk prediction

scenarios. The former directly enhances the proportion of samples in the minority class, while the latter factors in that misclassifying negatives is less detrimental than misclassifying positives [21].

To mitigate the risk of underfitting arising from the potential omission of vital information by undersampling techniques, algorithms such as EasyEnsemble and BalanceCascade [22] have been developed. These algorithms aim to minimize the probability of discarding crucial information during the undersampling process. EasyEnsemble combines the anti-underfitting capacity of boosting with the anti-overfitting capability of bagging. Conversely, to alleviate the risk of overfitting associated with oversampling, distance-based k-neighborhood methods for resampling are considered more effective. Notably, the synthetic minority oversampling technique (SMOTE) has garnered attention in recent years, particularly in credit scenarios characterized by an imbalance between “good and bad customers”.

In the realm of loan default prediction, researchers have utilized the SMOTE algorithm in various ways, emphasizing the criticality of information within the minority class. Studies suggest that SMOTE, or more boundary-point-oriented adaptive oversampling techniques like adaptive integrated oversampling, can yield superior results when modeling with such data [23]. Moreover, combining oversampling techniques with integrated learning has been proposed to mitigate overfitting risks. For instance, sampling combined with boosting methods and support vector machines, as well as a combination of adaptive integrated oversampling with support vector machines and boosting, have demonstrated promising results in empirical analyses [24].

Nonetheless, subsequent studies caution against excessively tightening criteria due to potential default risks, as rejecting numerous creditworthy users can significantly diminish bank earnings, sometimes surpassing losses incurred from a single defaulting user [11]. Over-reliance on oversampling techniques could exacerbate this inverse risk. However, this does not imply superiority of undersampling techniques, which exhibit distinct drawbacks, notably information loss from the majority class, particularly with clustering-based undersampling methods [25][26]. To harness the full potential of minority class samples while retaining information from majority class samples, comprehensive techniques combining oversampling and undersampling have emerged. Examples include SMOTE with Tomek links and SMOTE with edited nearest neighbors (ENNs), both of which have demonstrated enhancements in dataset quality and classifier performance [15]. In a comprehensive study conducted as early as 2012, ref. [27] designed a detailed examination of RTs. The study evaluated four undersampling, three oversampling, and one composite resampling technique across five datasets to ascertain the potential benefits for intelligent classifiers such as the multilayer perceptron (MLP) when using these techniques. The comparative analysis revealed that there is no one-size-fits-all solution with respect to the effectiveness of sampling techniques across all classifiers. However, it was observed that undersampling methods like neighborhood clean rule (NCL) and oversampling techniques like SMOTE and SMOTE + ENN consistently demonstrated stable performance. Notably, oversampling imparted a significant performance enhancement, particularly benefiting higher-performing intelligent classifiers.

On the other hand, prevailing class balancing experiments often strive to equalize the proportions of majority and minority classes, yet few studies have delved into addressing datasets exhibiting extreme imbalances. The IR, denoting the ratio of majority to minority samples, serves as a gauge for assessing the extent of class imbalance. Commonly used benchmark credit datasets typically exhibit IRs ranging from 2 to 10, such as the German credit dataset (IR: 2.33) and the Australia credit dataset (IR: 1.24), while certain private datasets may escalate to IRs of 10 to 30 [28]. Typically, larger sample sizes correlate with higher IRs. However, there exists no standardized criterion for defining extreme imbalance. An IR above 5 implies that merely 16.6% of positive samples are available, posing a considerable challenge for classifiers. Ref. [29] advocated for the use of gradient boosting and random forest algorithms to effectively handle datasets with extreme imbalance. Through experimentation with oversampling techniques, it was observed that an optimal class distribution should encompass 50% to 90% of the minority classes. In other words, it suffices to moderate the extreme imbalances to achieve a mild imbalance without necessitating an IR of 1. Conversely, ref. [18] employed simulation datasets to simulate varying IRs and found that higher IRs do not consistently lead to poorer classifier performance; rather, performance is significantly influenced by the feature dimensions of the dataset. Indeed, IR serves as one of the statistical features of the dataset, alongside feature dimension, dataset size, feature type, and resampling method, collectively impacting the final prediction outcome [28]. However, high IR alone does not inherently account for prediction difficulty; rather, it is the indistinct decision boundary stemming from too few minority class samples, overlapping due to resampling, and excessive noise that pose the primary challenges [15]. Thus, the primary objective of balancing techniques should focus on clarifying classification boundaries rather than merely striving for dataset balance. Ref. [15] echoes the sentiments of the aforementioned study, emphasizing the collective influence of IR on the efficacy of various RTs. Following a comparative analysis involving methods such as Tomek-link removal (Tomek), ENN, BorderlineSMOTE, adaptive integrated oversampling (ADASYN), and SMOTE + ENN, it was concluded that the complexity of RTs does not necessarily correlate

with their ability to address datasets with higher IR. Importantly, it was observed that no single RT emerged as universally effective across all classification and CI problems.

RTs proactively address the CI problem during the data preprocessing stage. While numerous studies propose resolving the CI problem through adjustments within machine learning classifiers or by integrating balancing strategies directly into ensemble models, recent advancements in algorithms such as eXtreme Gradient Boosting (XGBoost) [30], LightGBM, and CatBoost offer hyperparameters capable of fine-tuning the weights of positive samples. Even amidst imbalanced datasets, these algorithms enable the objective function to prioritize information gleaned from minority class samples. Furthermore, incorporating resampling techniques within ensemble learning to balance each training subset yields models with heightened robustness compared with classifiers solely adjusting sample weights. For instance, bagging classifiers and random forests can be augmented with balancing techniques to ensure a portion of minority class samples in each training subset [31]. To compare the effectivenesses of various classifiers, ref. [29] conducted experiments across five datasets, incorporating various IRs. The study evaluated the performances of classifiers such as logistic regression, decision tree (C4.5), neural network, gradient boosting, k-nearest neighbors, support vector machines, and random forest, considering positive sample proportions ranging from 1% to 30%. Results from the experiments revealed that gradient boosting and random forest exhibited exceptional performance, particularly when handling datasets with extreme IR. Conversely, support vector machines, k-nearest neighbors, and decision tree (C4.5) struggled to effectively manage the CI problem. In conclusion, the study suggests that ensemble learning methods, specifically boosting and bagging, outperform individual classifiers when addressing imbalanced credit datasets, highlighting their efficacy in handling CI challenges.

However, the effectiveness of solely relying on model weights to address the CI problem diminishes if RTs are not applied to the dataset beforehand [32]. Moreover, the embedding of resampling techniques within ensemble models significantly escalates computational costs, rendering it less efficient and more constrained when handling large datasets [4]. To address this issue, ref. [28] conducted a comprehensive comparison between various pairs of classifiers and RTs. Their objective was to identify dependable combinations of advanced RTs and classifiers capable of handling datasets with differing IR levels effectively. By conducting paired experiments involving nine RTs and nine classifiers, their findings revealed that the combination of RUS and random subspace consistently achieved satisfactory performance across most cases. Following closely behind was the combination of SMOTE + ENN and logistic regression. Interestingly, these results deviate from previous studies that tended to favor ensemble classifiers. Ref. [28] argue that even simple classifiers can achieve commendable performance, provided that suitable RTs are employed.

References

1. Henley, W.; Hand, D.J. A k-nearest-neighbour classifier for assessing consumer credit risk. J. R. Stat. Soc. 1996, 45, 77–95.
2. Abellán, J.; Castellano, J.G. A comparative study on base classifiers in ensemble methods for credit scoring. Expert Syst. Appl. 2017, 73, 1–10.
3. Tsai, C.F.; Wu, J.W. Using neural network ensembles for bankruptcy prediction and credit scoring. Expert Syst. Appl. 2008, 34, 2639–2649.
4. Andrés Alonso, J.M.C. Machine Learning in Credit Risk: Measuring the Dilemma between Prediction and Supervisory Cost; Banco de España: Madrid, Spain, 2020.
5. Ding, S.; Cui, T.; Bellotti, A.; Abedin, M.; Lucey, B. The role of feature importance in predicting corporate financial distress in pre and post COVID periods: Evidence from China. Int. Rev. Financ. Anal. 2023, 90, 102851.
6. Wang, L. Imbalanced credit risk prediction based on SMOTE and multi-kernel FCM improved by particle swarm optimization. Appl. Soft Comput. 2022, 114, 108153.
7. Moscato, V.; Picariello, A.; Sperlí, G. A benchmark of machine learning approaches for credit score prediction. Expert Syst. Appl. 2021, 165, 113986.
8. García, V.; Marqués, A.I.; Sánchez, J.S. Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction. Inf. Fusion 2019, 47, 88–101.
9. Haixiang, G.; Li, Y.; Shang, J.; Mingyun, G.; Yuanyue, H.; Gong, B. Learning from class-imbalanced data: Review of methods and applications. Expert Syst. Appl. 2016, 73, 220–239.
10. García, V.; Marqués, A.I.; Sánchez, J.S. An insight into the experimental design for credit risk and corporate bankruptcy prediction systems. J. Intell. Inf. Syst. 2015, 44, 159–189.

11. Niu, K.; Zhang, Z.; Liu, Y.; Li, R. Resampling ensemble model based on data distribution for imbalanced credit risk evaluation in P2P lending. Inf. Sci. 2020, 536, 120–134.
12. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. Artif. Intell. 2002, 16, 321–357.
13. Cui, T.; Li, J.; John, W.; Andrew, P. An ensemble based Genetic Programming system to predict English football premier league games. In Proceedings of the 2013 IEEE Symposium Series on Computational Intelligence (SSCI2013), Singapore, 16–19 April 2013; pp. 138–143.
14. Fiore, U.; De Santis, A.; Perla, F.; Zanetti, P.; Palmieri, F. Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. Inf. Sci. 2019, 479, 448–455.
15. Jiang, C.; Lu, W.; Wang, Z.; Ding, Y. Benchmarking state-of-the-art imbalanced data learning approaches for credit scoring. Expert Syst. Appl. 2023, 213, 118878.
16. Ding, S.; Cui, T.; Zhang, Y. Incorporating the RMB internationalization effect into its exchange rate volatility forecasting. N. Am. J. Econ. Financ. 2020, 54, 101103.
17. Ding, S.; Cui, T.; Zheng, D.; Du, M. The effects of commodity financialization on commodity market volatility. Resour. Policy. 2021, 73, 102220.
18. Zhu, R.; Guo, Y.; Xue, J.H. Adjusting the imbalance ratio by the dimensionality of imbalanced data. Pattern Recognit. Lett. 2020, 133, 217–223.
19. Caouette, J.; Altman, E.; Narayanan, P.; Nimmo, R. Managing Credit Risk: The Great Challenge for the Global Financial Markets, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2011; pp. 349–365.
20. Khan, A.A.; Chaudhari, O.; Chandra, R. A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. Expert Syst. Appl. 2024, 244, 122778.
21. Xia, Y.; Liu, C.; Liu, N. Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending, Electron. Commer. Res. Appl 2017, 24, 30–49.
22. Liu, X.Y.; Wu, J.; Zhou, Z.H. Exploratory undersampling for class-Imbalance learning. IEEE Trans. Syst. Man Cybern. Part B 2009, 39, 539–550.
23. Liu, B.; Chen, K. Loan risk prediction method based on SMOTE and XGBoost. Comput. Mod. 2020, 2, 26–30.
24. Zięba, M.; Tomczak, J.M. Boosted SVM with active learning strategy for imbalanced data. Soft Comput. 2015, 19, 3357–3368.
25. Ding, S.; Cui, T.; Wu, X.; Du, M. Supply chain management based on volatility clustering: The effect of CBDC volatility. Res. Int. Bus. Financ. 2022, 62, 101690.
26. Yen, S.J.; Lee, Y.S. Cluster-based under-sampling approaches for imbalanced data distributions. Expert Syst. Appl. 2009, 36, 5718–5727.
27. García, V.; Marqués, A.I.; Sánchez, J.S. Improving Risk Predictions by Preprocessing Imbalanced Credit Data. In Neural Information Processing; Huang, T., Zeng, Z., Li, C., Leung, C.S., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 68–75.
28. Xiao, J.; Wang, Y.; Chen, J.; Xie, L.; Huang, J. Impact of resampling methods and classification models on the imbalanced credit scoring problems. Inf. Sci. 2021, 569, 508–526.
29. Brown, I.; Mues, C. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Syst. Appl. 2012, 39, 3446–3453.
30. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016.
31. Ma, X.; Sha, J.; Wang, D.; Yu, Y.; Yang, Q.; Niu, X. Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. Electron. Commer. Res. Appl. 2018, 31, 24–39.
32. Kou, G.; Chen, H.; Hefni, M.A. Improved hybrid resampling and ensemble model for imbalance learning and credit evaluation. J. Manag. Sci. Eng. 2022, 7, 511–529.