Machine Learning Analysis of Borrower Risk Progression in Financial Lending Systems

Gayatri M. Chutake[1], Aryan S. Bute[2], Sanika S. Chaudhary[3], Adit D. Khair[4], Prof. Neha A. Kandalkar[5]
[1,2,3,4,5] *Department of Artificial Intelligence & Data Science PRIMIT&R, Amravati, Maharashtra, India*

***Abstract*—Credit risk assessment is one of the major steps in the loan approval process of banks and financial institutions, as incorrect credit approval decisions would result in loss in financial transactions and increase in non-performing assets. Classic methods for credit scoring is based on predefined rules or linear statistical methods, which limit their ability in modeling complex nonlinear applicant behavior found in real-world financial data. This work proposes a machine learning-based credit risk prediction framework that improves detection of potential loan defaulters.**

Credit risk evaluation has been one of the most crucial steps within the lending process undertaken by banks and other financial institutions since inappropriate credit risk evaluation could lead to loss of finances as well as increased non-performing assets within the financial agencies. The conventional credit scoring models have been known to apply the use of either rule-based techniques or linear models. The constraint with the former shall be viewed from the standpoint that the ability of the conventional credit scoring models would be impaired by the non-linear behaviors depicted by the credit borrowers.

The proposed framework embeds a wide range of data pre-processing tasks, such as handling missing values, processing categorical variables, and addressing the problem of class imbalance using the Synthetic Minority Over-sampling Technique. Several classification algorithms-for example, Logistic Regression, Random Forest, and XGBoost-are tuned and tested for performance by using appropriate measures to handle imbalanced data. Optuna is also used for hyperparameter tuning.

The result of the experiment with a publicly available credit dataset shows that the tuned XGBoost model performs better compared to other models for correct classification, which is able to achieve 97.4% accuracy in terms of the F1-score. The above mention outcome shows that the developed approach does have the efficacy to strike a balance between the default detection and the incorrect loans being approved. The present study underlines the feasibility of using machine learning approaches in efficiently performing credit risk analysis.

***Index Terms*—Credit Risk Prediction, Loan Default Detection, XGBoost, Optuna Optimization, SMOTE, Machine Learning**

I. INTRODUCTION

Credit risk can be described as the risk that a borrower will not be able to repay a loan as agreed, thus resulting in financial losses for lending organizations. Thus, an effective evaluation of credit risk has become a basic requirement for banks and financial institutions. The recent increased use of online lending platforms has further increased the volume of loan requests that need to be processed within a short time. Thus, there has been a high demand for an automated approach for the evaluation of credit risk to facilitate effective loan assessments.

The traditional methods of credit score calculations involve the use of rules or linear models. These methods may be easier to understand and implement, but they aren't very effective at identifying the relationships that may not be linear between variables of borrowers' characteristics and the probability of default. The real-world borrowers' dataset may include characteristics like non-linear patterns, interaction, and noise that aren't effectively handled by traditional models. [5]

Machine learning algorithms have recently appeared as an interesting approach to credit risk assessment because of their capacity to automatically detect patterns from previously seen data. Machine learning stuff can handle tough jobs, like figuring out how applicant details connect to whether they pay back loans or not. But credit data in real life has all these problems. Missing bits of info pop up a lot. Categorical variables make things tricky too. And class imbalance is a big one, where defaulters are way fewer than people who actually repay. [1], [2], [9]

This paper tries to fix those issues with a system based on machine learning for predicting credit risk. It starts with solid preprocessing to clean the data reliably. Then handles the imbalanced part in a way that makes sense. Evaluation of models is key, done robustly. The system looks at different classification models. It uses automated optimization for hyperparameters. The goal is better predictions overall. That way, high-risk borrowers get spotted more accurately. And fewer good applicants get wrongly turned down. I think that part about reducing rejections stands out, since it affects real people. It seems like without this, the models might just miss the mark on reliability.

II. METHODS AND MATERIALS

A. Dataset Overview

The sample used in this project is a publicly available set of previous loan applications that are usually utilized in credit risk-related research. It has documents of approximately 50,000 applicants who took loans. The records consist of a mixture of numerical and categorical data that is used to define the background, financial status and past behavior of the applicant to borrow funds.

The data includes significant information about the age, annual earnings, employment, the amount of the loan, and the purpose of it. It also contains the details concerning credit history and the way the applicant managed to make the repayments previously. The last outcome variable demonstrates the repayment or non-repayment of the loan. A 1 implies that the applicant was unable to repay the loan whereas 0 implies that the loan was repaid successfully.

Similar to the reality banking system, there are very few cases of loan default compared to the number of cases of loan repayment. Due to this, the structure of the dataset is imbalanced as there are fewer records of default as compared to the non-default records. This is a more realistic feature of the data that is fit to test credit risk prediction models.

Overall, this dataset reflects realistic lending scenarios and captures typical challenges in real-world credit evaluation tasks. In fact, it is suitable for testing and analyzing machine learning models representing credit risk prediction.

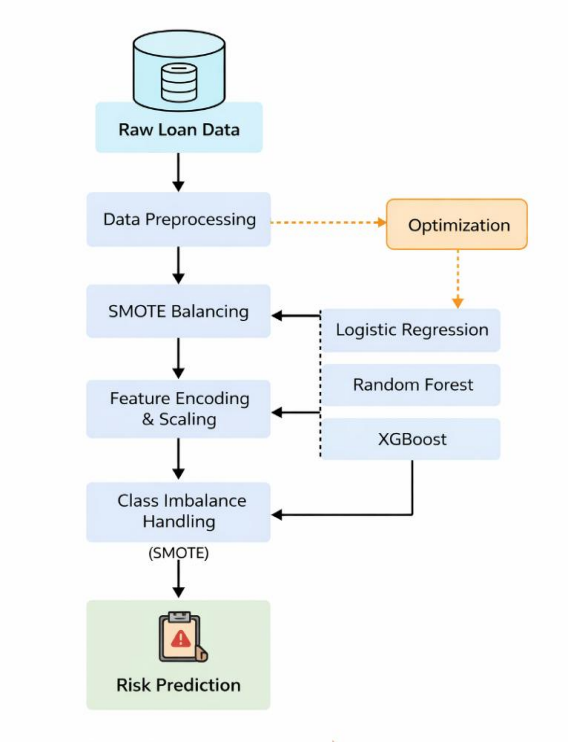

Fig. 1. Architecture of the proposed machine learning–based credit risk prediction framework

B. **Data Preprocessing**

The dataset was closely analyzed and preprocessed before any training in order to ensure reliable learning and consistent results. Financial datasets are often characterized by missing or inconsistent entries; such data, if not treated properly, may negatively affect model performance. In this work, median imputation was applied for missing values in numerical features; this method provides stable estimates and is resistant to outliers. For numerical representation of categorical variables, label encoding and one-hot encoding were used, chosen based on the characteristics and cardinality of each attribute [14], [15].

Standard scaling was performed for attributes like annual income and loan amount to reduce the effect of scale differences among numerical features. It also prevented a single feature from dominating the learning process because of its range. Since the classes for default and non-default are imbalanced, SMOTE [2], [9] was performed on the training data. The synthetic samples created for the minority class by SMOTE increased the sensitivity for defaults in the model and reduced the bias in the prediction.

Table 1. Data Preprocessing Components and Functions

Component	Function
Handling Missing Values	Handles null values using statistical imputation techniques
Categorical Encoding	Converts categorical features into numerical representations
Feature Scaling	Standardizes numeric attributes on a common scale
SMOTE	Balances the distribution between default and non-default classes
Train–Test Split	Splits a dataset into training and testing data.

C. Operational Flow

The overall workflow of the credit risk prediction system is as shown in Figure 2. The approach begins by loading historical data on loan applications, and checking for missing or inconsistent records. After data cleaning, preprocessing steps-encoding and scaling features—are performed to set the dataset ready for model training. After those steps, SMOTE is applied to mitigate the imbalance between defaults and non-defaults.

The preprocessed data is used in the training of different machine learning classifiers. Each model learns from the patterns in applicant-related attributes and makes corresponding predictions. Model performance is evaluated using metrics appropriate for imbalanced classification problems, rather than using accuracy alone. Based on the evaluation, the most consistent performing model is chosen.

Step-by-step protocol of the procedure will make the experiments constant and more accessible to replicate. It also ensures that the credit risk forecasts remain reliable in situations where the model is applied in actual loan evaluation scenarios.

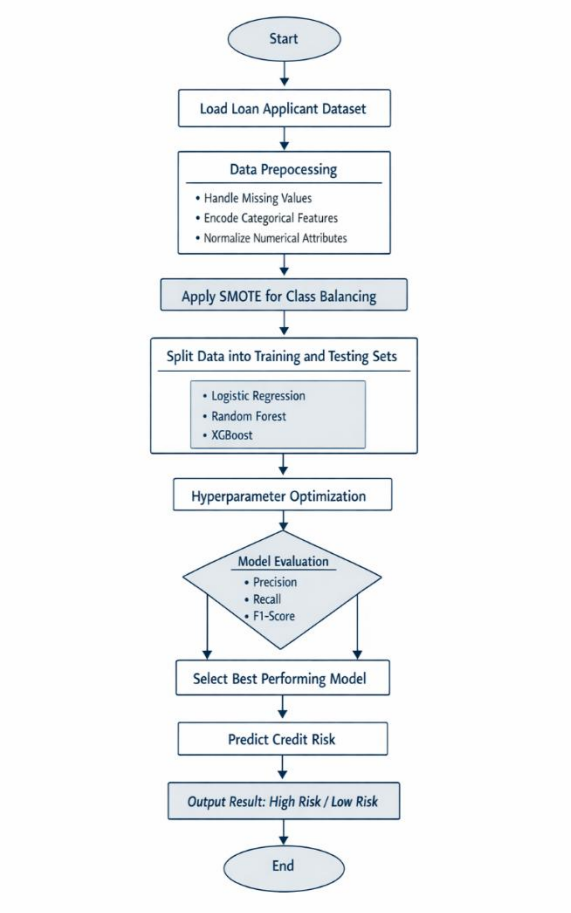

Fig. 2. Workflow of the proposed machine learning–based credit risk prediction system

D. Algorithmic Steps

The operation of the proposed system is guided by the following sequence of steps.

1. Load the dataset of loan applicants.
2. Address missing and inconsistent values.
3. Encode categorical attributes.
4. Normalize numerical features.
5. Balance the dataset by applying SMOTE.
6. Train Logistic Regression, Random Forest and XGBoost models.
7. Optimize model hyperparameters using Optuna.
8. Evaluate the models on precision, recall, and F1-score.
9. Generate the end predictions of credit risk.

E. Advantages of the Proposed Methodology

The various benefits that can be inferred out of utilizing a machine learning-based technique, that is, an approach that was adopted in its application in the present study, as compared to traditional credit appraisal means is the fact that, as a technique that

learns as it goes, it can easily recognize risky borrowers.

Ensemble learning methods lead to more credible prediction and this is especially the case when dealing with big financial data which are heterogeneous in nature. In addition, the improved performance of the model via automatic hyperparameter optimization shall improve the model and the model builders will not be burdened by the hyperparameter tuning. Discussing in particular the implementation, the model scalability is a good factor, as it will enable easier integration with the current banking systems to conduct real-time credit appraisal, thereby minimizing the biased decision-making process when approving and rejecting the applications.

F. Reference to Standards

The trend in which the proposed system is drawn conforms to good practice in financial risk analysis, in combination with the implementation of advice concerning machine learning in credit score systems. The implication of assessment measures taking into account the imbalance in the classes accompanied by a corresponding use of resampling procedures has been mentioned multiple times in the literature on financial analytics.

Through this knowledge about decision anchoring based on processes that are data-driven to provide consistency in various facets of the outcomes assessment makes the system applicable in making sure the machine learning is applied in the right way in lending scenarios. Meanwhile, it enables one to encourage elements of fairness to make the credit risk models work correctly.

III. RESULTS AND DISCUSSION

A. Model Implementation and Evaluation

The credit risk prediction system is developed on Python, as well as the standard machine learning libraries implementations. Following a preprocessing task on the data, three classification models are trained on the data set, including Logistic Regression, Random Forest, and XGBoost. The performance was measured by precision, recall, and F1-score since these metrics provide more reliable evaluation than accuracy in the case of imbalanced data. It proved to be a consistently better-performing model during experimentation with respect to the other classifiers. This superiority lies in its potential to capture nonlinear relationships and interactions among features. For enhancing this performance, hyperparameter optimization with Optuna was performed for better refinement of model parameters toward improved overall stability.

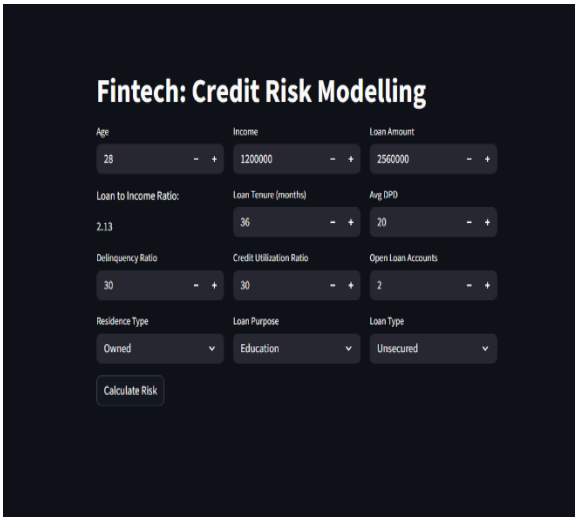

Fig. 3 User Interface for Credit Risk Prediction System

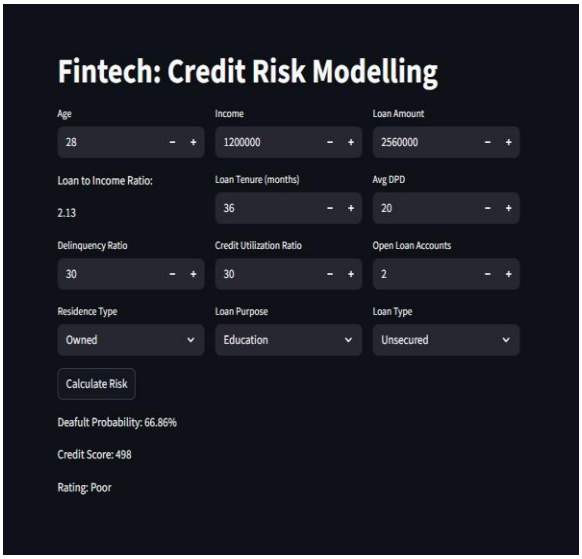

Fig. 4. Risk Assessment Output Showing Default Probability and Credit Rating

System Interface Description

The developed credit risk prediction system comes up with an interactive and user-friendly interface that is developed with the intention of imitating actual scenarios in fintech loan analyses. It is observed that

from Fig. 3, this interface has space for the entry of crucial borrower and loan parameters, which are age, annually earned income, loan amount, loan period, loan-to-income ratio, loan delinquency ratio, credit utilization ratio, average DPDS, number of existing loan accounts, type of residence, type of loan, and loan purpose.

These variables were chosen based on their relevance in credit risk assessment, as highlighted by literature in finance. The user interface of the model ensures dynamic handling of inputs, making it easier for the user to change inputs and reflect on the results.

Risk Prediction Output Explanation

When the input data is submitted, the system is capable of processing the data through the trained machine learning algorithm in order to predict the credit risk of the borrower. Referring to Figure 4, the credit risk analysis system is capable of providing the borrower with results in the form of the default probability, credit score, and credit rating category.

The default probability is expressed as a percentage value, which is the probability of default. The credit score is obtained by normalizing the output of the model to generate a standardized scoring system. The credit rating, which may be Poor, Fair, Good, or Excellent, gives a more understandable measure of creditworthiness.

B. Comparative Study

The developed system was benchmarked against the performance of conventional credit-scoring approaches documented in prior research to evaluate its efficacy. Conventional approaches conventionally rely on rule-based frameworks or linear statistical models, which offer limited adaptability when dealing with complex financial data.

Table 2 summarizes this comparison.

Feature	Proposed System	Traditional Methods
Model Type	Ensemble-based ML	Rule-based / Linear
Default Detection	High	Moderate
Handling Class Imbalance	SMOTE	Limited
Evaluation Metrics	Precision, Recall, F1-score	Accuracy
Scalability	High	Limited

This comparison illustrates that machine learning–based approaches handle complex financial data more effectively than traditional methods. Specifically, they show better results in identifying high-risk borrowers and handling class imbalance-issues of first-order importance in practical credit risk assessment [5], [7], [8].

C. Performance Analysis

The evaluation of the machine learning models used precision, recall, and F1-score, since these measures provide a better result for datasets that were imbalanced. The quantitative comparisons among the three models are presented in Table 3.

Table 3 presents a quantitative comparison.

Model	Precision	Recall	F1-score
Logistic Regression	0.91	0.89	0.90
Random Forest	0.95	0.94	0.94
XGBoost (Optimized)	0.98	0.97	0.974

XGBoost had the best optimized model performance, reaching an F1-score of 97.4%. That means high balance between correct identifications of loan defaulters and minimizing fake approvals. Random Forest also showed competitive performance but with relatively lower recall for the cases of default. Logistic Regression turned out to be consistent in performance but was not good enough to capture complex nonlinear relations in the data.

These findings indicate that the ensemble-based methods, specially the optimized XGBoost model, give better predictive power than traditional statistical models. Hence, they are more suitable for practical credit risk prediction tasks where the accuracy and reliability are both important [4], [16], [18].

To further assess the classification performance of the XGBoost model with the optimized parameters, the ROC curve was examined, presented in Figure 3. The

ROC curve reflects the trade-off that exists between the true positive rate and the false positive rate over different decision thresholds. An AUC of a higher value provides better separation between defaulting and non-defaulting borrowers, thus yielding effective risk-based decision-making from financial institutions.

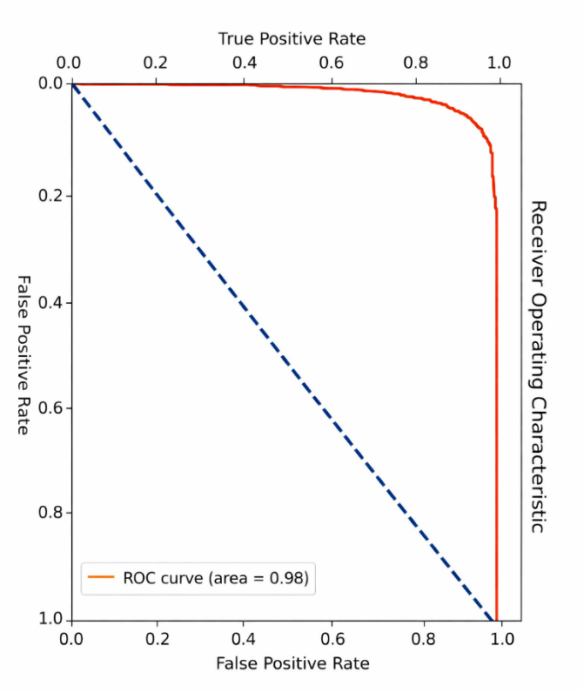

Fig. 5. Receiver Operating Characteristic (ROC) Curve for the Proposed Credit Risk Prediction Model

D. Applications and Practical Impact

The nature of credit risk prediction system that will be applied in carrying out this research is necessary in any specific form of financial service, such as personal loans, mortgage loans, credit card lending, etc.

Overall, financial institutions including the banks can quickly grant loans through prediction systems created based on the historical data associated with loans.

In the case of an operational environment, it can be used to reduce non-performing assets through an effective means of identifying high-risk borrowers at the right time and in a timely manner. It can also support the risk management team as it provides the insight hence assisting them to develop lending policies. It is, therefore, appropriate in situations where an institution wishes to improve its efficiency in operational activities by being able to fully manage credit risks.

E. Methodological Advantages

The suggested credit risk prediction system that relies on machine learning has several advantages over the traditional approach to credit risk prediction, where the rules are followed, or statistics are used, as will be explained in this paper. The proposed system can learn the information and find the hidden links between the attributes of the borrowers, which is not easy to find out by the traditional approaches.

The fact that the given methodology is efficient when addressing an imbalanced financial dataset is one of the most important benefits of the suggested approach. The cases of defaults in a real bank scenario represent very low percentages of the entire data. The use of the Synthetic Minority Over-sampling Technique (SMOTE) is used to balance the classes during training and, therefore, can contribute to the better efficiency of the model in detecting the high-risk borrowers. [2], [9].

Moreover, since ensemble learning models, such as Random Forest and XGBoost, can be also included, it can make the prediction model more stable and capable of generalization, which can be convenient when the data to work with is financial and can consist of noisy and high-dimensional variables. [4], [16], [20] In addition, the efficiency of the model can be optimized by the optimization of hyperparameters. [3].

The other advantage of the proposed system is that it is scalable and expandable. The system may be linked to the existing banking system to do real time credit assessment. The system can be re-trained as soon as new information on loans is received so as to accommodate the variation in the behavior of the customers. The system also helps in making decisions consistently and without bias thereby contributing to fair lending procedures and increases the openness of the financial decision-making by limiting the human factor in the loan approval process. Conclusion: The research methodology is a viable and practical answer to the existing credit risk management.

F. Challenges and Limitations

Despite the successes of the proposed system, it is impossible to achieve success without the quality of the data at hand. The model being trained on previous

data on a loan will mean that any unforeseen changes in the economic or the user behavior can affect the accuracy of the prediction system. It is explained by the fact that this is a typical weakness of data-based models that are applied to financial systems.

A second problem is that of interpretability of models. Big group models like the Random Forest and the XGBoost give a rather excellent degree of precision, but the process of decision-making is not consistently interpretable. This may render it hard to comprehend the logic behind some of the credit decisions. [12], [18].

The future study can be oriented to solving the issues through the combination of explainable AI and the use of other inputs in real-time like economic indicators.

IV. CONCLUSION

The research suggested a machine learning-based credit risk prediction method to increase the ability of the banks to identify default risk on loan payments. Through the power of pre-processing data, how to address the issue of imbalanced data, and machine learning methods applied to the real-life credit data, the suggested system has been determined to have potential prediction capabilities.

In experimental results, it is evident that an optimized XGBoost method would be best as opposed to other classical credit-scoring models because it can achieve an F1-score of as high as 97.4. This is a pointer that the technique is efficient, particularly when the risks that are taken in dealing with riskier default are balanced, as well as reducing false loan applications. The system is scalable, and it can be utilized in a practical manner in banking purposes. [1], [4], [5].

Overall, it can be stated that the findings of these studies prove the importance of machine learning in management of financial risks and in lending decisions. Moreover, in the context of system development in general, it is necessary to mention that such systems are oriented at offering improved mechanisms of control over non-performing assets/ properties, as well as lending decisions made with the reference to.

Further research will deal with the expansion of model transparency by using methods of explainable AI and the introduction of real-time economic measures. These additions might bring in the degree of trust, intelligibility and regulatory compliance to the concept of smart banking. [12], [18]

Note: AI-related technologies have been applied during literature clarification and technical support only in the process of forming the manuscript. The authors have effectively validated experiments, study design, analysis and the general findings.

REFERENCES

[1] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785–794.

[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[3] J. Bergstra, D. Yamins, and D. Cox, “Making a science of model search: Hyperparameter optimization in hundreds of dimensions,” Proc. 30th Int. Conf. Machine Learning (ICML), 2013, pp. 115–123.

[4] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[5] D. Hand and W. Henley, “Statistical classification methods in consumer credit scoring: A review,” Journal of the Royal Statistical Society, vol. 160, no. 3, pp. 523–541, 1997.

[6] S. Moro, P. Cortez, and P. Rita, “A data-driven approach to predicting the success of bank telemarketing,” Decision Support Systems, vol. 62, pp. 22–31, 2014.

[7] A. Lessmann, B. Baesens, H. V. Seow, and L. C. Thomas, “Benchmarking state-of-the-art classification algorithms for credit scoring,” European Journal of Operational Research, vol. 247, no. 1, pp. 124–136, 2015.

[8] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen, “Benchmarking state-of-the-art classification algorithms for credit scoring,” Journal of the Operational Research Society, vol. 54, no. 6, pp. 627–635, 2003.

[9] H. He and E. A. Garcia, “Learning from imbalanced data,” IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, Sept. 2009.

[10] S. Wang and X. Yao, "Diversity analysis on imbalanced data sets by using ensemble models," Proc. IEEE Symposium on Computational Intelligence, 2009, pp. 324–331.
[11] J. Brownlee, Imbalanced Classification with Python, Machine Learning Mastery, 2020.
[12] P. Cortez and M. J. Embrechts, "Using sensitivity analysis and visualization techniques to open black box data mining models," Information Sciences, vol. 225, pp. 1–17, 2013.
[13] F. Provost and T. Fawcett, "Reliable classification for imprecise environments," Machine Learning, vol. 42, no. 3, pp. 203–231, 2001.
[14] K. J. Cios, W. Pedrycz, R. Swiniarski, and L. A. Kurgan, Data Mining: A Knowledge Discovery Approach, Springer, 2007.
[15] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," Journal of Machine Learning Research, vol. 3, pp. 1157–1182, 2003.
[16] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119–139, 1997.
[17] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," Proc. 8th ACM SIGKDD Int. Conf., 2002, pp. 694–699.
[18] M. Kuhn and K. Johnson, Applied Predictive Modeling, Springer, 2013.
[19] R. B. Altman, "Credit scoring and its applications," IEEE Intelligent Systems, vol. 27, no. 2, pp. 90–95, 2012.
[20] J. Friedman, Greedy function approximation: A gradient boosting machine, Annals of Statistics, vol. 29, no. 5, pp. 11891232, 2001.