

Credit Risk Assessment using Machine Learning Techniques

Kavita Kolpe¹, Madhuri Suryavanshi², Rushikesh Magadum³, Aditya Marke⁴, Aditya Mashere⁵ and Siddhesh More⁶

¹⁻⁶Pimpri Chinchwad College of Engineering /Computer Engineering, Pune, India

Email: kavita.kolpe@pccoepune.org, madhuri.suryavanshi@pccoepune.org, rushikesh.magadum22@pccoepune.org, aditya.marke22@pccoepune.org, aditya.mashere22@pccoepune.org, siddhesh.more22@pccoepune.org

Abstract—Credit risk assessment is a critical aspect of finance, particularly for evaluating the financial health of individuals and companies. Companies often rely on investments and loans, making a meticulous analysis of their financial conditions, as reflected in financial statements like balance sheets, cash flow statements, and profit and loss statements, essential to avoid severe losses for investors. This analysis, however, is complex and requires a refined understanding of financial ratios such as working capital, total assets, retained earnings, EBIT, market value of equity, total liabilities, and sales. Machine learning algorithms, including Random Forest, Gradient Boosting, Logistic Regression, Neural Networks, and Support Vector Machines, offer robust solutions for predicting credit risks. Our review highlights Random Forest and Gradient Boosting as the most effective, balancing accuracy and performance, while also exploring the role of Explainable AI (XAI) techniques like SHAP and LIME in enhancing model interpretability. Furthermore, we address challenges in transparency, regulatory compliance, and the integration of alternative data sources, emphasizing the need for scalable, interpretable, and fair models to advance credit risk assessment practices.

Index Terms—Credit Risk Evaluation, Machine Learning Techniques, Random Forest, Gradient Boosting, XGBoost, LightGBM, SMOTE-ENN, SHAP Values, Feature Selection Techniques, Feature Extraction Methods, Explainable Artificial Intelligence (XAI), Ratio Analysis.

I. INTRODUCTION

Credit risk assessment is a very significant role in the financial sector, enabling institutions to assess the financial soundness of individuals and organizations before extending loans or investments. Traditionally, this was done through manual analysis of financial statements, including balance sheets, cash flow statements, and profit and loss statements. However, such assessments have become increasingly complex and voluminous, making it challenging to rely solely on traditional techniques and necessitating the use of machine learning techniques. Machine learning has modernized credit risk assessment by presenting tools for precision in predictions, and efficient and analytical evaluation of financial ratios relating to working capital, retained earnings, EBIT, and total liabilities. These algorithms include recent ones such as Random Forest, Gradient Boosting, Logistic Regression, Neural Networks, and Support Vector Machines, each being proposed and utilised with relative strengths and drawbacks. This review paper assesses developments in credit risk assessment through broad-scale analysis of available research into the area. In our paper, we review how machine learning techniques of

different forms achieve higher accuracies and better scalability to provide more acceptable results, then shed light on XAI's capabilities to promote better interpretability by techniques, including SHAP and LIME. We look forward to this analysis toward helping us understand the progress within credit risk assessment and also perceive the scope left toward our eventual goal of obtaining fair, scalable, and interpretable solutions.

II. LITERATURE SURVEY

Feature Enhanced Ensemble Modeling with Voting Optimization for Credit Risk Assessment (2024) presents a new multi-stage ensemble model designed to enhance credit risk prediction, especially for Small and Medium Enterprises (SMEs). Compared to other credit evaluation models, this study's analysis is more thorough because it uses both financial and non-financial data. By dynamically rating the relative relevance of several base models inside the ensemble, the voting-weight optimization inside the ensemble methodology is used in this study to improve predictive accuracy. To further make the model explainable, the authors employ SHapley Additive exPlanations (SHAP), a cutting-edge interpretability technique. SHAP makes it possible to narrow down exactly which specific elements—such as debt-to-income ratios, legal compliance, or other operational and financial variables—are influencing the model's projections. Financial decision-makers who need practical insights into credit risk characteristics will find this technique particularly relevant due to its balance between interpretability and precision. The research paper entitled Comparative Study of Ensemble Methods for Credit Risk Prediction, 2024, provides an in-depth comparison of ensemble learning methods, with a focus on XGBoost and Random Forest applied to credit risk prediction.

The evaluation makes use of two important metrics: interpretability and prediction accuracy. In order to further explore the reasons behind the model's prediction, SHAP is employed as an interpretation tool to describe the feature importance of both techniques. The authors, for instance, describe how variables like loan term, income stability, and credit history affect risk classification. By contrasting the advantages and disadvantages of XGBoost and Random Forest, the study offers practical guidance on how to choose the best models for various credit risk prediction situations. Actually, this comparison aids practitioners in striking a compromise between financial model explainability and accuracy. The authors address this widespread class imbalance in credit risk datasets, which results in predictions that are skewed toward the majority class, in their paper Effective Credit Risk Prediction Using Ensemble Classifiers with Model Explanation, 2024. In order to combat this, the study combines the potent ensemble algorithm XGBoost with the hybrid resampling technique SMOTE-ENN (Synthetic Minority Oversampling Technique with Edited Nearest Neighbors).

As a result, the model's overall accuracy has been maintained while its capacity to accurately classify minority classifications, such high-risk borrowers, has been much enhanced. In order to enhance the model's interpretability and enable stakeholders to comprehend how specific characteristics, including the borrower's income, prior defaults, and loan-to-value ratio, affect the projections, the authors also make use of SHAP. According to the study, the methodology is especially well-suited for financial applications, where maintaining regulatory compliance and fostering trust depend heavily on high accuracy and openness. According to the publication Credit Risk Assessment Using Random Forest and Gradient Boosting (2023), the use of these methods can help choose crucial variables for credit risk assessment analysis, like income and credit history. These ensemble approaches are appropriate for a comparatively basic dataset because they strike a balance between interpretability and accuracy. The study is conscious of a drawback, though, which is that it may perform worse in more dynamic and complex scenarios when dealing with complex feature relationships.

This is to be interpreted as an indication of requiring more complex modeling approaches or hybrid ensemble methods that would more accurately describe subtle relationships among variables. The paper Machine Learning for Enhanced Credit Risk Assessment: An Empirical Approach (2023) explores the performance of ensemble models, specifically XGBoost, for credit risk classification. The authors argue that XGBoost has a consistent lead over traditional machine learning models as it can deal with large data sets and data distributions that are imbalanced well. Therefore, a feature importance ranking study highlights importance, as ranking improves the interpretation of models besides pinpointing the critical factors influencing prediction for credit risks. The study shows that practically, it reveals implications where it becomes important due to scenarios, which have faced balancing accuracy and fairness on the side, as is happening with traditional models, to put them into proper balance using advanced techniques in modern-day assessments on credit risks.

The paper "Comparative Study of Machine Learning Algorithms for Credit Risk Prediction" from year 2020 conducts a broad-based comparison of several machine learning algorithms with respect to credit risk datasets, including Artificial Neural Networks and Random Forests. It is shown that both ANN and Random Forest perform better by using large datasets and can be put into practice organizations which deal with large customer

portfolios. It doesn't dive deeply into advanced ensemble techniques or modern interpretability tools like SHAP, which puts the study somewhat limited in terms of applicability to contemporary credit risk modeling. This paper offers very useful insights about how various algorithms perform but is shallow in what's required to address the complexities of credit risk prediction in today's financial landscape. The study entitled Boosting Techniques for Credit Scoring: A Comprehensive Analysis of XGBoost, AdaBoost, and GBDT by 2019, focuses on the role that boosting algorithms play in improving credit scoring accuracy.

According to the authors, the application of XGBoost proves efficient in minimizing the rates of misclassifications, especially when dealing with imbalanced datasets. Traditional methods have been proven to fail under such circumstances. In comparing XGBoost with AdaBoost and Gradient Boosting Decision Trees (GBDT), the results clearly indicate the robustness of XGBoost and its high scalability, demonstrating superiority over other algorithms. To conclude, this study has great significance for techniques used in boosting as they tend to deal effectively with complex data, thereby giving importance to a novel method in predicting credit risk. Collectively, these studies demonstrate the significant developments in credit risk assessment, in particular, the importance of ensemble methods, boosting techniques, and the development of SHAP and other interpretability tools. The incorporation of strategies to address imbalances in data and incorporate other non-financial variables signifies a shift towards better and more reliable approaches towards credit risk prediction. Such diverse methodologies involved in this research body clearly derive its essence from the fact that credit risk assessment has become a dynamic entity with an increased reliance on smart machine learning techniques.

III. DATASET DESCRIPTION

The dataset consists of data related to the financial, geographically, and sectoral position of 121,008 firms from an initial count of 48 attributes. Using data for only one year diminishes the count of attributes down to 16, focusing primarily on key financial metrics such as turnover, profitability, and leverage. The geographic description, sector categorization, and target variable, MScore or credit rating are additional. The dataset can be further reduced in dimensionality to roughly ten elements, keeping only the most important characteristics for predicting financial credit risk. A targeted and effective analysis utilizing machine learning models is ensured by this fine-tuning. The following 10 characteristics can be distinguished based on their applicability and context in predicting financial credit risk:

1. Turnover: Represents the company's annual revenue, crucial for evaluating its financial scale.
2. EBIT (Earnings Before Interest and Taxes): Indicates the company's operating profitability.
3. PLTax (Profit or Loss Before Tax): Reflects the company's profitability before tax liabilities.
4. Leverage: Debt-to-equity ratio, a key measure of financial stability and risk.
5. ROE (Return on Equity): Shows the company's efficiency in generating returns for shareholders.
6. Total Assets: Represents the company's overall financial strength and size.
7. MScore: The target variable, which is the credit rating to be predicted.
8. Country: Provides insights into the company's operating environment.
9. Sector 1: Captures the primary classification of the company's industry.
10. Region: Indicates the broader geographic area of the company's operations.

These attributes balance financial, geographic, and industrial dimensions to provide a comprehensive input set for predicting credit risk efficiently.

No	Company name	Turnover.2020
0	LENDLEASE S.R.L.	29458
1	PRICEWATERHOUSECOOPERS BUSINESS SERVICES SRL (...)	16731
2	EVISIO S.P.A.	48568
3	CASA SERVICE MACHINE	47999
4	PANFERTIL SPA	45948

	Turnover.2019	Turnover.2018	Turnover.2017	Turnover.2016	Turnover.2015
0	16716	9612	8097	7941.0	5600.0
1	16403	16843	12241	9252.0	9515.0
2	43039	34302	25791	19760.0	6941.0
3	43484	43043	41682	51267.0	52584.0
4	47336	45626	48222	57074.0	62263.0

	EBIT.2020	EBIT.2019	...	ROE.2018	ROE.2017	ROE.2016	ROE.2015
0	-1556.0	-4540	...	8.24	-146.65	60.76	-471.72
1	1838.0	841	...	61.42	-55.57	-127.29	-87.13
2	1661.0	1464	...	57.52	42.73	20.34	44.62
3	416.0	255	...	-17.24	0.71	2.89	6.45
4	44.0	713	...	-5.17	-6.74	0.03	-8.19

	TAsset.2020	TAsset.2019	TAsset.2018	TAsset.2017	TAsset.2016
0	49263	28268	15455	15992	13597.0
1	16550	16887	16468	10773	6697.0
2	13500	9620	7371	5432	4170.0
3	24978	25032	25729	21632	25403.0
4	36823	34659	36205	38423	41847.0

Figure 1. Dataset Overview

TABLE I: COMPARATIVE ANALYSIS

Sr. No.	Title	Year	Methodology	Strengths	Weaknesses
1	Feature Enhanced Ensemble Modeling with Voting Optimization for Credit Risk Assessment	2024	Multi-stage ensemble, Bagging-based oversampling, Voting-weight optimization, SHAP	- Combines non-financial behavioural data with financial data for improved predictions - SHAP ensures model transparency	- Complex and computationally demanding - Requires extensive data preprocessing
2	Comparative Study of Ensemble Methods for Credit Risk Prediction	2024	Random Forest, XGBoost, LightGBM, SHAP, AdaBoost, Smote-Enn	- Excellent for handling complex, high-dimensional data - Provides granular feature importance with SHAP	- High computational cost - Interpretability is challenging without SHAP
3	Effective Credit Risk Prediction Using Ensemble Classifiers with Model Explanation	2024	SMOTE-ENN, XGBoost, Random Forest, SHAP	- Handles imbalanced datasets well with SMOTE-ENN - SHAP ensures transparency and interpretability - High accuracy	- Computationally intensive - Overfitting possible with synthetic data generation from SMOTE
4	Credit Risk Assessment Using Random Forest and Gradient Boosting	2023	Random Forest, Gradient Boosting (XGBoost), Correlation Analysis	- Provides a balance of interpretability and predictive accuracy - Effective at identifying key credit risk factors	- Limited in capturing complex feature interactions - Prone to overfitting on noisy data
5	Machine Learning for Enhanced Credit Risk Assessment: An Empirical Approach	2023	Ensemble models (e.g., XGBoost, Random Forest, Gradient Boosting); 10 algorithms tested on large dataset using 5-fold cross-validation.	Extensive use of ensemble methods; Large dataset with over 2.5 million observations; Shows XGBoost outperforms traditional methods in credit classification.	Complexity of ensemble methods makes interpretability difficult; No detailed hyperparameter tuning; Focus on performance rather than explainability.
6	Machine Learning Algorithms for Credit Risk Assessment	2023	Algorithms: Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Neural Networks Pre-processing: Feature Engineering, Data Splitting	Random Forest and Gradient Boosting are top performers for high-dimensional data - Neural Networks capture complex patterns effectively	- Neural Networks lack interpretability - Regulatory compliance and transparency pose challenges with black-box models
7	A machine learning-based credit risk prediction engine system	2024	Stacked classifier model using Random Forest (RF), Gradient Boosting (GB), and XGBoost (XGB) with Information Gain (IG)-based feature selection	Outperforms individual models, handles class imbalance, applied across three diverse credit risk datasets.	High computational complexity, less focus on model interpretability
8	Machine Learning Algorithms for Credit Risk Assessment: An Economic and Financial Analysis	2023	Compares traditional credit scoring models with machine learning algorithms, including Logistic Regression, Random Forest, Gradient Boosting, and Neural Networks.	-Demonstrates the higher accuracy and predictive power of machine learning models, discusses economic benefits like cost reduction.	-Challenges with model interpretability and regulatory compliance, especially for complex models like Neural Networks.

IV. GENERAL FRAMEWORK DIAGRAM

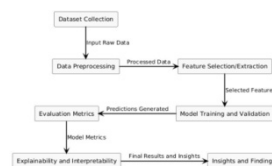


Figure 2. General Framework Diagram of Credit Risk Assessment

V. ALGORITHMIC SURVEY

A. Preprocessing

Standardization and Normalization

- i. Ensures that data features are on the same scale, improving model convergence and performance.
- ii. This is particularly important for algorithms sensitive to feature scales, like SVM and Neural Networks.

One-Hot Encoding

- i. Converts categorical variables into binary variables, ensuring all categories are equally represented in the data.
- ii. Helps models handle categorical data effectively, such as business type or credit rating categories.

Mean Imputation:

- i. Fills missing values in the dataset by replacing them with the mean value of the corresponding feature.
- ii. Prevents data loss and maintains the integrity of the dataset.

Variance Inflation Factor (VIF):

- i. Detects multicollinearity among independent variables by calculating how much the variance of a regression coefficient is inflated due to multicollinearity.
- ii. Reduces redundancy by removing highly correlated features.

SMOTE (Synthetic Minority Oversampling Technique):

- i. Addresses data imbalance by generating synthetic samples for the minority class (e.g., default cases).
- ii. Ensures that models can learn from a balanced dataset, improving classification performance for minority classes.

Outlier Detection:

- i. Identifies and removes data points that are significantly different from the majority of the data, reducing noise.
- ii. Helps in enhancing model accuracy by focusing on more representative data points.

Discretization:

- i. Continuous variables like credit applications and employment length were discretized into fewer categories to reduce complexity.

B. Feature Selection

SHAP (SHapley Additive exPlanations):

- i. Provides post-hoc interpretability by quantifying the contribution of each feature to the model's predictions.
- ii. Useful for understanding feature relevance in high-stakes decisions.

Principal Component Analysis (PCA):

- i. Reduces dimensionality by transforming original features into principal components.
- ii. Captures the most significant variance in the data for model input.

C. Feature Extraction

In the context of SME credit risk assessment, simply relying on financial data such as total borrowing or debt-to-income ratio is often inadequate. To address this, multi-feature enhancement is crucial. This section elaborates on how advanced models integrate diverse data sources to improve predictive accuracy.

Random Forest and Gradient Boosting

These ensemble methods not only perform classification but also help in extracting key features. During the tree-building process, features are evaluated based on their contribution to reducing impurity (e.g., Gini impurity in decision trees). The most relevant features, such as annual income, loan amount, and employment length, are extracted as part of the model building.

Correlation Analysis

This helps in identifying multicollinear features. Strongly correlated features (e.g., age and credit history length) are either combined or dropped to improve model performance.

D. Prediction Methods

Random Forest (RF):

Because of its resilience and capacity to manage datasets with several attributes, Random Forest is a popular

method. By constructing several decision trees and averaging their results, it lowers overfitting and boosts dependability. Because it offers feature relevance scores, which aid in identifying the most pertinent variables, RF is especially useful for feature selection. Compared to more sophisticated approaches, it has the drawback of not capturing complicated feature interactions, although being extremely interpretable and effective for simpler datasets.

Gradient Boosting Algorithms (XGBoost, LightGBM, GBDT):

The reason why Gradient Boosting algorithms like XGBoost, LightGBM, and GBDT are used is that they have higher predictive accuracy as well as efficiency in handling large, complex datasets. XGBoost has high speed with proper regularization techniques against overfitting, and LightGBM deals well with large data sets containing categorical features and achieves leaf-wise growth for efficient handling of large datasets. GBDT refines its predictions incrementally, which is very effective with structured data. These algorithms do an excellent job modeling non-linear relationships, but they are computationally expensive and less interpretable than more simple models.

Logistic Regression (LR):

Logistic Regression is one of the basic methods for credit risk assessment, especially as a benchmark for evaluating other algorithms. It is simple, interpretable, and effective in capturing linear relationships between features and outcomes. Even though it is limited in handling non-linear patterns, LR is still widely used due to its transparency and ease of implementation, which makes it a staple in financial applications.

Support Vector Machines (SVM):

SVM is powerful in separation of risky from non-risky customers, especially if the dimension of the space is high. The model uses the hyperplane in the classification of data and, hence, popular for handling a nonlinear relationship using kernel functions. This is highly recommended for datasets for which the classes are not directly separable but may be very time-consuming and less interpretive compared to simple models.

VI. CONCLUSION AND FUTURE SCOPE

Feature selection and extraction are crucial for improving predictive accuracy in credit risk evaluation using machine learning. Techniques such as Random Forest (RF) are best suited for smaller datasets because of their simplicity and interpretability, while Gradient Boosting methods (XGBoost, LightGBM) perform better with high-dimensional and complex data but with lower interpretability. SMOTE-ENN can effectively correct the class imbalance problem dominating credit risk scoring applications by oversampling synthetic samples and cleaning noisy data—a factor that should be taken into extreme caution so as not to cause overfitting. SHAP plays a critical role in improving the interpretability post-hoc since it provides valuable information regarding complex models and develops trust in the predictions.

Going forward, hybrid models combining techniques like XGBoost with SMOTE-ENN can be explored to address imbalanced datasets while ensuring high predictive accuracy. Advanced Explainable AI (XAI) frameworks will ensure that the trade-off between high predictive performance and interpretability of the results allows for ethical and transparent decision-making in financial contexts. Moreover, the deep learning and feature engineering approach with scalable, automated real-time solutions to credit risk assessment opens vast avenues for improvement in efficiency and reliability in this domain.

REFERENCES

- [1] Yang, Dongqi & Xiao, Binqing. (2024). Feature Enhanced Ensemble Modeling with Voting Optimization for Credit Risk Assessment. IEEE Access. PP. 1-1. 10.1109/ACCESS.2024.3445499.
- [2] T. Xu, “Comparative Analysis of Machine Learning Algorithms for Consumer Credit Risk Assessment”, TCSISR, vol. 4, pp. 60–67, Jun. 2024, doi: 10.62051/r1m3pg16.
- [3] Aruleba, Idowu & Sun, Yanxia. (2024). Effective Credit Risk Prediction Using Ensemble Classifiers with Model Explanation. IEEE Access. PP. 1-1. 10.1109/ACCESS.2024.3445308.
- [4] Tian, Zhenya & Xiao, Jialiang & Feng, Haonan & Wei, Yutian. (2020). Credit Risk Assessment based on Gradient Boosting Decision Tree. Procedia Computer Science. 174. 150-160. 10.1016/j.procs.2020.06.070.
- [5] Suhadolnik, Nicolas & Ueyama, Jo & Da Silva, Sergio. (2023). Machine Learning for Enhanced Credit Risk Assessment: An Empirical Approach. Journal of Risk and Financial Management. 16. 496. 10.3390/jrfm16120496.
- [6] Xu, Tianyi. (2024). Comparative Analysis of Machine Learning Algorithms for Consumer Credit Risk Assessment. Transactions on Computer Science and Intelligent Systems Research. 4. 60-67. 10.62051/r1m3pg16.

- [7] Zou, Y.; Gao, C. Extreme Learning Machine Enhanced Gradient Boosting for Credit Scoring. *Algorithms* 2022, 15, 149. <https://doi.org/10.3390/a15050149>
- [8] Xu, Feiyu & Uszkoreit, Hans & Du, Yangzhou & Fan, Wei & Zhao, Dongyan & Zhu, Jun. (2019). Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. 10.1007/978-3-030-32236-6_51.
- [9] Shi, S., Tse, R., Luo, W. et al. Machine learning-driven credit risk: a systemic review. *Neural Comput & Applic* 34, 14327–14339 (2022). <https://doi.org/10.1007/s00521-022-07472-2>
- [10] Luisa Roa, Alejandro Correa-Bahnsen, Gabriel Suarez, Fernando Cortés-Tejada, Maria A. Luque, Cristián Bravo, Super-app behavioral patterns in credit risk models: Financial, statistical and regulatory implications ,*Expert Systems with Applications*, Volume 169, 2021, 114486, ISSN 0957-4174, <https://doi.org/10.1016/j.eswa.2020.114486>.
- [11] Alessandro Bitetto, Paola Cerchiello, Stefano Filomeni, Alessandra Tanda, Barbara Tarantino, Machine learning and credit risk: Empirical evidence from small- and mid-sized businesses, *Socio-Economic Planning Sciences*, Volume 90, 2023, 101746, ISSN 0038-0121, <https://doi.org/10.1016/j.seps.2023.101746>.
- [12] Saxena, Aditya & Dungore, Parizad. (2021). Credit Risk Assessment Model for UAE's Commercial Banks: A Machine Learning Approach. <https://doi.org/10.1016/j.sxpe.2021.101764>