Journal of Advanced Artificial Intelligence
Volume 2 - No.1, August 2025

Credit Risk Prediction using Ensemble and Linear Machine Learning Models

Bukunmi Gabriel Odunlami
Department of Electrical and Computer Engineering,
New Jersey Institute of Technology
Newark, 07102, New Jersey, USA

Blessing Nwonu
Department of Mathematics
Temple University
Philadephia, 19122, Pennysylvania, USA

ABSTRACT

Predicting the likelihood of loan default remains a critical challenge in credit risk modeling, where data imbalance, high dimensionality, and nonlinear interactions often limit the effectiveness of traditional scoring techniques. This paper presents a machine learning pipeline for credit risk prediction using financial datasets. We evaluate six main classifiers—Logistic Regression, Gaussian Naive Bayes, Support Vector Machines, Random Forest, XGBoost, and LightGBM and a variant of two of the classifiers for further comparison. Models are benchmarked using accuracy, precision, recall, and the Kolmogorov–Smirnov statistic widely used in financial risk scoring. Our results indicate that ensemble methods combined with hybrid resampling techniques can consistently offer significant improvements in default risk separation without requiring dimensionality reduction methods, complex deep neural architectures or other black-box models. This makes them suitable for both regulated credit scoring environments and modern machine learning-driven financial applications.

Keywords

Credit risk, Ensemble model, Hybrid resampling, Supervised learning, Kolmogorov-Smirnov statistic

1. INTRODUCTION

The possibility of a borrower failing to meet repayment obligations directly impacts a financial institution's solvency, capital allocation, and operational continuity. Failures in risk assessment can lead to cascading defaults, systemic distress, and regulatory intervention. As global credit markets expand and digitize, the ability to accurately assess borrower risk becomes not only a regulatory necessity but also a competitive differentiator. Yet, the task of managing credit risk grows increasingly complex. Institutions today must contend with rapidly changing market conditions, borrower behaviors, emerging financial technologies, and a heightened regulatory environment. Modern credit evaluation is no longer limited to a borrower's declared income or collateral. It spans a broad set of attributes such as employment stability, credit history, payment behavior, loan terms, and increasingly leverages external data such as credit bureau reports, digital footprints, and macroeconomic signals. Financial institutions face mounting challenges, including data sparsity, model transparency requirements, cybersecurity risks, and the imperative to align credit decisions with evolving business strategies [7]. Traditional methods for assessing creditworthiness, such as the "5C" framework (Character, Capacity, Capital, Collateral, and Conditions), rely heavily on expert judgment and qualitative appraisal.

While these approaches provide interpretability and historical anchoring, they are now challenged by the volume, velocity, and variety of financial data in modern lending ecosystems. Similarly, scorecard-based models such as behavioural scorecard, though widely adopted due to their transparency and ease of regulatory acceptance, tend to underperform in nonlinear, high-dimensional, and imbalanced settings prevalent in contemporary credit data [10].

To meet these challenges, credit risk modeling has become a critical axis around which both strategy and compliance revolve, particularly under international standards such as Basel II and III [5]. Credit institutions are progressively integrating machine learning into their risk assessment workflows. This study contributes to the growing body of literature by developing and evaluating a comprehensive, reproducible pipeline for credit risk prediction using structured loan-level data. Our core contributions are threefold:

(1) We design a full pipeline that includes preprocessing, feature engineering and model training using both interpretable and high-performance learners.
(2) We demonstrate that integrating an ensemble learning algorithm with a hybrid resampling strategy can yield strong classification performance, eliminating the need for explicit dimensionality reduction.
(3) We conduct detailed performance diagnostics to assess not only predictive accuracy but also the robustness and calibration of the models across class-imbalanced data.

2. RELATED WORK

Credit risk modeling has evolved significantly over the past two decades, shifting from traditional statistical approaches to more flexible and accurate machine learning frameworks. Logistic regression and other linear discriminant models have long been foundational tools due to their interpretability and compliance with regulatory standards. However, these models often fall short in

handling high-dimensional feature spaces, nonlinearity, and the class imbalance that typically characterizes loan datasets [11]. To address these limitations, recent studies have focused on the application of supervised and unsupervised machine learning approach to credit scoring. [2] systematically compared several commonly used ML classifiers, including Random Forest, Support Vector Machines, and boosting methods across European credit data. Their results showed that ensemble models consistently outperformed traditional linear models. [12] examined peer-to-peer lending data and confirmed the strength of boosting algorithms such as XGBoost and LightGBM in terms of both predictive power and stability under imbalance.

The role of preprocessing and transformation has also been a focal point in recent literature. The works in [16] and [17] evaluated credit scoring models with and without Weights of Evidence (WoE) encoding. Their findings suggest that WoE transformations may enhance model performance depending on the underlying classifier but are not universally beneficial. [20] explored the use of hybrid resampling techniques such as synthetic minority over-sampling technique (SMOTE) and edited nearest neighbors methods (ENN) in combination with tree-based learners. They reported that LightGBM with SMOTEENN and the principal component analysis method yielded the highest KS statistic in their benchmarks, outperforming deep learning models. [4] stressed the need for standardized pipelines in credit risk prediction research, advocating for consistent preprocessing, feature selection, and validation methods that improves reproducibility. Our study is designed in line with these best practices, incorporating hybrid resampling, dimensionality reduction, and a unified evaluation framework across multiple classifiers.

3. DATA PREPROCESSING AND MODELING PIPELINE

The prediction task is formulated as a supervised binary classification problem using a structured loan-level data obtained from a real-time financial database. The dataset includes borrower demographic features (such as age, employment length, income, and home ownership status), loan metadata (amount, intent, and grade), and a historical risk label indicating default or repayment. Loans are graded from A to G based on internal creditworthiness criteria, with higher grades corresponding to higher expected default risk and interest rates. See Figure 1. Also, Table 1 shows a detailed description of our dataset.

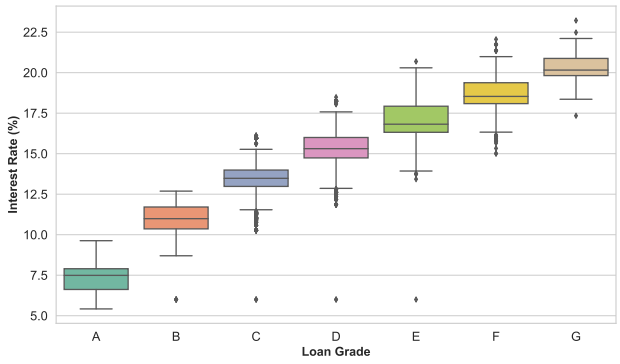

Fig. 1: Loan grade vs interest rates

3.1 Feature correlation analysis

One of the challenges with modeling financial dataset to assess fraud and borrower risk using machine learning models is the complex inter-feature relationships. Understanding feature correlation is thus essential for uncovering hidden data patterns and to determine proper modeling approach for the task. The feature relationship of our dataset is shown in Figure 2. Using Pearson correlation coefficients (ranging from -1 to $+1$) in Equation (1), we evaluate the linear relationships among all numeric variables in the dataset.

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{1}$$

where X_i and Y_i are the individual sample points, $\bar{X}$ and $\bar{Y}$ are the sample means of X and Y respectively, and n is the number of observations. This coefficient quantifies the strength and direction of the linear relationship between two continuous variables, taking values between -1 (perfect negative correlation) and $+1$ (perfect positive correlation) with 0 indicating no linear correlation. The key observations are enumerated as follows:

(1) **High Positive Correlation:** Age and credit history length show a very strong positive correlation of approximately $r = 0.86$, indicating that older individuals generally have longer credit histories.

(2) **Moderate Positive Correlation:** Loan amount and the percentage of income allocated to the loan exhibit a moderate correlation ($r \approx 0.57$), suggesting that larger loans tend to represent a greater share of the borrower's income. Similarly, loan interest rate and default status correlate at $r \approx 0.34$ which means that higher interest rates may be associated with an increased risk of default. Default status and loan-to-income ratio also show moderate correlation ($r \approx 0.38$) and this further emphasizes the importance of borrower affordability.

(3) **Weak or Negligible Correlation:** Several variables such as income and interest rate or age and interest rate, show negligible correlation (near zero) and this implies minimal direct linear association or potentially more complex nonlinear patterns.

(4) **Negative Correlation:** A moderate negative correlation ($r \approx -0.25$) is observed between income and loan-to-income ratio, reflecting the intuitive relationship that higher-income individuals tend to have a smaller proportion of their income committed to loan repayments.

These insights are necessary for both feature engineering and model diagnostics. Strongly correlated features may introduce multicollinearity if modeled directly without regularization or dimensionality control. Moreover, the presence of several moderate or weak linear correlations with the target variable suggests that linear models may be insufficient to capture the underlying structure of the data. The implication of this is that even though this is a binary classification problem, linear models would be limited in performance as the interaction of most of the features are not linearly seperable. In contrast, ensemble methods such as LightGBM and XGBoost are well-suited to exploit complex, nonlinear, and hierarchical relationships across features [13]. Their ability to model such interactions without extensive feature engineering further justifies their use in this paper.

Table 1. : Summary of Financial Dataset Attributes

Feature Name	Data Type	Description
person_age	Numerical	Age of the borrower
person_income	Numerical	Reported annual income of the borrower
person_home_ownership	Categorical	Home ownership status (e.g., rent, own, mortgage)
person_emp_length	Numerical	Employment length in years
loan_intent	Categorical	Purpose of the loan (e.g., personal, education, medical)
loan_grade	Categorical	Credit grade assigned to the loan (A to G)
loan_amnt	Numerical	Amount of the loan requested
loan_int_rate	Numerical	Interest rate assigned to the loan
loan_status	Categorical	Target variable indicating loan outcome (0 = non-default, 1 = default)
loan_percent_income	Numerical	Ratio of loan amount to annual income
person_default_on_file	Categorical	Indicator of prior default on file
person_cred_hist_length	Numerical	Length of borrower's credit history
Term	Categorical	Duration of the loan (e.g., 36 months, 60 months)
Address State	Categorical	U.S. state of the borrower's residence
Debt-to-Income Ratio	Numerical	Ratio of borrower's monthly debt payments to income
Revolving Balance	Numerical	Total credit revolving balance
Total Accounts	Numerical	Total number of open credit lines

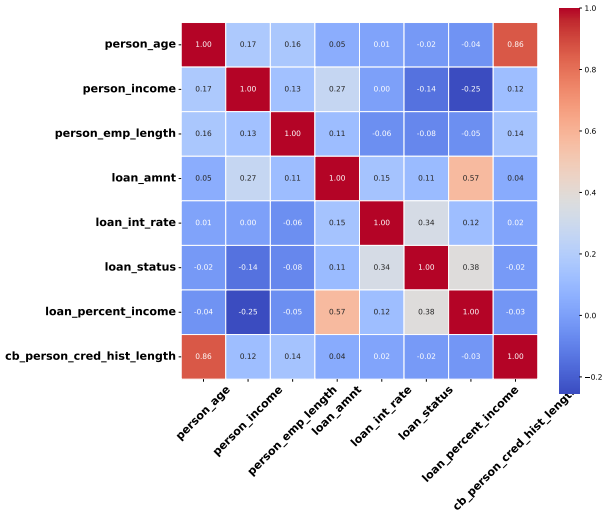

Fig. 2: Feature correlation heatmap

3.2 Handling missing values and encoding categorical variables

A few numerical features with missing values, such as *person_emp_length* and *loan_int_rate*, were imputed using mean substitution. Formally, if x_i denotes a valid observation, then the missing entries $\hat{x}$ are replaced by:

$$\hat{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2}$$

This strategy assumes missing data are random and helps preserve dataset size without introducing artificial bias. Categorical features were also transformed into numerical format to enable compatibility with machine learning algorithms:

(1) *Ordinal Encoding*: This is applied to `loan_grade`, mapping credit grades `A--G` to ordinal integers 0 through 6.

(2) *Binary Encoding*: Binary features such as those representing prior default of a loan are converted such that 'Y' to 1 and 'N' to 0.

(3) *One-Hot Encoding*: This is applied to other nominal variables such as address, `loan_intent` and `person_home_ownership` e.t.c., with one category dropped to prevent multicollinearity.

3.3 Feature scaling and cross validation strategy

All numerical features were standardized to zero mean and unit variance using z-score normalization:

$$x_{\text{scaled}} = \frac{x - \mu}{\sigma} \tag{3}$$

where μ and σ represent the empirical mean and standard deviation of the feature, respectively. This ensures that all variables contribute equally to the model and accelerates convergence during optimization. To estimate generalization performance and reduce variance due to random train-test splits, we adopted a 5-fold stratified cross-validation approach. The dataset is partitioned into five equal-sized subsets denoted as $D_1, D_2, \ldots, D_5$, ensuring that the proportion of default and non-default classes is approximately preserved in each fold. For each iteration $i \in \{1, 2, 3, 4, 5\}$, one fold D_i is held out as the validation set, and the remaining four folds $\bigcup_{j \neq i} D_j$ are used for training. Model performance is recorded on the held-out fold, and this process is repeated for all five folds. The final performance metric is computed as the average across all iterations:

$$\bar{M} = \frac{1}{5}\sum_{i=1}^{5} M_i \tag{4}$$

where M_i denotes the performance metric (e.g., accuracy, precision) computed on the i-th fold. This stratification guarantees that each fold reflects the overall class distribution, which is particularly important for imbalanced datasets and it reduces the risk of biased model evaluation and improves the robustness of the comparative analysis across classifiers.

4. EVALUATION METRICS

To evaluate the performance of the classification models for credit risk prediction, we adopt both standard and domain-specific metrics. These include accuracy, precision, recall, and the KS statistic. Let TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives respectively.

(1) Accuracy: This measures the proportion of correctly classified instances among all samples:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{5}$$

(2) Precision refers to positive predicted values and it quantifies the proportion of correctly predicted positive instances out of all predicted positives:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

(3) Recall is used to measure the proportion of actual positives that are correctly identified:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

(4) KS Statistic This is a non-parametric test statistic that quantifies the maximum distance between the cumulative distribution functions (CDFs) of the positive and negative classes. Formally, it can be expressed as:

$$\text{KS} = \max_x |F_{\text{pos}}(x) - F_{\text{neg}}(x)| \tag{8}$$

where $F_{\text{pos}}(x)$ and $F_{\text{neg}}(x)$ denote the empirical CDFs of the scores for the positive (default) and negative (non-default) classes respectively. A higher KS value indicates better separation between the two distributions and is widely used in credit scoring to assess model discriminatory power [15].

5. PREDICTION MODELS AND RESULTS

In this paper, we implemented six main classifiers and a variant of two of the classifiers for further comparison. These statistical machine learning approaches are well-known with provable analytical guarantees and have been used across several prediction tasks in machine learning. We refer the interested reader to well-cited sources such as [9, 1]. In what follows, we will discuss each of the prediction algorithm in details with specific choice of hyper-parameters where applicable. Our proposed workflow is summarized:

Algorithm 1 Credit Risk Classification Using Ensemble and Linear Models

Credit dataset $\mathcal{D}$ Predicted credit risk labels $\mathcal{P}$

Step 1: Data Preprocessing

1. Handle missing values in numerical features using mean imputation.
2. Encode categorical variables using ordinal and one-hot encoding.
3. Standardize features using Z-score normalization.
4. Split dataset into training subset $\mathcal{D}_{\text{train}}$ (20%) using stratified sampling.

Step 2: Model Training and Evaluation

5. Evaluate the following models using 5-fold stratified cross-validation:
   - (i) Logistic Regression via SGD
   - (ii) Naive Bayes
   - (iii) SVM via SGD (hinge loss)
   - (iv) SGD with Modified Huber Loss
   - (v) XGBoost
   - (vi) Random Forest
   - (vii) LightGBM (baseline)
   - (viii) LightGBM + PCA
   - (ix) LightGBM + SMOTEENN
   - (x) LightGBM + SMOTEENN + PCA

6. Evaluate model performance with defined metrics

5.1 Logistic Regression (SGD classifier)

Logistic regression is a widely used linear classification algorithm that models the probability of a binary outcome using the logistic sigmoid function. It is particularly valued for its ease of implementation, interpretation and probabilistic output [19]. In this study, we implemented logistic regression using stochastic gradient descent (SGD), a first-order optimization algorithm well-suited for large datasets and high-dimensional input spaces. To promote generalization and mitigate overfitting, L2 regularization is used. The learning rate was indirectly controlled through a small regularization parameter with value $\alpha = 0.0001$. The model was trained for up to 1000 iterations, which was empirically sufficient for convergence, and a fixed random seed was used to ensure the reproducibility of results.

5.2 Naive Bayes (GaussianNB)

The Gaussian Naive Bayes variant assumes that continuous features follow a normal distribution. This model is fast and serves as a good benchmark, especially when the assumptions of feature independence and Gaussian distribution are approximately met. No manual hyperparameter tuning was applied. Although we did not formally test for normality, financial features like income and interest rate are typically right-skewed in practice [14]. The relatively lower performance of the Naive Bayes model may stem from violations of its assumption that continuous features are normally distributed.(see also [18])

5.3 SVM (SGDClassifier with Hinge and Modified Huber Loss)

SVM was implemented using a linear model optimized through SGD with hinge loss. L2 regularization helped reduce the risk

of overfitting, and a small regularization strength $\alpha = 0.0001$ supported gradual convergence. In order to strike a balance between the sharp decision boundaries of hinge loss and the stability of squared loss, we also used the *modified Huber loss* as an alternative loss function in this classifier. The modified Huber loss is defined as:

$$\ell(z) = \begin{cases} \frac{1}{2}(1-z)^2 & \text{if } z \geq -1 \\ -4z & \text{if } z < -1 \end{cases} \quad \text{where } z = y \cdot f(x) \qquad (9)$$

This piecewise formulation offers smooth differentiability, convexity, and stronger gradients for misclassified samples compared to the standard hinge loss expressed as
As shown in Figure 3, the modified Huber loss penalizes incorrect classifications more aggressively when the margin z is confidently wrong and applies a squared loss when predictions fall within the decision margin. This makes it robust to label noise and suitable for large-margin classification in noisy or overlapping class settings.

$$\ell_{\text{hinge}}(z) = \max(0, 1-z) \qquad (10)$$

Despite these theoretical advantages, our empirical results indicate

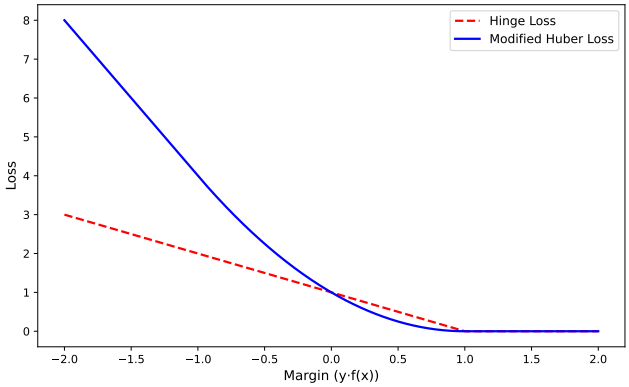

Fig. 3: Comparison of hinge and modified Huber loss functions.

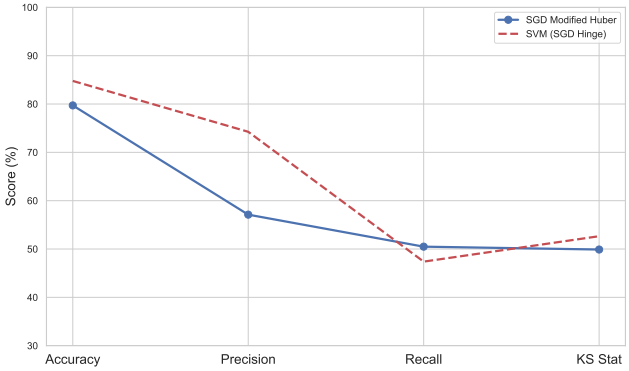

Fig. 4: Performance comparison of SGD variants across accuracy, precision, recall, and KS statistic.

that the SGD classifier trained with modified Huber loss yielded the weakest performance in terms of the KS statistics. This suggests that the model struggled to cater for the underlying distributional differences between the classes. This discrepancy stem from over-sensitivity to outliers andno the inherent smoothness of the loss not aligning well with the structure of the credit risk data. See Figure 4.

5.4 Extreme Gradient Boosting (XGBoost)

Gradient Boosting is a powerful ensemble technique that builds models sequentially, with each new model attempting to correct the errors made by its predecessors. Unlike bagging methods that train multiple trees in parallel, gradient boosting focuses on additive model optimization where learners are trained one after the other to minimize a loss function.
Formally, given a training dataset $\{(x_i, y_i)\}_{i=1}^N$, the method starts with an initial model $F_0(x)$, and builds the ensemble as:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \qquad (11)$$

where $h_m(x)$ is the weak learner (typically a shallow decision tree) trained to predict the negative gradient of the loss function at iteration m, and γ_m is the learning rate controlling the contribution of each learner.

In this study, we have used two extensions of the standard gradient boosting classifier. XGBoost is an optimized version of the standard gradient boosting classifier with regularization, tree pruning, and parallelized training. The model was configured with 100 boosting rounds (n_estimators=100), a learning rate of 0.1 to balance convergence speed and generalization, and a tree depth of 3 to prevent overfitting. XGBoost delivered strong performance across all evaluation metrics as presented in Table 2.

5.5 Random Forest Classifier

Random Forest is also an ensemble learning technique that constructs a multitude of decision trees during training time and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees. It belongs to the family of bagging methods where multiple base learners are trained in parallel using different subsets of the training data and features.
Introduced by [3], the Random Forest Classifier combines the predictive power of many weak learners to form a robust model that balances data fitting and accuracy. Each tree in the forest is trained on a bootstrapped sample from the training set, and at each split in the tree, a random subset of features is considered. This randomization reduces the correlation among individual trees, thereby increasing the generalization capability of the ensemble.
Mathematically, let $D = \{(x_i, y_i)\}_{i=1}^N$ be the training dataset, Random Forest constructs T decision trees $\{h_1, h_2, ..., h_T\}$, each trained on a bootstrap sample D_t drawn from D. At inference time, the final prediction $\hat{y}$ is given by majority voting:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), ..., h_T(x)\} \qquad (12)$$

for classification tasks. In the context of credit risk prediction, Random Forest is particularly useful due to its ability to handle high-dimensional data and automatically estimate feature importance. It effectively captures nonlinear relationships between input variables and the target class (e.g., ‘default‘ or ‘non-default‘). Furthermore, it is resilient to noisy features and less prone to overfitting compared to individual decision trees.
In this paper,we have tuned the hyperparameters of the RF via grid search, including the number of trees, maximum tree depth, and minimum samples per split.

Table 2. : Model Evaluation Summary (Metrics by Model)

Model	Accuracy	Precision	Recall	KS Statistic
Logistic Regression	0.845	0.685	0.554	0.578
Naive Bayes	0.822	0.579	0.666	0.569
SVM (Hinge Loss)	0.848	0.743	0.474	0.527
SGD (Modified Huber)	0.797	0.571	0.505	0.499
XGBoost	0.932	0.958	0.718	0.723
Random Forest	0.920	0.933	0.681	0.716
LightGBM	0.929	0.945	0.717	0.727
LightGBM + PCA	0.890	0.823	0.629	0.638
LightGBM + SMOTEENN + PCA	0.945	0.949	0.952	0.890
LightGBM + SMOTEENN	**0.961**	**0.979**	**0.950**	**0.934**

5.6 Light Gradient Boosting Machine (LightGBM)

Unlike XGBoost which uses level-wise tree growth to maintain balanced trees, LightGBM employs a best-first growth strategy that allows it to converge faster. It is optimized for speed and memory efficiency using histogram-based binning and exclusive feature bundling [8]. The hyperparameters chosen for LightGBM include 100 estimators, a learning rate of 0.1, and a maximum tree depth of 6 or 10. These were selected to balance model complexity, convergence speed, and generalization performance. To investigate the impact of dimensionality reduction on our credit data, PCA was applied to reduce the feature space while retaining 95% of variance of the dataset. Further, given the imbalance between default and non-default classes, we applied a resampling and class imbalance method. The SMOTE method generates synthetic samples of the minority class (in this case, defaulted loans) by interpolating between existing minority class instances which leads to increase in their representation in the training set. Also, the use of the ENN approach here subsequently removes potentially mislabeled or noisy examples from the majority class to refine the class boundary [6]. This hybrid method combines over- and under-sampling to enhance class separability in training data. This was intended to simplify the model structure without significantly affecting predictive performance. In addition, we employ SMOTEENN combined with PCA to investigate if this will result in the highest KS statistics for our credit data.

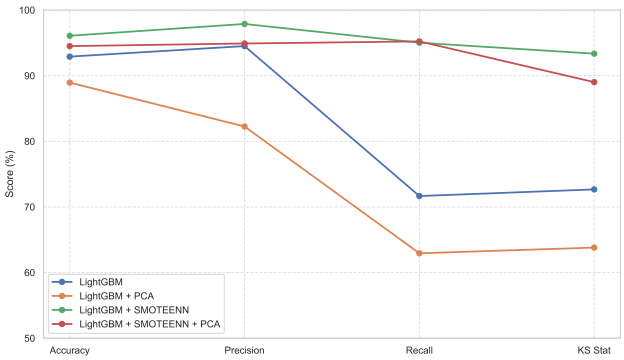

Fig. 5: Performance comparison of LightGBM variants including PCA and SMOTEENN.

While this method enhanced class balance and led to a modest improvement in the KS statistic, the introduction of artificially generated defaulter records raises concerns about the interpretability and authenticity of the predictions in real-world applications. In light of this, we adopted a combination of LightGBM and SMOTEENN without dimensionality reduction to exploit both robust ensemble learning and hybrid resampling. This has not been explicitly explored in many prior literature for credit risk prediction. A comparison of these variants of LightGBM is given in Figure 5. It is important to note that the integration of dimensionality reduction and hybrid resampling techniques in the LightGBM variants could have been extended to the other ensemble methods for a broader comparison. However, our focus on LightGBM was motivated by its superior performance, as the standard LightGBM model already achieved the highest KS statistic among all evaluated models.

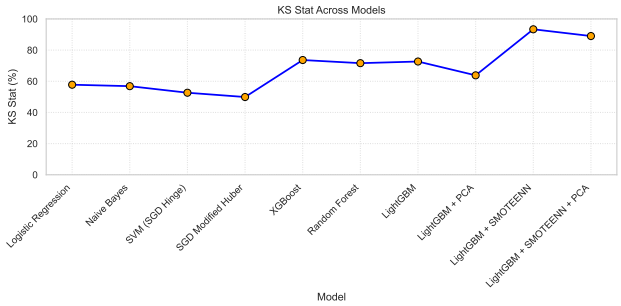

Fig. 6: KS statistics

Figure 6 presents a comparative analysis of the KS statistic across the different classification models. Among the baseline models, ensemble methods such as XGBoost, Random Forest, and LightGBM consistently outperformed linear classifiers, with LightGBM achieving the highest KS score. Notably, the introduction of SMOTEENN significantly improved the discriminatory power of LightGBM, as evidenced by the sharp increase in the KS value. The combination of SMOTEENN and LightGBM yielded the best result.
Overall, the performance of the models implemented in this paper is summarized as shown in Table 2. It is also possible to evaluate the trade-offs between the true positive rate and false positive rate across our various classification thresholds. This can be done using the receiver operating characteristic curve (ROC).

As shown in Figure 7, among the models, LightGBM and XGBoost attained near-perfect classification ability both yielding an Area Under the Curve (AUC) of 0.98. This is followed closely by

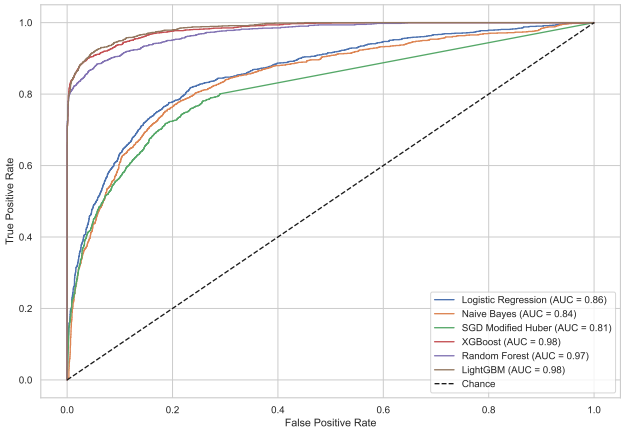

Fig. 7: ROC Curve Comparison across classifiers.

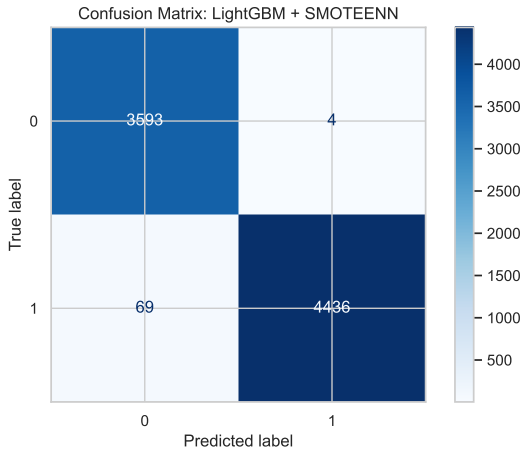

Fig. 8: Confusion Matrix of LightGBM + SMOTEENN.

Random Forest with an AUC of 0.97. These results emphasize the strength of ensemble-based methods especially boosting techniques in learning complex nonlinear relationships critical to credit default classification. Conversely, Logistic Regression and Naive Bayes achieved moderately high AUC scores. The SGD classifier with Modified Huber loss, however, registered the lowest AUC value which indicates a comparatively weaker classification capacity for highly nonlinear data. These comparative results validate the superior ability of ensemble learners to distinguish between creditworthy and non-creditworthy instances.

For our best performing model, Figure 8 shows the confusion matrix. This hybrid approach yielded an optimal balance between sensitivity and specificity. These results demonstrate an exceptionally low number of both false positives and false negatives, confirming that the model can reliably distinguish between default and non-default cases. The integration of SMOTEENN with LightGBM thus enhances the ability to generalize by addressing class imbalance while preserving decision boundary clarity.

6. DISCUSSION

This study set out to investigate whether machine learning models particularly ensemble techniques, can effectively predict credit default risk from structured loan data. The results clearly indicate that it is possible to build a model that distinguishes between borrowers who are likely to default and those who are not, with high accuracy and stability. In practical terms, our best-performing model was able to correctly classify over 96% of the test cases, with precision, recall, and KS statistic values exceeding 0.93. These metrics suggest not just strong predictive accuracy, but also a good balance between identifying defaulters and avoiding false alarms. The AUC of 0.98 further reinforces this, showing that the model consistently ranks risky borrowers higher than safe ones. In other words, the model doesn't just fit the data, it generalizes well across unseen examples within our cross-validation framework.

Several insights are generalizable beyond the specific dataset used. First, the integration of hybrid resampling with tree-based ensemble models is a robust and transferable strategy for handling imbalanced credit data. Many loan datasets contain far fewer defaults than non-defaults, and this imbalance often leads to biased or ineffective models. Our approach demonstrates that applying SMOTEENN before training an ensemble classifier significantly improves class separation and model calibration. This insight can benefit industrial credit scoring projects facing similar imbalance challenges.

Second, we show that interpretable, non-deep models like LightGBM and Random Forest can match or exceed the performance of more complex models. This is especially important for financial institutions operating under regulatory constraints, where model transparency is just as critical as accuracy. These models offer explanations for their decisions and thus enables credit analysts to trust and validate the outcomes. Further, the pipeline we developed—combining data preprocessing, stratified cross-validation, hyperparameter tuning, and diagnostic evaluation is modular and reproducible. It can be adapted to other types of credit products or datasets with minimal changes such as small business lending, peer-to-peer credit, or microfinance platforms.

While we used only one dataset, the consistency of our results across multiple model variants (e.g., with and without dimensionality reduction) and metrics suggests that the findings are robust. Future work could further test this generalizability by applying the same pipeline to new datasets, including different time periods or borrower populations.

7. CONCLUSIONS

This paper presented a structured approach to credit risk prediction using both linear and ensemble-based machine learning models on a financial institution loan-level data. Through a unified pipeline comprising preprocessing, feature transformation, model training, and validation, we comparatively evaluated different algorithms across standardized metrics including accuracy, precision, recall, and the KS statistic. Ensemble methods generally outperformed linear baselines under these evaluation metrics. We demonstrated the use of dimensionality reduction and hybrid resampling with ensemble classifiers which have receive limited attention as an interpretable alternative for handling class imbalance in prior literature. This approach presents a practical trade-off between model complexity and performance. The consistency of our result across multiple model variants and metrics suggests that the findings are robust and generalizable. In addition to comparative

benchmarking, we performed ROC analysis and confusion matrix evaluation to better understand predictive trade-offs and class-level performance, especially in minority class recall. Future work could extend this study by using formal optimization methods for hyperparameter tuning and investigating the stability of the proposed hybrid pipeline across different datasets and temporal windows.

8. REFERENCES

[1] Jung Min Ahn, Jungwook Kim, and Kyunghyun Kim. Ensemble machine learning of gradient boosting (xgboost, lightgbm, catboost) and attention-based cnn-lstm for harmful algal blooms forecasting. *Toxins*, 15(10), 2023.

[2] Luca Bitetto, Francesco Bonacina, Stefano Moramarco, Ugo Moscato, and Eva Pagani. A comparative analysis of supervised learning algorithms for credit scoring: evidence from european data. *Socio-Economic Planning Sciences*, 89:101535, 2023.

[3] L Breiman. Random forests. *Machine Learning*, 45:5–32, 10 2001.

[4] W. et al. Chang. Application of machine learning in credit risk prediction: a comprehensive review and evaluation. *Economic Modelling*, 102:105579, 2021.

[5] Aslı Demirgüç-Kunt, Enrica Detragiache, and Thierry Tressel. Banking on the principles: Compliance with basel core principles and bank soundness. *Journal of Financial Intermediation*, 17(4):511–542, 2008.

[6] Gazi Husain, Daniel Nasef, Rejath Jose, Jonathan Mayer, Molly Bekbolatova, Timothy Devine, and Milan Toma. Smote vs. smoteenn: A study on the performance of resampling algorithms for addressing class imbalance in regression models. *Algorithms*, 18(1), 2025.

[7] Evangelos Kalapodas and Mary Thomson. Credit risk assessment: A challenge for financial institutions. *IMA Journal of Management Mathematics*, 17, 01 2006.

[8] R. Kavitha, Rupa Shiva Dharshini V, and Priyadharshini M. Performance comparison of xgboost and lightgbm gradient boosting algorithms in predicting cervical cancer risk. In *2024 International Conference on Computing and Data Science (ICCDS)*, pages 1–6, 2024.

[9] Rakesh Kumar, Meeta Chaudhry, H. K. Patel, Navin Prakash, Abhinav Dogra, and Sunil Kumar. An analysis of ensemble machine learning algorithms for breast cancer detection: Performance and generalization. In *2024 11th International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 366–370, 2024.

[10] Stefan Lessmann, Bart Baesens, Hsin-Vonn Seow, and Lyn C. Thomas. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124–136, 2015.

[11] Yu Li. Credit risk prediction based on machine learning methods. In *2019 14th International Conference on Computer Science & Education (ICCSE)*, pages 1011–1013. IEEE, 2019.

[12] Yi Liu, Menglong Yang, Yudong Wang, Yongshan Li, Tiancheng Xiong, and Anzhe Li. Applying machine learning algorithms to predict default probability in the online credit market: Evidence from china. *International Review of Financial Analysis*, 79:101971, 2022.

[13] V. Z. Marmarelis, D. C. Shin, D. Song, R. E. Hampson, S. A. Deadwyler, and T. W. Berger. Dynamic nonlinear modeling of interactions between neuronal ensembles using principal dynamic modes. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3334–3337, Aug 2011.

[14] Mehrdad Naderi, Farzane Hashemi, Andriette Bekker, and Ahad Jamalizadeh. Modeling right-skewed financial data streams: A likelihood inference based on the generalized birnbaum–saunders mixture model. *Applied Mathematics and Computation*, 376:125109, 2020.

[15] David Powers and Ailab. Evaluation: From precision, recall and f-measure to roc, informedness, markedness correlation. *J. Mach. Learn. Technol*, 2:2229–3981, 01 2011.

[16] Modisane B. Seitshiro and Seshni Govender. Credit risk prediction with and without weights of evidence using quantitative learning models. *Cogent Economics & Finance*, 12(1):2338971, 2024.

[17] Vandana Sharma, Amit Singh, Ashendra Kumar Saxena, and Vineet Saxena. A logistic regression based credit risk assessment using woe bining and enhanced feature engineering approach anova and chi-square. In *2023 12th International Conference on System Modeling Advancement in Research Trends (SMART)*, pages 499–507, Dec 2023.

[18] Merve Veziroğlu, Erkan Eziroğlu, and İhsan Bucak. *Performance Comparison between Naive Bayes and Machine Learning Algorithms for News Classification*. 01 2024.

[19] Andrew Worster, Jerome Fan, and Afisi Ismaila. Understanding linear and logistic regression analyses. *CJEM*, 9:111–3, 03 2007.

[20] C. Yu, Y. Jin, Q. Xing, Y. Zhang, S. Guo, and S. Meng. Advanced user credit risk prediction model using lightgbm, xgboost and tabnet with smoteenn. *Risks*, 12(1):174, 2024.