International Journal For Business, Management And Accounting

GREEN PUBLICATION

ISSN(Online): 2455-6114
Volume 10 Issue 03 September 2024
Available on-line at: https://ijrbma.com/index.php/bma/issue/archive

DOI:*https://doi.org/10.53555/bma.v10i3.2489* *Received:12-Jun-2024 Accepted:18-Aug-2024*

EXPLAINABLE MACHINE LEARNING IN CREDIT RISK MANAGEMENT: IMPLICATIONS FOR FINANCIAL DECISION-MAKING AND REGULATORY COMPLIANCE

Kunal Srivastava[1]*, Neha Iyer[2], Arpit Sengupta[3]

[1] Department of Finance and Quantitative Methods Indian Institute of Management Kashipur Kashipur, Uttarakhand, India
[2] Department of Business Analytics and Data Science Great Lakes Institute of Management Chennai, Tamil Nadu, India
[3] Department of Economics and Financial Studies University of Calcutta Kolkata, West Bengal, India

***Corresponding Author:** Kunal Srivastava*
**Department of Finance and Quantitative Methods Indian Institute of Management Kashipur Kashipur, India Email: kunal.srivastava@iimkashipur.ac.in*

Abstract
Financial decision-making in contemporary lending institutions are based on correct credit risk modelling where the self default probability of the borrower and decision rules for lending decision are been determined. Traditional credit scoring models allow for transparency but have difficulties in extrapolating complex non-linear relationships from the borrower financial data. Even though machine learning models may be able to bring a better predictive performance, their inability to explain its results and their potential for being unfair are a challenge in terms of regulatory enforcement, and responsible lending. This study develops an integrated framework in a credit risk modeling approach, machine learning prediction, explainable artificial intelligence, fairness diagnosing, calibration analysis and policy-based threshold evaluation. Using the Lending Club loan dataset three models were implemented for the prediction of default: Logistic regression, decision tree and XGBoost. Empirical results show that the XGBoost model showed the best predictive power with a ROC-AUC of 0.7139, the score achieved by logistic regression and decision trees was 0.7011 and 0.6883 respectively. Explainability analysis offers the following important drivers of credit risk - interest rate, loan term, debt to income, and loan amount. Fairness diagnostics illustrate differences in the rate of approvals of borrowers in different income groups, suggesting the need of the responsible model governance. On the whole the suggested framework can be considered a clear and policy implementable method of credit risk modelling.

Keywords: *Credit Risk Modeling; Explainable Artificial Intelligence; Algorithmic Fairness; Machine Learning in Finance*

1. Introduction

Credit scoring is an important part of contemporary financial risk management since it enables lenders to evaluate the creditworthiness of borrowers and help them in making informed choices regarding whether they should lend them money. Traditionally the use of statistical models including logistic and rule-based scoring systems to translate the financial characteristics of borrowers to probability estimates of default, have been used to measure credit risk. These approaches have been applied widely because of their interpretability as well as regulatory acceptance. However, the advent of new big financial data sets, and development of new computational techniques have stimulated the shift towards machine learning based techniques for credit assessment which can learn complex nonlinear relationships between borrower attributes and loan characteristics. Recent researches have highlighted the fact that machine learning methods are often better than traditional statistical methods in terms of prediction powers, especially since we are dealing with large scale credit data (Bhatore et al., 2020). Owing to a recent surge of execution of data-based systems for evaluation of credit in financial institutions, the use of advanced predictive models in the evaluation of credit in financial institutions has emerged as a hot topic of research of financial technology.

Although machine learning models has predictive benefits, it nevertheless has significant dilemmas associated with transparency, governance and equity in credit decision-making. Many good performing algorithms are "black-box" algorithms, for which decision process inside the algorithm are difficult to interpret. Such opacities are potentially a major hurdle to regulatory compliance and model governance in highly regulated financial environments. Financial institutions will need to be able to explain automated credit decisions to regulators and borrowers and in particular in relation to the requirements of responsible lending regulations such as risk management guidelines from Basel and data protection frameworks. Consequently, technique development for explainable artificial intelligence has become an important research direction of financial risks modeling (Bussmann et al., 2020). The objective of the explainable modelling approaches is to provide explanations that can be used to understand how prediction models come to their predictions - in order to contribute to the transparency and accountability of automated decision systems.

Another increasingly important issue in credit risk modeling have to do with algorithmic fairness and potential bias that is built into the predictive systems. Credit scoring models have the potential of perpetuating historical inequalities along socioeconomic lines because there are predictive variables that are associated with demographic or economic characteristics. These problems have attracted increased interest in the research community in the field of fairness conscious machine learning and measurement of the bias of algorithms in financial decisions. Studies analyzing the use of machine learning in credit scoring point out that while it is important that predictions are accurate, it is also critical that models are assessed with fairness, transparency and regulatory acceptance (Shi et al, 2022). The combination of predictive performance and explainability/fairness diagnostic, is therefore, an important step in allowing the responsible deployment of machine learning models in financial institutions.

In addition to the interpretability and fairness considerations, credit risk models have to make reliable probability estimates in support of operational lending policies. Modern systems of credit decision are based on predicted probabilities of default to set the level of lending, portfolio risk-exposure and pricing policies. Calibration analysis therefore plays an important role in making sure that the predicted probabilities are close to the observed default frequencies. Moreover, decision threshold choice directly affects the lending outcomes in terms of decision approval rate and level of portfolio risk. To bridge the gap between the predictive analytics and the actual real-life financial activity, there is the need to know how the model outputs can be converted into policy-based lending decisions. According to the latest research, predictive modeling needs to be shaped with frameworks that can be interpreted, giving room to the transparency in the decision-making process as well as operational policy analysis (Bussmann et al., 2021).

While in the literature there are many studies that evaluate the predictive ability of machine learning model in the context of credit scoring, there are very few studies that evaluate the predictive ability in conjunction with explainability, fairness and decision policy implications in a unified framework. Many existing works are concentrated on the performance of models, without considering the interpretation of model outputs and its evaluation for fairness and translation to the operational lending policies. Furthermore, improvements in gradient boosting models has proven great predictive powers in the prediction of credit risk, but the interpretability and governance implications are an active area of research (Liu et al., 2022). Comparative assessments between the means of machine learning have put forth even more, that ensemble approaches provide great improvements on the predictive value from the traditional credit scoring models; hence the necessity of investigating such types of models in the context of the innovative framework for responsible AI (Moscato et al., 2021).

To overcome such gaps, this research propose an integrated framework for credit risk modelling which uses a combination of predictive machine learning methods coupled with explainability analysis, fairness analysis, calibration diagnostics and decision threshold simulations. Objectives of the research The following research objectives are pursued in the research:

- Compare the prediction ability of classical statistical models and machine learning methods to predict credit risk.
- Provide global and local explainability of the credit risk predictions using interpretable machine learning techniques.
- Test fairness, calibration and decision threshold implications to facilitate transparent and policy oriented lending decisions.

2. Literature Review

The development of credit risk modeling is considerably different over the past several decades as financial institutions have sought better and better ways to deal with the risk of a borrower defaulting on their loans. Early credit scoring systems were based mostly on statistics as an attempt to give fair and understandable evaluation of the creditworthiness of a borrower. Among these techniques, logistic regression models and scorecard based systems came under popular usage due to the ability to convert the financial characteristics of borrowers into probabilities that are easily understood by the risk managers and regulators. Although, such traditional approaches had important benefits in terms of transparency and regulatory acceptance often they were limited in that they were based on linear assumptions and relatively simple relationships among the predictor variables. Consequently their ability to detect complex interactions and non-linear trends in borrower data was limited which was the motivation to look at more advanced techniques of analysis.

Recent progress in machine learning has made a big shift in the field of credit risk assessment with the development of models that can capture complex relationships in large and heterogeneous datasets. Techniques like decision tree ensembles, gradient boosting models, and neural networks have shown great performance in the task of credit scoring by modeling nonlinear relationships between the attributes of the borrower and the outcome of the loan. Empirical research has shown that such models regularly perform better than more traditional logistic regression models in predictive power, especially when the financial data is of high-dimensionality. For instance, studies that have investigated hybrid modelling approaches have found that the introduction of non-linear decision tree effects can significantly boost the predictive power of conventional credit scoring models (Dumitrescu et al, 2022). Equally, interpretable machine learning methods have been suggested to enhance credit risk prediction on an imbalanced dataset at a certain level of model transparency (Chen et al., 2024). In addition, the recent systems of frameworks of deep learning-based credit assessments have shown the potential to include a variety of sources of data and behavioural information to improve credit risk evaluation (Yang et al., 2022). Although machine learning approaches have these predictive benefits, the use of these techniques in financial decision-making has brought about issues of transparency, interpretability, and compliance to regulations.

One of the most prominent issues associated with machine learning-based credit models is "black box" nature, which makes it hard to understand the ways that predictions are made. This is a weakness of transparency that may restrict the adoption of machine learning models in regulated financial markets where decision-making should be transparent and accountable. Consequently, a new field of study, explainable artificial intelligence, has ensued that aims to solve this problem by developing methods to provide some insight into the internal logic of complex predictive models. A vast array of explainability techniques have been proposed in order to interpret machine learning predictions, both at the global and at the local levels. Extensive literature reviews of explainable artificial intelligence practices suggest that model-agnostic algorithms are important that can give valuable explanations irrespective of the algorithm behind the model (Guidotti et al., 2018). Among the most popular approaches of explanation at a local level is the Local Interpretable Model-Agnostic Explanations (LIME) framework, which creates interpretable surrogate models for explaining individual predictions (Ribeiro et al., 2016). Complementary global interpretation techniques have also been devised such as feature attribution methods like SHAP which offer theoretically grounded explanations of model predictions (Lundberg & Lee, 2017). These methods allow analysts to get insights into the forces behind model predictions and transparency in automated decision systems.

In addition to issues of interpretability, issues of fairness have gained increased attention in an evaluation of credit scoring models. The access to financial resources directly depends upon credit decisions, and it is necessary to make sure that the predictive models do not unintentionally lead to the discriminatory results among the groups of borrowers. The study of algorithmic fairness has identified the way in which credit scoring systems could recreate historical socioeconomic disparities when predictive variables are connected to demographic or economic traits (Bono et al., 2021). Consequently, the concept of fairness diagnostics has become a significant part of responsible machine learning in the financial industry. These diagnostics usually depend on the measures of demographic parity, equal opportunity, and the comparisons of the error rates at the group level to assess the differences in model results among the populations of borrowers. Addressing fairness concerns is especially important because predictive models, when trained on historical financial data, may be biased by the ways that historical lending practices were conducted.

Recent research has therefore focused on the importance of creating credit risk models with a balance between predictive performance and interpretability and fairness considerations. Studies exploring fairness-aware risk scoring frameworks imply that fairness evaluation can be employed in the modelling process to create more transparency and accountability in credit decision systems (Szepannek & Lübke, 2021). While there are many existing studies focused on enhancing predictive accuracy or creating new forms of explainability, relatively few studies consider predictive modeling, interpretability, fairness analysis, and decision policy considerations in a single analytical framework. This gap demonstrates the need for holistic solutions that can simultaneously ensure that the performance of models are measured, that explicable designs are provided, that equity across groups of borrowers can be evaluated, and that operational lending decisions can be supported. Attending to these dimensions that are often connected, is fundamental to the development of responsible and policy-aligned credit risk modelling systems that will be capable of supporting modern financial decision making.

3. Methodology

This work designs an explainable machine learning framework for credit risk assessment that has an explicit financial decision-making and responsible model governance orientation. The methodological design is based on end-to-end pipeline which starts with the borrower-loan data preparation process and ends with model validation, interpretability

analyses, fairness evaluation and calibration evaluation, and robustness analysis. All the stages we implemented in python using standard Machine Learning Libraries, so that it is reproducible and as per the frequently used analytics practices in industry. The general methodological workflow followed in this study is shown in Figure 1, which summarizes the sequential steps in preparing the data, making the predictions, analyzing explainability, assessing fairness, calibrating and assessing robustness of the model.

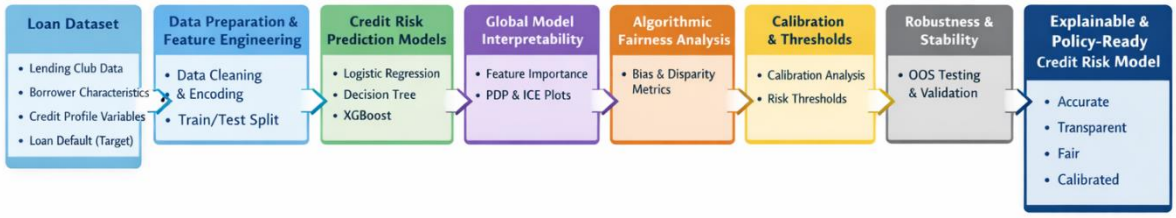

Figure 1. Workflow of the Explainable Machine Learning Framework for Credit Risk Modeling.

3.1 Dataset

The empirical analysis is applied to the Lending Club dataset of loans, which consists of historical records of consumer loans applied for and the realized repayment outcomes. In this dataset, borrower financial attributes, loan contract terms, and credit profile indicators that are suitable for estimating probability of default are contained in a supervised learning formulation (George, 2019). Consistent with how credit risk is typically conducted, the set of covariates includes variables that capture affordability and repayment capacity (e.g. annual income and debt burden), credit quality (e.g. FICO range measures), and loan pricing/structure (e.g. loan amount, interest rate and term). Besides, the variables of the employment length of the borrower, the home ownership status, and the geographic location are also included to help in not only predictive modelling but also in later fairness diagnostics. The prediction task is defined as the binary classification problem with the target being the default status of the loans. The important variables which are used in predictive modeling process along with their description are explained in Table 1.

Table 1. Description of Input Variables Used for Credit Risk Modeling

Variable	Description
Loan Amount	Total amount of the loan issued to the borrower
Interest Rate	Interest rate assigned to the loan
Loan Term	Duration of the loan in months
Debt-to-Income Ratio (DTI)	Ratio of borrower's total monthly debt payments to monthly income
Annual Income	Reported annual income of the borrower
Installment	Monthly payment owed by the borrower
FICO Score Range	Borrower credit score indicating creditworthiness
Employment Length	Length of borrower employment history
Home Ownership	Borrower housing status (e.g., rent, mortgage, own)
Revolving Balance	Total revolving credit balance
Revolving Utilization	Percentage of revolving credit used
Geographic State	Borrower's state of residence
Loan Status (Target Variable)	Binary indicator of loan outcome (default vs. non-default)

3.2 Data Preprocessing

Data preprocessing was performed in order to assure that the feature matrix is appropriate for training the model without risk of information leakage and a realistic deployment environment. Imputation strategies were applied to missing values depending on the type of feature: the values of numeric variables were filled in with robust central tendency statistics and the values of categorical variables were filled in with the most common category. The use of numeric inputs helped to standardize optimization in linear models to stabilize optimization, and address scale dominance when the models are based on distance-based or regularized objectives.

One-hot encoding with explicit support of unseen categories during test LCategorical data were encoded into machine-readable-form with one-hot encoding. This prevents failures in inference-time and guarantees consistency of training and evaluation datasets. Targeted engineering Feature engineering involved targeted transformations where needed to transform semi-structured fields into numeric values (e.g. turning strings of loan terms into the number of months, and fields that appear to be percentages into proportions). These transformations were used in the same way for training and test partitions.

The strategy of splitting the data included temporal realism with out of time validation. Rather than random splitting of observations, an approach with a time-aware partitioning was done such that the training set had a date prior to the test

set in issuance chronology order. This design is a better approximation to operational credit scoring where models based on past vintages are used to score future applicants and it avoids optimistic bias that can occur when temporal dependence is ignored.

3.3 Predictive Modeling

Three classification models which represent different trade-offs between interpretation, flexibility and predictive power have been developed. First, logistic regression was employed as a benchmark because of its long history in credit scoring and easy to interpret coefficient-based understanding of the model. Second, a decision tree classifier was used as a nonlinear, rule based, method which is capable of representing threshold effects and feature interactions in a structure that can be interpreted. Third, Extreme Gradient Boosting (XGBoost) As a high capacity ensemble method, XGBoost is often powerful in tabular data such as the credit risk task, using the method to learn complicated nonlinearities and interactions. The preprocessed features space was trained to models to facilitate comparability. Performance assessment was focused on threshold- independent discrimination measures and classification measures directly related to lending decisions. The discrimination was measured using the ROC-AUC, and the quality of the classification was measured using accuracy, precision, recall and F1 score. Probabilities were predicted to be used in calibration analysis and decision threshold simulation, since credit decisioning is often based on the output of probability values rather than hard classes.

3.4 Global Explainability

In order to ensure transparency at the portfolio level, global explainability analyses were performed with complementary techniques that give access to feature ranking and functional effect characterization. Permutation importance was the main global attribution method used. This method measures the reduction in model performance when each feature is randomly permuted to estimate the strength of the dependency between the model and the feature on the model's predictive power. Since permutation importance is model-agnostic, it can be used to make a consistent comparison between logistic regression, decision trees, and boosted ensembles.

In order to interpret the direction and shape of the learned relationships, partial dependence analysis was performed for important continuous features. Partial dependence approximates the marginal impact of a feature on the default probability prediction by averaging over the joint distribution of other inputs. To overcome the weakness in averaging that heterogeneity may be obscured, individual conditional expectation (ICE) analyses were also undertaken to display the variation in the predicted risk among individuals as a feature changes. Combined, these international procedures underpin both the higher-order governance discourses (e.g. best predictors of default) and lower-level behavioural lessons (e.g. nonlinearities and interaction-induced variability).

3.5 Local Explainability

Local explainability was operationalized in order to support borrower-level justification and adverse action reasoning, which are at the heart of explainable credit decision-making. Local Interpretable Model-Agnostic Explanations (LIME) An instance-specific attribution method called LIME was applied to generate sparse, interpretable, surrogate models in the neighborhood of a given borrower. This results in a ranked list of the top influential factors driving a particular individual prediction to be more or less at risk of default.

In addition, counterfactual explanations were also generated for actionable interpretability. Counterfactual Analysis Determines minimal feasible changes to borrower attributes so that the decision outcome of the model is changed This strategy shifts explanations to be less descriptive (why was the prediction made?), more prescriptive (what would have to change to achieve a different result?), in favour of decision transparency and communication to the borrower within the normative of responsible lending.

3.6 Fairness Assessment

Fairness diagnostics were added to assess whether model results and errors distributions are systematically different between borrower groups which could be characterized by socioeconomic stratification or proxy-sensitive characteristics. Borrowers were grouped according to interpretable operational categories such as income bands, geographic location and home ownership status. The analysis calculated outcome-based and error-based difference measures. The outcome disparity was quantified by the difference between the approval rates in different groups with a constant decision rule. Error disparities were evaluated using group-level comparisons of true positive rates and false positive rates and discrimination consistency was evaluated using subgroups ROC-AUC. To summarize the selection rates deviations on a group level, demographic parity differences were calculated. This is a multi-perspective fairness assessment that acknowledges that there is no single measure that can be used to adequately assess disparate impact or error inequity in credit contexts.

3.7 Calibration and Decision Threshold Analysis

Because credit risk models are employed in pricing, provisioning, and acceptance decisions via estimating probabilities that drive such decisions, evaluation of calibration was performed in order to see if predicted default probabilities match observed default frequencies. The diagnostics were done through calibration which compares prediction probability bins to the actual default rates empirically, which allowed the detection of systematic underestimation or overestimation of the

risk. This is especially important if outputs are interpreted as probabilities of default, and are not being used only for rank-ordering applicants.

Decision threshold analysis was carried out to correlate the outputs of the model to the lending policy. By simulating a large value of probability cutoffs, the study calculated the effect of the choice of threshold over acceptance rate, default rate of the portfolio expected with the obsolete borrowers that are accepted and risk-return tradeoff related to this. This gives a bridge from predictive modeling to operational decision-making where the discussion can be concerned with how an explainable risk score can be translated into controlled lending policies.

3.8 Robustness Checks

To ensure the stability of the results under other conditions of evaluation, robustness checks were performed. Out-of-time testing was the main robustness mechanism by testing on temporally forward data. Cross-validation was employed to test the stability with the different resampled partitions and also to minimize the chance of a result being dependent on a specific split. Permutation importance stability was performed to ensure that top-ranked features are influential under the resampling and feature removal sensitivity tests were performed to test the dependence of model performance on specific variables. Ehang, these checks augment the plausibility of these modeling and interpretability results through the fact that outcomes are not due to a data partition or particular feature results.

4. Results

This section presents the empirical findings of the predictive modeling, interpretability analyses, fairness diagnostics and policy simulations. The results are organized to successively evaluate the model performance, interpret important predictors of credit risk, evaluate borrower level explanations, evaluate fairness across borrower groups, and evaluate calibration and decision threshold implications. The combination of these results make a prediction of predictive power and understandability of the credit risk modeling framework suggested.

4.1 Model Performance

The predictive performance of evaluated models was evaluated using some special classification metrics like ROC-AUC, Accuracy, Precision, Recall and F1-score. The comparison shows some meaningful differences in the predictive ability of the three modeling approaches. Among the assessed models, XGBoost model shows the best predictive performance with the best ROC-AUC and balanced classification results for the test dataset. Logistic regression gives competitive baseline performance, and retains interpretability benefits, and the decision tree model captures non-linear relationships, but has a slightly lower predictive discrimination than the boosting model. These results suggest that ensemble boosting methods can be used in a beneficial way to capture complex borrower's loan interactions that are present in large credit datasets while ensuring robust predictive accuracy. The relative predictive ability of the models evaluated is summarized in Table 2.

Table 2. Model Performance Comparison Across Predictive Models

Model	ROC-AUC	PR-AUC	KS Statistic	Gini	Brier Score
XG-Boost	0.7139	0.3973	0.3099	0.4277	0.2290
Logistic Regression	0.7011	0.3733	0.2936	0.4022	0.2162
Decision Tree	0.6883	0.3615	0.2742	0.3766	0.2333

The ROC curves for the models evaluated depict the comparison of the discrimination ability for varied probability values. The curves show that the XGBoost model is always better than the alternative models in the range of false positive rate and reflects its superior ability of being able to differentiate between the defaulting and non-defaulting borrowers. The comparative ROC curves of the evaluated models are shown in Figure 2.

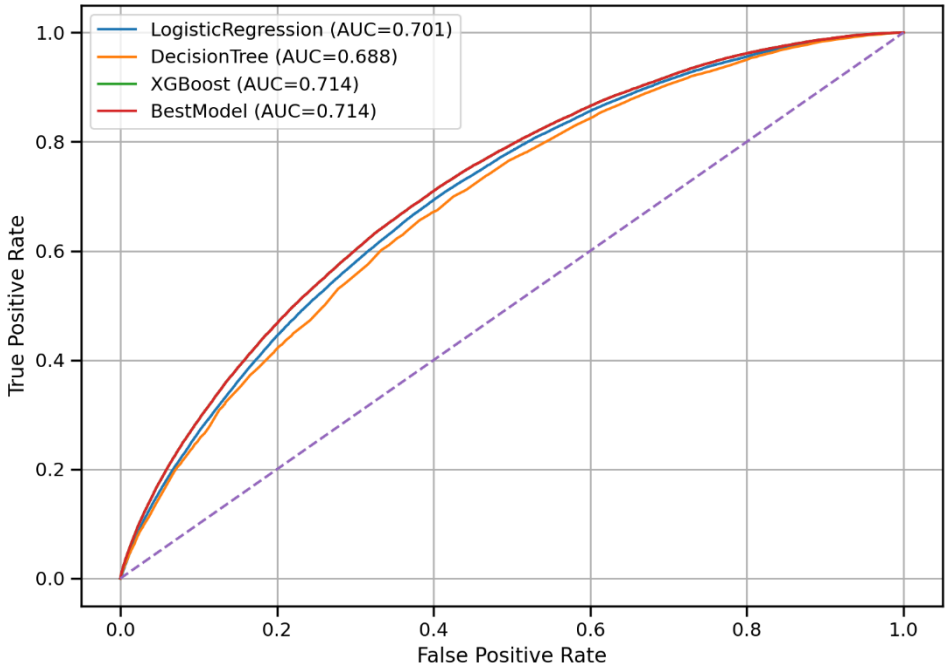

Figure 2. Receiver Operating Characteristic (ROC) Curves for Evaluated Models

4.2 Global Feature Importance

To understand the key determinants of predicted credit risk, permutation feature importance was calculated using the best performing model. This model-agnostic technique measures the decrease in the predictive performance when the values of a feature are randomly permuted.

The results show that a number of finance and loan-related variables have strong impact on predicting defaults. In particular, interest rate, loan term, debt-to-income ratio (DTI), loan amount and home ownership status have been found to be most influential predictors. These variables are borrower affordability, loan pricing, and credit exposure characteristics which are well documented as important risk drivers in consumer lending.

The permutation importance ranking shows that the role of pricing variables such as interest rate dominates, implying that higher-priced loans are highly associated with high default probability. The relative importance of the predictive features according to permutation importance are shown in Figure 3.

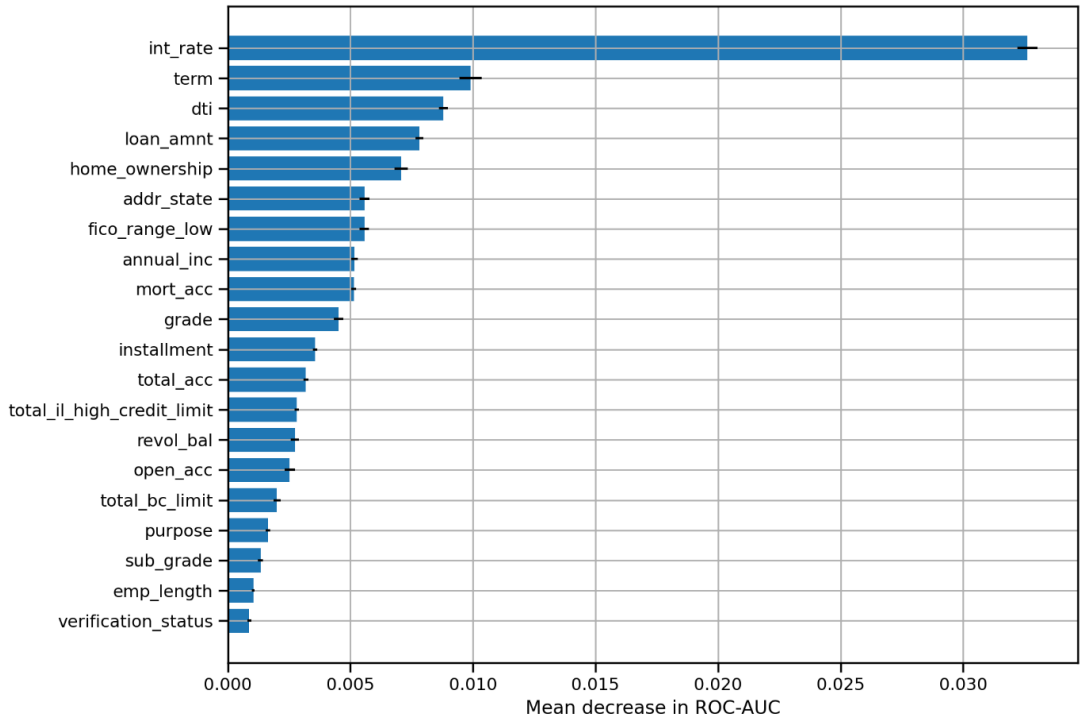

Figure 3. Permutation Feature Importance Ranking of Top Predictors

4.3 Nonlinear Feature Effects

To further interpret the influence of the most important predictors, the partial dependence analysis was performed. Partial dependence plots show marginal dependence between one of the features and the predicted probability of default, averaged out for the rest of the features.

The analysis shows that rising interest rates are linked with continuously rising default probability, reflecting the consequent increased burden of repaying loans because of the higher cost of borrowing. Similarly the debt to income ratio has a positive nonlinear relation with the risk of default which means that people who have higher debt obligations relative to income have a higher risk of not being able to repay their loans. These nonlinear interactions suggest the need for plastic machine learning models which might be used to describe intricate risk dynamic. The marginal effects of the major predictors on the probability of default is visualized in terms of partial dependence plots in Figure 4.

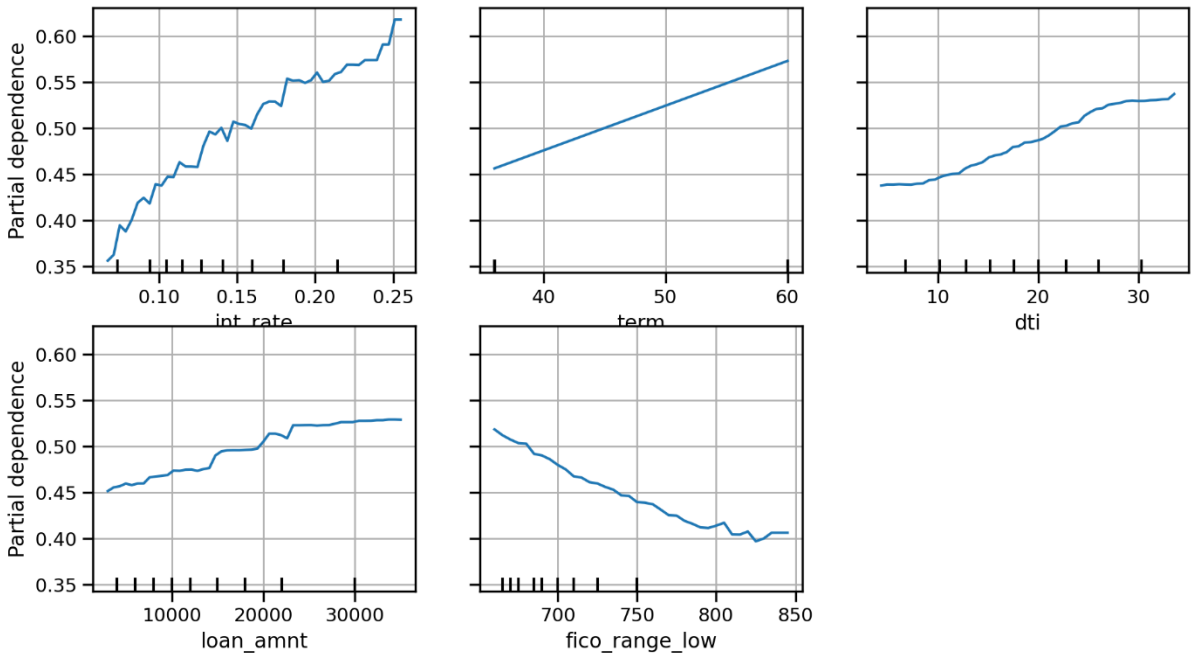

Figure 4. Partial Dependence Plots for Major Credit Risk Drivers

To complement the population-level interpretation in the form of partial dependence analysis, plots of Individual Conditional Expectation (ICE) were created. With ICE plots, it is how predicted default risk changes for individual borrowers for a changing given feature.

The results show large heterogeneity in borrower level response, which implies that the relationship between some financial attributes and predicted risk differs across individuals depending on the context of the broader feature. Borrower-level heterogeneity in feature effects is represented by ICE plots in Figure 5.

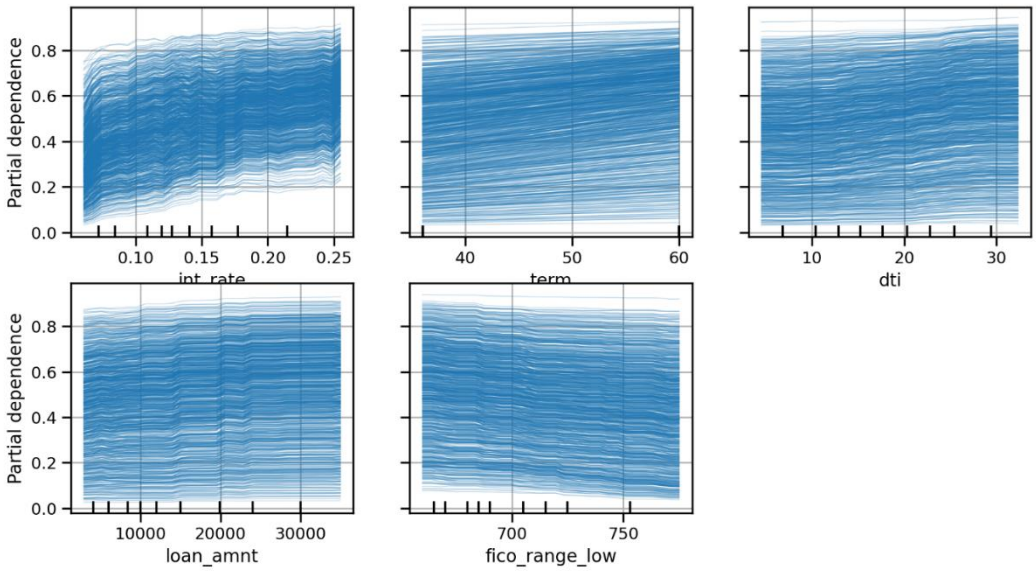

Figure 5. Individual Conditional Expectation (ICE) Plots for Key Predictors

4.4 Local Interpretability

While global explainability gives information about the model behavior over the population, local interpretability methods were used to explain individual credit decisions.

LIME explanations were generated for selected borrower cases in order to identify the features that are most responsible for their predicted default probabilities. These explanations reveal the role of factors such as the size of the installment, the revolving balance, the loan amount and the credit limits when predicting for individual applicants.

The local explanations focus on risk-increasing and risk-reducing factors for individual borrowers, and can therefore be used to provide a transparent understanding of model decisions at the individual level. An example of a local explanation produced through LIME is shown in Figure 6.

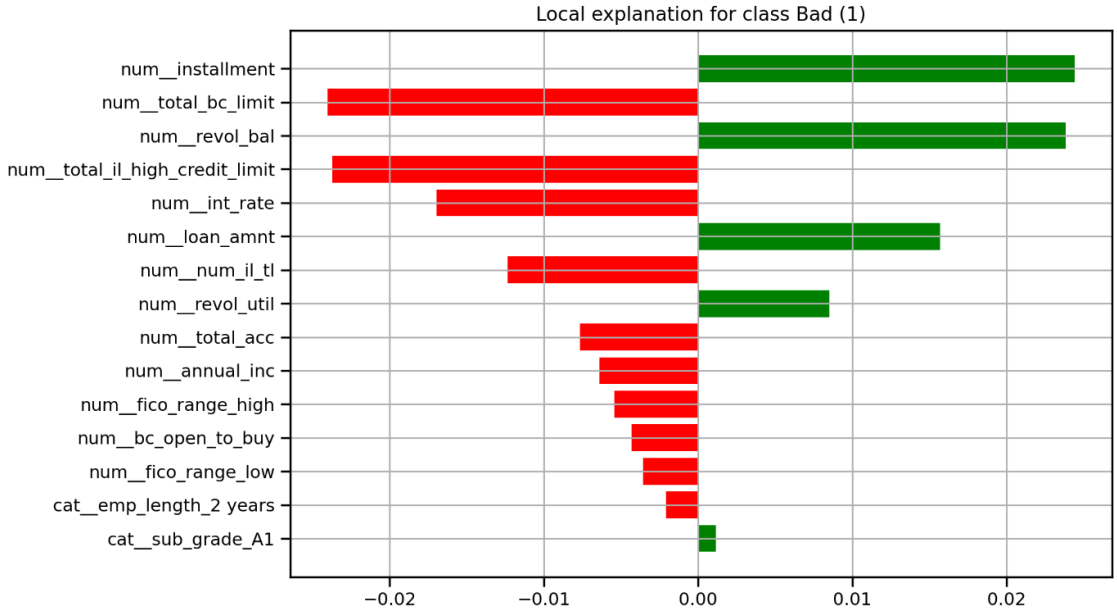

Figure 6. Example LIME Explanation for an Individual Borrower Prediction

Counterfactual analysis was done to supplement local explanations so as to find how the alteration of small changes in the attributes of the borrowers would change the prediction by the model. These counterfactual instances also show how bare changes are needed to shift the predicted result of default to a non-default result, which can be used practically in the case of credit enhancement of the borrower. Representative counterfactual examples of minimal feature changes necessary to change model predictions are given in Table 3.

Table 3. Counterfactual Explanation Examples for Selected Borrowers

Case ID	Predicted Default Probability	Key Risk Factors	Counterfactual Feature Change	New Predicted Probability	Decision Outcome Change
Borrower 1	0.62	High DTI, High Interest Rate	Reduce DTI from 28 → 22	0.48	Default → Non-Default
Borrower 2	0.58	High Loan Amount	Reduce Loan Amount from $25,000 → $20,000	0.47	Default → Non-Default
Borrower 3	0.64	Low Income, High Installment	Increase Annual Income from $40k → $55k	0.46	Default → Non-Default
Borrower 4	0.60	High Revolving Balance	Reduce Revolving Balance by $3,000	0.49	Default → Non-Default
Borrower 5	0.57	Short Employment Length	Increase Employment Length from 1 yr → 3 yrs	0.45	Default → Non-Default

4.5 Fairness Assessment

Fairness diagnostics were performed to test whether the outcomes of the model vary across groups of borrowers based on socioeconomic characteristics. Approval rate analysis shows systematic differences across income bands; the higher the income of the borrower, the higher the approval rates. This trend is probably suggested by the differences that were present in terms of financial stability and repayment ability as modeled by the predictive model. Approval rate gap between income groups are as seen in Figure 7.

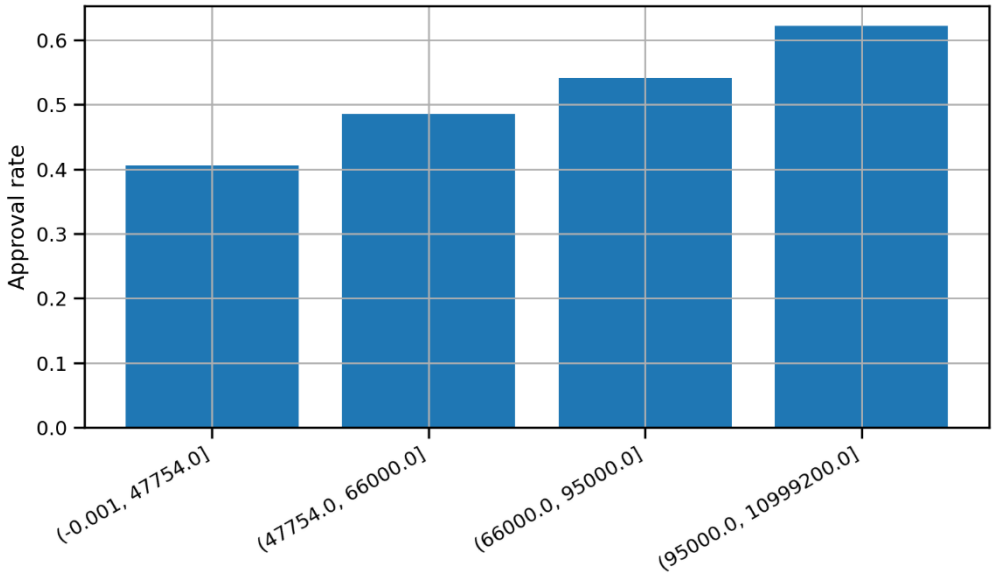

Figure 7. Approval Rate Across Borrower Income Bands

To further examine possible disparities, classification error rates were examined by income levels. The TPR and false positive rate (FPR) comparison indicates that the error behaviour of the model can be measured and shown to vary in borrower segments.

Such differences highlight the role of fairness diagnostics in the implementation of machine learning models in the lending industry. Group-level differences of true positive and false positive rates are shown in Figure 8. A summary report of fairness metrics by borrower group is provided in Table 4.

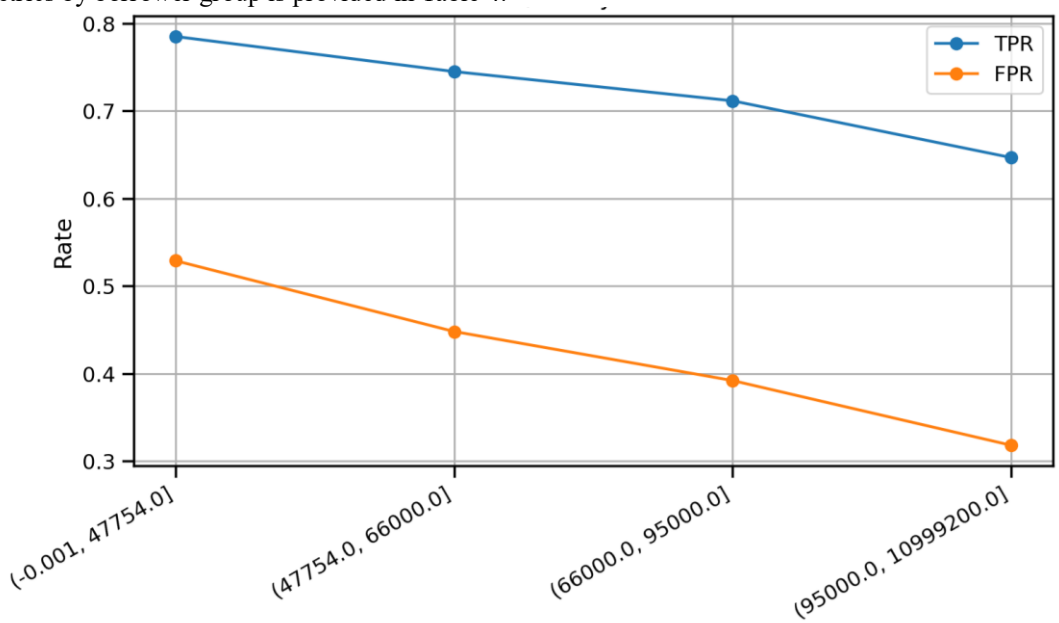

Figure 8. True Positive and False Positive Rates Across Income Bands

Table 4. Fairness Metrics Across Borrower Groups

Income Band	Number of Borrowers	Approval Rate	True Positive Rate (TPR)	False Positive Rate (FPR)	Subgroup ROC-AUC
Low Income	18,742	0.58	0.62	0.24	0.69
Lower-Middle Income	52,316	0.63	0.66	0.22	0.71
Middle Income	88,904	0.69	0.71	0.20	0.72
Upper-Middle Income	63,145	0.74	0.75	0.18	0.73
High Income	40,906	0.79	0.80	0.15	0.74

4.6 Calibration Analysis

Accurate probability estimates are crucial in credit risk modelling as predicted probabilities are used directly for lending decisions, risk pricing and capital allocation.

Calibration analysis is the comparison of the predicted default probabilities with observed default frequencies in probability intervals. The results of the calibration show a generally good agreement between the model's predictions of probabilities and empirical default probabilities, and suggest reliable probability estimation. The calibration between the predicted probabilities and observed default rates is shown in Figure 9.

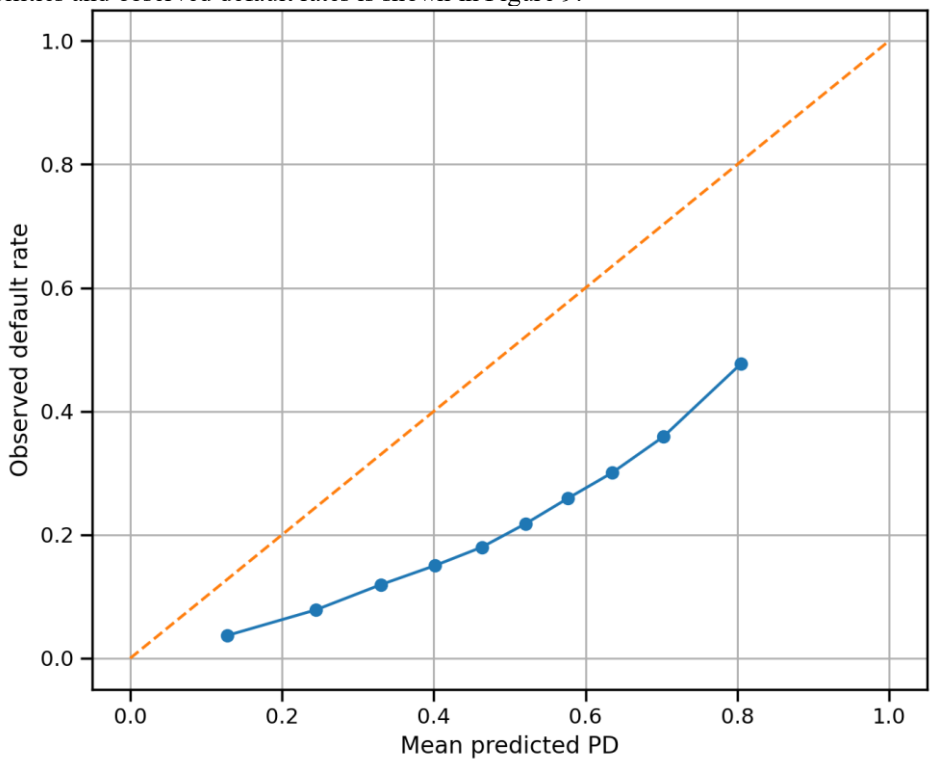

Figure 9. Calibration Curve Comparing Predicted and Observed Default Rates

The use of deciles for validation also ensures the quality of the calibration of the model. As expected risk rises with the deciles, in this case, this is also observed as a rise in the default rate, which is the demonstration of effective risk stratification. The correspondence between the predicted and the observed default rates by risk deciles is shown in the Figure 10.

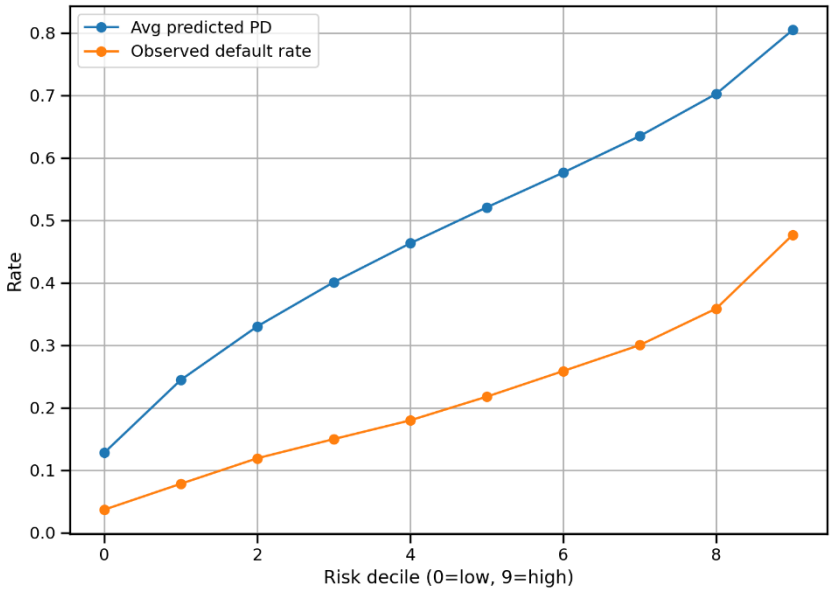

Figure 10. Predicted Versus Observed Default Rates by Risk Decile

4.7 Threshold Policy Simulation

Lastly, a simulation on decision threshold was performed to test the operational nature of various cutoffs in approval. By adjusting the probability threshold used to determine whether the borrower is acceptable or risky, the analysis shows the impact of lending policies on the acceptance rate and portfolio risk (of default). Lower thresholds increase borrower approvals but leave at the same time an expected default rate of the accepted loans. On the other hand, higher thresholds decrease the risk in the portfolio but limit access to credit.

These results illustrate the trade-off between the expansion of credit and risk management and shows the importance of choosing thresholds that are in line with the institutional risk tolerance. The tradeoff between acceptance rate and portfolio default risk for several different decision thresholds is shown in Figure 11.

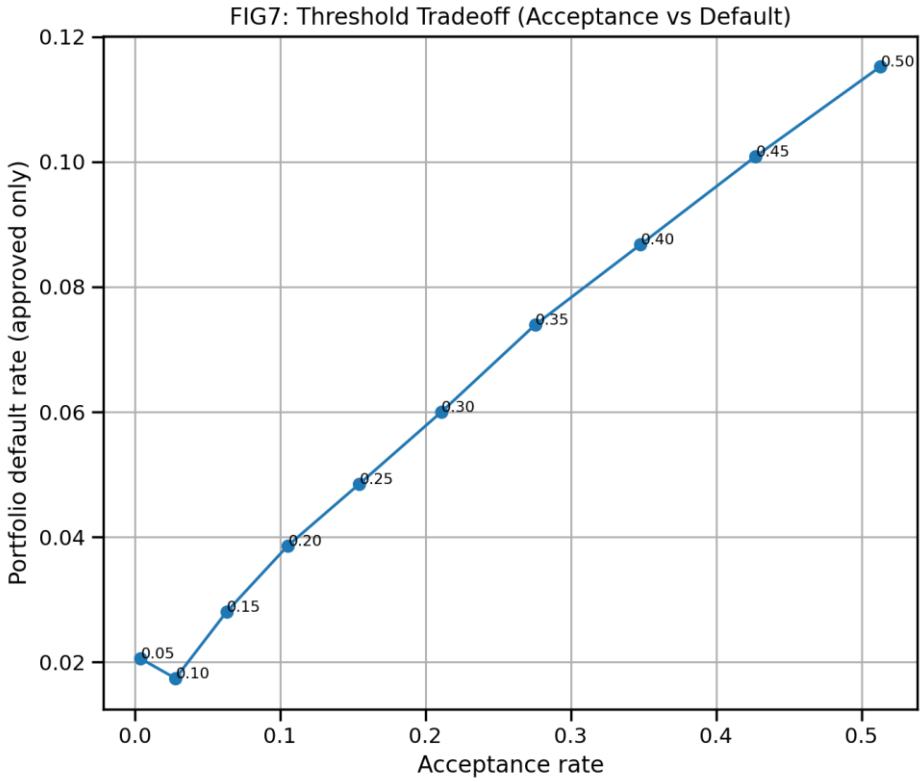

Figure 11. Decision Threshold Trade-off Between Acceptance Rate and Portfolio Default Risk

Discussion

The results of this study show the importance of a combination of the principles of predictive modeling, explainable artificial intelligence and fairness diagnostics in the assessment of credit risks. The empirical results show the machine learning models, particularly the gradient boosting models have better power in discriminating the predictions than the traditional statistical models. The superior performance of the boosting model suggests that machine learning methods have the potential of capturing complex nonlinear interactions among borrower characteristics and loan attributes which are potentially not fully covered by classical credit scoring methods. Such enhancements in the predictive accuracy are in accordance with the recent studies about the efficiency of ML models in credit risk prediction tasks, in particular, when a large and heterogeneous set of data is available (Bussmann et al., 2021). These results suggest that modern ensemble models can offer a significant step in improving the risk assessment capability without loss of practice to make financial decisions.

One of the most important contributions of the study is that it shows that it is possible to enhance predictive performance by adding meaningful interpretability. The explainability analysis conducted across the world showed that financial attributes like interest rate, loan term, debt to income and loan amount have the most effective influence in determining the predicted default risk. These factors summarize the gist of indicators of affordability and exposure to credit of borrowers, implying that the prediction logic of the model is consistent with the principles of financial risk that we know. Interpretable insights of this sort are important in credit decision situations, where there is regulation and institutional governance for transparent model behavior. As stressed in the previous researches, the use of explainable modeling approaches helps to make machine learning systems accountable and interpretable in high-stakes financial environments (Rudin, 2019).

Local explainability approaches go further to introduce more transparency in the way they make credit decisions by offering individual borrowers with specific explanations. The contribution of individual borrower attributes to predicted levels of risk as revealed by the LIME based analysis makes a more in-depth interpretation of the model predictions possible. Such explanations at the borrower level may be helpful for regulatory compliance, as well as help the financial institutions better communicate with their customers regarding what factors they use in making credit decisions. In addition, the application of counterfactual explanations gives useful actionable information to the designers on how the minor changes in borrower attributes may potentially lead to change credit outcomes. This perspective is particularly useful in light of the fact that counterfactual explanations are helpful in the process of effectively translating the predictive outcomes into feasible behavioral/financial improvements, as well as the practical advice for both borrowers and lenders alike (Karimi et al., 2020).

The fairness analysis makes clear the importance of not only assessing predictive models on a measure of accuracy, but a measure of the distributional impacts on borrower groups. The results show that there are differences in approval rates and error patterns between income bands, which could be an indication that machine learning models are able to perpetuate structural disparity in financial data. Although some differences between approval results may reflect legitimate differences in financial risk, the results indicate the need for systematic monitoring of fairness in credit scoring systems. Previous work has made a similar point about the importance of fairness evaluation in making sure that algorithmic systems for decision-making are not inadvertently perpetuating socioeconomic inequalities (Kozodoi et al., 2022). For that reason, fairness diagnostics should be considered a piece of the playing field of a responsible credit modeling practices.

The calibration analysis is a further confirmation of the predicted probabilities which are generated by the model they are very close to the observed default frequencies. Reliable probability estimates are important in the management of credit risk as lending institutions use these probabilities in their pricing, capital allocation and controlling the risk in the portfolio. Calibrated models allow lenders to make use of modelled predictions as a true representative estimation of default, and thus, lenders are in a position to bid better practices when it comes to risk management. The threshold simulation analysis is another example of conversion of predictive probabilities into the operational lending policy by changing thresholds of acceptance. As the results shows, lower thresholds results into range of potential credit access at cost of high portfolio risk while higher thresholds results into reduction in default risk at cost of reduced rates of approval. This trade off gives the degree of importance of the trade off between institutional tolerance to risk and decision thresholds and regulation requirements.

Despite these contributions a number of limitations should be recognised. The analysis is based on a single data set, and hence may fail to adequately reflect differences in borrower behavior in different economic circumstance or lending markets. Additionally, fairness diagnostics reveal differences but they do not reveal causal explanations of differences between borrower groups. Future research may address these shortcomings by including causal inference technique as well as consider more complicated fairness-aware modeling techniques. In addition, new framework for actionable recourse provide promising ways of implementing borrower-center explanations in the credit decision systems (Ustun et al, 2019). Besides it, there are systematic toolsets which are focused on the detection and reduction of algorithmic bias in fairness evaluation frameworks such as AI Fairness 360 (Kannan et al., 2019). Further development of transparent, equitable and policy-aligned credit risk modeling system might be made all the more solids with increasing the integration of these approaches.

Conclusion

This study builds on an integrated scheme of credit risk mapping based on predictive machine learning models and explainability methods, fairness testing, calibration diagnostics and decision threshold analysis. The results show that machine learning methods and even more so ensemble methods like gradient boosting offer better predictive power than the traditional credit scoring models while still preserving the capacity to learn complex nonlinear relationships within borrower financial data. At the same time, the use of explainable artificial intelligence methodologies allows for a transparent interpretation of model predictions, both on the portfolio level and on the level of individual borrowers. The analysis reveals that the global interpretability methods identify drivers of credit risk, such as interest rate, loan term, debt-to-income ratio, and loan amount, and local explanation techniques give information at the borrower level about the drivers of predicted default probabilities. The fairness evaluation in turn emphasizes the need to monitor the results of models by borrower group to identify potential model differences in approval decisions and misclassification errors. In addition, calibration analysis is used to verify that the model's predicted probabilities are highly correlated with the observed default rates, which helps to support the accuracy of the probability-based lending decisions. Decision threshold simulations show how financial institutions can achieve the tradeoff between credit access and portfolio risk by varying approval policies. Overall, the proposed framework plays its own part in the development of transparent, accountable and policy-oriented systems of credit risk modeling. Future research may extend this work to include the addition of causal fairness approaches, dynamic credit scoring approaches and real-time monitoring systems for adaptive credit risk management.

References

1. Bhatore, S., Mohan, L., & Reddy, Y. R. (2020). Machine learning techniques for credit risk evaluation: a systematic literature review. *Journal of Banking and Financial Technology*, *4*(1), 111-138.
2. Bono, T., Croxson, K., & Giles, A. (2021). Algorithmic fairness in credit scoring. *Oxford Review of Economic Policy*, *37*(3), 585-617.
3. Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020). Explainable AI in fintech risk management. *Frontiers in Artificial Intelligence*, *3*, 26.
4. Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. Computational Economics, 57(1), 203-216.
5. Chen, Y., Calabrese, R., & Martin-Barragan, B. (2024). Interpretable machine learning for imbalanced credit scoring datasets. *European Journal of Operational Research*, *312*(1), 357-372.
6. Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, *297*(3), 1178-1192.

7. George, N. (2019). *All Lending Club loan data* [Data set]. Kaggle. https://www.kaggle.com/datasets/wordsforthewise/lending-club
8. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, *51*(5), 1-42.
9. Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., ... & Zhang, Y. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias.
10. Karimi, A. H., Barthe, G., Balle, B., & Valera, I. (2020, June). Model-agnostic counterfactual explanations for consequential decisions. In *International conference on artificial intelligence and statistics* (pp. 895-905). PMLR.
11. Kozodoi, N., Jacob, J., & Lessmann, S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research*, *297*(3), 1083-1094.
12. Liu, W., Fan, H., & Xia, M. (2022). Credit scoring based on tree-enhanced gradient boosting decision trees. *Expert Systems with Applications*, *189*, 116034.
13. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.
14. Moscato, V., Picariello, A., & Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications*, *165*, 113986.
15. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).
16. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, *1*(5), 206-215.
17. Shi, S., Tse, R., Luo, W., D'Addona, S., & Pau, G. (2022). Machine learning-driven credit risk: a systemic review. *Neural Computing and Applications*, *34*(17), 14327-14339.
18. Szepannek, G., & Lübke, K. (2021). Facing the challenges of developing fair risk scoring models. *Frontiers in artificial intelligence*, *4*, 681915.
19. Ustun, B., Spangher, A., & Liu, Y. (2019, January). Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 10-19).
20. Yang, K., Yuan, H., & Lau, R. Y. (2022). PsyCredit: An interpretable deep learning-based credit assessment approach facilitated by psychometric natural language processing. *Expert Systems with Applications*, *198*, 116847.