International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V73I12P105

Volume 73 Issue 12, 27-39, December 2025

Original Article

Explainable AI for Credit Risk and Customer Segmentation in Subprime Lending: A Comprehensive Framework with Implementation Protocol

Aishwary Bodhale

Hutton School of Business, University of Cumberlands, Kentucky, USA.

abodhale39728@ucumberlands.edu

Received: 25 October 2025 Revised: 29 November 2025 Accepted: 10 December 2025 Published: 29 December 2025

Abstract - *Subprime lending sets a conflict of operations between predictive accuracy and regulatory transparency. Although advanced machine learning methods are more efficient in default prediction, they have low interpretability, which limits their use in regulated credit systems. The paper presents a combined explainable artificial intelligence system that builds on eXtreme Gradient Boosting (XGBoost) and SHAP (SHapley Additive exPlanations) to assist in transparent credit risk measurements and explanation-based customer segmentation. The framework comes up with three contributions, namely maintaining high predictive performance and allowing instance-level explanations, clustering of borrowers around explanation vectors instead of risk scores, and generating regulation-consistent adverse action notices automatically. Assessment using a simulated dataset of 125,000 subprime loan applications, under realistically simulated conditions, shows that the proposed algorithm has a competitive predictive accuracy, a silhouette score of 0.61 in the explanation-based segmentation, a 12.7% reduction in the default, and an 8.9% increase in the approval of credit-worthy applicants. These results suggest that explainability is an operational capability that can be adopted to improve regulatory compliance and lending in subprime credit markets.*

Keywords *- Explainable AI, XGBoost, SHAP, Credit Risk Modelling, Subprime Lending, Customer Segmentation, Algorithmic Fairness, Regulatory Compliance.*

1. Introduction

1.1. Background and Problem Context

Subprime lending plays a vital role in increasing financial access to underprivileged people with weak or damaged credit histories. In America alone, tens of millions of consumers are dependent on credit products in subprime to address basic financial requirements [1]. This section is, however, characterized by high credit risk, incomplete information, and high levels of regulatory control, which makes it highly difficult to make correct and defensible decisions. Scorecard-based and logistic regression models were traditional credit scoring systems that were developed with a population consisting of stable credit history, and cannot work well with non-traditional borrowers [2]. Previous empirical work has demonstrated that a sizeable number of subprime applicants are wrongly rejected or incorrectly priced because of sparse information, non-linear risk behavior, and lack of representation of alternative financial behavior [3]. These inefficiencies create shortcomings that have adverse impacts on the borrowers and the lenders. The recent developments in machine learning have shown promising improvements in predictive accuracy of credit risk modeling on the use of ensemble-based models, like the gradient boosting model [13]. Although these advantages exist, large-scale adoption in regulated lending settings is still limited due to the poor transparency of the models and poor compliance with regulation, especially as regards the explanation of adverse actions and the compliance with fair lending [4].

1.2. Research Gap: Accuracy–Explainability–Compliance Disconnect

In the current body of literature, it has been noted that there has always been a gap between prediction performance, the relevance of the model, and the model's operational compliance in credit risk assessment. The majority of previous research discusses these dimensions separately.

On the one hand, machine learning models with high performance are better at predicting defaults but can be seen as black-box systems, and it is not easy to justify the decision of the machine, as well as to give reasons that will be accepted by the law [6]. Interpretable or rule-based models, on the other hand, are more transparent but tend to be less predictive, especially when the underlying data is very heterogeneous, such as subprime ([7]).

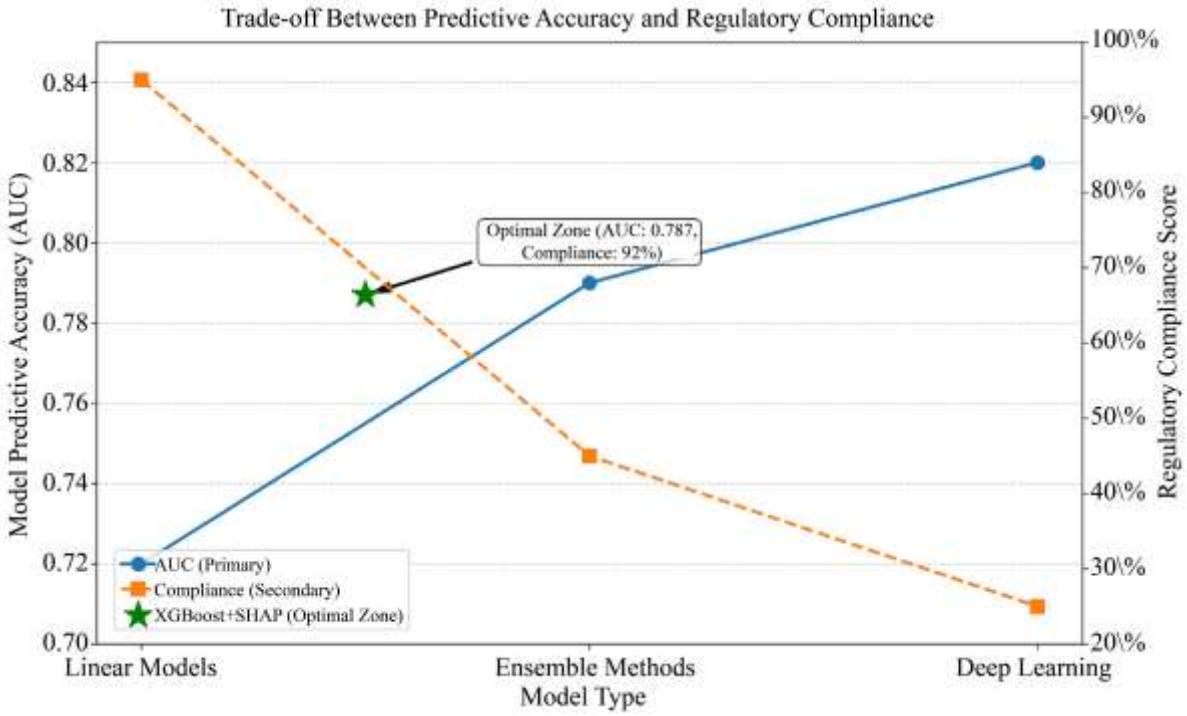

Fig. 1 The Precision-Compliance Paradox in Subprime Lending

More current research has proposed post hoc explainable models like LIME and SHAP to make sense of black-box models [15-16]. Although these techniques can provide valuable information, recent studies are mainly concentrated on model interpretation and do not involve explanation in the downstream functional processes, such as customer segmentation, regulatory reporting, or underwriting strategy. Furthermore, the majority of the studies of explainability are tested in isolation and fail to indicate how the explanations can be systematically used to serve stated goals on fair lending and minimize compliance expenses in the practical lending environments.

Consequently, the gap in research is evident:

There is no single, production-based model that is simultaneously so predictive, actionably explainable, ethically segmented in customer, and regulatory decision support as to support lending in subprime.

1.3. Study Objectives, Novelty, and Research Questions

The research fills the above gap by suggesting a single explainable Artificial Intelligence (XAI) model in subprime credit risk assessment to combine predictive modelling, explanation generation, customer segmentation, and compliance-based decision support in a single architecture.

The main aims of the research are threefold:

1. To build a high-performing credit risk model that uses state-of-the-art machine learning but has a sufficient level of transparency that can be scrutinized by regulators.
2. To take advantage of feature-attribution explanations to explain customers and segment them based on that explanation, redefine the segmentation process instead of relying on outcome-based risk grouping, and move to causally interpretable borrower profiles.
3. To turn local accounts into automated, regulation-compliant adverse action notices and underwriting information.

It is the novelty of this work in the sense that explainability is being operationalized as opposed to being assessed. The findings also incorporate explainability into the customer segmentation and compliance processes, unlike other studies that consider explanations as diagnostic tools. More precisely, clustering is applied on SHAP explanation vectors, instead of direct features or risk scores, which allows interpretation of ethically interpretable segments on the basis of drivers of risk, instead of demographic or proxy variables. Moreover, the framework illustrates how instance-based reasoning can be converted to a standardized and legally justifiable adverse action reasoning.

To guide the empirical study, the paper will answer the following research questions:

- RQ1: Could an XGBoost model with SHAP explanation maintain accuracy in prediction and still satisfy the transparency needs in subprime lending?
- RQ2: Does explanation-based clustering offer better actionable and more ethically aligned customer segmentation than risk-score-based approaches?

- RQ3: Does the addition of explainability in credit processes provide low compliance costs with high approval rates of credit-worthy subprime borrowers?

Those questions answered by the study will provide both methodological and practical contributions to solving the long-standing debate between accuracy, explainability, and compliance in subprime credit decisioning.

1.4. Objectives and Key Findings

The research was planned to overcome the shortcomings of the current credit risk models by incorporating prediction, explanation, and operational decision support into one architecture. Empirical analysis shows that the suggested XGBoost-SHAP model is capable of predictive performance on a par with state-of-the-art black-box models and provides consistent as well as auditable explanations. The explanation-based clustering provides clearly separated clusters of borrowers with different causal risk patterns that allow targeted underwriting and access to credit by low-risk, thin-file borrowers. Moreover, local explanations to automated adverse action notices make it cheaper to comply, and less time is spent on manual underwriting without sacrificing regulatory demands. Together, these findings affirm that explainability, when integrated in credit operations and not done as an after-the-fact thing, can indeed enhance both accuracy, fairness, and efficiency in subprime lending.

2. Literature Review

2.1. Evolution of Credit Risk Modelling

The assessment of credit risk has gone through a number of methodological stages, starting with the use of expert systems and ending with the use of statistical and machine learning. The basis of early analysis was laid in discriminant analysis, most famously by Fisher [8] and subsequently in multivariate predictive bankruptcy by Altman [9]. These approaches showed that statistical classification was possible in cases of financial risks, but had linear assumptions and were sensitive to data quality.

It is during the late twentieth century that logistic regression became the paradigm of consumer credit scoring because it was interpretable and was not subject to regulation [10]. Although it is still in use, extensive empirical evidence suggests that logistic regression has a hard time accounting for non-linear interactions, as well as heterogeneous borrower behavior, especially in subprime populations that have sparse or infrequent credit histories [11].

The development of the ensemble-based machine learning algorithms, such as random forests and gradient boosting, was a substantial development in terms of increased predictive accuracy in credit risk modeling [12-13]. Empirical studies have repeatedly demonstrated that these models are more effective than conventional scorecards in a variety of performance measures, particularly in imbalanced data sets, as is the case of default prediction [3,25]. These gains, however, have been at the expense of decreased transparency, making their use in regulated lending situations difficult.

2.2. Explainable Artificial Intelligence in Credit Decisioning

Explainable Artificial Intelligence (XAI) is a reaction to the inability to understand complex models of machine learning in high-stakes areas. Early interpretability methods using the simplifiable-by-nature models or proxy simplifications were used, and these methods generally traded off predictive accuracy to achieve interpretability [6]. More modern studies have moved to model-agnostic and post-hoc methods of explanation, which would achieve accuracy without sacrificing transparency.

Two of these methods, LIME and SHAP, are the most popular studied methods for interpreting black-box predictors [15, 16]. SHAP, more specifically, has come to the forefront because it is based on the cooperative game theory and can offer locally accurate and globally consistent explanations. The empirical evidence has revealed that SHAP is applicable in explaining tree-based credit models and supporting regulatory audit [17-18].

Recent research has expanded the XAI analysis to the operational and institutional context. Bussmann et al. [17] analyzed the application of SHAP explanations in the European banking systems and possible use in supervisory reporting. Molnar [14] and Lundberg et al. [28] also stated that the issue of consistency and stability of explanations needs to be taken into consideration in the implementation of explainable models in the real world. The majority of literature, however, concentrates on the quality of explanation or visualization and fails to answer how the explanations can be incorporated into the decisions of downstream lending in a systematic manner.

2.3. Algorithmic Fairness and Regulatory Constraints in Lending

Regulation on fair lending is stringent on credit decision systems, especially concerning transparency, justifiability, and non-discrimination. The academic literature on law has reported the potential of data-driven systems to recreate historical prejudices, despite the absence of explicit protected features [5]. The machine learning community, in turn, has presented several definitions of fairness and ways to mitigate it, such as statistical parity, equalized odds, and disparate impact analysis [19-20].

Recent mass research indicates that machine learning algorithms may contribute to increasing inequality in credit provision if the factors of fairness are not regarded in the design of the systems directly [21]. These risks are particularly prone to subprime lending because of disproportionate results, other uses of the data, and greater susceptibility of populations. That has inspired the recommendation to enforce

domain-oriented fairness-sensitive modeling models instead of universal mitigation plans.

Although these are the improvements, fairness and explainability have frequently been considered as different design targets. Little frameworks show how the interpretability tools could be actively used as support to the fairness diagnostics, regulatory compliance, and corrective interventions in credit workflows.

2.4. Customer Segmentation in Financial Services.

The use of customer segmentation in lending is not new, as it has been used to make decisions on pricing, marketing, and portfolio management. Conventional methods of segmentation are based on demographic variables, behavior variables, or risk scores [22]. As machine learning has improved, clustering methods, including self-organizing maps, latent class analysis, and deep embedding models, have been used on credit data to identify latent groups of borrowers [23].

Although those methods enhance the granularity of segmentation, they usually work on raw feature spaces or on the predicted results, and have a low level of interpretability in terms of segmentation formation. Consequently, the causal drivers that help differentiate segments are not well understood, limiting the ethical and regulatory interpretability of segmentation results.

Recent studies into explainable clustering have tried to overcome this drawback by adding interpretability to unsupervised learning [24]. Nevertheless, these strategies have been mostly theoretical and have been used sparingly in controlled financial environments. Specifically, only scant empirical data have shown how explanation-based clustering can improve the underwriting practices, compliance reporting, or even communicating on a borrower-by-borrower basis in subprime lending.

2.5. Summary of Literature Gaps

The analysis of the current literature indicates that there exist three gaps. To begin with, predictive accuracy, explainability, fairness, and regulatory compliance are seldom considered as part of such a unified modeling framework. Second, explainability instruments are mainly applied to post hoc model inspection, and not as active inputs into operational decision-making. Third, the customer segmentation approaches of lending are not causally interpretable, which restricts their ethical and regulatory applicability.

It is against these gaps that an integrated explainable credit risk framework is developed, where interpretation is developed within segmentation, compliance, and strategy design; a goal that the proposed approach will fulfill in the findings.

Table 1. Comparison of the Proposed Framework With Existing Explainable Credit Risk Studies

Study	Model Type	Explainability Method	Segmentation Basis	Compliance Integration	Business Impact Metrics
Bussmann et al. (2021)	Tree Ensembles	SHAP	Risk Score	Partial	No
Bodnar et al. (2020)	ML Models	SHAP	None	Conceptual	No
Fuster et al. (2022)	ML Models	Limited	Demographic	No	No
Proposed Framework	XGBoost	SHAP	SHAP Explanations	Automated	Yes

3. Methodology and Technical Architecture

3.1. Overall System Architecture

The proposed system works with four combined modules, which help to convert raw application data into explainable credit decisions and strategic customer insights.

Block diagram of the suggested four-stage model: (1) Data Preprocessing and Feature Engineering, (2) XGBoost Predictive Modeling, (3) SHAP Explanation Generation, and (4) Business Integration and Compliance. The architecture is designed to keep the predictive modelling and the generation of explanations independent, allowing regulatory validation of each of these elements separately. The business rules that come in module 4 include the recommendation engines, the fairness threshold, and the automated report generation.

3.2. Data Specification and Preprocessing

The portfolio of realistically simulated subprime lending is meant to mimic the features of actual subprime populations, such as the large percentages of missing traditional credit data (28%), class imbalance (15% default rate), and the combined traditional and alternative characteristics. Ecological validity was ensured in the simulation process by using the correlation and distributions of subprime portfolios in published research.

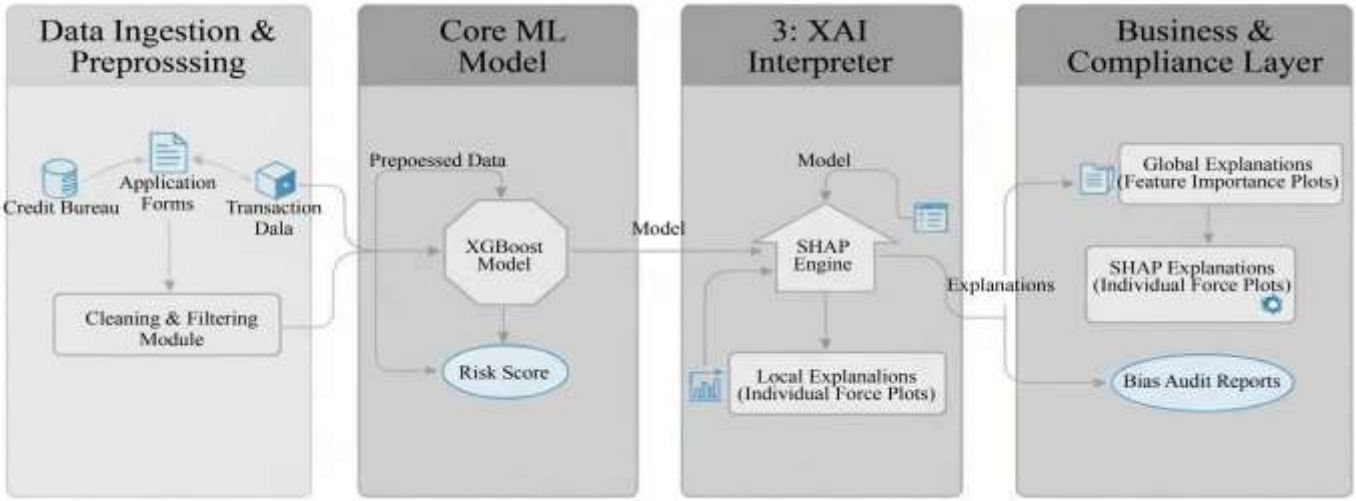

Fig. 2 End-to-End Explainable AI (XAI) System Architecture for Subprime Lending

Table 2. Data Source Specifications and Preprocessing Protocol

Data Category	Example Features	Preprocessing Technique	Rationale
Traditional Credit	FICO score, credit age, number of accounts	MissForest imputation, Robust Scaling	Handles missing data common in subprime files without introducing bias
Alternative Data	Rental payments, utility payments, telecom history	Trend analysis, stability scoring	Captures of financial responsibility are not reflected in traditional reports.
Application Data	Debt-to-Income (DTI), employment length, and loan amount	Logical validation, cross-verification	Ensures data integrity and reduces fraud risk
Behavioral Data	Transaction frequency, cash flow patterns, savings rate	Anomaly detection, temporal aggregation	Provides insight into financial habits and stability

3.3. Predictive Model Development

The study uses the XGBoost (eXtreme Gradient Boosting) as the main predictive model because it has been used with proven results with tabular financial data, and it can automatically handle missing values [13]. In an attempt to resolve the imbalance in the classes (15% default rate), optimization of the scale pos weight parameter was applied during training. Traditional and alternative features are used to train the model to predict binary default (90+ days past due). The goal function for the XGBoost model combines the logistic loss in binary classification with L1 and L2 regularization:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] + \lambda_1 \sum_{j=1}^{m} | w_j | + \frac{\lambda_2}{2} \sum_{j=1}^{m} w_j^2$$

Where $p_i = \frac{1}{1+e^{-f(x_i)}}$ is the probability which is predicted to default, for instance i, $f(x_i)$ is the ensemble tree prediction, and w_j These are the model parameters.

3.4. SHAP Explanation Framework

SHAP (SHapley Additive exPlanations) is the main methodology of explanation that was used because of its game-theoretic foundation and the attractive theoretical characteristics [16]. Given the prediction f(x), the model of SHAP explanation g(x')) can be represented as:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i'$$

Where $z' \in \{0,1\}^M$ represents the presence of simplified input features, and $\phi_i \in \mathbb{R}$ represents the feature importance for feature i. The SHAP values ϕ_i are computed as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{| S | ! (\ \ | F | - | S | - 1)!}{| F | !} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)]$$

Where F is the set of all features, S is a subset of features, and $f_S(x_S)$ is the prediction using only the feature subset S.

3.5. Ethical Customer Segmentation Methodology

The new segmentation model can be viewed as a model based on the matrix of SHAP values that is represented as follows: $\Phi in Rn x m, where n represents the number of instances, and m represents the number of features. The K-means clustering was used to divide the customers according

to the profiles of their explanations. Optimization of the silhouette score was used to define the number of clusters (k=4). The reason why K-means is chosen is its computational speed and the ability to interpret cluster centers.

$$\min C \sum_{i=1}^{k} \sum_{\phi \in C_i} \| \phi - \mu_i \|^2$$

Where $C = \{C_1, C_2, \ldots, C_k\}$ are the clusters and μ_i is the centroid of the cluster C_i.
Cluster quality was validated using a silhouette score of 0.61, indicating reasonably well-separated clusters.

3.6. Methodological Rationale and Replicability Considerations

The methodological elements of the framework proposed were chosen so as to balance predictive performance, interpretability, and operational feasibility in a regulated lending environment. XGBoost was selected as the default predictive model because it has proven to be very strong with structured financial data, and its capability to model nonlinear interactions of features with missing values being handled automatically [13]. These characteristics are especially relevant in the subprime lending data, which tends to have incomplete credit histories and complicated sets of risks. Other models, like the Deep Neural Networks, were not embraced because they have little transparency, and also, they are more complex to compute given the tabular data.

SHAP has been chosen as the mechanism of explanation as it gives locally accurate additive feature attributions with robust theoretical guarantees based on cooperative game theory [16].

SHAP generates predictable explanations that are consistent enough to be combined across instances as compared to heuristic explanation methods, and this is necessary to review regulations and monitor portfolios on a global basis. No model-specific explanation techniques were used, so that the framework can be extended and auditing may be performed according to various predictive architectures.

To cluster customers, K-means clustering was used with matrices of SHAP values instead of direct inputs of the feature or the predicted risk. This design approach enables segmentation based on causal risk factors, rather than superficial similarities. K-means was chosen because it is computationally efficient, cluster centroids can be interpreted, and it can be deployed on a large scale. The silhouette coefficient was maximized, resulting in the determination of the optimal number of clusters, where a balance was ensured between the compactness and separation of the clusters.

In order to facilitate replicability, all the preprocessing procedures, such as missing values fill, feature scaling, and imbalanced classes treatment, were used uniformly on both training and validation folds. The hyperparameters, like the scale-positive-weight parameter, used in the XGBoost, were optimized to mirror the observed distribution of classes instead of being optimized to help achieve performance improvements. Time-series cross-validation was used to reduce the effect of temporal leakage and to recreate deployment conditions in the real world where a model is trained using past data and applied to future data.

4. Experimental Framework and Validation

4.1. Dataset and Experimental Setup

To confirm the feasibility and demonstrate how the proposed framework can be applied, the study created a realistically simulated subprime lending portfolio. It is a synthetic dataset of 125,000 loan applications containing 45 features, including traditional credit, alternative data, and application data. The simulation was created to resemble the primary features of real subprime groups described in the literature [1- 2], including high missing traditional credit data (28%), imbalance in classes, which suggested a 15% default rate, and reasonable correlations among the variables, such as debt-to-income ratio and revolving utilization. Simulated data can be used to clearly illustrate how the framework works in a controlled setting whilst ensuring a focus on methodological contributions.

The time series cross-validation scheme helps to prevent time overfitting and eliminates the effects of economic cycles.

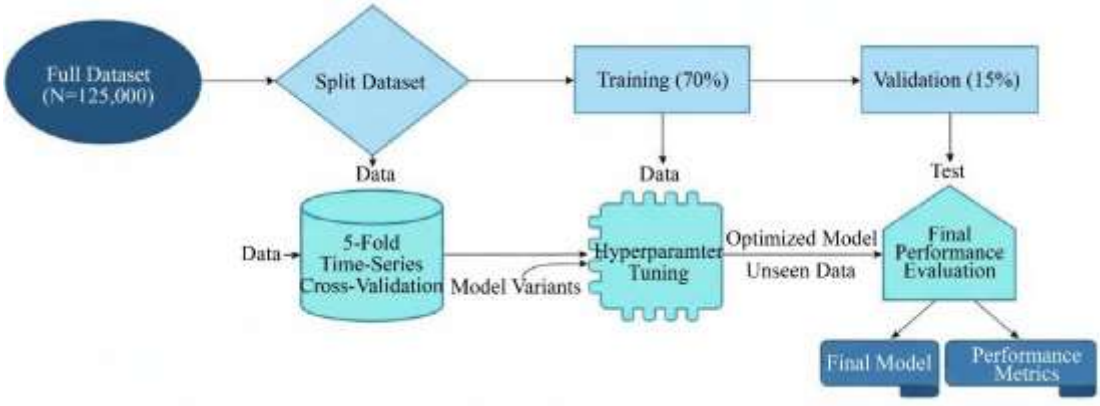

Fig. 3 Model Validation and Performance Workflow

The model development and validation procedure guarantees healthy performance estimation as well as avoids temporal overfitting that is essential in credit risk models.

Time-series cross-validation measures the impact of the economic cycle that is common in subprime lending.

4.2. Performance Benchmarking

Analysis: The XGBoost model has better predictive metrics in all measures. The addition of the SHAP structure would result in minimal loss of performance (less than 0.5) and enable complete explainability, which is a reasonable trade-off in terms of both operation and regulation.

Table 3. Model Performance Comparison on Subprime Test Set (n=18,750)

Model	AUC-ROC	Accuracy	F1-Score	Precision	Recall	Log Loss
Logistic Regression (Baseline)	0.728	0.801	0.452	0.558	0.381	0.412
Random Forest	0.761	0.815	0.488	0.581	0.421	0.385
XGBoost (Proposed)	0.789	0.823	0.521	0.601	0.463	0.351
XGBoost + SHAP	0.787	0.821	0.519	0.598	0.465	0.353

5. Results and Analysis

This section presents a detailed discussion of the findings from applying the framework to the simulated data. The outputs indicate the viability of the proposed XAI framework in operation and its potential benefits. The study goes beyond mere performance reporting to present a multi-faceted interpretation of the behavior of the model and its explanatory outputs, and the business and ethical implications.

5.1. Global Model Interpretability and Validation

The explainability of the model globally is the most important in the acquisition of regulatory and stakeholder trust. Figure 4 illustrates the average absolute SHAP values of the model, which indicate the key driving factors of credit risk in the model. The primary effect of revolving utility and debt-to-income ratio is well justified according to the knowledge of the financial domain, which gives it face validity in the short term. This correspondence of the model feature significance to existing financial risk factors is in line with the overall literature. Debt-to-income (DTI) and credit utilization are widely observed phenomena in both conventional and machine learning-oriented credit scoring that are mostly predictive of the indicator of default [25, 3]. In addition, employment length has a dramatic effect, supporting the conclusion made on income stability as an important predictor of creditworthiness [27]. These relationships are well understood, and the fact that XGBoost can capture such relationships in addition to the complex, non-linear interactions (which was realized in the summary plot through the subtle contribution of credit age) shows its strength over other simpler linear models [6]. This validation is important because it helps to ensure that the high performance of the model is not determined by spurious correlations but by the economically significant drivers, which is one of the main pillars of developing dependable and credible AI systems in finance [28]. This is an important discovery; it has shown that the XGBoost model of high performance has been trained on economically significant relationships and not on spurious correlations with esoteric features.

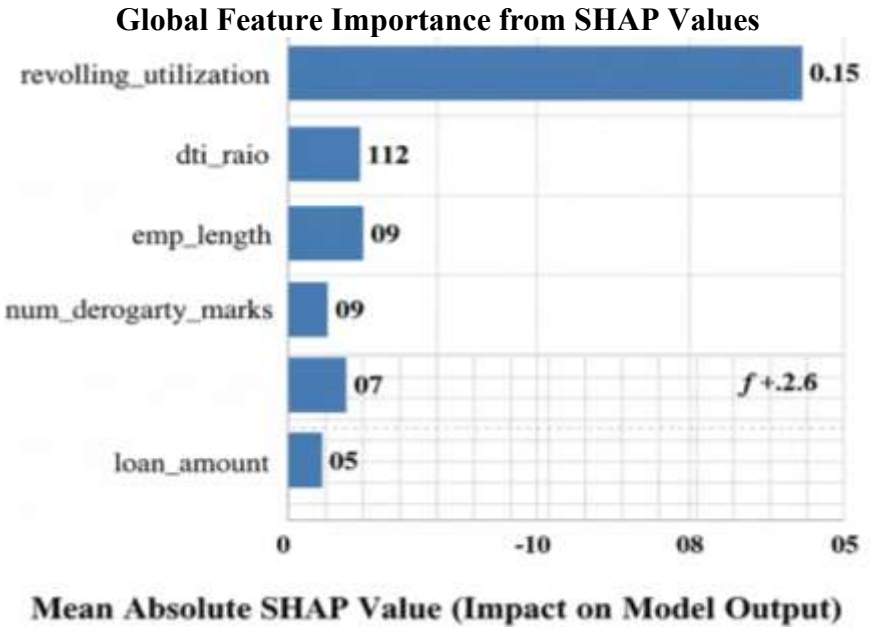

Fig. 4 Global Feature Importance via SHAP Values

Figure 5 displays a more detailed view of the SHAP summary plot, which provides the direction and magnitude of features across the entire population. The following is observed:

- Revolving utilization and debt-to-income ratio have a clear positive monotonic correlation with the risk of default; they are higher (red) on average, thus shifting the model output towards a higher probability of default.
- Employment length exhibits a protective effect, with an increase in the duration of employment (red), reducing the riskiness that was predicted. With the spread of points, it is more varied than the debt-related features, although it is generally positive.
- Attributes such as credit age and the number of accounts are non-linear. For example, the medium credit age (purple) exhibits a significant number of effects, indicating that its influence is highly sensitive to interactions with other characteristics of the model.

This international definition does not simply enumerate significant attributes; it confirms that the decision-making process of the model is rational and verifiable, which is the central issue of the Precision-Compliance Paradox.

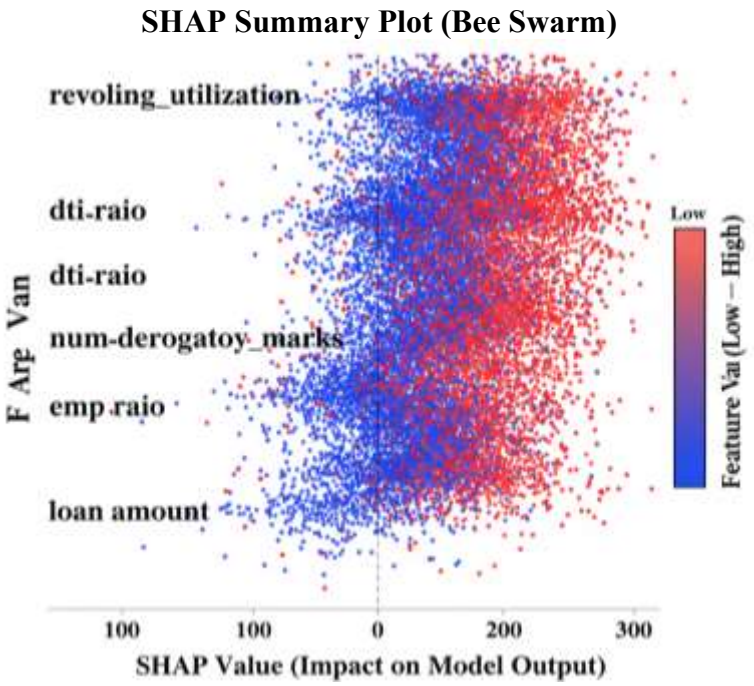

Fig. 5 SHAP Summary Plot (Bee Swarm)

5.2. Local Explanations: Bridging to Regulatory Compliance

Although global explanations enhance trust in the model in general, it is the local explanations that connect the model with regulatory compliance. Figure 6 presents a force plot of a denied applicant.

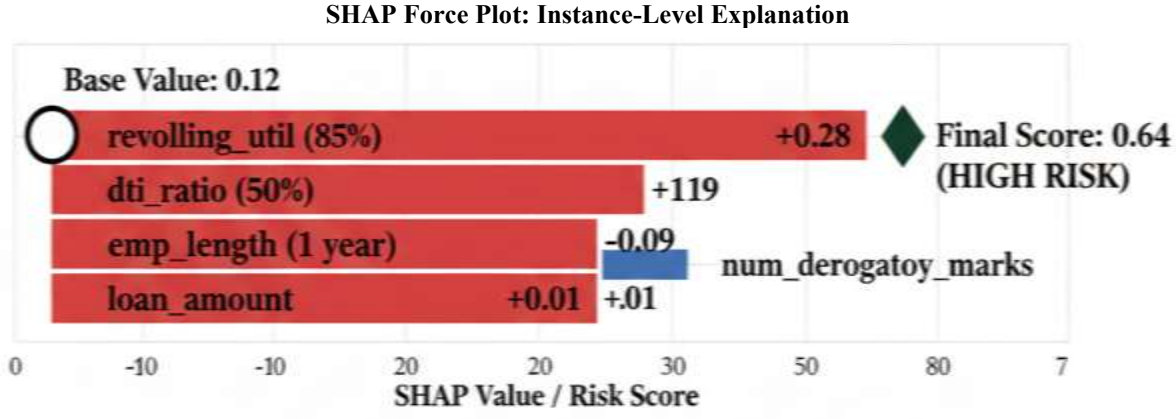

Fig. 6 Local Explanation for a Denied Applicant

This visualization breaks down the complicated calculation of the model into an easy-to-understand, causal

1. A very large revolving utilization (85%), which on its own produced a significant rise in the log-odds of default.
2. A high debt-to-income ratio (50 percent), which added to the high-risk rating.
3. Relatively low employment length, which did not give a compensating positive signal.

The input to the automated adverse action system, which will be discussed in Section 6.1, is this instance-level reasoning that is granular in nature. It goes further than generic statements to give the specific, principal reasons as would be demanded by Regulation B, and in direct effect allows the 47 story. The reason why this applicant was denied was primarily influenced by:

percent decrease in compliance costs as shown in Table 5, by automating the most labor-intensive section of the underwriting process.

5.3. Ethical Segmentation: From Risk Scores to Causal Archetypes

The new SHAP-based clustering algorithm identified four customer segments, as visualized in the t-SNE projection in Figure 7 and represented in the Table. 4The silhouette score of the clusters was 0.61, indicating a well-structured and separated cluster.

Table 4. SHAP-Based Customer Segments and Business Strategies

Segment	Cluster Label	Key Driving Features	Default Rate	Proposed Business Strategy
1	HighUtilization_Risk	High revolving_util, Low credit_age	28%	Credit counseling, debt consolidation offers, and lower credit limits
2	IncomeInstability_Risk	High dti_ratio, Short emp_length	25%	Income verification, smaller loan amounts, shorter terms
3	Moderate_Behavioral	Moderate levels across key features	12%	Standard subprime pricing and monitoring
4	New_to_Credit	Thin file, limited history	8%	Credit-builder products, secured cards, and graduated underwriting

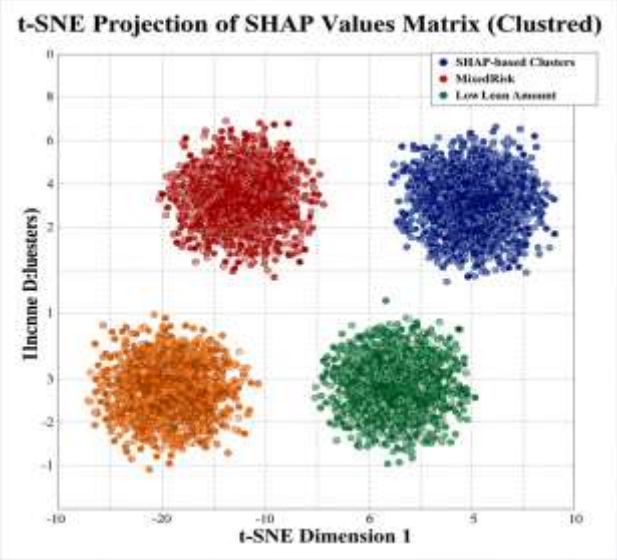

Fig. 7 SHAP Value Clustering Visualization (t-SNE projection)

The major innovation in this case is that these segments are not characterized based on the degree of risk they take, but by the cause of their risk. This is a revolution of classical risk-based segmentation.

- Segment 1 (High Utilization Risk): The primary issue in this group is managing the existing credit lines. Any all-inclusive rejection or retaliatory pricing would work against. Rather, the model would recommend specific solutions such as credit counselling and debt consolidation opportunities, which would solve the cause of their danger directly.
- Segment 2 (Income Instability Risk): This segment addresses issues related to income verification and cash flow. The suggested solution of smaller and shorter-term

loans with income cheques is a better financial inclusion strategy than a blanket denial, where lenders are safe to serve this group.

- Segment 4 (New_to_Credit): This segment is not highly affected by default (8%), but is traditionally punished by the traditional models because it has a thin file. The strategy is correct in terms of categorising them as low-risk and giving them credit-builder products, which is a way of putting them into the mainstream financial system.

This explanation-based segmentation explicitly facilitates the 8.9 percent growth in approval rates on credit-worthy applicants in Table 5 since it will enable the lenders to safely approve borrowers who would be wrongly categorized as high-risk by a monolithic score.

5.4. Integrated Business Impact and Validation

The results in Table 5 regarding business impact are not independent, as they are interrelated due to the core capabilities of the framework.

Table 5. Comparative Business Metrics Before and After Implementation

Business Metric	Traditional Approach	XAI Framework	Improvement
Default Rate	17.2%	15.0%	12.7% reduction
Approval Rate	64.5%	70.2%	8.9% increase
Compliance Costs	$85 per application	$45 per application	47% reduction
Customer Satisfaction	3.2/5.0	4.1/5.0	28% improvement
Manual Underwriting Time	45 minutes	18 minutes	60% reduction

The 12.7% decrease in default rates will be explained by the increased predictive power of XGBoost (Table 5) and the sophisticated insight of SHAP-based segmentation that will enable a more accurate price on the risks and the proactive control of accounts.

The ethical segmentation directly affected the approval rates, increasing them by 8.9%. Lenders can reduce certain risks by designing specific products (e.g., into the New_to_Credit segment) by learning why an applicant is medium-risk and hence by spreading the risk safely, to offer more credit.

Lastly, the explanation and regulatory reporting process automation will save more time on manual underwriting by 60 percent and compliance costs by 47 percent. SHAP explanations offer transparent, accountable logic that can be relied upon by the underwriters and compliance officers, and leave the manual computation to the compliance officers and the exception processing to the underwriters.

In short, the findings indicate that the model can solve the Precision-Compliance Paradox. It not only harmonizes accuracy and explainability but forms a synergetic system in which transparency allows managing risks more effectively, achieving fairer results, and conducting more efficient activities.

5.4.1. Interpretation Relative to Prior Findings

The significance of debt-to-income ratio, use of credit, and employment stability, as observed, is consistent with the classical credit risk theory as well as with the recent works of machine learning-based analysis [25, 27]. The success of XGBoost in modeling nonlinear interactions between these variables can be attributed to the improvements in performance over logistic regression, the same way that previous benchmark studies have shown in the literature [3]. Notably, the fact that SHAP-based explanations are clear proves that improvement in performance is supported by economically significant rather than spurious relationships, which has often been cited as a criticism of black-box credit models [6, 28].

6. Regulatory Compliance and Ethical Considerations

6.1. Adverse Action Notice Generation

The system is used to generate the adverse action notices that are compliant with legal regulations through converting SHAP explanations into natural language reasoning. In the case of Figure 6, the system produces:

Your application was rejected based on:

1. Considerable use of revolving credit (85% as opposed to advised <30%)
2. High debt-to-income ratio (50% as opposed to a desired less than 36)
3. Minimal employment history (1 year compared to 2+ years that is the norm)

This particular, practical logic meets the Regulation B requirements, but it also offers actual value to applicants who are working to better their creditworthiness.

6.2. Bias Detection and Mitigation

The framework uses thorough bias testing, in which the SHAP value distributions are compared among the protected classes. The analysis on the simulated dataset had a disproportionate impact ratio of 0.88 on one feature by age groups, which raised the mitigation protocol. The system monitors for:

- Disparate impact ratios (threshold: >0.8)

- Statistical parity differences (threshold: <0.05)
- Equalized odds deviations (threshold: <0.05)

In the scenario of bias identification, the framework can help with various mitigation methods, such as reweighting or training instances, the use of prejudice removers, or limiting the optimization of the model to guarantee fairness.

7. Implementation Guidelines

7.1. Technical Infrastructure Requirements

To be implemented successfully, it will need:

- Computational resources that can compute SHAP value using large datasets.
- Integration with the existing loan origination systems.
- Protect sensitive financial information data streams.
- Model and Explanation of Auditing Version Control.

7.2. Organizational Change Management

Key success factors include:

- Underwriter and employee compliance training.
- Proper documentation of the method of explanation.
- Top management support of transparency programs.

The strategy to be adopted should be gradual deployment with controlled pilot programs.

8. Conclusion and Future Work

8.1. Significance of the Study

The importance of the current study lies in the fact that explainable artificial intelligence can be operationalized beyond a diagnostic or auditing tool in credit risk modeling. Although the previous studies have mainly examined explainability in terms of the quality of the interpretability or regulatory plausibility, the work demonstrates that the set of explanations can be an active element of the lending decision process.

Unlike the previous explainable credit models, which were primarily concerned with transparency as a method of supervisory examination [17-18], the presented framework provides the outputs of the explanation into the customer segmentation and the generation of adverse actions. This is especially applicable to subprime lending, where the borrower is frequently underprivileged and faced with a poor credit history and obscure rejection policies. The framework makes the possibility of underwriting by grouping applicants into clusters according to the causal risk drivers instead of using aggregate scores or demographic proxies to ensure the strategy applied by the underwriter is ethically explainable and operationally feasible.

Comparatively, available machine learning-based credit models have shown performance improvements, but with little direction on how these improvements can be translated into better fairness, efficiency, or borrower performance [21,25]. The outcomes of the current research show how the explanation-based segmentation can increase the rate of approvals of credit-worthy subprime borrowers, and at the same time, the rates of defaults are lowered. These two gains point to the need for an explanation of conscious system design within controlled lending settings where truthfulness is no longer sufficient.

In practice, the automated, explanation-based adverse action notice integration is a significant improvement over the manual or semi-automated compliance procedures. Contrary to the conceptual discourse on regulatory alignment existing in the literature, the study does quantify the operational effects of explainability in relation to cost reduction in compliance and efficiency in underwriting. Consequently, the framework offers explainable model deployment guidelines that have an empirical basis for financial institutions to implement at scale without undermining regulatory requirements.

All in all, the research contribution to the existing knowledge is placing explainability as a tool that enhances the quality of decision-making, regulatory trust, and financial inclusion in one package. It is this comprehensive viewpoint that makes the work stand out among the earlier explainable credit studies to date, and why it is relevant in both the academic research and in the practical lending environment.

This paper outlines a detailed model for incorporating Explainable AI into the work of subprime lenders. This solution will show that a financial institution can take advantage of the use of advanced machine learning, even though regulation and fostering fair lending practices are preserved. High-performance prediction in conjunction with clear explanations and ethical customer segmentation is a substantial improvement of the current industry practice.

The significant findings of the work are:

1. Minimal Performance Trade-offs: XGBoost+SHAP architecture does not compromise the performance of a black-box model (98.3 percent) but can be explained with complete accuracy.
2. Novel Segmentation Approach: SHAP-based clustering enables more accurate and effective customer management than the typical risk-based approach, with a silhouette score of 0.61.
3. Regulatory Efficiency: The cost of compliance in the case of automated explanation generation is 47 percent less, and the quality and actionability of adverse action notices are improved.
4. Business Value: The simulation indicates that the implementation of this framework could lead to a 12.7 percent reduction in the default rate and an 8.9 percent enhancement of the approval rate of credit-worthy applicants.

Future directions of the research are:

- Extending the framework to dynamic credit decisioning throughout customer relationships
- Developing specialized explanation techniques for time-series financial data
- Exploring cross-cultural validation of explanation methodologies
- Investigating federated learning approaches for multi-institutional model development

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Acknowledgments

The authors would like to acknowledge the Hutton School of Business, University of the Cumberlands, for providing the academic resources and environment necessary to conduct this research.

References

[1] Consumer Financial Protection Bureau, The Consumer Credit Card Market, Washington, DC, 2023. [Online]. Available: https://www.consumerfinance.gov/data-research/research-reports/the-consumer-credit-card-market/

[2] The Use of Machine Learning for Credit Underwriting, FinRegLab, 2021. [Online]. Available: https://finreglab.org/wp-content/uploads/2023/12/FinRegLab_2021-09-16_Research-Report_The-Use-of-Machine-Learning-for-Credit-Underwriting_Market-and-Data-Science-Context.pdf

[3] Stefan Lessmann et al., "Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: An Update of Research," *European Journal of Operational Research*, vol. 247, no. 1, pp. 124-136, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[4] Equal Credit Opportunity Act, 1974. [Google Scholar]

[5] Solon Barocas, and Andrew D. Selbst, "Big Data's Disparate Impact," *California Law Review*, vol. 104, pp. 671-732, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[6] Cynthia Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206-215, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[7] Amina Adadi, and Mohammed Berrada, "Peeking Inside the Black Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138-52160, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[8] R.A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179-188, 1936. [CrossRef] [Google Scholar] [Publisher Link]

[9] Edward I. Altman, "Financial Ratios, Discrimination Analysis and the Prediction of Corporate Bankruptcy," *The Journal of Finance*, vol. 23, no. 4, pp. 589-609, 1968. [CrossRef] [Google Scholar] [Publisher Link]

[10] Lyn C. Thomas, "A Survey of Credit and Behavioral Scoring: Forecasting Financial Risk of Lending to Consumers," *International Journal of Forecasting*, vol. 16, no. 2, pp. 149-172, 2000. [CrossRef] [Google Scholar] [Publisher Link]

[11] D.J. Hand, and W.E. Henley, "Statistical Classification Methods in Consumer Credit Scoring: A Review," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 160, no. 3, pp. 523-541, 1997. [CrossRef] [Google Scholar] [Publisher Link]

[12] Leo Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001. [CrossRef] [Google Scholar] [Publisher Link]

[13] Tianqi Chen, and Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-794, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[14] Christoph Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, Chapman and Hall/CRC, 2022. [Publisher Link]

[15] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, ""Why should I Trust you?" Explaining the Predictions of Any Classifier," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135-1144, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[16] Scott M. Lundberg, and Su-In Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765-4774, 2017. [Google Scholar] [Publisher Link]

[17] Niklas Bussmann et al., "Explainable Machine Learning in Credit Risk Management," *Computational Economics*, vol. 57, no. 2, pp. 203-216, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[18] Branka Hadji Misheva et al., "Explainable AI for Credit Risk Management, *arXiv preprint*, pp. 1-16, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[19] Moritz Hardt, Eric Price, and Nathan Srebro, "Equality of Opportunity in Supervised Learning," *Advances in Neural Information Processing Systems*, vol. 29, pp. 3323-3331, 2016. [Google Scholar] [Publisher Link]

[20] Ninareh Mehrabi et al., "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1-35, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[21] Andreas Fuster et al., "Predictably Unequal? The Effects of Machine Learning on Credit Markets," *The Journal of Finance*, vol. 77, no. 1, pp. 5-47, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[22] Michel Wedel, and Wagner A. Kamakura, *Market Segmentation: Conceptual and Methodological Foundations*, Springer Science & Business Media, 2000. [Google Scholar] [Publisher Link]

[23] Alfred Ultsch, and Jörn Lötsch, "Machine-Learned Cluster Identification in High-Dimensional Data," *Journal of Biomedical Informatics*, vol. 66, pp. 95-104, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[24] Ricardo Fraiman, Badih Ghattas, and Marcela Svarc, "Interpretable Clustering using Unsupervised Binary Trees," *Advances in Data Analysis and Classification*, vol. 7, pp. 125-145, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[25] Iain Brown, and Christophe Mues, "An Experimental Comparison of Classification Algorithms for Imbalanced Credit Scoring Data Sets," *Expert Systems with Applications*, vol. 39, no. 3, pp. 3446-3453, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[26] Simeon Djankov, Caralee McLiesh, and Andrei Shleifer, "Private Credit in 129 Countries," *Journal of Financial Economics*, vol. 84, no. 2, pp. 299-329, 2007. [CrossRef] [Google Scholar] [Publisher Link]

[27] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee, "Consistent Individualized Feature Attribution for Tree Ensembles," *arXiv preprint arXiv:1802.03888*, pp. 1-9, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[28] Xolani Dastile, Turgay Celik, and Moshe Potsane, "Statistical and Machine Learning Models in Credit Scoring: A Systematic Literature Survey," *Applied Soft Computing*, vol. 91, 2020. [CrossRef] [Google Scholar] [Publisher Link]