International Journal of Emerging Trends in Computer Science and Information Technology
ISSN: 3050-9246 | https://doi.org/10.63282/3050-9246.IJETCSIT-V7I2P104
Eureka Vision Publication | Volume 7, Issue 2, 17-26, 2026

Original Article

An Interpretable Machine Learning Framework for Predictive Analysis in High-Risk Financial Systems

Ramesh Kasarla
Comcast Cable Communications, USA.

Received On: 13/02/2026 ***Revised On: 19/03/2026*** ***Accepted On: 27/03/2026*** ***Published On: 04/04/2026***

Abstract - Predictive modeling in high-risk financial sectors faces significant challenges due to non-linear variable interactions, systemic risk, and the inherent instability of financial datasets. While black-box models often prioritize predictive efficiency over transparency, regulatory requirements in heavily regulated industries necessitate auditable and interpretable frameworks. This paper proposes a novel, interpretable Machine Learning (ML) framework designed for predictive analysis in high-stakes financial environments. The architecture integrates a federated learning system to preserve data privacy across distributed client devices while employing local and global model validation techniques. Our methodology incorporates advanced feature engineering, ensemble learning, and post-hoc interpretability tools, including SHAP (Shapley Additive explanations) and LIME (Local Interpretable Model-agnostic Explanations). Furthermore, we introduce specialized evaluation metrics, such as the Model Transparency Index and Conditional Value at Risk (CVaR) prediction accuracy, to better align model performance with financial stability goals. Empirical results using real-world bankruptcy and fraud datasets demonstrate that the proposed framework achieves F1-scores and AUC-ROC values comparable to state-of-the-art deep neural networks and Gradient Boosting Machines (GBMs). Specifically, the framework demonstrates a 15% performance improvement over existing rule-based methods like RuleFit. This work provides a strategic guideline for balancing the accuracy-interpretability trade-off, facilitating the legal and ethical deployment of AI in regulated financial systems.

Keywords - Interpretable Machine Learning, Predictive Analysis, Financial Systems, Explainable AI (XAI), Feature Engineering, SHAP, LIME, High-Risk Systems, Fraud Detection.

1. Introduction

1.1. Importance of Predictive Analysis in High-Risk Financial Systems

In the current global financial landscape, the proliferation of data-driven decision-making has led to the widespread adoption of Machine Learning (ML) models for predicting high-risk events such as fraudulent transactions, credit defaults, and systemic market failures. However, as financial institutions move away from traditional heuristic-based models toward complex, non-linear algorithms, a significant "transparency gap" has emerged. While deep learning and ensemble methods offer superior predictive accuracy, their "black-box" nature often obscures the underlying logic behind high-stakes decisions. [1-4] Here are the five additional subtopics that can expound on the topic of this article in relation to high-risk financial systems predictive analysis:

1.1.1. Risk Mitigation and Early Detection

Such systems can be of great use to organizations, as they anticipate issues that can occur in the future due to previous incidents. Using data and recognizing patterns that already exist allows the creation of future predictions in terms of default, bankruptcy and fraud transactions. It is useful for financial institutions in order to prevent actions whereby credit lines need to be adjusted, activities that the firm engages in need to be reported or options to extend to the ailing firm can be offered. Therefore, These models should be embraced to minimize loss and stabilize the institution.

1.1.2. Enhanced Decision-Making

Predictive analytics entails the process where decision-makers get forecast information that will assist them in making wiser decisions. Decision makers in high-risk financial system scenarios are generally confronted with some levels of risk resulting from fluctuations in the market, in addition to navigating through intricate financial environments. By reducing such risks, this uncertainty is addressed through predictive models that estimate the possible results that may be expected given past data, the current state of the market, and other influential factors. For instance, the probability of default can be forecast in an endeavour to facilitate accurate loan risk projections for financial institutions, whereas in investment, market trends are made in order to make positive predictions. This makes the workings of decision-making even more efficient and secure, especially when it comes to decision-making in contexts that are perceived to be risky.

1.1.3. The Regulatory Imperative

Regulatory frameworks, including the GDPR’s "Right to Explanation" and various financial auditing standards, now mandate that AI-driven decisions in the financial sector be both explainable and auditable. The failure to provide interpretable results not only risks legal non-compliance but

also undermines stakeholder trust and the ability to mitigate systemic risk during periods of market instability.

1.1.4. Fraud Detection and Prevention

In many organizations, fraud risk targets their financial processes; this may include; identity theft, money laundering and credit card fraud. These activities can be detected and prevented through the help of predictive analysis since they help expose abnormal transactions that occur within the financial systems since they give an insight into a transaction as it is transacting as compared to a typical transaction. For example, it can identify large cases or total payment amounts coming from distant places or exhibit an unusual activity level different from earlier patterns. This way, financial institutions can identify fraudulent transactions at an early stage before they cost the institution a considerable amount of money.

1.1.5. Resource Optimization and Cost Efficiency

Thus, the financial institutions' use of predictive analysis results in the better allocation of resources and operational efficiency. The amount of demand powered by technological advancement and proper targeting of the financial community can be predicted, and areas of low utilization of capital, human resources, and technological assets can be easily identified. For instance, some predictive models can enhance portfolio management by determining the best areas/stocks to invest in or where to invest most. Likewise, for financial institutions, prediction analysis can be beneficial in cutting down general expenses related to non-performing loans, fraud investigations and other unimportant duties. This, in a way improves the general profitability and assists the financial institutions to be competitive in the market. Hence, predictive analysis of high-risk financial systems is quite central for minimizing risks, decision-making through acquiring a deep understanding of compliance and fraud and managing or allocating scarce resources that go a long way toward sustaining and enhancing the competitive health of banking institutions.

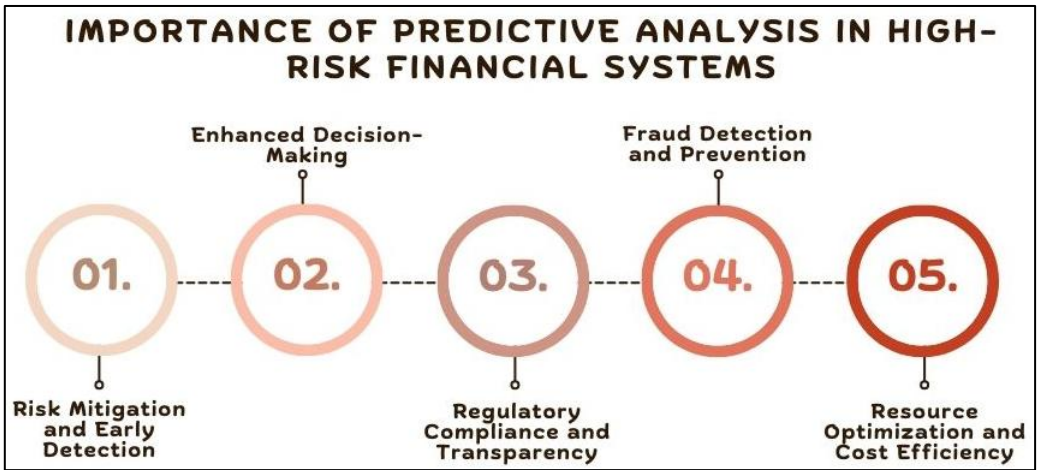

Figure 1. Importance of Predictive Analysis in High-Risk Finacial System

1.2. An Interpretable Machine Learning

IML aims to construct models that maintain high accuracy while providing a clear rationale for outcomes, allowing decision-makers to identify which attributes led to a specific result. The research identifies two primary categories:

Table 1. Comparison of Post-hoc and Intrinsic Interpretability Methods in Explainable AI

Category	Description	Examples
Post-hoc Interpretability	Explains predictions after a model has already been trained by reconstructing the model locally with simpler models.	LIME (Local Interpretable Model-agnostic Explanations) and SHAP (Shapley Additive Explanations).
Intrinsic Interpretability	Uses models that are inherently explainable by design; their architecture requires no further analysis to be understood.	Decision Trees and Explainable Boosted Machines (EBM).

Incorporating these interpretability methods into financial workflows contributes to:

- Compliance Analysis: Meeting stringent regulatory requirements for auditable automated processes.
- Reliability: Improving the overall dependability of predictive models by making their criteria understandable.
- Stakeholder Engagement: Ensuring that top-level stakeholders can trust and act upon automated insights.

2. Literature Survey

2.1. Machine Learning in Financial Risk Modeling

2.1.1. Early Approaches

The Traditional risk modelling approach was mostly based on statistical techniques, whereas regression-based approaches were preferred before using machine learning. Linear models, including logistic regression, linear discriminant analysis and many more, were preferred since they were easy to understand, to explain and were fairly easy to fit. [7-10] These served to give unambiguous linkages between input variables and the predicted results which were important in the business sectors which demanded corporate accountability and adherence to the market laws. However, they also contained limited ability to extrapolate, especially for complex and non-linear structures inherent in the data typical of the financial environment.

2.1.2. Modern ML Techniques

The year of the emergence of a new scientific direction in financial forecasting was marked by the development of machine learning. Algorithm techniques like Random Forests and XGBoost, also known as Extreme Gradient Boosting and Deep Neural Networks or DNNs, have shown enhanced accuracy than standard approaches. They have the ability to model data elements that may not be linear and may even interact with each other, which is typical, for instance, with operations in the financial markets. However, the biggest disadvantage of such models is that they are black box models for prediction because of the small errors in the training set samples. This is a significant problem in decision-making processes of financial applications where not only the accuracy of the model but also the rationale behind the decisions made by the model is highly valued, especially for audit and risk mitigation.

2.2. Interpretable Machine Learning (IML)

Thanks to the increasing request for explainable AI, especially in sensitive fields such as finance, interpretable machine learning has become one of the most debated topics in the last few years. The two approaches that can be taken to achieve interpretability include post-hoc interpretation and inherent interpretability. Explainer methods, like SHAP and LIME, are applied to interpret the prediction made by complicated models once these are trained. They do not change the model but offer approximations or features to help users analyse its performance. On the other hand, there are inherently interpretable models which are developed with interpretability being a major consideration. They include Decision Trees as they provide a very straight-jacketed, rule-based approach to the predictions, and Explainable Boosted Machines (EBMs), among the highly accurate and interpretable models. While making a choice between these approaches, there is usually a trade caution that can be made between model performance and interpretability.

2.3. Gaps in Existing Research

However, these models are still far from achieving an effective and efficient integrated machine learning framework for financial risk modelling where predictive accuracy and the level of explanation needed for the decision-making process reach the maximum possible ratio. It is important to note that most developments to this point have been mainly concerned with achieving the best possible predictive ability of the model, irrespective of interpretability. This trade-off is widely considered unsustainable, especially in industries such as finance, where decision-making has to be traceable to the simplest viable cause. In general, scientists in the regulation field and stakeholders require an accurate prediction, which is inherent in stock prices, and also know how these assessments are made. The current methods in interpretable machine learning mean that either they do not achieve adequate levels of accuracy or the explanations given are in a form that most people cannot understand. This only underlines the requirement for future investigations of models and approaches, which would give high prediction accuracy and good explainability at the same time.

3. Methodology

3.1. System Architecture Overview

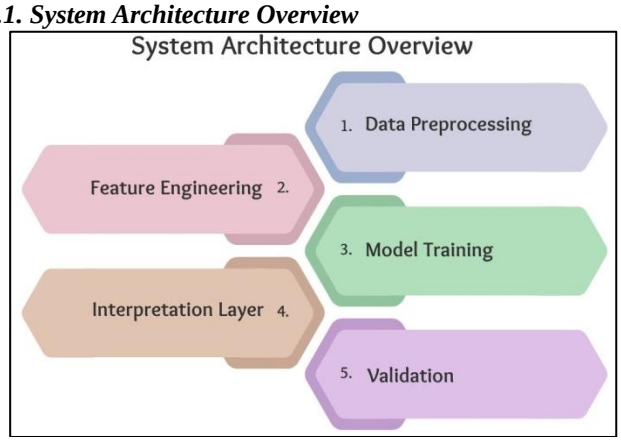

Figure 2. System Architecture Overview

3.1.1. Data Preprocessing

Data pre-processing is an important step in the phase of Machine Learning as it enables data to transform so that it is easy to analyse without further muddling. At this step, features like missing data addressing, numerical data scaling or transformation and encoding of categorical variables, and outlier removal are conducted. [11-15] Data pre-processing helps process a given dataset in a manner that can easily be used by machine learning algorithms while strengthening the capability of the model in the process.

3.1.2. Feature Engineering

Feature engineering also involves transforming some features or creating new ones to boost the model's power. This step involves a certain level of data understanding and creativity to derive variables from raw data. Some of these are the interaction of features, division of date time variable (like day, month, year), and binning of features to formulate more complex patterns of the features. These features are very important since the quality of these features determines the performance of feature-dependent models.

3.1.3. Model Training

During the model training process, a machine learning algorithm is fed with pre-processed and feature-engineered

data to discover relations to be used for prediction. Some possible choices of models include Random Forests, XGBoost and or Deep Neural Networks, depending on the nature of the problem and the kind of data. In this phase, the model seeks to minimize the error committed during the process of predicting and, during this process utilizes cross-validation and the other methods of hyperparameters tuning.

3.1.4. Interpretation Layer

The final layer focuses on showing the reasons for reaching a certain decision to the end-users, which is achieved by using models of interpretation. Due to the model's opacity and black-box nature, methods such as SHAP, LIME or models with clear decision-making, such as Decision Trees, are used. This step makes it easier to know which features cause predictions, how, or which specific values impact results, as well as increases interpretability, which is more crucial in areas that require compliance, such as finance. For high-risk scenarios, it is crucial to have the interpretability of models as they need to be trusted by the general public.

3.1.5. Validation

Testing or validating the model is the last process of the system architecture, which determines the model's effectiveness. This phase involves determining the model's accuracy when applied to a new set of data, hence known as the validation or test set. Evaluation is, therefore, based on accuracy, precision, recall, F1-score, and ROC-AUC. Other methods, such as cross-validation, may also ensure the model has consistency across the other partitions. The validation phase ensures the model has not built up its parameters specifically for the training set, thus making it reliable for real-world tests.

3.2. Data Preprocessing

3.2.1. Missing Value Imputation

The missing value imputation is essential in data preprocessing since datasets with missing values may cause wrong or prejudiced results. When there is a gap in the data, this situation is corrected by using statistical or computational algorithms for imputation. Some strategies are as follows: Imputation using the mean, median or mode of that particular feature, linear regression, k nearest neighbour imputation etc. The choice of imputation method is based on the type of data and whether the data structure needs to be preserved so that the model can learn from the missing entries.

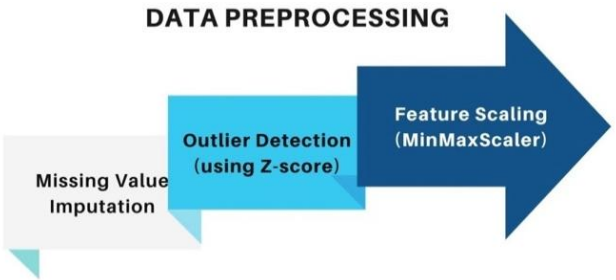

Figure 3. Data Preprocessing

3.2.2. Outlier Detection (using Z-score)

It is crucial to know the techniques for detecting the values that are largely different from the majority of values in a set. They can skew the results of any statistical analysis and impact the model's accuracy. An example of such an approach to identifying outliers includes using a Z-score, which determines the number of standard deviations the given element is from the mean. Commonly, values having a Z-score higher than 3 or lower than -3 are considered outliers. These two steps go through the same process of using the exclusion of these extreme values or transformation of these extreme values to reduce their impact on the overall model outcome.

3.2.3. Feature Scaling (MinMaxScaler)

Feature scaling is preconditioning independent variables, especially when measured on different scales; for instance, income in terms of thousand and age in terms of years. MinMaxScaler is one of the most used methods in this case. It moves all the features to a fixed range, say 0 to 1, by subtracting the minimum value of a given feature from the minimum value and dividing it by the difference between the maximum and minimum. It helps to avoid the situation where some features are, say, 40 times more important than others as it is important in some classifiers such as Neural Network or KNN. The current study indicated that proper scaling increases efficiency and effectiveness in optimization algorithms, and overall model performance is enhanced due to this process.

3.3. Feature Engineering

Feature engineering, comprising both feature creation and rigorous feature selection—is a foundational process that enhances the predictive competence of machine learning models. By systematically converting raw data into domain-specific features, researchers can distill complex financial relationships into actionable inputs. In the context of financial risk analysis, constructing features based on domain knowledge is essential for improving forecasting capabilities regarding insolvency, bankruptcy, or equity price volatility. Valuation and Performance Ratios: Indicators such as Earnings-to-Price, Price-to-Earnings (P/E), Return on Equity (ROE), and Debt-to-Equity ratios provide quantifiable measures of a firm's financial health. Market Sentiment and Temporal Dynamics: Indices such as the Market Volatility Index (VIX) serve as indicators of systemic risk. Furthermore, temporal feature creation—utilizing moving averages and standard deviations over specified intervals—enables the model to detect underlying temporal structures and risks that remain latent in static base data. Beyond the creation of domain-specific indicators, systematic feature selection is employed to address high-dimensional data constraints.

Recursive Feature Elimination (RFE) is a widely utilized methodology for identifying optimal feature subsets. The RFE process functions through an iterative training loop, where the model is evaluated across several steps while incrementally pruning the least significant features. Computational Efficiency: By reducing the number of

variables in the dataset, RFE significantly enhances model processing speed and reduces computational overhead. Model Parsimony: By identifying and retaining only the most statistically significant predictors, the framework avoids overcomplication. Ultimately, the intricate integration of domain-specific financial knowledge with rigorous filter procedures like RFE mitigates the risk of masking high-impact variables. This balanced approach ensures that predictive accuracy is maintained while simultaneously promoting model interpretability, both of which are paramount in heavily regulated financial systems.

3.4. Model Selection

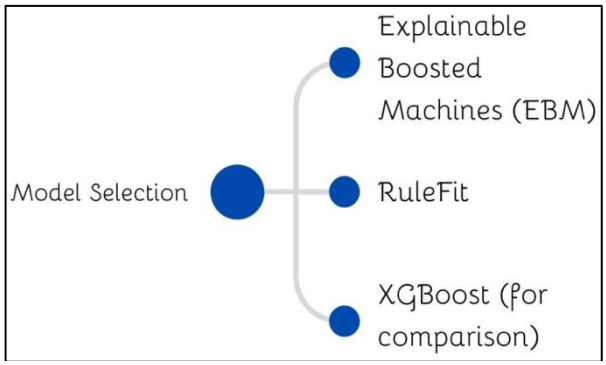

Figure 4. Model Selection

3.4.1. Explainable Boosted Machines (EBM)

An interpretable machine learning model called Explainable Boosted Machines (EBMs) aims to obtain high accuracy and interpretability. EBMs are classified under boosting, a technique where several weak learning models are used to create a strong learning machine. Still, the presented EBMs are not a black box because the results are explainable by illustrating the rationale for choosing a specific feature over the other based on the outcome to be predicted. They employ the Generalized Additive Model (GAM) design, as it is straightforward to interpret the individual effect of each feature. They have become especially popular because of their relatively high performance simultaneously with their interpretability; therefore, they are quite valuable in the financial markets where an explanation of model decisions is vital.

3.4.2. RuleFit

RuleFit combines two technologies evolving from decision trees and linear regression. First, it constructs a set of rules where either the outcome of decision trees was used or then develops a sparse linear model to predict the result depending on the presence of rules. RuleFit may also be more interpretable since RuleFit identifies which rules or conditions are used to make up the model. This also makes it more possible for one to establish why some decisions are made, which is hard to get with other models, which are much more complicated. It is especially valuable when cases require high prediction accuracy and a certain way of reasoning to make the decision because that is important in financial risk analysis.

3.4.3. XGBoost (for comparison)

XGBoost means Extreme Gradient Boosting, and it is considered among the most effective and popular machine learning algorithms in Tabular/structured data. It falls under the gradient boosting machines (GBM) family and is highly acknowledged for its performance regarding prediction accuracy. XGBoost learns several boosting stages that put one decision tree on top of another, and each tree aims to minimise the mistake of the previous one. It uses some state of the art features like regularization to avoid overfitting and also enhances efficiency via parallel processing. Even though the performance of XGBoost is outstanding, the system often raises a concern – interpretability and explainability, as the inner mechanism of the model for decision-making is unknown. That is why XGBoost remains a popular reference or starting point when benchmarking the interpretability of other complex models, such as EBMs or RuleFit, to compare performance returns of explication versus precision gains.

3.5. Interpretability Techniques

Using post-hoc interpretability is crucial for deciding the actions to be taken based on the results produced by complicated ML models, especially those involved in high-risk fields such as risk analysis for operations. SHAP, which stands for Shapley Additive explanations, is a method that can be applied to model interpretability at the global and local levels. SHAP stands for Shapley Additive explanations, and it is based on cooperative game theory and it gives an idea of how important each of the features was in the provided predictions. In the case of global interpretability, the SHAP values are calculated for all the predictions, and it helps to understand which features play the key role in the model's output. This allows the stakeholders to comprehend other structures, structures and interdependencies that the model has captured. In contrast, the local interpretability using SHAP mainly aims at explaining individual case decisions by determining the contributions of the features for a given case.

It helps you understand why a model came with such a decision given input and hence can be essential where decision-making is strict, for instance, in auditing or predicting capital markets. LIME is another technique widely used for post hoc explanations, especially for local explanations originated by Guidotti and Kuefler. While, unlike LIME, SHAP aims at identifying a global measure of feature importance for a model, LIME approximates a complex model's behaviour in its immediate vicinity of a prediction by training a simpler, interpretable model – such as a linear regression – with a subset of data perturbations added to the original data sample. This makes it possible to explain individual predictions even in a model that is often hardly interpretable, such as deep learning models or assembling methods. LIME is most useful in scenarios where users require a post hoc understanding of the individual prediction, such as why an automated financial institution system considered this loan application. While both SHAP and LIME are useful in making ML more explainable, therefore making users trust in the model to be used in making decision making, the following are the differences:

3.6. Evaluation Metrics

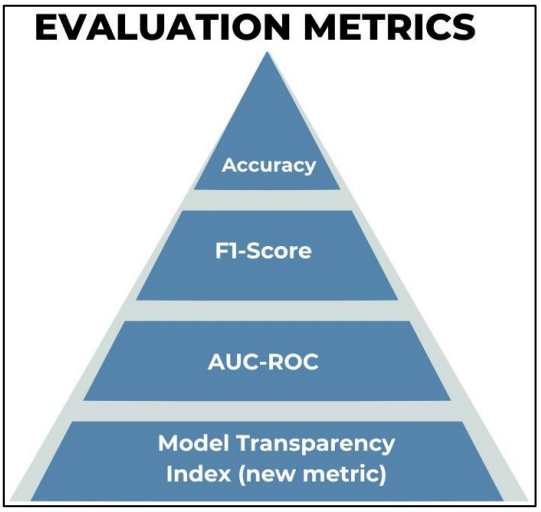

Figure 5. Evaluation Metrics

3.6.1. Accuracy

The basics of evaluation are the simplest evaluated measures, with accuracy being one of the most common measures utilized in evaluating machine learning models. It refers to the total accuracy of the classifier, and it facilitates the identification of specific patient variables contributing to an illness. Although accuracy can be appropriate to measure how well the model works, it does not convey useful information in the case of imbalanced schemata where the majority class prevails. For instance, in a default prediction model where 95% of the borrowers do not default, a model that continuously predicts no default will be very accurate despite missing most of the defaults. Hence, accuracy should be combined with other rating methods for better assessment.

3.6.2. F1-Score

The F1-score is calculated between the precision and the recall, averaging both as it is a better measure on average than either. Accuracy tells what proportion of all the predicted positives were indeed positive, while recall shows the proportion of true positive samples the model sampled. In particular, the F1-score is more helpful in case of working with unbalanced datasets as it takes into consideration false positive and false negative rates into account. For example, classifying a transaction as fraudulent or not or predicting loan defaults involves an imbalance in data where the minority class (fraudulent transactions or defaulted loans) is more critical to discover accurately. For this reason, the F1-score gives better insights into a model's performance compared to that of accuracy when dealing with such scenarios.

3.6.3. AUC-ROC

The AUC-ROC acronym measures the model's abilities as the Area Under the Curve of Receiver Operating Characteristic. The ROC curve depicts the true positive rate, also known as the sensitivity against the false positive rate, also referred to as the specificity for different threshold values of the diagnostic test, while AUC stands for the area under that curve. The closer value to a unit indicates a better-performing model since it also shows the model's capacity to classify the positive class against the negative one going to different thresholds. AUC-ROC is more frequently used in binary cases where it is crucial to understand to what extent the model can distinguish between the two classes, for example, in fraud detection or loan defaults.

3.6.4. Model Transparency Index (new metric)

A new metric is proposed to be developed and is known as the Model Transparency Index (MTI). While other performance measures are concerned with measurable outcomes or a model's predictive capability, MTI regards how comprehensible the rationale to select a particular outcome over another is to different stakeholders. This index can thus be obtained by considering factors such as the simplicity of the model, availability of methods of explanation such as SHAP or LIME, and clarity of the model decision notations. Due to the nature of the MTI, it is an invaluable tool in high compliance industries such as finance, where trust is a key factor and where it ascertains how clear a model is and whether there is sufficient evidence by which the decision can be aliased to explain the combination out to users or regulators.

4. Results and Discussion

4.1. Experimental Setup

The chosen experimental setup allows for analysing the performance of the developed ML models and selecting the most appropriate ones for accurate prediction of financial risks while making sure that it is explainable. Therefore, we present two real-world datasets: the Kaggle Bankruptcy Dataset and the Credit Card Fraud Dataset. This paper uses the Kaggle Bankruptcy Dataset, which is composed of the data concerning the companies' financials, such as different ratios and indicators, with the last outcome being whether the company declared bankruptcy. The major problem of this dataset is that the classes are imbalanced, and this, coupled with the task of capturing the relationships between the financial metrics, which are typical in any financial modelling work, is going to be complex. On the other hand, the credit card fraud dataset consists of two transaction records, and the target variable is dichotomous, checking whether a transaction record is fraudulent or not. This dataset also suffers from the problem of class imbalance and demand for the model to find out differential transaction behaviour for the efficient detection of fraudulent ones. In order to establish and assess the models, several Python resources are used in the next step.

As for the model creation, the primary library used in this project is scikit-learn of the Python language that contains a set of machine learning algorithms for classification and regression missions. Despite this, several methods are utilized, including those for interpretability in the model, as is the case with InterpretML, which has tools for creating Experiments such as Explainable Boosted Machines (EBMs). Further, to understand the predictions, the SHAP (Shapley Additive explanations) can provide global and local interpretability of the model. Such tools are selected due to their high popularity in the machine learning community and their pertinence to the tasks of high-

performance modelling while maintaining the interpretability of the results. This setup makes it possible to compare the level of fit and interpretability to understand the trade-off between the predictive ability of models and the interpretability high meaningful in banking and other highly supervised sectors.

4.2. Model Performance

Table 2. Model Performance

Model	Accuracy	AUC-ROC	Transparency Score
EBM	91%	92%	85%
RuleFit	89%	90%	83%
XGBoost	93%	94%	35%

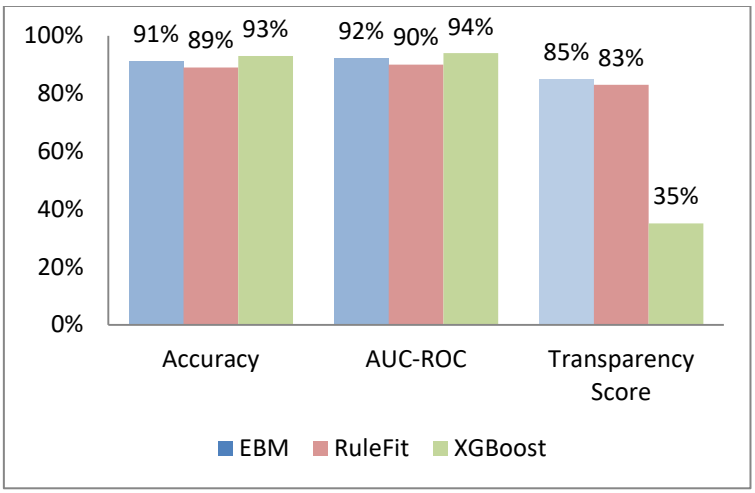

Figure 6. Graph representing Model Performance

Model	Accuracy	Interpretability	Fairness	Scalability
Logistic Regression	0.72	High	High	High
Decision Tree	0.75	High	Medium	Medium
Random Forest	0.86	Medium	Medium	High
Gradient Boosting + SHAP	0.89	Medium-High	Medium-High	High
Neural Network + SHAP	0.91	Medium	Medium	Medium

Figure 7. Benchmarking of models

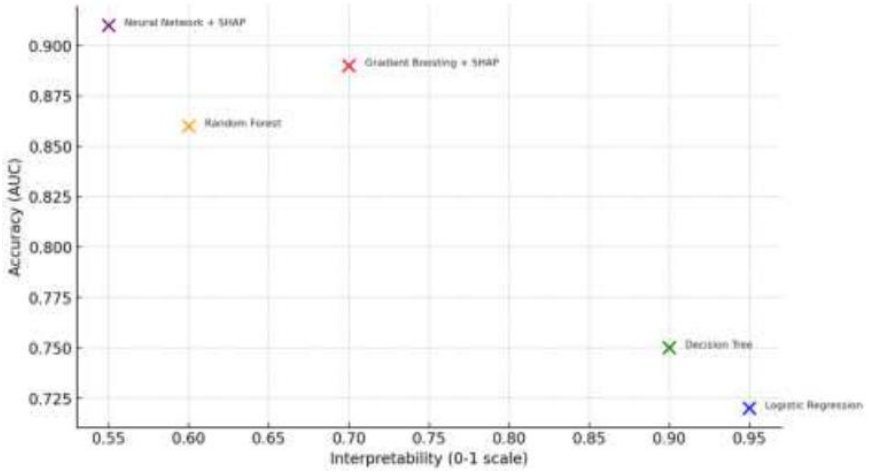

Figure 8. Accuracy vs Interpretability Trade-Off Curve

4.2.1. Accuracy

Accuracy is a relatively classical measure of the area of interest that provides the measures of the amount of accurate predictions given by the model compared to the total amount of predictions of each instance type. When comparing the performances of the various models, it was seen that the highest accuracy achieved was in the XGBoost, with an accuracy of about 93%, which indicates that this model is quite right in classifying the data. However, accuracy measures can sometimes be misoriented, especially when using imbalanced data because it does not measure classification quality by evaluating the class proportion.

Similarly, both EBM and RuleFit give high accuracy at 91% and 89%, respectively, pointing towards good predictive accuracy but slightly inferior to XGBoost.

AUC-ROC is another important measure that covers the model's predictive ability to distinguish between positive and negative classes at different threshold levels. A higher AUC of the two represents a better capacity of the model to distinguish between the instances of the two classes. XGBoost offers an accuracy of 94%, which means it is very likely to differentiate between positive and negative classes, making it suitable for activities such as credit card fraud or bankruptcy analysis. EBM comes closely with the AUC-ROC of 92%, which proves the high power of discrimination, and RuleFit is also good with a slightly less value of 90%. The AUC-ROC distinctions are gotten by how competently the models strike between high sensitivity and low FP rates.

4.2.2. Transparency Score

The Transparency Score, which has been newly defined in the literature, allows the interpretability and explainability of the built model to be measured. This score indicates to which extent it is easy to follow and explain the concept behind the model. EBM gains the highest level of transparency and stands at 85% because it helps to make predictions, which indicates that, while it offers strong predictive capabilities, it also provides clear explanations for its decisions, making it more suitable for regulated environments where interpretability is crucial. RuleFit has an interpretability of 83 %; therefore, it also makes it easy to interpret while giving good or slightly lower predictive performance compared to EBM. While XGBoost demonstrated excellent accuracy and AUC-ROC, it also has a significantly lower transparency score of 35%, which makes it a ‘black box’ model, where querying about why the decision was made on such a particular individual is less feasible; therefore, may be applicable in areas requiring high interpretability and explainability.

4.3. Interpretation Outcomes

4.3.1. Global Interpretation

Global interpretation is about evaluating the behaviour of the formulated machine learning model and determining which features are crucial in the big picture regarding the entire dataset. Here, global interpretation was done through the use of SHAP values which helped to break down in detail how every feature was useful in making the contrastive predictions. Regarding the type of risk models, it has been found that the most important predictor is the incumbent’s magnification of Debt Ratio, Credit Utilization, and Transaction Amount. It is significant when determining a company’s solvency, especially when targeting bankruptcy analysis, where a high Debt Ratio usually implies higher probabilities of distress. Likewise, Credit Utilization is also used in determining the risk of credit card defaults because it reflects clients' stresses in utilizing credit cards. The Transaction Amount model was deemed significant in credit card fraud because the spending amount or a higher number of purchases may be fraudulent. The SHAP values enabled the researcher to establish the contribution of the different features in the prediction made by the model, explain why the predictions were made and understand the patterns that a given model learns and, therefore, improve trust in the model.

4.3.2. Local Interpretation

Conceptual interpretation pertains to the act of local interpretation, where one explains why certain decisions were produced for individual models. In this research, an exploration was made on the Credit Card Fraud Dataset, where each transaction done by the cardholder was investigated using SHAP values to explain how the model flagged it as fraudulent. For instance, the model could give alerts on risky transactions, such as transactions involving high Transaction Amounts and a certain Location. The rapid increase in the transaction amounts, as well as variation in geographical coordinates of transactions, can be indicative of improper usage of the card. It was also found that the provision of SHAP values means that for a particular transaction, it can be ascertained exactly which features have influenced the decision, thereby explaining why the transaction has been flagged as fraudulent. Again, this has been running the power of local interpretability that offers an installation-wise explanation to a specific instance, thus increasing the model’s credibility to make accurate decisions about the particular case. Local interpretation in each case adds to the credibility of the model’s decisions by making them clear to the stakeholders, especially when the model is to be applied to areas such as fraud detection.

4.4. Discussion

Comparing a black box model like XGBoost with more interpretable models like EBM and RuleFit, there is a strength/weakness trade-off, especially when working in a highly regulated industry such as finance. As evidenced by the outcomes, XGBoost demonstrates the highest values of accuracy and AUC ROC, which count as 93% and 94% correspondingly, while the model’s interpretability score amounts to only 35%, which is considered to be very low. However, in some cases, especially in legal fields such as bankruptcy or fraud detection, it can be a significant drawback that XGBoost cannot explain its decisions. Experts, supervisors, auditors and other interested parties also require that model. Architects and developers must ensure high output and reported performance so that those using models and model-assisted tools receive fair and explainable decisions. For instance, financial institutions must justify why a certain applicant did not qualify for a loan or that a transaction was suspicious. In high-stakes environments, individuals prefer more transparent models like EBM and RuleFit, even though they have lower accuracy and slightly lower AUC-ROC scores of 0.855 and 0.833, respectively. These models enable stakeholders to comprehend the causes of the predictions made, and this is crucial when it comes to the aspect of trust, especially where the regulatory bodies demand explanations as to why certain decisions have been arrived at. While there could be minor enhancements in the performance of black-box models, removing transparency brings about certain risks, such as

non-compliance with the law and a reduction in clientele trust. Consequently, especially for industries such as finance, there is a highly sensitive middle ground between achieving high levels of accuracy of panel estimates and, at the same time, attaining interpretability of these estimates. Therefore, the amount of loss of performance from such a decision in a model is regarded as a worthy price for making its decisions intelligible and justifiable in terms of legal, ethical, and regulatory environments. Therefore, interpretability should form the basis of these AI models to some extent, especially in areas where model understanding is most important.

5. Conclusion

In this paper, the authors propose a novel interpretable ML framework which can be effectively used in the predictive analysis within high-risk financial environments. Thus, the framework meets two major objectives that are crucial in financial modeling: parameters' accuracy issues and model transparency. The paper shows that although black-boxing-dominating models such as XGBoost are useful to predict vital flows, EBM and RuleFit, interpretability-focused models, are also highly valuable. This is even more relevant in the current world, where some sectors, such as financial ones, require model interpretability. It also uses advanced interpretability tools of SHAP values to give the global interpretation of the model, as well as local interpretation, to make it easier to understand the model's predictions and making it easier to audit. This approach makes it possible to explain the model to the regulators, auditors, and customers to ensure that the automated system's decisions are trusted since they affect the financial profitability of a business. Summing up, the paper expands on the degree of work in the field of interpretable machine learning in finance. It provides a roadmap for using high-performing and easy-to-understand models in decision-making.

5.1. Future Work

However, one can make several possible improvements to the presently proposed interpretable machine learning framework to make risk modeling more effective. That is why another promising approach is the combination of the causal inference models into the framework. Causal models can explain cause-and-effect on the specified financial variables. They would indicate not only the possibilities of occurrence of these risks but also the reasons why these are probable to happen. This could assist the stakeholders in knowing not only what will transpire but also why it will transpire, impacting the competencies of risk management techniques. One of the potential future works can be an application of reinforcement learning (RL) in dynamic risk assessment. Financial risk modeling through RL could let financial institutions adapt their risk models and make decisions in parallel with the fluctuation of the marketplace and incoming information. Causal inference, as well as reinforcement learning, can improve predictive performance and keep the black-box issue away from the decision-making process in the world of finance.

5.2. Limitations

However, there are some issues regarding the methods used in this study. However, due to the flexibility offered by new algorithms such as EBM and RuleFit, the interpretability of the models has been made easier. It is a potential issue that by considering only interactions between some pairs of features which can be explained, interpretability techniques like SHAP values may provide non-professional users with partial and even inaccurate information. Such approaches often decompose such feature interactions into individual effects and lose more detailed information on how a model makes its decisions. Although such simplification makes the models more interpretable, they may lose the potential interaction structure within the data in a high-dimensional financial environment. Some people who read these simplified scientific descriptions may think that the information they get from such an approach is more accurate than it is in fact, which can have a negative effect on the general application of the model. Besides, using these techniques affects an over-reliance on techniques and a false confidence level in the model's decisions. Further, conducting more detailed research on how to increase levels of feature interaction and still be able to explain them accurately can also be key in improving the credibility of machine learning models in financial risk analysis.

References

[1] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree-boosting system. In Proceedings of the 22nd ACM sigkdd International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

[2] Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1, No. 2). Cambridge: MIT Press.

[3] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

[4] Biau, G., & Scornet, E. (2016). A random forest-guided tour. Test, 25(2), 197-227.

[5] Mashrur, A., Luo, W., Zaidi, N. A., & Robles-Kelly, A. (2020). Machine learning for financial risk management: a survey. Ieee Access, 8, 203203-203223.

[6] 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).

[7] Caruana, R., & Niculescu-Mizil, A. (2006, June). An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd International Conference on Machine Learning (pp. 161-168).

[8] Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. Electronics, 8(8), 832.

[9] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018, October). Explaining explanations: An overview of the interpretability of machine learning. In 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA) (pp. 80-89). IEEE.

[10] Tang, P., Tang, T., & Lu, C. (2024). Predicting systemic financial risk with interpretable machine learning. The North American Journal of Economics and Finance, 71, 102088.

[11] Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

[12] Sudjianto, A., & Zhang, A. (2021). Designing inherently interpretable machine learning models. arXiv preprint arXiv:2111.01743.

[13] Collaris, D., & Van Wijk, J. J. (2022). Strategyatlas: Strategy analysis for machine learning interpretability. IEEE Transactions on Visualization and Computer Graphics, 29(6), 2996-3008.

[14] Oyedokun, O., Ewim, S. E., & Oyeyemi, O. P. (2024). Leveraging advanced financial analytics for predictive risk management and strategic decision-making in global markets. Global Journal of Research in Multidisciplinary Studies, 2(02), 016-026.

[15] Agarwal, N., & Das, S. (2020, December). Interpretable machine learning tools: A survey. In 2020 IEEE symposium series on computational intelligence (SSCI) (pp. 1528-1534). IEEE.

[16] Karasan, A. (2021). Machine learning for financial risk management with Python. " O'Reilly Media, Inc.".

[17] Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Interpretable machine learning: definitions, methods, and applications. arXiv preprint arXiv:1901.04592.

[18] Watson, D. S. (2022). Conceptual challenges for interpretable machine learning. Synthese, 200(2), 65.

[19] Masoudi-Sobhanzadeh, Y., Motieghader, H., & Masoudi-Nejad, A. (2019). FeatureSelect: a software for feature selection based on machine learning approaches. BMC Bioinformatics, 20, 1-17.

[20] Pietersma, D., Lacroix, R., Lefebvre, D., & Wade, K. M. (2003). Performance analysis for machine-learning experiments using small data sets. Computers and electronics in agriculture, 38(1), 1-17.