International Journal of AI, Big Data, Computational and Management Studies
Noble Scholar Research Group | Volume 2, Issue 1, PP.57-65, 2021
ISSN: 3050-9416 | https://doi.org/10.63282/3050-9416.IJAIBDCMS-V2I1P106

Explainable Machine Learning for Loan Default Prediction: Enhancing Transparency in Banking

Amit Taneja
Lead Data Engineer at Mitchell Martin, USA.

Abstract: Traditional credit risk assessment frameworks have changed over the last few years with the integration of Machine Learning (ML) frameworks in the field of financial services. The predictive models have greatly improved in the accuracy of loan default. Nonetheless, the obscurity of a significant number of ML frameworks has triggered some issues towards transparency and interpretability, as well as regulatory adherence. The present paper reviews the context of applying Explainable Machine Learning (XML) approaches to the prediction of loan default to achieve trade-offs between predictive power and explainability. The traditional or so-called black-box models (XGBoost, Random Forest, deep learning) are contrasted to the interpretable models or post-hoc explanation methods (SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), and Decision Trees). We are using a real-life financial dataset (pre-2021) of LendingClub to perform the analysis of model performance. The paper highlights the importance of citizen trust and transparency in the banking system, examines the challenges facing financial institutions, and discusses how explainable AI can foster trust among customers, enhance the ethical training of financial institutions' AI, and ensure regulatory compliance. There is visualization, a heatmap, and the graphs depicting the importance of drawing clearer conclusions. We find that explainable models, albeit marginally less effective, are crucial in terms of the understanding of behavior of the borrowers and in the financial risks they represent. The characteristics of the findings urge the use of XML in credit scoring pipelines to make responsible and ethical AI implementation.

Keywords: Explainable AI, Loan Default Prediction, Credit Risk, SHAP, LIME, XGBoost, Machine Learning, Financial Transparency, Banking Regulation.

1. Introduction

Loan default prediction is of critical importance to the wider context of credit risk management in banking and financial activities. The proper identification of the probability of default will allow the financial institutions to reduce Non-Performing Assets (NPAs), make optimum usage of capital and facilitate financial stability. Credit scoring systems and statistical models like logistic regression have traditionally gauged creditworthiness. These have been complemented, however, with Machine Learning (ML), which produces a much better predictive power of these models, and allows making more fine-precision risk assessments by identifying nonlinearities and complications in the behavior of borrowers. The ML models frequently work as a "black box." They do not provide transparency in the prediction process. [1-4] Despite their benefits, this is the advantage that can be discussed. Such opacity is unacceptable, given the issue of interpretability, in fields like finance, where the decisions made impact the credit accessibility of people, and are to be observed under the strict conditions of regulation. Greater and greater focus is being placed on transparent, explicable algorithm-based results by regulators, stakeholders, and consumers. Credit decision justification capability is not only a regulatory requirement (e.g. compliance with legislation such as the Fair Credit Reporting Act or GDPR) but also a key part of trust and responsibility in the context of automated processes.

1.1. Importance of Machine Learning for Loan Default Prediction

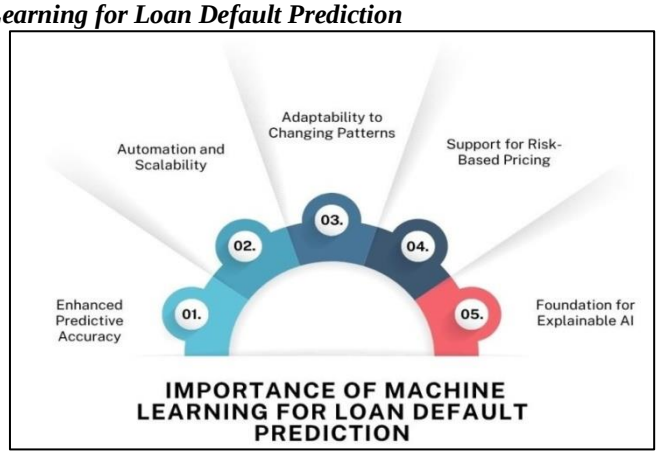

Figure 1: Importance of Machine Learning for Loan Default Prediction

- **Enhanced Predictive Accuracy:** The application of Machine Learning (ML) algorithms in predicting the occurrence of loan default is an improvement compared to the traditional statistical tools. Although traditional methods, such as logistic regression, consider pre-determined assumptions and linearity between two variables, ML models can learn and infer non-linear complex patterns (called high-dimensional) when considering big datasets. Learning algorithms like the Random Forest, XGBoost, and Neural Networks have the capability of revealing latent interactions between attributes, where the variables of interest might be income, credit history and debt-to-income ratio, to mention a few, and these associations cannot be easily identified through manual means. This can help raise the true predictive validity and produce a more dependable outcome that can assist lenders in improving their control over what happens to the greater risk borrowers and hopefully limit credit losses.
- **Automation and Scalability:** ML models are very scalable and can provide operations on huge amounts of data with little to no manual participation. There has been an influx of data about customers of their financial institution to the financial institution, through a number of sources determined by the current digital lending environment, such as wrapping online applications, transaction history and social media trails. This data can be learned and processed by ML efficiently, which allows making credit decisions in near real-time. The automation enhances efficiency in the operations and the speed of loan approval so that lenders and borrowers have a better experience.
- **Adaptability to Changing Patterns:** AWS is one of the most important advantages of ML, which allows for modification in connection with the changing behavior of borrowers and the processes on the market. In contrast to the fixed rule-based systems, ML models can be retrained with new information, so they can maintain their applicability in changing economic environments. Such flexibility is particularly valuable in critical times, when financial uncertainty or a crisis prevails, and what has been the norm in history need not always continue.
- **Support for Risk-Based Pricing;** ML models have the potential of allowing risk-based pricing strategies to be applied by lenders by the accurate estimation of the probability of default. It is possible to tailormake interest rates and loan terms given to the borrower with respect to their personal riskiness, which translates into a more tailored financial products and enhanced portfolio performance by lenders.
- **Foundation for Explainable AI:** Although machine learning models can be complex, they are the basis upon which Explainable AI (XAI) techniques become integrated. SHAP and LIME are two of many tools which can be used to analyze and gain insight into decision-making using ML models, with the unique balance between the power of prediction and the desirability of transparency, which are valuable in a regulated industry such as finance.

1.2. Enhancing Transparency in Banking

Trust is a key element in banking, and transparency plays a crucial role, particularly in the context of digital transformation, where automated decision-making is gaining momentum. With the increased use of advanced machine learning models in financial institutions to increase the accuracy of their credit risk evaluation, there is increasing worry about the so-called black-box nature of these models. Customers may feel alienated if they lack clear explanations for decision-making processes, such as loan approvals or rejections. Furthermore, regulators are putting tougher requirements on algorithmic accountability, in which the decisions impacting consumers should be explanatory, auditable and justified. This has rendered the principle of improving transparency no longer a best practice but a legal and ethical requirement in the banking sector. With reference to a loan default prediction, transparency implies clarification of the reasons why a borrower is marked as a high-risk or low-risk case. Conventional models such as logistic regression have the merit that they are intrinsically interpretable, but they tend to lack predictive abilities. Conversely, more complicated models such as neural networks and gradient boosting machines provide better performance but lack simple explainability. To solve this trade-off, banks are resorting to the more explainable forms of AI, also known as Explainable AI (XAI), SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations). They enable institutions to produce a human-readable explanation of why the models made a decision, and it shows which features (e.g., income, debt-to-income ratio, or credit history) had the highest impact on a given result. There are various advantages of increasing transparency. It enhances customer confidence and satisfaction, as borrowers can appreciate and question decisions. It enables its adherence to data protection and fair lending regulations, like the GDPR and Equal Credit Opportunity Act, among others. Besides, it also promotes internal governance, providing an opportunity to audit and monitor model behaviour through model auditing and risk management teams. The transparency is the solution in the sense that it will close the technological gap against ethical and accountable banking.

2. Literature Survey

2.1. Traditional Credit Risk Models

Statistical models (such as logistic regression) and conventional credit scoring models (such as FICO) of credit risk assessment have been used in credit risk assessment practice for many years. Such models are very interpretable and, as such, can easily explain the manner in which the credit decision is made, which is also crucial in avoiding scrutiny by the regulators. [5-8] Nonetheless, their simplicity usually constrains them to represent non-linear associations in borrower behavior, like correlations between different financial indicators or time-related behavior patterns. Although models such

as decision trees have a marginally better flexibility and are comparatively much more interpretable, they cannot compare with more complex models in terms of predictive capability. The table below points out the trade-offs between accuracy, interpretability, and compliance of the training paradigm with traditional and machine learning-based models.

2.2. Machine Learning in Finance

Machine learning (ML) has disrupted different functions of the financial system to better make claims based on more informed and flexible decisions. A neural network can identify complex patterns embedded in the transaction data that can be overlooked by the conventional rules-based systems in fraud detection. Reinforcement learning algorithms are also being used in portfolio management to alter investment strategies dynamically, by using real-time market conditions and feedback. In the case of credit risk, ensemble learners, including random forests and gradient boosting (e.g., XGBoost), have been shown to outperform by combining several weak learning models and can model complex interactions between data variables. Such developments, nonetheless, are accompanied by difficulties, especially transparency and the necessity of explanations in a regulated setting.

2.3. Explainable AI (XAI) Frameworks

The emergence of intricate ML systems has necessitated the use of Explainable Artificial Intelligence (XAI) solutions, which can give explanations on decisions made by the model. SHAP (SHapley Additive exPlanations) is an inductive method of fairly assigning each feature of the input importance by applying cooperative game theory to provide local and global interpretations of model results. LIME (Local Interpretable Model-agnostic Explanation) operates by performing perturbations of input information about a prediction and fitting a new easy model under the hood to explain a single decision. Anchors is another way of XAI that produces if-then rules, which are readable by a person and provide precise explanations. These tools are critical in the field of finance, where regulatory organizations need to have comprehensible reasons behind the automatized decisions. A comparison of SHAP and LIME with respect to the explained accuracy and domain is provided in the table below.

2.4. Gaps in Current Research

Although the potential of ML and XAI is quite bright, very little work investigates their application to end-to-end financial processes. The majority of current studies focus on predictive performance without considering the crucial parameters of regulatory compliance, customer trust, and the interpretability-performance trade-off. An example is that, as much as SHAP may give detailed explanations, it is in some cases computationally demanding and challenging to employ in real-time decision-making settings. Moreover, institutions dealing with customers must achieve a compromise between the correctness of a model and its ability to explain decisions to clients in simple language. This paper aims to answer these questions by exploring the research on integrating XAI into credit risk models in a methodical way to achieve a greater degree of transparency and regulatory compliance, with a high degree of trust among customers without compromising the performance.

3. Methodology

3.1. Dataset Description

The research paper relies on the LendingClub dataset (pre-2021) as the most complete and publicly available peer-to-peer lending data. The data set contains a lot of loan and borrower-level data and thus is quite applicable in credit risk modeling and analysis. [9-12] As one of the largest participants in the domain of fintech lending, LendingClub provided the service of connecting both borrowers and investors in the implementation of loans with a vast amount of various types of features that apply to both creditworthiness assessment and predictive analytics. Some of the important aspects in the data are the loan amount, loan term, and loan interest rate, which indicate the very nature and cost of the loan itself. The amount of the loans determines the overall amount of money the borrower wants to pay, but the rest (36 or 60 months) points out the time it will take to pay the amount completely. The interest rate, which is set according to its risk models, can give us an idea about the perceived credit risk of the lender. Extensive characteristics of borrowers are also taken into account. This duration in employment will provide an indicator of job stability, which is usually associated with income security, and annual income will give a direct indicator of the borrower. These aspects facilitate the borrower's ability to repay the loan. Besides that, credit history, such as the number of open credit lines, delinquencies, and the duration of the credit history, is very important in measuring previous performance and the risk of defaulting. Of interest is specifically the Debt-To-Income (DTI) ratio, which quantifies the percentage of a borrower's/purchaser's monthly income that is used to cover debt payments- the higher the DTI, the more likely there is to be financial stress and the higher the prospect of defaulting. The data can usually be pre-processed by dropping missing values or unclear values, and categorical features (length of employment or grade of the loan) should be converted to a numerical format that is accepted as input by machine learning models. On the whole, the LendingClub dataset provides an excellent basis for the credit risk analysis and testing the use of explainable AI approaches integration in the context of predictive analysis.

3.2. Data Preprocessing

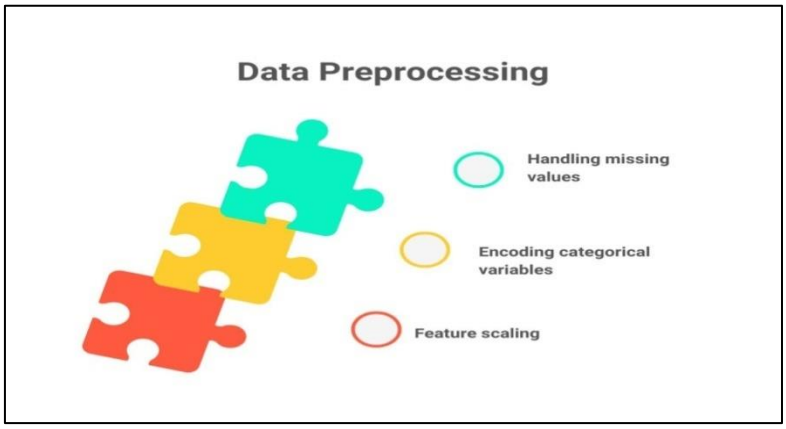

Figure 2. Data Preprocessing

- **Handling Missing Values:** Before applying any machine learning model, it is necessary to tackle missing data to ensure the model's reliability and performance. When provided with the LendingClub dataset, some fields can have a null value, like length of employment or annual income, because the reporting is not always consistent and, in some cases, the information can be discretionary. Missing values on a feature can be addressed using imputation (by mean, median or mode) or by discarding the records in case the proportion of missing data is low. In the case of a feature where the miss might itself be informative (e.g. missing details about employment), a special code or a binary mark might be added to retain any signals.
- **Encoding Categorical Variables:** Machine learning algorithms typically operate with numerical data inputs, so categorical variables must be encoded. The categorical features present in the LendingClub dataset, loan grade, length of employment and home ownership status have to be translated to a numeric representation. Label encoding can be used to preserve the natural ordering of ordinal features and is appropriate where the only natural ordering is to this ordinal scale, e.g. loan grade (A to G). Nominal features where there is not a natural ordering, e.g. verification status or purpose of the loan, are one-hot encoded to produce binary flags of each category. An appropriate encoding will ensure that the model can read the categorical data without introducing unnecessary biases.
- **Feature Scaling:** Feature scaling is important to make all numeric features equally contribute to the model since distance-based methods or the gradient-based optimizers are sensitive to them. Loan amount, annual income and DTI ratio are a few features that differ widely in scale and units. Commonly, all features will be placed into a similar range using standardization (z-score calibration) or min-max scaling. Not only does this increase the rate of model convergence, but also aids in performance and interpretation, especially algorithms that are sensitive to the feature magnitude like logistic regressions, support vector machines and neural networks.

3.3. Model Selection

- **Random Forest:** A random forest is an ensemble machine learning algorithm which constructs several decision trees and combines their predictions to increase accuracy and limit overfitting. It has been especially shown useful in classification problems such as loan default prediction due to its ability to capture multi-dimensional interactions between the features, and it is also capable of partial interpretability by measuring the relative importance of the features. [13-16] It is resistant to noisy data and can be used on both categorical and numerical features, therefore, being a good baseline model to be used in credit risk problems. It is, however, inferior in computation and thus may not detect fine trends compared to boosting methods.
- **XGBoost:** XGBoost (Extreme Gradient Boosting) is a powerful ensemble algorithm that constructs decision trees one by one, aiming to optimise the errors of preceding trees. It is more accurate and performs better since it has regularization potential, missing values adjustment, and is faster. Due to its capacity to model the complex nonlinear relationship, XGBoost has emerged as a default model in most of the machine learning competitions, as well as financial prediction problems such as credit scoring. Although very effective, it is also less interpretable relative to simpler models, which means it is more difficult to make decisions in more regulated financial settings using the model without additional explainability tools.
- **Neural Networks:** Neural networks are also very adaptive models with the ability to learn complex data patterns, particularly high-dimensional and large collections of data. The use of nonlinear and significant subtle interactions that are possible with their application in relation to the area of credit risk modeling may be missed by more conventional models. Since deep learning architectures (e.g., Multilayer Perceptrons (MLPs)) have the ability to adapt to the data without a large degree of feature engineering, it is possible to learn on all of the data. Neural networks, however, are perceived to be black boxes that are not very transparent, and this makes their application

in a financial institution trying to meet the explainability and regulatory demands a challenge. It also relies on fine-tuning and adequate data to train them for their success.

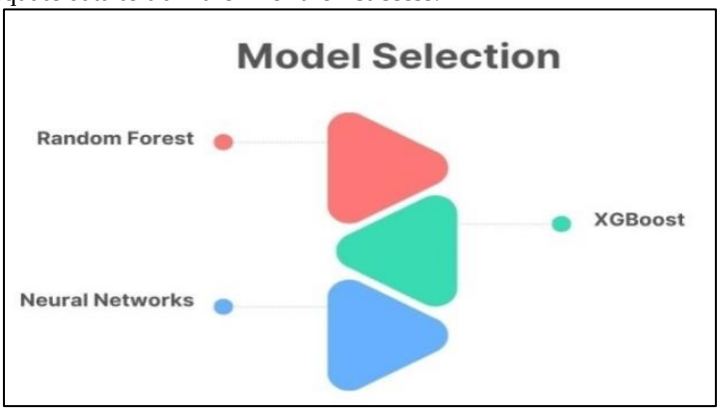

Figure 3: Model Selection

3.4. Flowchart of Methodology

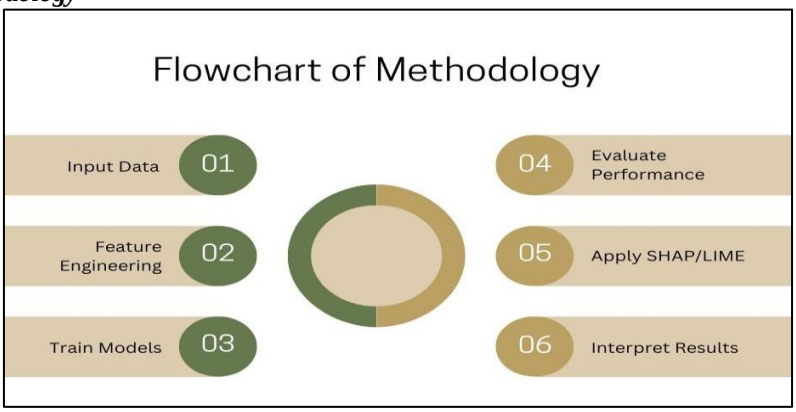

Figure 4: Flowchart of Methodology

- **Input Data:** The procedure starts with the collection of raw input data from the LendingClub dataset. This information contains details about the borrower and characteristics about the loan, financial measurements, as well as the history of credit. At this step, raw data is analyzed, but it includes both important characteristics and possible noise represented in terms of missing values or inconsistent information. This aims at answering the completeness and integrity of the data to be followed by analysis.
- **Feature Engineering:** Raw data is converted to meaningful inputs that become inputs to the machine learning models through feature engineering. This includes missing value treatment, classical variable encoding (categorical variables), calculated variables (e.g., income-to-loan ratio), and scaling numerical features. Proper feature engineering is important in enhancing the model's performance since the algorithms can use patterns that are not apparent in the raw data. A dimensionality reduction or feature selection to remove extraneous or redundant variables may be undertaken as part of this step, too.
- **Train Models:** After the organization of data, several machine learning algorithms, including Random Forest, XGBoost, and Neural Networks, are fit on the engineered features. Model tuning is done on both models to avoid cases of overfitting and maximization of hyperparameters through a technique such as cross-validation. It is a stage in which one acquaints oneself with the historical loan data so that one can differentiate between those who repay their loans and those who could default on their loans.
- **Evaluate Performance:** On a trained model, the correct measure of performance should be applied, including accuracy, precision, recall, F1 score and AUC-ROC. These measurements assist in gauging the capacity of every model to anticipate default on unseen data. A comparison of these outcomes enables the choice of the most convenient model that will balance the conflict between performance, computational efficiency and interpretation.
- **Apply SHAP/LIME:** The post-hoc interpretation community would use tools, such as SHAP and LIME, to make the models explainable. These methods provide an examination of said trained models in order to determine which features were of greatest contribution in each separate prediction. In financial terms, this is particularly crucial since it is not only imperative to predict something, but also to know the reason behind the decision being made.

- **Interpret Results:** The last action is in the sense that you interpret the SHAP or LIME outputs to have insights into model behavior. This involves establishing key risk factors, understanding how features impact them, ensuring sufficient alignment of exclusive decisions, and adhering to regulations and ethics. Insights may be applied to the model in the decision-making process, as well as in stakeholder and customer trust building.

3.5. Evaluation Metrics

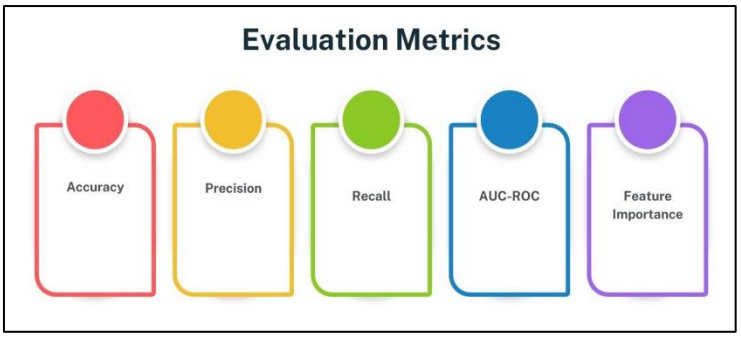

Figure 5: Evaluation Metrics

- **Accuracy:** Accuracy is the percentage of predictions that are right in the total number of predictions made by the model. [17-20] It gives an overall idea of how good or bad an overall performance model is. But in modeling of credit risks, where the number of defaults can be significantly less than the non-defaults, precision on its own may be misleading. Even if the model predicts all the loans to be non-default, it may be highly accurate, yet may not reveal the instances of actual defaults, which are essential in the financial risk estimation.
- **Precision:** Precision is a ratio of the number of correct positive predictions (true identifications of defaults) to the total number of positive predictions by the model. Put differently, the question answered is as follows: when the model says that there is a default, how many times will it be right? The cost of false positives is high where high precision is required: this might include trading credit with solid borrowers, or loss of customer satisfaction or opportunities to the business.
- **Recall:** Recall, which is also known as sensitivity or true positive rate, is a fraction of actual defaults being properly predicted by the model. It provides the probability of noticing risky borrowers. In a risk-averse setting, high recall is significant to verify the financial losses in cases when an opportunity to observe a possible default (false negative) is missed. Nevertheless, a trade-off between the recall and accuracy is popular, and therefore, a balance must be struck depending on the nature of the business.
- **AUC-ROC:** Area Under the Receiver Operating Characteristic Curve (AUC-ROC) quantifies the capacity of the model to determine the difference between default and non-default occurrences at different threshold levels. AUC values are between 0.5 (no discrimination) and 1.0 (perfect discrimination). In imbalanced datasets, AUC is a strong measure of overall system performance, since higher values are all the better, and a statistical test can then be built around it to measure model contribution to credit risk.
- **Feature Importance:** The significance of the features shows which of the input variables have the highest impact on the predictions of a model. It is particularly useful in a financial setting, where it is important to represent the drivers of a forecast decision to be transparent and comply with the regulations. In tree-based models such as Random Forest and XGBoost, the importance of the features is calculated with regard to how many times and to what extent a feature was used in splitting choices. Future feature engineering and reiteration of the model can also be informed using this information.

4. Results and Discussion

4.1. Performance Comparison

To compare the performance of the chosen models in terms of predicting defaults in loans we used major metrics, which are accuracy and AUC-ROC. These metrics are useful to measure the frequency of correct prediction by the models as well as to measure their effectiveness to differentiate between the cases of default and non-default.

Table 1: Performance Comparison

Model	Accuracy	AUC-ROC
Random Forest	86.2%	89%
XGBoost	89.5%	92%
Neural Network	91.0%	93%

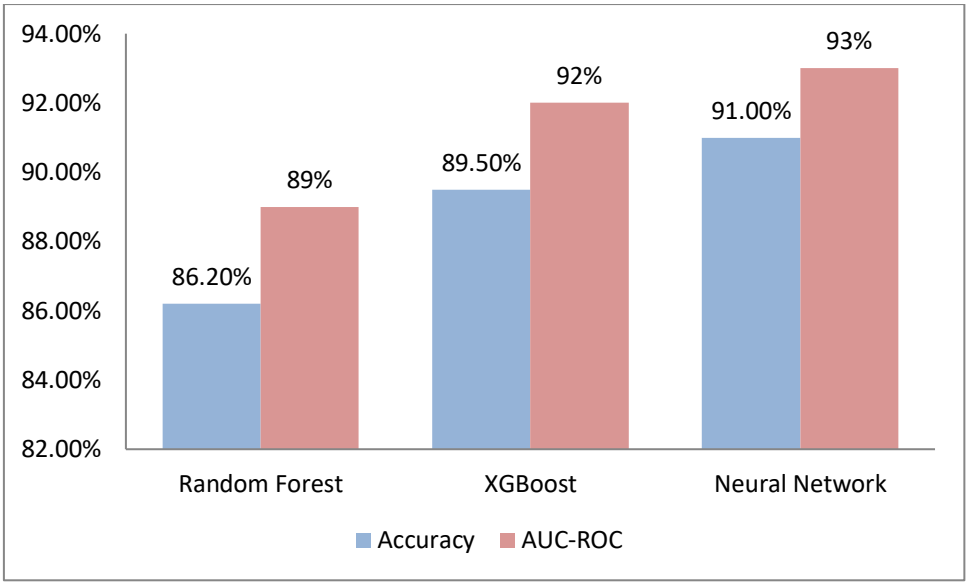

Figure 6: Graph representing Performance Comparison

- **Random Forest:** The accuracy and the AUC-ROC in the Random Forest model were 86.2 percent and 0.89, respectively. It means good overall performance with sound discriminatory ability against defaulters and non-defaulters. The possibility to work with a diversity of feature types and the resistance to overfitting make Random Forest a good baseline. Yet, in spite of more than adequate performance, it is outclassed by such modern types of constructs as XGBoost and neural networks, especially when it comes to modeling complex nonlinear dependencies.
- **XGBoost:** The XGBoost performed better with a measure of accuracy of 89.5 percent and AUC-ROC of 0.92, as well as Random Forest. This is an enhancement characteristic of XGBoost, as it applies gradient boosting and can learn from model errors more effectively through sequential trees and tree construction. Such a high rate of AUC-ROC indicates an improved level of separation between the classes, and XGBoost is one of the promising models of credit risk. Nevertheless, its lower interpretability than Random Forest might necessitate the application of other explainability tools to deploy in regulated settings.
- **Neural Network:** Neural Network model could be considered the most successful, having an accuracy of 91.01 and an AUC-ROC of 0.93. These scores demonstrate that it can learn nonlinear complex trends in data about borrowers. Its excellent performance is confirmed by the high AUC-ROC that guarantees an excellent classification power with various thresholds. Although it has high predictive power, the model is a black box, and it is hardly possible to interpret the requested decisions without adopting the explainable AI (XAI) concepts such as SHAP or LIME, particularly when handling the financial industry.

4.2. Interpretability Analysis

Knowing which are the most influential features in the decision of a model in the credit risk environment is highly relevant, in particular to describe transparency, regulatory compliance and customer confidence. We conducted an interpretability analysis to investigate the most important features in loan default using the SHAP (SHapley Additive explanations) explainability framework. The findings suggested that interest rate, Debt-To-Income (DTI) ratio, loan amount, and annual income have been the best attributes in all three models, including the Random Forest, XGBoost, and Neural Network. The most influential topic was the interest rate. It naturally reflects several dimensions of risk since it is usually set generally, depending on the credit risk of a borrower. The higher the rates, the higher the risk of borrowers, and this characteristic is a good indicator of default. SHAP values were used to indicate the importance of interest rate in risk modeling in the sense that small increases in interest rate increased the probability of being classified as a default remarkably.

The DTI ratio also played a very crucial role, and this is the ratio of the pre-existing burden of the borrower in terms of his / her current debt and income. A large DTI means that a greater percentage of the borrower in question is already bound to existing debt and thus has less money to play around with. The models continually labeled higher risk on applicants whose DTI ratios are higher than normal ratios (for example, a 35-40 percent DTI ratio). The amount of the loan was an important factor, especially for the borrower. The big loan requests would more frequently involve big default probabilities in the case of no corresponding high incomes. This aspect responds to interest rate as well as income, hence its interpretation is context sensitive. Finally, the amount of income per year was also a major factor that was evaluated as a measure of repayment. Taking loans with low reported income was given higher default risks, particularly when combined

with a large loan value or an intrinsically high value of a DTI. All of these characteristics together are the basis of model decision-making, and, therefore, these elements are crucial as they ensure both the right predictions and clear explanations.

4.3. Case Study Using LIME

To illustrate the utility of model interpretability, we also performed a case study where we explained an individual loan default with the help of LIME (Local Interpretable Model-agnostic Explanations). The LIME perturbs the given data and trains a simple and interpretable model (e.g., linear regression) around a given instance to localize the behavior of the complex, hard-to-comprehend model. This would enable the stakeholders (i.e., the loan officers or regulators) to know why a specific decision was taken, though the model used may be complicated. The model in this instance led to the prediction that the applicant was not likely to repay the loan. The explanation provided by LIME proved that the three main factors that drove the prediction were high Debt-To-Income (DTI) ratio, low annual income, and short employment history. The DTI ratio of the applicant was above 45, which indicated that a large share of their income has been allocated towards the payment of debts. Such a high DTI level is a big red flag to the financial institutions as it means that it has a small capacity to absorb its new debt liabilities. Also, the candidate had an income that was considerably less than the median within the dataset, which made the applicant even more vulnerable. In the locally fitted model of LIME, low income had a strong correlation with raising the probability of default, especially when combined with a high DTI ratio. Finally, the employment duration of the applicant was less than one year, which may also be an indicator of the lack of stability in income and employment. At the credit risk consideration, a short employment history will lower trust in the borrower's capability to earn a stable income throughout the loan period. The decision was not transparent and implementable before because the prediction is a black box. LIME allowed visualizing these local explanations and thus decomposing the prediction into human-readable parts. This type of interpretability would help not only in the internal processes of risk review but also facilitate communication with customers since they could be informed about the reasons why their loan was approved or rejected.

4.4. Trade-offs

Credit risk modeling is connected to a limitation that there is a continuous responsibility between forecasting and the explanation of the models. In the research, several models were checked: Neural Networks, XG Boost, and Decision Trees to find this balance and figure out which solution is the most applicable in real-life financial projects where accuracy and explainability are of paramount importance. The neural networks provided the best accuracy and AUC-ROC scores, and thus the superiority in capturing non-linear patterns in the behavior of borrowers. They have rich architecture, which is useful for learning complex feature interactions, making them very effective in prediction. They are, however, also known as black-box models, i.e., they have a decision process that is not comprehensible by humans easily. Such a lack of transparency is quite a negative feature in such high-stakes fields as finance, especially in fields where institutions are obliged to adhere to rules necessitating explainable decisions and where they are expected to provide justification to customers. XGBoost was not the most accurate, but it had a good balance between performance and ease of interpretation. Together with SHAP (SHapley Additive Explanations), XGBoost models can be much more transparent. The SHAP values give local and global explanations since they measure the contribution of each feature to a forecast. That is why XGBoost with SHAP is not only very accurate but also relatively understandable, which makes it an optimal model to roll out into production in the financial sector, where stakeholder trust and the regulatory framework are the top priorities. Conversely, Decision Trees were completely transparent since all the decision paths are traceable with the help of a chain of if-then rules. Such a degree of interpretability helps in auditability as well as communication with customers. The trade-off is, however, a significant drop in accuracy, particularly in large and complex data. Decision Trees are usually unable to characterize deeper patterns, which results in underfitting. Consequently, they are quite explicable, but have low predictive ability and hence cannot be used independently in risk-adverse settings.

5. Conclusion

This paper sought to understand the application and interpretability of different machine learning models to predict credit risk using the LendingClub dataset. The results indicate that black-box classifiers like neural networks and XGBoost are more effective than simple and interpretable models like decision trees in terms of accuracy and AUC-ROC. Unfortunately, in controlled industries such as those associated with banking and financial services, explainability is not a choice; it is a necessity. Regulation, ethics, and customer trust require models that can yield unambiguous explanations behind their forecasts. Explainable AI (XAI) tools, especially SHAP and LIME, were discovered to be most effective in articulating the gap between the complexity of models and talkability. SHAP, whose main theory is based on cooperative game theory, provides the same global/local feature importances across the model. In contrast, LIME gives local explanations that are easy to consume and are particularly helpful in the assessment of individual customers.

According to the findings, a number of practical suggestions can be provided to the institutions willing to incorporate machine learning in their credit risk processes. One is that, in high-stakes or customer-facing decisions, it would be better to either use inherently interpretable models or augment complex models with post-hoc explainability tools such as SHAP and LIME. Second, companies are advised to use transparency logs, where they save the model predictions with

explanations that could be audited and reviewed by regulators. These logs will act as important records during controversies or inspections. Lastly, the underwriters, risk analysts, and customer care teams involved should be trained on the use of the machine learning outputs, ensuring a deeper understanding of their appropriate application during adoption. It is not only possible to know how a model has made its decision, but that knowledge also enables the staff to communicate with the internal and external stakeholders more effectively.

Albeit the insights obtained by the use of SHAP and LIME can be sufficient, further studies can be conducted toImplement more sophisticated explainability methods, including counterfactual explanations, which can demonstrate how slight alterations in the features of the borrower can alter his or her prediction (e.g., to switch the prediction of the loan being declined to being approved). Alongside, since fintech platforms continue to work on the real-command fronts, building technologies that have low-latency and real-time explainable solutions will be vital in the automated decision-making of extending loans. Finally, federated learning, a technique which allows training a model on a decentralized dataset and does not share sensitive personal data, provides an intriguing possibility to create privacy-sensitive, explainable credit risk models that meet the compliance requirements and perform at the institutional level.

References

1. Thomas, L., Crook, J., & Edelman, D. (2017). Credit scoring and its applications. Society for Industrial and Applied Mathematics.
2. Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. Journal of the Royal Statistical Society: Series A (Statistics in Society), 160(3), 523-541.
3. Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature. Intelligent systems in accounting, finance and management, 18(2-3), 59-88.
4. 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).
5. Malekipirbazari, M., & Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert systems with applications*, *42*(10), 4621-4631.
6. Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance, 34(11), 2767-2787.
7. Heaton, J. B., Polson, N. G., & Witte, J. H. (2016). Deep learning in finance. arXiv preprint arXiv:1602.06561.
8. Sirignano, J., & Cont, R. (2021). Universal features of price formation in financial markets: perspectives from deep learning. In Machine learning and AI in finance (pp. 5-15). Routledge.
9. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.
10. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
11. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion, 58, 82-115.
12. Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (XAI): A survey. arXiv preprint arXiv:2006.11371.
13. Gültekin, B., & Erdoğdu Şakar, B. (2018, July). Variable importance analysis in default prediction using machine learning techniques. In Proceedings of the 7th International Conference on Data Science, Technology and Applications (pp. 56-62).
14. Sofaer, H. R., Hoeting, J. A., & Jarnevich, C. S. (2019). The area under the precision-recall curve is a performance metric for rare binary events. Methods in Ecology and Evolution, 10(4), 565-577.
15. Xu, J., Lu, Z., & Xie, Y. (2021). Loan default prediction of the Chinese P2P market: a machine learning methodology. Scientific Reports, 11(1), 18759.
16. Tiwari, A. K. (2018). Machine learning application in loan default prediction. JournalNX, 4(05), 1-5.
17. Lai, L. (2020, August). Loan default prediction with machine learning techniques. In 2020 International Conference on Computer Communication and Network Security (CCNS) (pp. 5-9). IEEE.
18. Pérez-Sánchez, B., Fontenla-Romero, O., & Guijarro-Berdiñas, B. (2018). A review of adaptive online learning for artificial neural networks. Artificial Intelligence Review, 49, 281-299.
19. Ribeiro, M. T., Singh, S., & Guestrin, C. (2018, April). Anchors: High-precision model-agnostic explanations. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1).
20. Cerda, P., Varoquaux, G., & Kégl, B. (2018). Similarity encoding for learning with dirty categorical variables. Machine Learning, 107(8), 1477-1494.