Explainable Artificial Intelligence in Credit Scoring

Maria Michelaki†
University of Nicosia, Computer Science Department, Nicosia, Cyprus
mariamixelaki95@yahoo.gr

ABSTRACT

Credit scoring is the process of evaluating the likelihood that a borrower will default on a loan based on historical data. Traditionally, interpretable models such as logistic regression, linear regression, and decision trees have been used because they offer transparency and ease of interpretation. However, with the rise of black-box models like Random Forests, Gradient Boosting Machines (GBM), and Deep Neural Networks (DNN), the need for explainability has become more pronounced (Baesens et al., 2016).

In credit scoring, interpretability is critical for compliance with regulatory standards (e.g., GDPR, Equal Credit Opportunity Act) and for maintaining fairness, transparency, and trust with end-users. Hence, Explainable Artificial Intelligence (XAI) methods have been developed to provide insights into the decision-making process of complex models (Guidotti et al., 2019).

1. Introduction

Credit scoring is a statistical approach used to estimate the likelihood that a loan applicant, existing borrower, or counterparty will default or become delinquent. Credit scores serve as indicators of an individual's creditworthiness.

In sectors such as finance incorrect predictions can have severe consequences for human lives. Therefore, understanding how AI systems arrive at their decisions is crucial. As AI continues to expand into these critical areas, the inability to comprehend complex models poses a significant challenge. Traditional methods such as regression and decision tree algorithms are inherently interpretable and easy to explain, particularly when they rely on a limited number of features. In contrast, models like deep neural networks, random forests, and gradient boosting machines are often regarded as black-box models. These techniques incorporate numerous features and intricate transformations, making it difficult to understand the relationship between input features and the target variable. Additionally, bias may infiltrate the model through the data itself. Standard performance metrics, such as accuracy, do not always reflect the true nature of a model's predictions. Relying solely on accuracy is insufficient for building trust and deploying models in real-world scenarios.

As reliance on intelligent systems continues to increase, there is a growing need for more transparent and interpretable models. The ability to explain a model's decisions has become essential for building trust and deploying Artificial Intelligence (AI) systems, especially in critical domains. Explainable Artificial Intelligence (XAI) encompasses a range of machine learning (ML) techniques designed to help human users understand, trust, and work with these models more effectively. Take a machine learning model designed for credit risk assessment as an example. If the model predicts that an individual is a high-risk borrower, an explanation might be: "The borrower is classified as high-risk due to a history of missed payments on previous loans." Such an explanation enables end-users to understand the rationale behind the prediction and make informed decisions accordingly.

2. Explainable AI Methods for Credit Scoring

In Explainable Artificial Intelligence (XAI), different methods for explaining model predictions can be categorized based on three key criteria:

2.1 Applicability to models

2.1.1 Model-Agnostic Methods: Methods that can be applied to any type of model (e.g., neural networks, decision trees, random forests). They operate independently of the model's internal structure. These methods work by analyzing the input-output behavior of the model, making them versatile and adaptable to different types of models.

2.1.2 Model-Specific Methods: Methods designed for specific types of models or algorithms. These methods leverage the internal structure and characteristics of the model. For example, TreeSHAP is an optimized version of

SHAP for tree-based models like Random Forests and Gradient Boosting Machines.

2.2 When the explanation is generated

2.2.1 Intrinsic Methods (Directly from the Model): Explanations are inherently part of the model structure. In other words, the model itself is designed to be interpretable, such as a linear/logistic regression or a decision tree model.. These models are simpler and structured in a way that allows humans to understand the relationship between inputs and outputs without additional tools or methods. The advantages of these methods are that they are simple and easy to interpret and there is no need for additional tools to generate explanations.

2.2.2 Post hoc Methods (After the Prediction): Explanations are generated *after* the model has made its predictions, often for black-box models that aren't intrinsically interpretable. These methods analyze the input-output relationship to provide explanations without altering the model. Example of these methods are SHAP, LIME, and Partial Dependence Plots. They are applicable to any model, including complex black-box models and can provide global or local explanations, but might be approximate and not fully accurate.

2.3 Which concerns the explanation

2.3.1 Global Explainability Methods: Global methods explain the overall behavior of the model.

Permutation Feature Importance (PFI)
Proposed by Breiman (2001), PFI measures the importance of each feature by calculating the change in model performance when the feature values are permuted. This method is model-agnostic and provides insights into which features influence the credit score predictions.

Partial Dependence Plots (PDP)
PDPs (Friedman, 2001) visualize the relationship between one or two features and the model's predictions. In credit scoring, PDPs can show how features like interest rates or income levels affect the probability of default.

SHAP (SHapley Additive exPlanations)
SHAP (Lundberg & Lee, 2017) provides a unified framework for feature importance based on game theory. SHAP values offer consistent and locally accurate explanations for individual predictions. For credit scoring, SHAP can identify how features like payment history or credit utilization impact a borrower's score.

2.3.2 Local Explainability Methods: Local methods provide explanations for individual predictions.

LIME (Local Interpretable Model-agnostic Explanations)
LIME (Ribeiro et al., 2016) approximates black-box models with interpretable surrogate models (e.g., linear regression) for a specific instance. In credit scoring, LIME can explain why a particular applicant was classified as high-risk or low-risk.

Anchor Explanations
Anchor (Ribeiro et al., 2018) generates "if-then" rules that serve as sufficient conditions for predictions. For example, an anchor might state: "If the applicant's credit mix is high and there are no delayed payments, then the borrower is low-risk."

3 Data

The data used for the project is the Credit Score Classification dataset found on Kaggle. The dataset consists of 100.000 rows and 28 different features. It provides information for predicting credit scores using various financial and personal attributes. The dataset includes features related to:

- Financial behavior: Attributes like payment delays, number of credit cards, and credit mix.
- Loan history: Indicators of past loan repayments, number of loans, and credit utilization.
- Demographics: Features such as age and other personal information.

You can explore it here.

4 Methodology

The main objective is to build a predictive model that can classify individuals into different credit score categories (e.g., Bad, Standard, Good) based on the available features. With the help of XAI methods, this can assist lenders in making informed decisions regarding credit approvals and risk management.

4.1 Data Cleaning

I first observed that the 100.000 rows contained information for 12.500 customers of the bank. Each customer has 8 entities for months from January to August. I decided to maintain this structure on the data and to not drop rows so as I can create a new dataset, made of aggregations for each variable for every customer from the original dataset. Most variables on the dataset contained missing values, and erroneous ones, such as age = -500, monthly balance = 33333333333 etc. To handle this problem, I kept the value that appears in majority for every 8 lines or I replaced the Nan or error value with the above or below one for each customer. Finally, I created a new CSV file for the clean data.

4.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is crucial for understanding a dataset's structure, patterns, and anomalies before building models. It helps in Identifying data distributions (Understand how numerical/categorical features are distributed), visualize correlations or trends between variables and identify important variables and ways to refine them. In this notebook, I found out the unbalanced nature of the target variable and the important correlation between most of the features. I also figured out the variables that are unrelated to our classification problem, such as Month or SSN. Most importantly, I understood the relationship between some features and the Credit Score, so as I can do sanity checks on my model predictions. For example, as the number of loans and the number of credit inquiries increases, the credit score drops. Also, customers with lower interest rates and lower delays tend to have good credit score.

4.3 Feature Engineering

Feature engineering improves model performance by transforming raw data into meaningful inputs. I created a new feature, Debt to income ratio, to help the model better capture patterns. I tried different two different ways to encode categorical variables. In the 'DL' methods notebook I used ordinal encoding for two variables and converted nominal variables to dummy variables with One-Hot Encoding. In the 'ML methods' notebook I used label encoder from the sklearn.preprocessing library.

4.3.1 Dimensionality Reduction

Correlation-based Feature Selection (CFS) is a feature selection technique that selects subsets of features that are highly correlated with the target variable but have low correlation with each other. For classification problems, the correlation is calculated using the Symmetrical Uncertainty measure, which considers the mutual information between the features and the target variable. From the Correlation matrix of the selected features I observed that all the attributes are correlated with each other and the vast majority of them are correlated with the target variable, so I did not use this method.

PCA Principal Component Analysis is a dimensionality reduction technique used to simplify datasets while retaining as much variance as possible. It transforms the data into a new coordinate system where the greatest variance lies along the first axis (principal component), the second greatest variance along the second axis, and so on. I printed the first 16 PCAs and the Scree plot that helps visualize how much variance is explained by each component. Since the Cumulative Explained Variance is only 53% for the optimal number of PCAs (2), the dimensionality reduction achieved by PCA may result in a significant loss of information. Ideally, PCA should retain at least 85-90% of the variance, so the dataset may not be well-suited for PCA.

Good feature engineering often leads to more accurate and efficient models.

4.4 Data Balancing

4.4.1 Undersampling for Data Balancing for the Machine Learning Methods.

Undersampling is a technique used to address class imbalance in a dataset by reducing the number of samples in the majority class so that it is balanced with the minority class. This can help machine learning models perform better when the dataset is skewed towards one class. Cluster Centroids replace the majority class with centroids of clusters formed within the majority class.

4.4.2 Oversampling for Data Balancing for the Deep Learning Methods.

SMOTE (Synthetic Minority Oversampling Technique) can be an effective way to handle imbalanced datasets. SMOTE generates synthetic examples for the minority class by interpolating between existing samples, thereby increasing the representation of the minority class without duplication.

4.5 Evaluation Metrics

Choosing suitable evaluation metrics is crucial because different problems require different ways to assess model performance. For example, accuracy works for balanced datasets, but precision, recall, or F1-score are better for imbalanced data. A high accuracy score can mask poor performance in minority classes. In critical areas like healthcare or finance, appropriate metrics help evaluate fairness and reliability. Choosing the right metric ensures you optimize the model for its intended purpose. The metrics I used are Accuracy, Precision, Recall, F1-score and Confusion Matrix plot that provides a matrix showing True Positives, True Negatives, False Positives, and False Negatives for each class, which helps identify which classes the model struggles with. Furthermore, I used Logarithmic Loss (Log Loss) to evaluate the uncertainty of the model's predictions by considering predicted probabilities and Gini Index that is commonly used in credit scoring as a measure of rank-ordering ability. Type II error is more costly in credit scoring, since Type II (False Negative ratio) predicts a borrower as a good borrower but in actual fact he or she is a bad borrower. Area Under the ROC Curve (AUC-ROC) Evaluates the trade-off between True Positive Rate (TPR) and False Positive Rate (FPR) for each class.

4.6 Machine Learning Models

The following models were tested and evaluated:

	Accuracy %	Precision %	Recall %	F1-Score %	Type II Error %
Decision Tree Classifier	75.44	75.25	75.44	75.31	6.16
Random Forest Classifier	83.12	83.56	83.12	82.85	8.10
Random Forest Classifier tuned	83.71	84.46	83.71	83.48	8.44
Logistic Regression	78.90	78.98	78.90	78.28	8.86
SVM	80	80.97	80	79.40	11.22
kNN	77.30	78.04	77.30	76.96	8.69
GaussianNB	79.66	81.56	79.66	79.23	11.22

Table 1 Performance of Machine Learning Models

Random Forest Classifier with parameters obtained with Hyperparameter tuning brought the greatest evaluation scores.

4.7 Deep Learning Models

I performed the next experiments on the original dataset. I started with a simple model with 1 hidden layer and evaluation metrics close to 70%, then added more layers and a dropout layer to prevent overfitting that increased the precision to 73%. Continued by adding Batch Normalization layers in the model to stabilize training by normalizing inputs to each layer and Early Stopping to monitor the validation loss. I then used Class Weights to deal with imbalanced data, but it did not improved the performance of the model. Class weights modify the loss function, giving more importance to misclassifications of the minority class. Next, I tried SMOTE and experimented with different Batch sizes, number of Epochs and models to achieve the following maximum evaluation metrics: 'Accuracy (%)': '78.78', 'Precision (Weighted, %)': '79.27', 'Recall (Weighted, %)': '78.78', 'F1-Score (Weighted, %)': '78.81', 'Log Loss (%)': '50.94', 'Gini Index (%)': '82.18', "Type II Error (Predicted 'Good' but Actually 'Bad/Standard')": 1124, 'Type II Error Percentage': '5.62'.

I then used the dataset with the aggregations for each customer, to find out that the models perform almost the same but with a lot less running time, minimizing the computational cost.

4.8 XAI methods

Post modelling interpretability tests can help in the understanding of the most important features of a model: how those features affect the predictions, how each feature contributes to the prediction, and how sensitive the model is to certain features. I have used the following model Agnostic methods for both local and global explanations: SHAP, LIME, ELI5, Anchors, and Partial Dependence Plots.

5 Design Considerations for Implementing XAI

Bias and Fairness: Ensuring that XAI methods do not reinforce biases in the training data remains a challenge. Bias detection and mitigation are crucial in credit scoring.

Scalability: XAI methods like SHAP can be computationally expensive for large datasets.

Human-Centric Explanations: Future research should focus on generating explanations that are easily understandable by non-technical users.

6 Conclusion

XAI methods such as SHAP, LIME, and Anchors play a crucial role in making credit scoring models interpretable and transparent. As AI adoption in credit risk assessment grows, leveraging these methods helps ensure trust, compliance, and fairness. Continued research and development in XAI will further improve the interpretability of complex models and their applications in critical sectors like finance.

REFERENCES

[1] Baesens, B., et al. (2016). *Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS*.

[2] Lundberg, S. M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. Advances in Neural Information Processing Systems.

[3] Ribeiro, M. T., et al. (2016). *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.

[4] Wachter, S., et al. (2017). *Counterfactual Explanations Without Opening the Black Box*. Harvard Journal of Law & Technology.

[5] Guidotti, R., et al. (2019). *A Survey of Methods for Explaining Black Box Models*. ACM Computing Surveys.

[6] Xolani Dastile, et al. (2019*). Statistical and machine learning models in credit scoring: A systematic literature survey.* Applied Soft Computing Journal.

[7] Dwivedi, et al. (2023). *Explainable AI (XAI): core ideas, techniques and solutions.* ACM Computing Surveys