XAI Final project – Dec 2025

Javier Rojo Llorens

Problem and Dataset

Problem:

The objective of this project is to study a classification problem between two classes where a machine learning model (RandomForestClassifier) predicts whether a user will miss out next month's payment or not, given some historical financial and demographic information of the user. Beyond achieving good predictive performance, the central focus is on understanding and explaining the model's decisions using Explainable Artificial Intelligence (XAI) techniques.

In this setting, predictions on their own are not enough. Stakeholders need to understand why the model makes a particular prediction, which variables influence the decision, and whether the model is learning meaningful patterns or relying on spurious correlations. This makes the problem a good fit for an XAI case study, since explanations can help validate the model, identify potential issues, and increase confidence in its predictions.

Dataset: [1]

The dataset consists of records from credit card clients, with each row representing an individual customer. It contains 25 variables that describe demographic information, credit limits, recent billing activity, repayment behavior over the previous six months, and the default outcome (if they paid next month or not).

Customer characteristics such as age, gender, education level, marital status, and assigned credit limit give some background context. Financial behavior is also captured through monthly repayment status indicators, bill statements and actual payment amounts, showing how each client has managed their credit over time.

Stakeholder:

The main stakeholders in this scenario would be financial institutions and credit analysts who use models to support credit-related decisions. Since these decisions can have real consequences for customers, it is important to understand why the model predicts that a client will miss a payment. Explainability helps reveal which factors drive the predictions, check that the model is using sensible information rather than spurious patterns and build trust in its outputs.

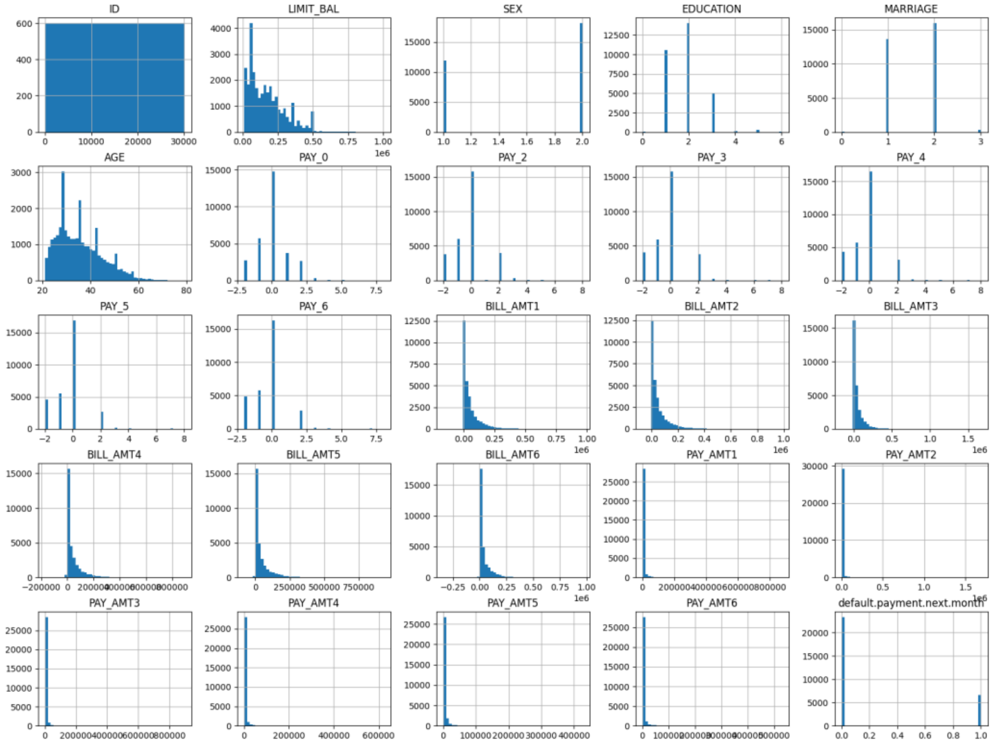


Figure 1: Dataset variable distribution, as we can observe there are different scales and everything has been turned into numerical. SEX: Gender (1 = male, 2 = female).

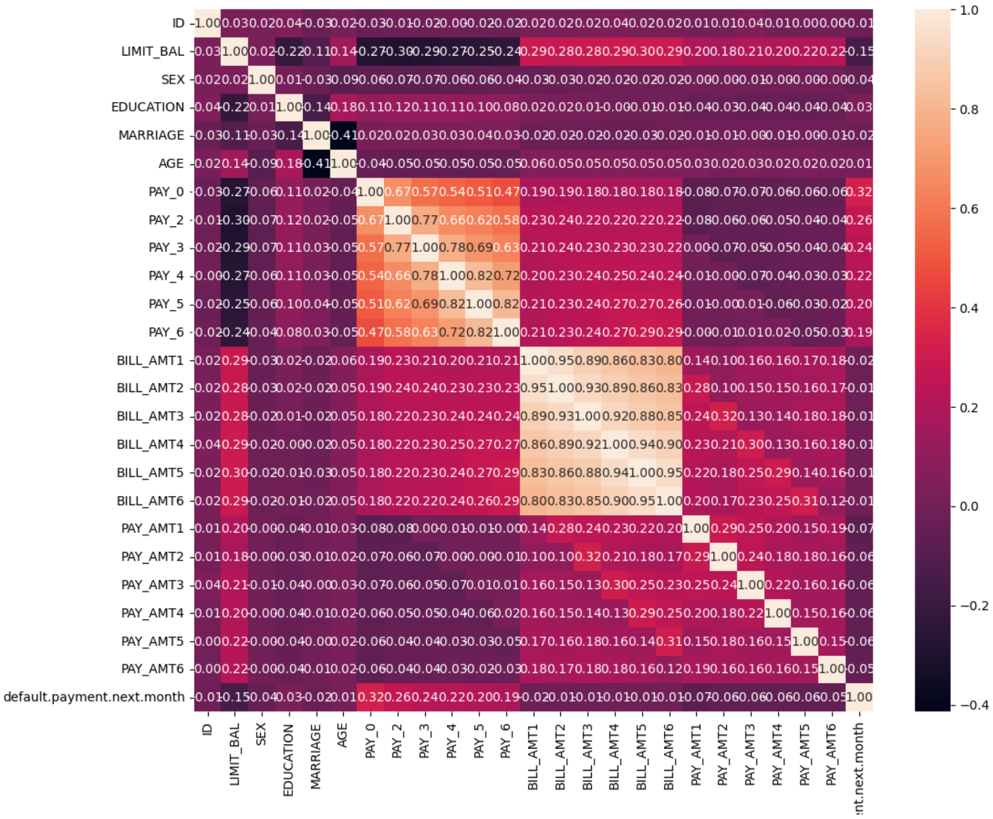


Figure 2: Correlation between features of the dataset, this will give us some better understanding of some of the metrics used.

Models and Evaluation

Model choice:

The dataset used in this project was obtained from Kaggle and includes an accompanying baseline solution [2]. Rather than proposing a new modeling approach from scratch, this work builds on the initial model provided with the dataset, which uses a Random Forest classifier.

This choice was maintained deliberately to focus the analysis on model explainability rather than model selection or architecture optimization. The Random Forest serves as a strong and commonly used baseline for tabular credit risk data, making it a suitable candidate for studying and comparing different explainability techniques.

The model is trained using a standard train–validation split with a fixed random seed to ensure reproducibility.

Evaluation:

Model performance is measured on the validation set using standard classification metrics. Accuracy is reported to measure overall correctness, while the confusion matrix is used to analyze the distribution of true positives, true negatives, false positives, and false negatives. This breakdown is particularly useful in a credit default setting, where different types of errors may have different practical implications.

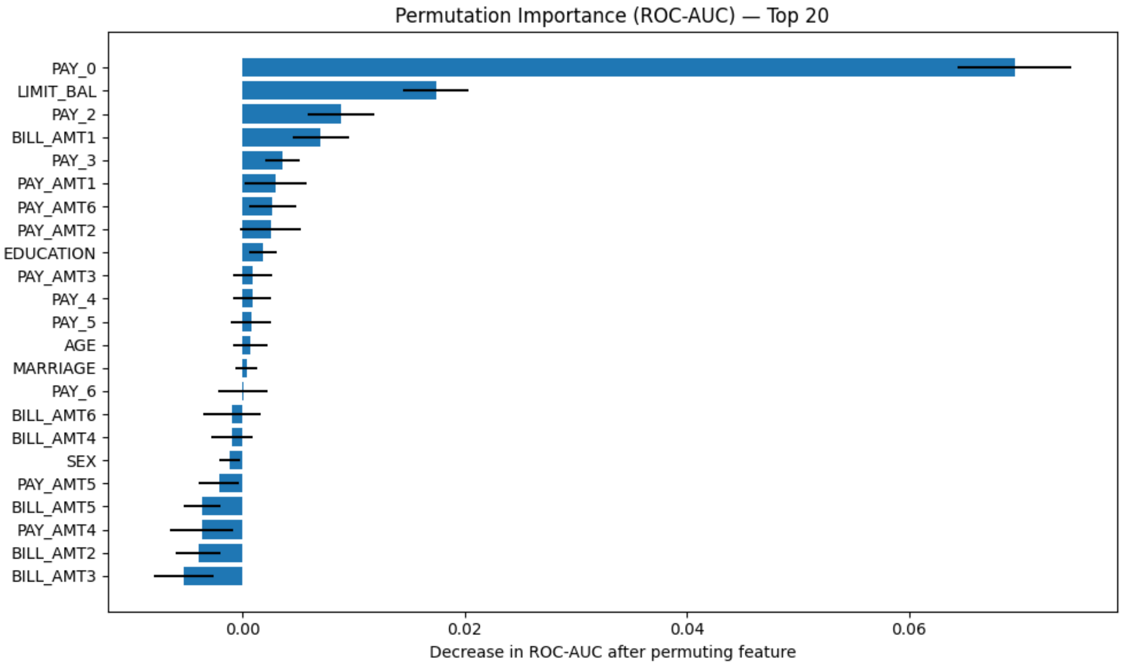
In addition, the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) is also computed. In credit risk prediction, ROC-AUC is often more informative than accuracy because it measures the model's ability to distinguish between defaulting and non-defaulting clients independently of any threshold. Since default datasets are typically imbalanced and the cost of false positives and false negatives is asymmetric, accuracy alone usually is misleading.

Together, these metrics provide a balanced view of both overall predictive performance and the model's behavior under different decision scenarios, forming a solid basis for the subsequent explainability analysis.

Explainability Techniques

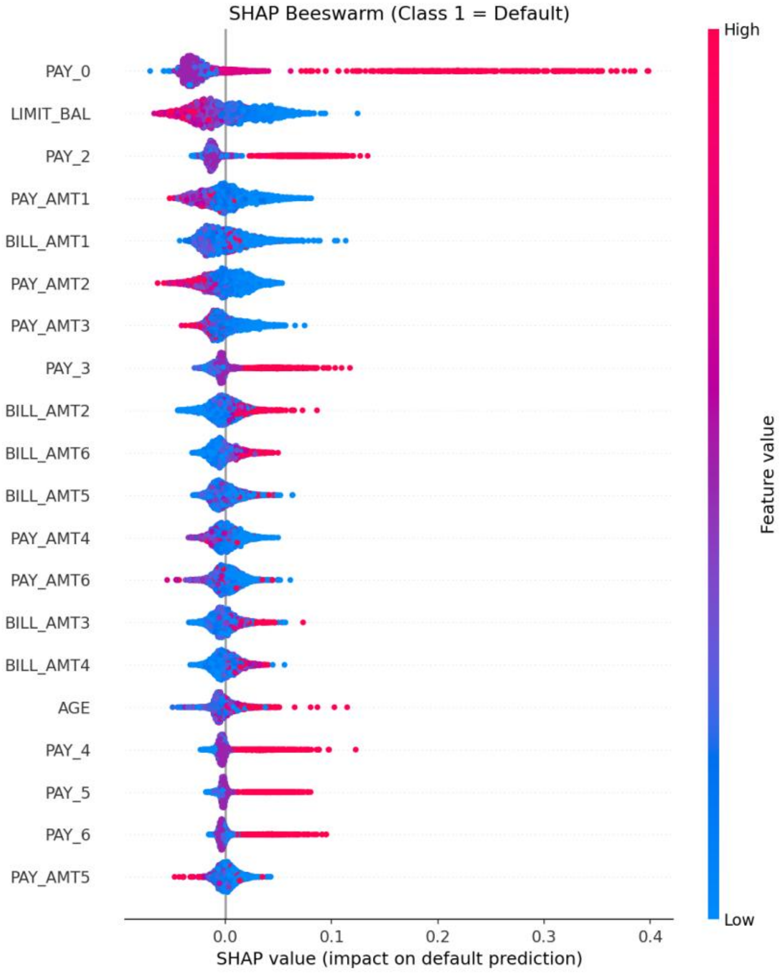
To analyze and interpret the behavior of the Random Forest classifier, three complementary XAI techniques are applied: Permutation importance using the aforementioned ROC-AUC metric. SHAP beeswarm plots for global explanations, LIME for local explanations, and sanity checks to evaluate the reliability of the explanations. Together, these methods provide insight at different levels of the model's behavior.

Global: Permutation Importance



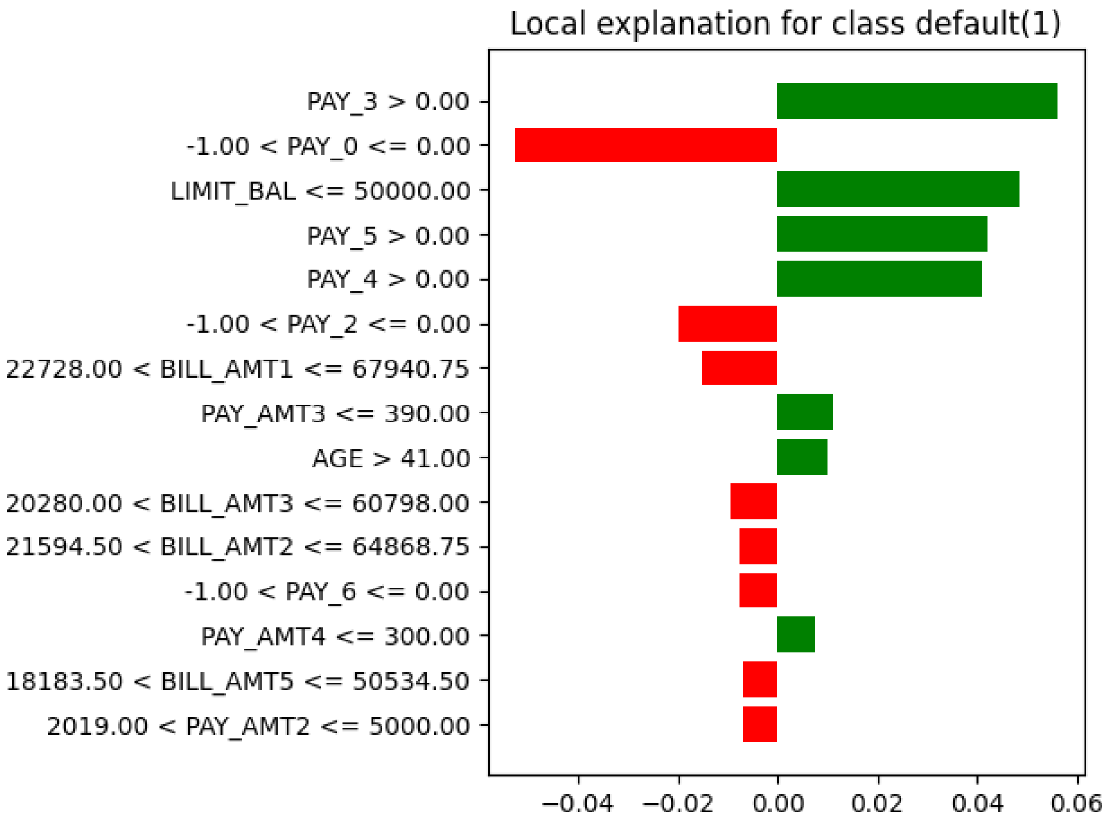
Permutation importance is a global, model-agnostic explainability technique that quantifies how much a model's performance degrades when the relationship between a feature and the target is disrupted. In this study, permutation importance based on ROC-AUC reveals a strong reliance on the most recent repayment status (PAY_0), while other correlated repayment history variables contribute less individually. This highlights both the model's dependence on short-term signals and the presence of feature redundancy, motivating further analysis and feature engineering.

Global: SHAP Bee swarm plot



The SHAP summary plot reveals that the model relies mostly on the most recent repayment status (PAY_0), with higher delays strongly increasing default risk. While earlier repayment history variables also contribute in the expected direction, their impact is comparatively weaker, indicating a short-term bias in the model's decision process. Credit limit and payment amounts exhibit coherent effects aligned with financial intuition, whereas demographic variables mostly play a minor role, except age. Overall, the analysis suggests that the model prioritizes immediate repayment signals over longer-term behavioral patterns, motivating the study of the different XAI metrics to understand if the model could be calibrated better.

Local: Lime



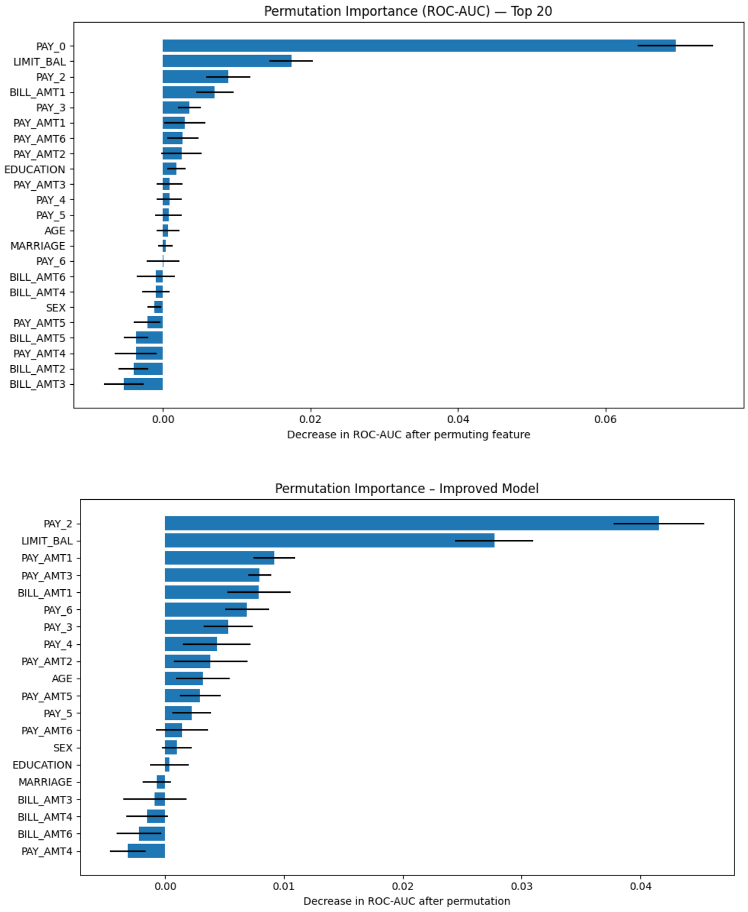
This is the local explanation of why the model predicted default for this specific client, approximating the model to a linear and more interpretable one.

While the global explainability methods we have studied show that the most recent repayment status (PAY_0) is the strongest driver of default risk on average, local explanations show how this is also dependent on the value, and if PAY_0 is as expected (lower than 0), individual decisions may rely more heavily on earlier repayment stats. This apparent discrepancy highlights the complementary nature of global and local explainability and underscores the importance of instance-level analysis.

However, we see some examples do follow the global trends, deciding almost unilaterally the prediction of the model. [3]

Sanity check: Drop column

Highly correlated features may be misrepresented in feature importance metrics, as a single feature can account for most of the observed importance while acting as a proxy for a group of redundant variables. To assess whether the dominant feature provided unique predictive information, a drop-column sanity check was performed. The model was retrained without the most important feature and evaluated using ROC-AUC. The resulting performance degradation was limited, indicating that the removed feature did not carry exclusive information and that other correlated variables compensated for its absence. This confirms that the observed dominance was partly driven by feature redundancy rather than by unique predictive signal.



Features dropped	ROC-AUC score
None	0.762
[PAY_0]	0.731 (-0.031)
[PAY_0, PAY_2]	0.723 (-0.039)

Figure X: Permutation importance as calculated with the model vs the real impact when the features are dropped from the dataset and the ROC-AUC is recalculated.

Actionable Use of Explanations

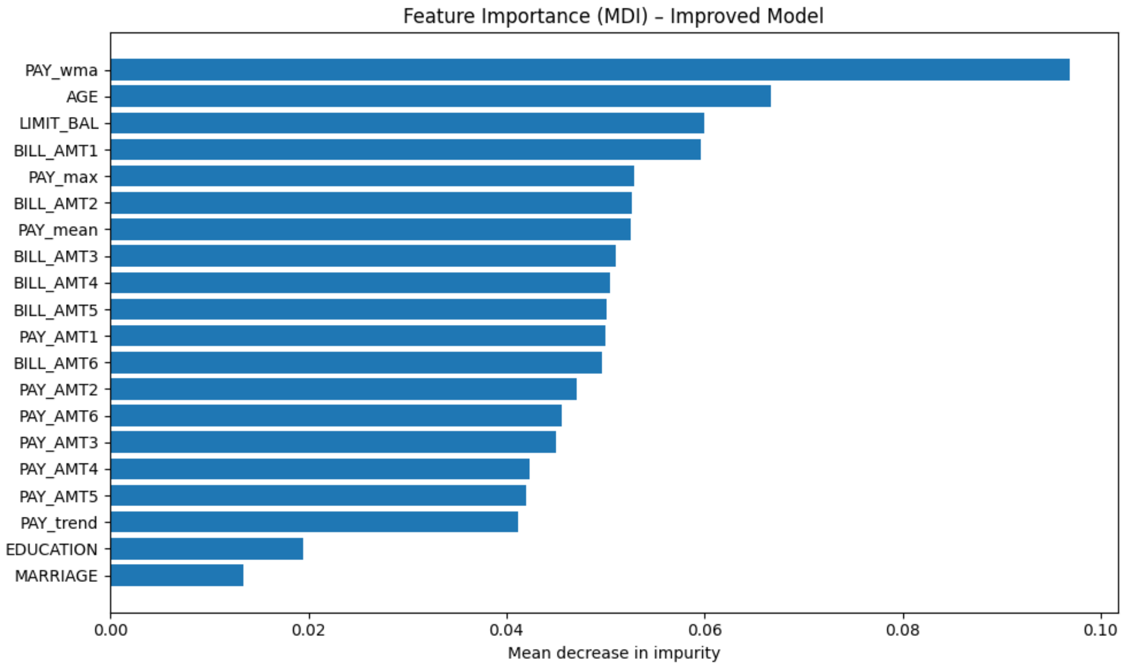
Explainable AI techniques were used not only to interpret the baseline model but to actively guide its improvement. Initial global explanations revealed an excessive reliance on short-term repayment indicators, while correlation analysis showed strong redundancy among month-level repayment variables. These findings motivated a shift from individual repayment statuses toward more stable representations of long-term behavior.

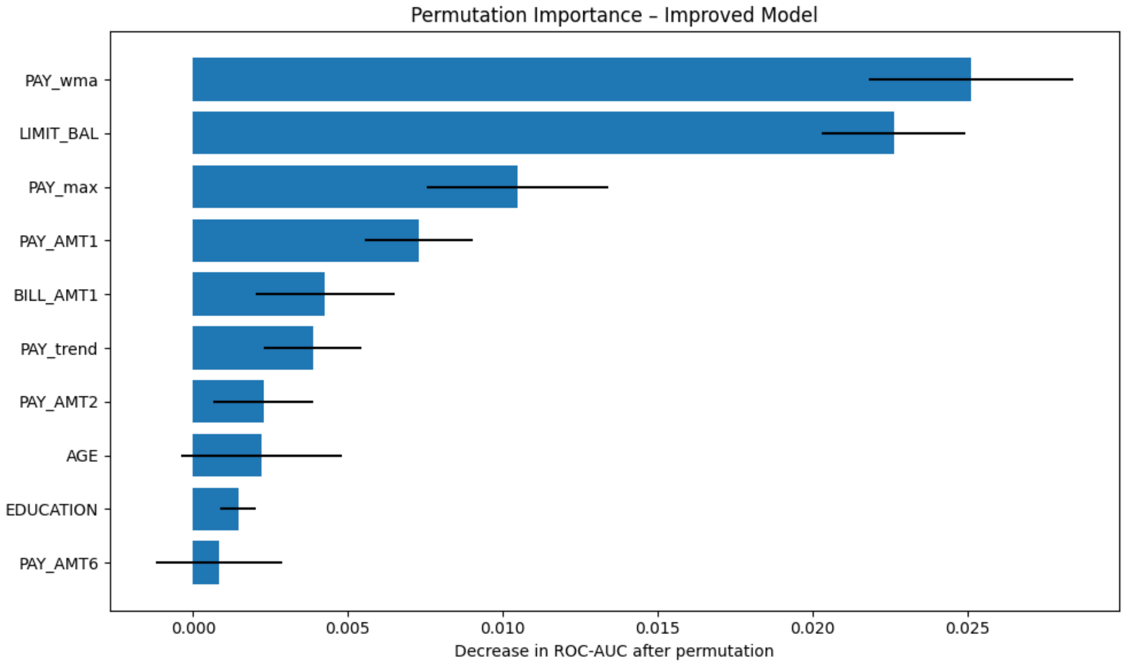
Based on these insights, several aggregated repayment features were introduced. PAY_mean captures the average repayment delay across multiple months, providing a robust summary of overall payment behavior. PAY_max encodes the most severe historical delay, reflecting extreme risk events that are particularly relevant in credit default prediction. PAY_trend measures the difference between recent and past repayment behavior, allowing the model to distinguish between improving and deteriorating financial trajectories. In addition, a moving average of repayment status was introduced to smooth short-term fluctuations while preserving recent behavioral patterns, offering a compromise between responsiveness and stability.

The model was retrained using these aggregated features in place of the original month-level repayment variables. Evaluation with ROC-AUC showed only minimal changes in performance 0.7634 (+0.0014), while permutation importance analysis revealed a more balanced and meaningful distribution of feature contributions. This is the most important improvement of the model. The new metrics, especially the moving average metric, more than make up for the lost individual categories, and it is more robust than PAY_0, which adds much noise and variability to the possible predictions.

Overall, explainability metrics were used as actionable tools to identify redundancy, guide feature engineering, and perform informed feature selection. This process led to a model that is more interpretable and robust, while slightly improving predictive performance.

Evaluation and Critique of XAI





Overall, several explainability methods were used to analyze the model at global and local levels. Permutation importance and global SHAP were largely consistent in identifying repayment-related features as the main drivers of default risk. However, impurity-based feature importance (MDI) sometimes ranked continuous variables highly despite having limited impact on predictive performance (as was the case for AGE), highlighting known biases of this method. Local explanations provided by SHAP and LIME generally agreed on influential features for individual predictions, although LIME was observed to be less stable due to its reliance on local approximations and discretization.

To validate the reliability of the explanations, multiple sanity checks were performed. A drop-column test showed that removing highly ranked features led to only moderate performance degradation, confirming that some importance rankings were influenced by feature redundancy rather than unique predictive information.

Despite their usefulness, the explainability methods have a lot of limitations. Feature importance measures can be misleading in the presence of correlated variables, and SHAP values do not imply causal relationships. Global explanations may also obscure instance-level variability, while local explanations can be over-interpreted if presented without broader context.

For stakeholders, there is a risk of interpreting explanations as deterministic rules rather than probabilistic contributions. Therefore, XAI results should be interpreted cautiously and supported by multiple methods and sanity checks.

Also, going back to these figures we can see an initially strange behaviour, where age seems to be a feature of high importance but not when it comes to permutation. mpurity-based feature importance assigns high relevance to AGE due to its continuous nature and frequent use in decision tree splits. However, permutation importance shows that permuting AGE has little effect on ROC-AUC, indicating that while AGE is used internally by the model, it does not provide unique predictive information. This highlights the known bias of impurity-based importance towards continuous features and supports the use of permutation importance for feature selection.

Conclusion

This project demonstrated the practical use of explainable AI techniques in the context of credit card default prediction using a Random Forest model. Multiple explanation methods were applied to understand model behavior and to assess the reliability of feature importance rankings. The analysis revealed strong redundancy among repayment history variables and an over-reliance on short-term indicators, which was confirmed through correlation analysis and drop-column sanity checks.

Based on these findings, the model was redesigned using aggregated repayment features that summarize average behavior, extreme risk, temporal trends, and weighted moving averages. This approach reduced multicollinearity and short-term bias while preserving the main predictive signal. The final model achieved comparable ROC-AUC performance and relied on a more stable and interpretable feature set. Overall, the work illustrates how XAI can support informed model refinement while also highlighting the importance of cautious interpretation, particularly in high-stakes applications such as credit risk assessment.

References

Pizarroso et al. 2023, Explainable Artificial Intelligence (XAI) techniques based on partial derivatives with applications to neural networks. https://repositorio.comillas.edu/xmlui/handle/11531/85986

Dataset: https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset

Initial model:

G. Preda, “Default of Credit Card Clients – Predictive Models,” *Kaggle Notebook*. [Online]. Available: https://www.kaggle.com/code/gpreda/default-of-credit-card-clients-predictive-models#Check-the-data