Credit Risk Prediction Using Machine Learning

Sherynne Derasse, Raphaël Paviet, Basile Minard

December 2025

Contents

List of Figures

List of Tables

1 Introduction and Business Case

Credit risk management is a central concern for financial institutions, as lending activities inherently expose banks and investors to the risk of borrower default. Accurately assessing this risk at the time of loan origination is crucial to ensure portfolio stability, optimize capital allocation, and limit potential financial losses. In recent years, the growing availability of borrower-level data has made machine learning techniques particularly attractive for enhancing traditional credit scoring approaches.

This project focuses on the prediction of the probability of loan default using data provided by Lending Club, a peer-to-peer lending platform that connects individual borrowers with investors. Each observation in the dataset corresponds to a loan issued to a borrower and contains information available at the time of loan approval, such as loan characteristics (amount, interest rate, maturity), borrower financial indicators (income, employment status), and credit history variables.

The business objective of this project is to develop a credit risk scoring model capable of estimating the likelihood that a borrower will default on a loan. In this context, a default is defined as a loan that has been classified as *Charged Off* or *Default*, while non-defaulted loans correspond to fully repaid or currently active loans. From a banking perspective, such a model supports decision-making processes by allowing institutions to better screen borrowers, adjust lending conditions, and manage expected losses.

The problem is formulated as a supervised binary classification task. Given a set of borrower and loan features observed at origination, the goal is to predict a binary target variable indicating whether the loan will default. Particular attention is paid to the asymmetric nature of misclassification costs in a banking context, where failing to identify a high-risk borrower (false negative) can result in significantly higher losses than incorrectly rejecting a low-risk applicant.

This project adopts a machine learning approach aligned with real-world banking practices. Beyond predictive performance, emphasis is placed on model robustness, handling of imbalanced data, and interpretability, which are essential requirements for credit risk models deployed in financial institutions. The overall objective is not only to achieve strong statistical performance but also to provide insights that are meaningful and actionable from a financial risk management perspective.

2 Dataset Description

2.1 Data Source

The dataset used in this project was obtained from Kaggle and originates from Lending Club, a US-based peer-to-peer lending platform. Lending Club acts as an intermediary between borrowers seeking personal loans and investors willing to fund those loans in exchange for interest income. In this setting, investors are exposed to credit risk, which depends on the borrower profile and the loan characteristics available at origination.

The dataset contains historical information on loans issued through Lending Club over the period 2007–2018 (depending on the version of the dataset), including the loan status and the latest payment information. It is provided in a tabular format with 2 260 701 entries and 151 variables, covering borrower socio-economic attributes, credit history indicators, and loan contract terms.

2.2 Target Variable and Business Meaning

The key outcome variable is the loan status, which reflects whether the borrower repaid the loan successfully or experienced repayment difficulties. For the purpose of credit risk modelling, this project focuses on the probability of default. In practice, a default event is defined using loan statuses such as *Charged Off* and *Default*. Other statuses (e.g., *Late*, *In Grace Period*) can be considered as indicators of repayment distress and are useful for exploratory analysis.

From a banking perspective, this target variable is directly linked to expected loss and portfolio risk management. Predicting defaults at origination can support lending decisions, pricing policies, and risk-based allocation of capital.

2.3 Features Overview

The available features include:

- **Loan contract terms**: loan amount, term (36 or 60 months), interest rate, installment amount.
- **Borrower financial and employment information**: annual income, employment length, home ownership status, verification status.
- **Credit history variables**: delinquency counts, number of credit inquiries, revolving utilization, total accounts, public records, and FICO-related variables.
- **Application and administrative fields**: application type (individual/joint), geographic fields (state and zip code prefix), and other operational variables.

Overall, the dataset provides a rich representation of borrower risk at origination. However, it also contains practical challenges commonly observed in real credit datasets, such as missing values, heterogeneous variable types (numerical and categorical), and an imbalanced distribution of the target variable (defaults being less frequent than non-defaults). These aspects motivate careful preprocessing and robust evaluation strategies in the subsequent sections.

3 Exploratory Data Analysis

The exploratory data analysis (EDA) phase aims to provide a first understanding of the structure, quality, and economic meaning of the Lending Club dataset. From a credit risk perspective, this step is essential to identify key patterns, detect potential data issues, and assess whether the available information is consistent with known risk drivers in lending activities.

3.1 Target Variable Distribution

The first observation concerns the distribution of the target variable, corresponding to loan default. The dataset exhibits a strong class imbalance, with a large majority of loans classified as non-defaulted (fully paid or currently active) and a significantly smaller proportion of defaulted loans (charged off or default).

This imbalance reflects a realistic lending environment, where most borrowers repay their loans, but it also introduces methodological challenges for machine learning models. In particular, accuracy alone becomes a misleading performance metric, as a naïve model predicting only non-defaults would achieve high accuracy while failing to identify risky borrowers. This observation motivates the later use of metrics such as ROC-AUC, recall, and precision, which are better suited to credit risk applications.

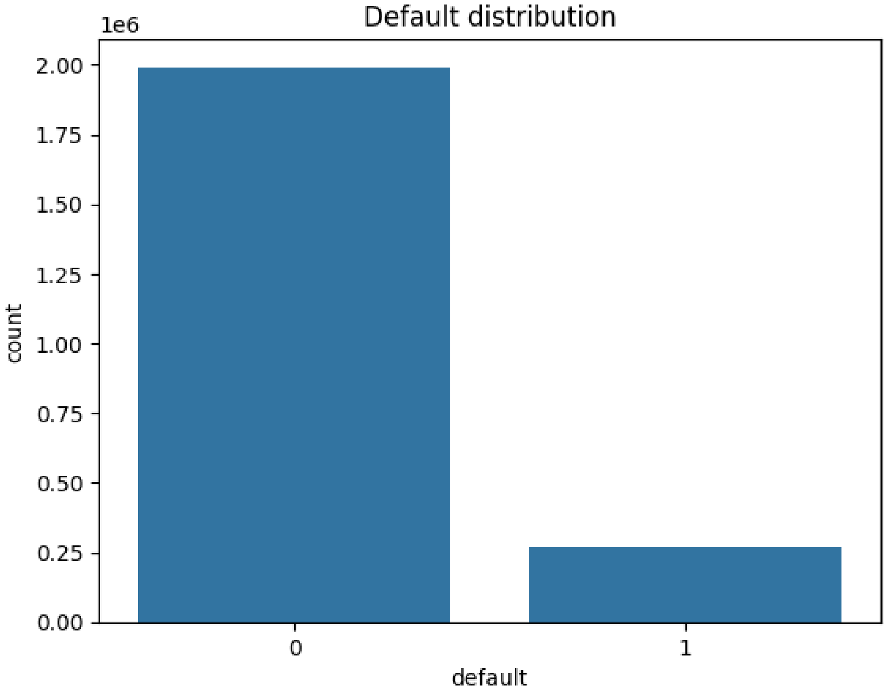

Figure 1: Distribution of the target variable showing a strong class imbalance between defaulted and non-defaulted loans.

As shown in Figure 1, the dataset exhibits a strong class imbalance, with defaulted loans representing a minority of observations.

3.2 Analysis of Key Financial Variables

Several core financial variables were examined to assess their relationship with loan performance. Among them, the interest rate, loan term, loan amount, and debt-to-income ratio emerge as economically meaningful indicators of credit risk.

Loans with higher interest rates tend to be associated with a higher incidence of default, which is consistent with risk-based pricing practices in banking. Similarly, loans with longer maturities (60 months) exhibit a higher proportion of adverse loan outcomes compared to shorter-term loans (36 months). Longer maturities increase exposure to macroeconomic uncertainty and borrower income volatility, which naturally elevates default risk.

Debt-to-income ratios also display higher values on average for defaulted loans, suggesting that borrowers with heavier debt burdens relative to income are more vulnerable to repayment difficulties. These observations align well with standard credit risk theory and confirm the relevance of the dataset for predictive modelling.

3.3 Correlation Analysis

A correlation analysis was conducted on numerical variables to identify relationships between features and potential multicollinearity issues. While no single variable exhibits an extreme linear correlation with default, several variables related to credit utilization, interest rate, and delinquency history show moderate associations.

This result highlights an important characteristic of credit risk modelling: default events are typically driven by a combination of factors rather than a single dominant variable. Consequently, multivariate models are required to capture nonlinear interactions and complex dependencies between borrower attributes.

Additionally, the correlation analysis reveals groups of highly correlated variables, particularly among balance-related and utilization metrics. This insight informs subsequent preprocessing steps, including feature selection and dimensionality reduction, to mitigate redundancy and improve model stability.

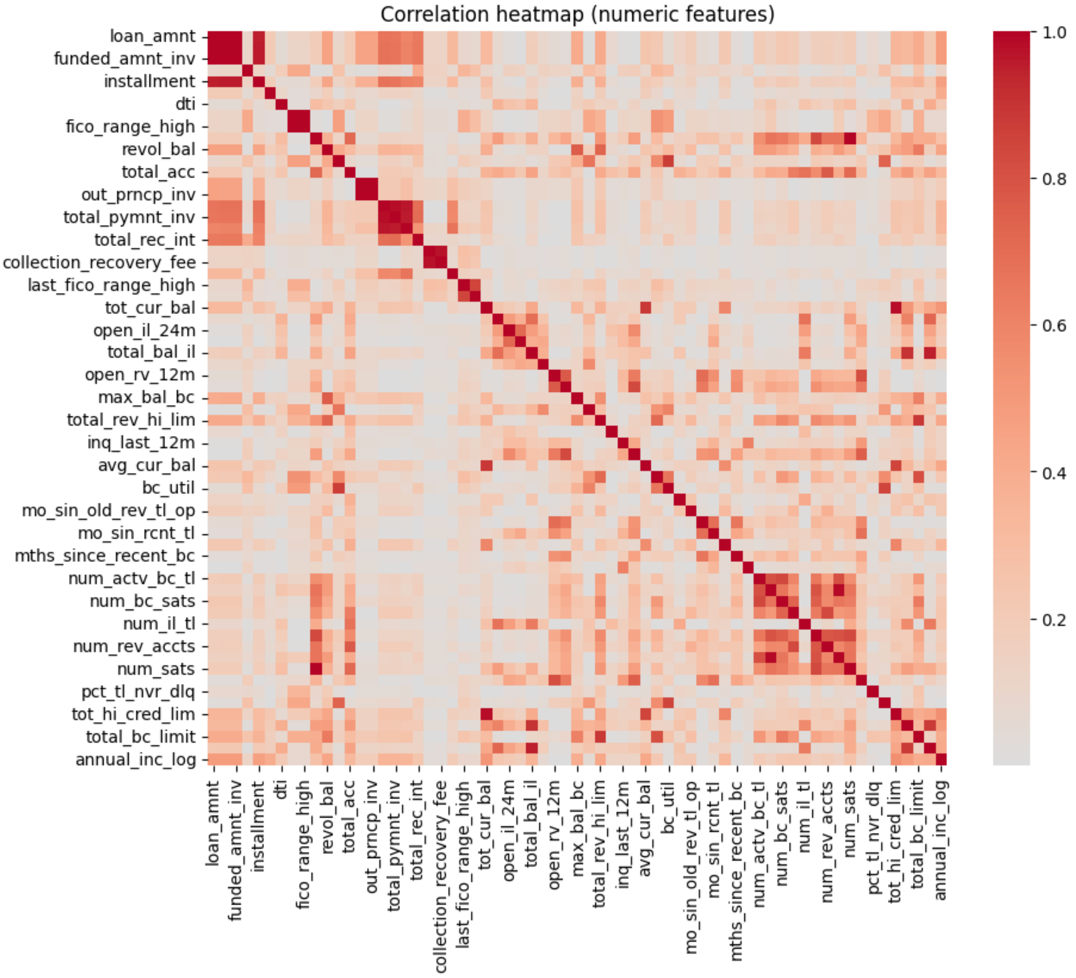

Figure 2: Correlation heatmap of selected numerical features, highlighting groups of correlated credit-related variables.

3.4 Economic Interpretation and Risk Patterns

Beyond purely statistical observations, the EDA confirms several economically intuitive patterns. Higher-risk loans tend to combine multiple unfavorable characteristics, such as elevated interest rates, longer maturities, higher leverage, and weaker credit histories. Importantly, these patterns are observable at loan origination, which is critical for building a predictive model usable in real-world lending decisions.

Overall, the exploratory analysis validates both the quality and the relevance of the dataset for credit risk modelling. It also provides a strong economic foundation for the preprocessing and modelling choices developed in the subsequent sections.

4 Problem Formalization

From a machine learning perspective, the credit risk prediction task addressed in this project is formulated as a supervised binary classification problem. Let $X \in R^p$ denote the vector of borrower and loan characteristics observed at the time of loan origination, and let $Y \in \{0, 1\}$ represent the target variable, where $Y = 1$ corresponds to a defaulted loan and $Y = 0$ to a non-defaulted loan.

The objective is to learn a decision function $f(X)$ that estimates the conditional probability of default:

$$P(Y = 1 \mid X),$$

using historical loan data. This probabilistic formulation is particularly relevant in a banking context, as it allows institutions to rank borrowers by risk and to define acceptance thresholds or pricing strategies based on estimated default probabilities rather than binary predictions alone.

A key characteristic of the problem is the strong class imbalance observed in the dataset, with default events representing a minority of observations. In addition, the cost of misclassification is asymmetric: failing to identify a borrower who will default (false negative) typically leads to higher financial losses than incorrectly classifying a low-risk borrower as risky (false positive). As a result, model evaluation cannot rely solely on accuracy.

Consequently, performance is assessed using metrics better suited to credit risk applications, including the Area Under the Receiver Operating Characteristic Curve (ROC-AUC), which captures the model's ability to rank borrowers by risk, as well as precision and recall on the default class. Particular attention is given to recall, as it reflects the model's capacity to detect high-risk borrowers.

Finally, all modelling choices are made under the constraint that input variables must be available at loan origination, ensuring that the resulting models remain applicable in real-world lending decisions and do not rely on post-default information.

5 Data Preprocessing

The data preprocessing stage aims to transform the raw Lending Club dataset into a clean, consistent, and model-ready format while preserving the economic meaning of the variables. In a credit risk context, particular care must be taken to avoid information leakage and to ensure that all features used for prediction are available at the time of loan origination.

5.1 Data Cleaning and Quality Checks

Initial preprocessing steps include the removal of duplicate observations and the verification of variable types. Several variables were identified as identifiers or administrative fields (such as loan IDs or URLs) and were excluded, as they do not carry predictive information. Additionally, features containing post-default information, such as recovery amounts or payment histories observed after loan issuance, were removed to prevent data leakage.

5.2 Missing Value Treatment

The dataset contains a significant number of missing values, which is typical of real-world credit data. Missing value treatment was performed using economically motivated strategies. Numerical variables were imputed using the median, which is robust to outliers and preserves the central tendency of skewed distributions. Categorical variables were imputed using the most frequent category.

This approach ensures consistency across observations while avoiding the introduction of unrealistic values. Variables with an excessive proportion of missing values and limited economic relevance were discarded.

5.3 Outlier Treatment

Several financial variables, such as annual income, loan amount, and revolving balances, exhibit heavy-tailed distributions with extreme values. Outliers were handled using an interquartile range (IQR)-based approach combined with economic reasoning. Extreme values beyond reasonable thresholds were capped to limit their influence on model estimation, while preserving the overall structure of the data.

This step improves model stability, particularly for linear models, without removing economically meaningful high-risk observations.

5.4 Feature Encoding

Categorical variables were transformed into numerical representations suitable for machine learning models. Nominal variables were encoded using one-hot encoding, while ordinal variables were mapped to ordered numerical scales where a natural ordering existed. This encoding strategy allows models to exploit categorical information without imposing artificial numerical relationships.

5.5 Feature Scaling

Numerical features were scaled using standardization to ensure comparable ranges across variables. Feature scaling is particularly important for models sensitive to variable magnitude, such as logistic regression and neural networks. Tree-based models were evaluated both with and without scaling, as they are generally invariant to monotonic transformations.

5.6 Train-Test Split Strategy

To evaluate model performance, the dataset was split into training and testing subsets using a stratified approach, preserving the proportion of defaulted and non-defaulted loans in both sets. This strategy ensures that model evaluation reflects the original class imbalance and provides a reliable estimate of out-of-sample performance.

Overall, the preprocessing pipeline balances statistical robustness, economic interpretability, and practical deployment constraints, forming a solid foundation for the subsequent modeling phase.

6 Models Presentation

This section presents the baseline predictive model used for credit risk estimation, namely logistic regression. The objective is twofold: first, to establish a transparent and interpretable benchmark model commonly used in banking applications; second, to highlight the importance of proper data validation through the detection and correction of data leakage.

6.1 Baseline Model: Logistic Regression

Logistic regression is a standard tool in credit risk modeling due to its simplicity, interpretability, and probabilistic output. The model estimates the probability of default as a logistic transformation of a linear combination of borrower and loan characteristics available at origination.

An initial logistic regression model was trained on the preprocessed dataset using a temporal train-test split. The resulting performance was exceptionally high, with an out-of-time validation ROC-AUC close to 0.999. While such a result may appear desirable at first glance, it is unrealistic for a complex real-world credit dataset and strongly suggests the presence of data leakage.

6.2 Detection of Data Leakage

Data leakage occurs when explanatory variables contain information that would not be available at the time of prediction, thereby allowing the model to implicitly "learn the future." In a lending context, this typically involves post-issuance variables related to loan repayment or recovery.

From an economic perspective, the model should rely exclusively on borrower characteristics and loan terms observed at origination, such as income, debt-to-income ratio, interest rate, or credit score. Variables describing repayment behavior after loan issuance violate temporal causality and artificially inflate predictive performance.

To empirically identify potential leakage variables, a univariate ROC-AUC analysis was conducted for each numerical feature against the default indicator. Variables exhibiting near-perfect or abnormal discriminative power were flagged as candidates for leakage.

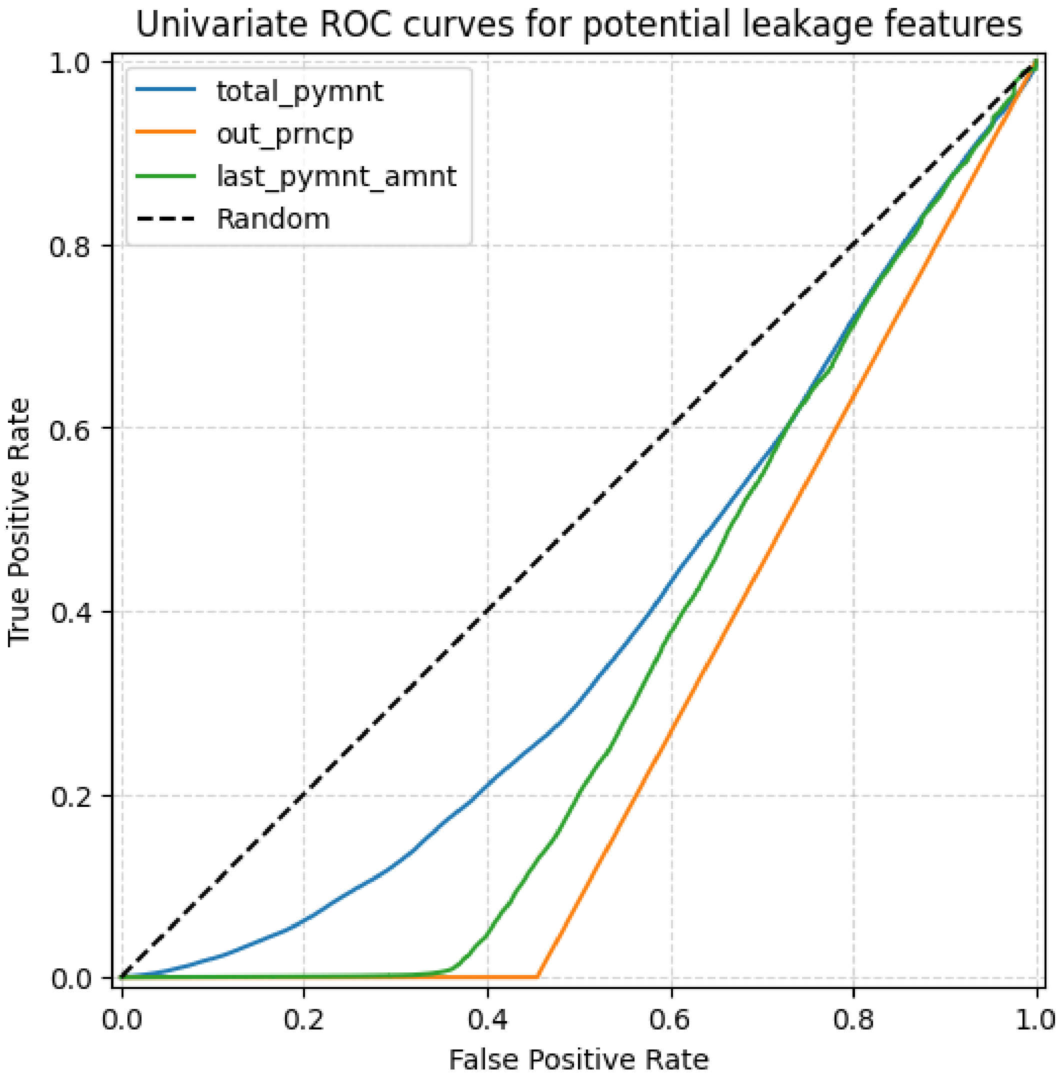

Figure 3: Univariate ROC curves for selected numerical features, highlighting variables with abnormal discriminative power indicative of post-issuance information leakage.

6.3 Leakage Analysis and Feature Removal

The univariate analysis revealed that variables such as *total_pymnt*, *out_prncp*, *last_pymnt_amnt*, and certain collection-related features exhibited strong predictive power due to their direct relationship with loan repayment outcomes.

Although some of these variables showed inverse correlations with default (AUC below 0.5), their economic interpretation remains post-outcome rather than pre-issuance. Consequently, even when statistically valid, these features cannot be used in a predictive model intended for loan approval decisions.

To preserve temporal causality and ensure realistic deployment conditions, all identified post-issuance variables were removed from the dataset before retraining the model.

6.4 Model Performance After Leakage Correction

After removing leakage-prone features, the logistic regression model was retrained using the same temporal validation framework. As expected, performance metrics decreased to more realistic levels, with an out-of-time ROC-AUC close to 0.96. This reduction confirms that the initial near-perfect performance was driven by information leakage rather than genuine predictive power.

The corrected model exhibits a strong ability to rank borrowers by risk, while revealing a significant class imbalance effect. In particular, recall on the default class remains high, indicating that the model successfully identifies most risky borrowers. However, precision is relatively low, reflecting a tendency to generate false positive alerts.

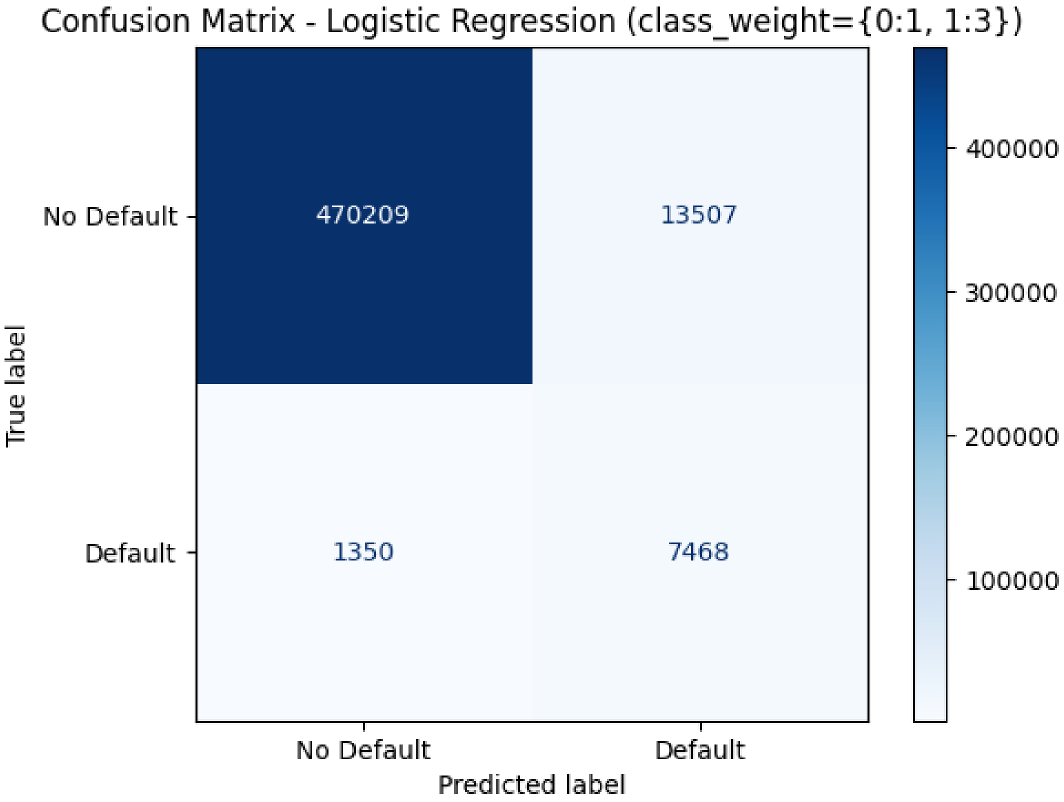

Figure 4: Confusion matrix for the logistic regression model after leakage correction, illustrating the trade-off between default detection and false positive rates.

6.5 Class Imbalance Handling and Risk Trade-Off

To address the imbalance between defaulted and non-defaulted loans, class weighting was introduced into the logistic regression pipeline. By assigning a higher weight to the default class, the model shifts toward a more conservative risk detection strategy.

This adjustment improves the balance between precision and recall, reducing the number of false positives while maintaining a high detection rate of actual defaults. From a banking perspective, this behavior is desirable in early warning or pre-screening systems, where missing high-risk borrowers is typically more costly than generating additional alerts.

Overall, the logistic regression model provides a strong and interpretable baseline. While it demonstrates excellent discriminatory power, its calibration limitations motivate the exploration of more flexible models in subsequent sections.

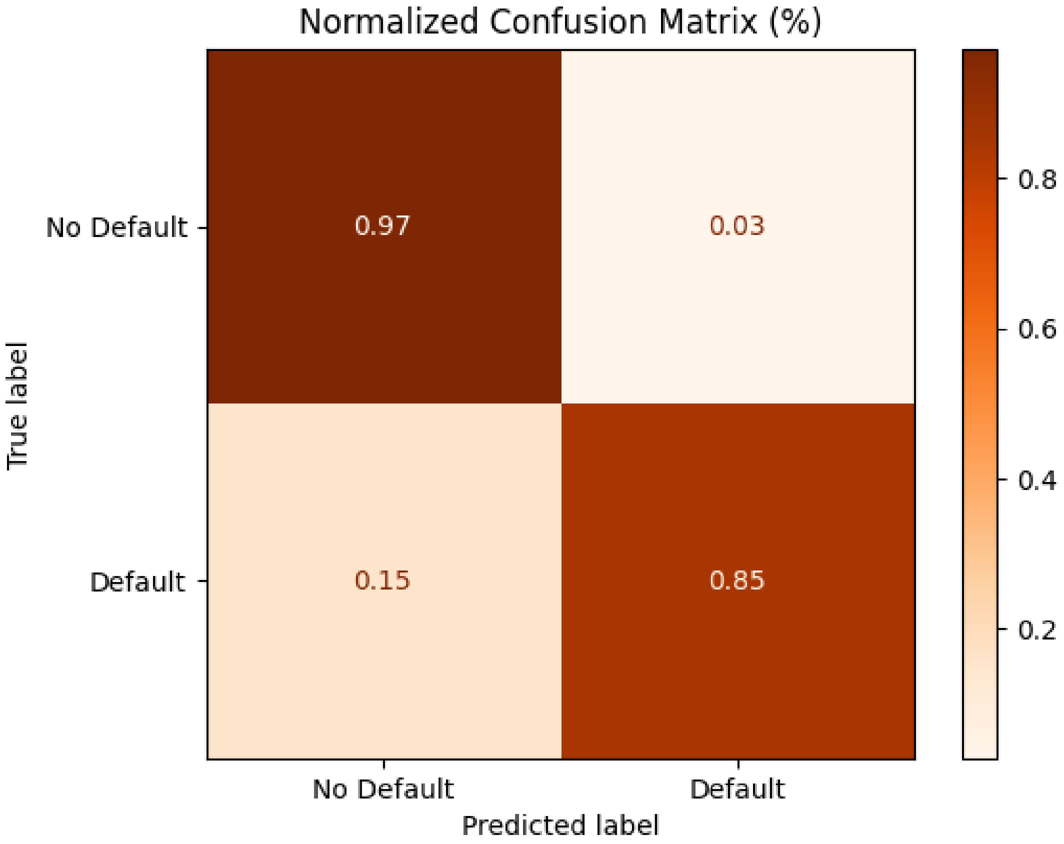

Figure 5: Normalized confusion matrix for the class-weighted logistic regression model, showing improved balance between precision and recall for the default class.

7 Handling Practical Challenges

Building a credit risk model suitable for real-world deployment involves addressing several practical challenges that go beyond raw predictive performance. In particular, temporal stability, class imbalance, and robustness to changing economic conditions are critical requirements in a banking context. This section details the strategies implemented to address these challenges and ensure reliable, economically meaningful predictions.

7.1 Temporal Stability and Time-Based Validation

In credit risk modeling, it is essential that a model trained on historical data remains valid when applied to future loan vintages. A model that performs well on randomly shuffled data but fails on future periods is unsuitable for operational use.

To evaluate temporal robustness, a strict three-way time-based split was implemented:

- Training set: loans issued up to 2016,
- Validation set: loans issued in 2017,

- Out-of-time test set: loans issued in 2018.

This approach closely replicates real deployment conditions, where models are trained on past information and used to score new borrowers. Unlike random splits, it prevents information leakage across time and provides a realistic assessment of generalization performance.

Across different regularization strengths, the logistic regression model exhibits highly consistent ROC-AUC scores between the validation and out-of-time test sets (approximately 0.966 on 2017 data and 0.957 on 2018 data). This stability indicates that the model does not overfit to specific historical patterns and generalizes well across loan vintages. A moderate regularization parameter ($C = 0.1$) was retained, as it ensures an optimal balance between predictive performance and stability, in line with regulatory expectations for credit risk models.

7.2 Class Imbalance and Cost-Sensitive Learning

A major challenge in credit risk modeling is the strong imbalance between non-defaulted and defaulted loans. While default rates represent approximately 14.7% of observations in the training set, they fall below 2% in the 2018 out-of-time test set. Without corrective measures, models tend to favor the majority class, leading to poor detection of risky borrowers.

To address this issue, cost-sensitive learning was implemented through class weighting in the logistic regression model. By assigning a higher weight to the default class, the model becomes more sensitive to risky profiles and prioritizes the detection of potential losses.

From an economic perspective, this behavior is desirable. In lending decisions, failing to identify a borrower who will default typically incurs higher costs than incorrectly flagging a creditworthy borrower. The class-weighted model therefore adopts a conservative risk posture consistent with early warning systems and credit pre-screening applications.

7.3 Synthetic Oversampling with SMOTE

As an alternative imbalance-handling strategy, the Synthetic Minority Oversampling Technique (SMOTE) was applied to the training data. SMOTE generates artificial default observations by interpolating between existing minority-class samples, allowing the model to better learn default-related patterns.

The SMOTE-based model achieves very high recall on defaulted loans (approximately 89%), indicating strong detection capability. However, this improvement comes at the cost of a substantial decrease in precision, resulting in a large number of false positive alerts.

In operational terms, such a model would reject an excessive number of creditworthy borrowers, making it impractical for deployment despite its protective nature. Consequently, SMOTE is deemed overly aggressive for this application, and class weighting is preferred as a more balanced and operationally viable solution.

7.4 Non-Linear Models and Model Robustness

While logistic regression provides interpretability and regulatory transparency, it is inherently limited to linear decision boundaries. To capture non-linear relationships and interactions between borrower characteristics, tree-based ensemble models were also evaluated.

A Random Forest classifier was trained using the same preprocessing pipeline and class weighting strategy. This model achieves a higher out-of-time ROC-AUC (approximately 0.968) and provides a significantly improved balance between precision and recall for the default class. The results indicate that Random Forest captures complex risk patterns while maintaining reasonable alert levels.

Additionally, a gradient boosting model (XGBoost) was implemented with explicit handling of class imbalance via the `scale_pos_weight` parameter. XGBoost achieves strong discriminative performance, with an out-of-time ROC-AUC comparable to Random Forest. However, its predictions remain highly conservative, prioritizing recall over precision and generating a large number of false alerts.

7.5 Economic Interpretation and Model Selection

The comparative analysis highlights a fundamental trade-off in credit risk modeling: maximizing default detection versus limiting false positives. Logistic regression, when properly regularized and class-weighted, provides a robust and interpretable baseline that aligns well with regulatory and operational constraints.

Random Forest improves predictive balance and captures non-linear risk structures, making it attractive for operational decision systems. XGBoost further enhances default detection but may be less suitable for direct deployment due to excessive alert rates.

Overall, these results demonstrate that model selection should not rely solely on performance metrics. Economic costs, business objectives, interpretability requirements, and deployment constraints must all be considered to ensure that the chosen model delivers real value in a banking environment.

7.6 Summary of Practical Takeaways

- Time-based out-of-time validation is essential to assess real-world model robustness.
- Temporal drift can strongly affect operational precision even when AUC remains stable.
- Class imbalance must be handled with economically motivated strategies; class weights provide a better operational trade-off than synthetic oversampling.
- Non-linear models improve risk discrimination but must be evaluated in light of alert costs and interpretability constraints.

8 Results and Model Comparison

We compare multiple models under the same OOTV protocol (train $\leq$ 2016, test = 2018). Performance is assessed using: (i) **AUC** for ranking quality, (ii) **Recall (Default)** to capture risky borrowers (loss avoidance), (iii) **Precision (Default)** to limit false alerts (business/operational cost), (iv) **False Positives (FP)** and **False Negatives (FN)** as direct operational measures.

8.1 Core Models (Baseline Comparison)

Table **??** summarizes the main models (excluding SMOTE, which was discarded).

Table 1: OOTV performance comparison (test year 2018).

Model	AUC	Prec (Def)	Rec (Def)	F1 (Def)	Acc	FP
Logistic Regression (cw)	0.9592	0.356	0.847	0.501	0.970	13 507
Random Forest (balanced)	**0.9683**	**0.384**	0.825	**0.524**	**0.973**	**11 695**
XGBoost (scale_pos_weight)	0.9662	0.210	**0.906**	0.341	0.937	30 058

Interpretation. XGBoost achieves strong recall but at the cost of a very high number of false positives (30k+), which is operationally expensive and would reject many good clients. Logistic Regression provides a strong, explainable baseline. However, **Random Forest offers the best overall compromise**: highest AUC, best F1, and the lowest FP count among the top-performing models. For a production credit scoring setting, this balance is typically preferred.

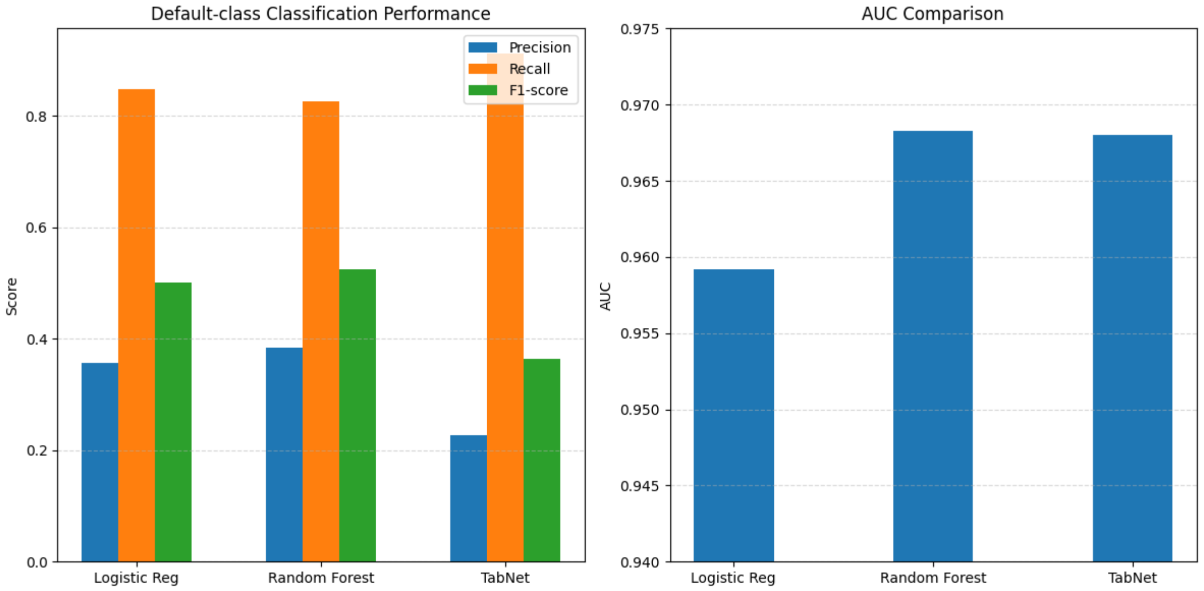

Figure 6: Default-class precision, recall and F1-score comparison. Random Forest provides the best balance between default detection and alert volume, while TabNet prioritizes recall at the expense of precision.

8.2 Innovative Model: TabNet (Deep Learning on Tabular Data)

TabNet is a deep learning architecture designed for tabular datasets. Unlike standard neural networks, TabNet uses **sequential attention** and learns sparse masks to focus on the most informative variables at each decision step. This allows it to capture non-linear interactions while retaining some interpretability through attention masks.

Empirically, TabNet behaves as a **more conservative** model: it captures more true defaults (high recall) but triggers substantially more false alerts (lower precision). This makes it attractive for early-warning monitoring, but less optimal for direct loan approval decisions.

8.3 Bank-Oriented Comparison: Random Forest vs TabNet

To translate model metrics into operational terms, we compare both models directly in terms of true defaults detected (TP), defaults missed (FN), and false alerts (FP).

Table 2: Operational comparison between Random Forest and TabNet (OOTV 2018).

Bank Criterion (Default class)	**Random Forest**	**TabNet**	**Best**
True defaults detected (TP)	7,275	**8,035**	TabNet (+10.4%)
Defaults missed (FN)	1,543	**783**	TabNet (-49.2%)
False alerts (FP)	**11,695**	27,435	Random Forest

Interpretation. TabNet substantially reduces missed defaults, but it more than doubles the false alert volume. This leads to a trade-off: **TabNet is preferable for surveillance / early-warning**, whereas **Random Forest is better suited for lending decisions** where false positives directly reduce business volume.

8.4 Explainability Benchmark: Decision Tree (and Governance)

For governance and interpretability, we train an optimized Decision Tree (GridSearch on a stratified subset due to memory constraints). The tuned tree provides transparent rules and highlights the primary drivers of default risk (e.g., low FICO history, short credit history, larger loan amounts, longer terms). However, despite strong recall, Decision Trees tend to over-alert and remain costly operationally.

8.5 Dimensionality Reduction (PCA) as a Stress Test

We evaluate PCA combined with Random Forest to test whether risk information can be compressed. While PCA reduces dimension, it significantly harms recall (collapse of default detection), which is unacceptable in credit risk because missed defaults translate directly into unexpected losses. PCA is therefore excluded from the final deployment.

8.6 Final Model Ranking and Recommended Strategy

Table 3: Final model comparison and recommended operational usage (OOTV 2018).

Model	**AUC**	**Rec (Def)**	**Prec (Def)**	**F1 (Def)**	**Recommended role**
Random Forest	**0.9683**	0.825	**0.384**	**0.524**	Production credit scoring (best performance / cost balance)
TabNet	0.9680	**0.911**	0.227	0.363	Early-warning system / risk surveillance (maximize default capture)
Logistic Regression (cw)	0.9592	0.847	0.356	0.501	Regulatory-compliant baseline and benchmarking model
Decision Tree (optimized)	0.9609	0.902	0.199	0.326	Governance and explainability reference
RF + PCA	0.9402	0.285	0.693	0.404	Not recommended (recall collapse)

Recommended operational strategy.

- **Primary production scorer**: Random Forest (best compromise between losses avoided and alert cost).

- **Early-warning system**: TabNet (maximize risk capture, accept higher alert volume).
- **Model governance and explainability**: optimized Decision Tree + Logistic Regression baseline.

Key business insights. Across explainable models and rule extraction, the most consistent drivers of default risk include: (i) weak credit score history (FICO-related variables), (ii) short credit maturity (recent `earliest_cr_line`), (iii) larger loan exposure and longer terms, (iv) behavioral signals of credit stress captured through credit activity variables.

Limitations and future improvements.

- **Probability calibration**: strong AUC does not guarantee calibrated PDs; calibration curves and Platt/Isotonic calibration could improve decision thresholds.
- **Threshold optimization**: moving beyond 0.5 to optimize a bank-specific cost function (FN cost $\gg$ FP cost).
- **Drift monitoring**: default rate drift suggests the need for periodic re-training and stability monitoring.
- **Compute constraints**: TabNet training time is significant; GPU or batch scoring is preferable.

Final conclusion. Random Forest is selected as the **production-ready credit risk scorer**. It provides strong ranking performance while maintaining a manageable false alert volume. TabNet is retained as an advanced option for early-warning monitoring, and the Decision Tree provides an essential interpretability reference for governance and validation.

9 Conclusion

9.1 Business-Oriented Conclusion (Banking Perspective)

This project aimed to build a credit risk scoring model suitable for real-world banking deployment, where predictive performance must be balanced with economic costs, regulatory constraints, and operational feasibility.

From a business standpoint, the results clearly demonstrate that **model selection cannot rely solely on a single performance metric such as AUC**. While all tested models achieve strong discriminative power, their operational behavior differs substantially once precision, recall, and alert volumes are taken into account.

The **Random Forest model emerges as the most appropriate production-ready scorer**. It offers the best compromise between:

- **Risk detection** (high recall on defaults),
- **Control of false alerts** (reasonable precision),
- **Overall stability across time**, validated under a strict out-of-time framework.

In a lending context, this balance is critical. Excessive false positives lead to the rejection of creditworthy borrowers and lost business opportunities, while excessive false negatives expose the bank to unexpected credit losses. Random Forest achieves a pragmatic middle ground, making it particularly well suited for **loan approval, pricing, and credit policy enforcement**.

More conservative models such as **TabNet** demonstrate excellent default detection capability and are highly valuable for **early-warning systems and portfolio surveillance**, where the priority is to flag emerging risks rather than to make immediate lending decisions. Conversely, simpler models such as **logistic regression**, while slightly less powerful, remain highly relevant in regulated environments due to their transparency and ease of governance.

Overall, this work reflects the reality of modern credit risk management: **there is no universally optimal model**, but rather models that are optimal for specific business objectives. The final strategy therefore naturally combines:

- Random Forest for **operational credit scoring**,
- Decision Trees for **interpretability and governance**,
- TabNet for **advanced risk monitoring and early detection**.

9.2 Technical Conclusion (Modeling and Implementation Perspective)

From a technical perspective, this project highlights the importance of **rigorous modeling discipline** when working with large-scale, real-world financial datasets such as LendingClub.

Several methodological choices proved decisive:

- **Strict temporal validation (OOTV)** was essential to avoid over-optimistic performance estimates and to properly assess model generalization under distribution shifts.
- **Data leakage detection and removal** was a key step, as initial near-perfect AUC values revealed the presence of post-origination variables that artificially inflated performance.
- **Class imbalance handling** required careful calibration. While SMOTE improved recall, it generated unacceptable levels of false positives. Cost-sensitive learning via class weighting provided a more operationally robust solution.
- **Model diversity** (linear, tree-based, ensemble, and deep learning) allowed us to assess trade-offs between interpretability, flexibility, and computational cost.

The experimental results confirm that:

- Linear models provide strong baselines and temporal stability but struggle with non-linear interactions.
- Tree-based ensembles (Random Forest, XGBoost) capture complex patterns and deliver superior performance on tabular credit data.
- Deep learning architectures such as **TabNet** offer promising performance and interpretability through attention mechanisms, but at the cost of higher computational complexity and operational overhead.

From an implementation standpoint, the project also illustrates practical constraints encountered in applied machine learning:

- Memory limitations required subsampling strategies for hyperparameter tuning.
- Execution time became a relevant factor when comparing deep learning and ensemble methods.
- Model interpretability tools (Decision Trees, feature importance, confusion matrices) were indispensable to validate economic consistency.

In conclusion, this work demonstrates that **robust credit risk modeling is not merely a coding exercise**, but a careful integration of data engineering, statistical rigor, economic reasoning, and business constraints. The final models are both technically sound and economically meaningful, making them suitable foundations for real-world credit decision systems.

10 References

References

[1] LendingClub Corporation. *LendingClub Loan Data.* Available at: https://www.lendingclub.com, 2007–2018.

[2] LendingClub kaggle. *LendingClub Loan Data.* Available at: https://www.kaggle.com/datasets/adarshsng/lending-club-loan-data-csv,

[3] LendingClub kaggle. *LendingClub risk metrics.* Available at: https://www.kaggle.com/code/janiobachmann/lending-club-risk-analysis-and-metrics,

[4] Arik, S. O., Pfister, T. TabNet: Attentive Interpretable Tabular Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[5] Article on Kaggle *Explication of what is TabNet.* ,. Avaible at : https://www.kaggle.com/code/enigmak/tabnet-deep-neural-network-for-tabular-data,

[6] implementation TabNet GitHub. *tabnet GitHub.* Available at: https: https://github.com/dreamquark-ai/tabnet,