MAS ETH in AI and Digital Technology

Predicting Risk of Customer Loan Default

AI-supported Loan Default Prediction for Faster and Fairer Credit Decisions

Loan Pirates

Frank Roth, Linda Carmichael, Christina Ntenekou

Spring 2026

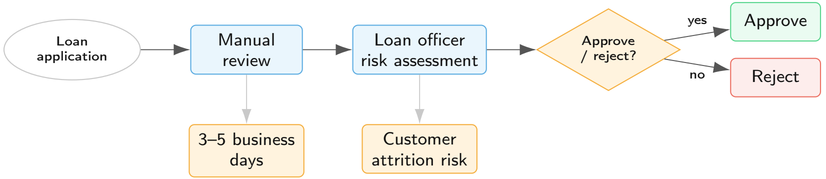
MAS ETH in AI and Digital Technology
Predicting Risk of Customer Loan Default
AI-supported Loan Default Prediction for Faster and Fairer Credit Decisions
Loan Pirates
Frank Roth, Linda Carmichael, Christina Ntenekou
Spring 2026

Section 1

Business Context

Section 1
Business Context

Current Process



The current SwissCredit process is manual, slow, and costly: more than 50,000 applications per month, 3–5 business days per decision, roughly 35% attrition among otherwise qualified applicants, and more than CHF 85 operational cost per application.

2026-06-11

Loan Pirates
└Business Context
└Current Process

Current Process

The current SwissCredit process is manual, slow, and costly: more than 50,000 applications per month, 3–5 business days per decision, roughly 35% attrition among otherwise qualified applicants, and more than CHF 85 operational cost per application.

Business Challenge

- **Time to market:** How can the process be accelerated?
  The digital-native competitors can approve loans within hours, while SwissCredit's manual process takes several days.
- **Risk discipline:** How do we keep the process running efficiently?
  Faster decisions must not increase portfolio default rates beyond the current 20% threshold.
- **Regulated AI context:** How do we remain AI-compliant?
  Creditworthiness evaluation is a high-risk use case that requires transparency, oversight, documentation, and bias monitoring.
- **Stakeholder trade-offs:** How do we keep Stakeholders happy?
  Risk, compliance, retail banking, IT, loan officers, and customers optimize for different outcomes.

Business Challenge

- **Time to market:** How can the process be accelerated?
  The digital-native competitors can approve loans within hours, while SwissCredit's manual process takes several days.
- **Risk discipline:** How do we keep the process running efficiently?
  Faster decisions must not increase portfolio default rates beyond the current 20% threshold.
- **Regulated AI context:** How do we remain AI-compliant?
  Creditworthiness evaluation is a high-risk use case that requires transparency, oversight, documentation, and bias monitoring.
- **Stakeholder trade-offs:** How do we keep Stakeholders happy?
  Risk, compliance, retail banking, IT, loan officers, and customers optimize for different outcomes.

Objectives and Key Results

- **Accelerate approvals:** Reduce review and decision time from 3–5 days to under 30 minutes per application.
- **Protect portfolio quality:** Keep default rates below the current 20% threshold while improving process efficiency.
- **Support fair decisions:** Exclude sensitive attributes (eg race) from training and retain them only for fairness auditing.
- **Enable explainable decision support:** Provide global and local explanations, audit logs, and human review for uncertain or high-impact cases.

Objectives and Key Results

- **Accelerate approvals:** Reduce review and decision time from 3–5 days to under 30 minutes per application.
- **Protect portfolio quality:** Keep default rates below the current 20% threshold while improving process efficiency.
- **Support fair decisions:** Exclude sensitive attributes (eg race) from training and retain them only for fairness auditing.
- **Enable explainable decision support:** Provide global and local explanations, audit logs, and human review for uncertain or high-impact cases.

Key Stakeholders

Stakeholder	Primary Concerns	Presentation Success Metrics
CEO / Board	Regulatory compliance, competitiveness	Time to market, audit, customer retention
Chief Risk Officer	Model accuracy, portfolio quality	AUC, recall, default-rate control
Retail Bank Head	Customer satisfaction, processing speed	Decision latency, churn reduction
Compliance Officer	Regulatory adherence, audit readiness	Auditable oversight, human review
IT Director	Reliability, integration, lifecycle	API uptime, monitoring, CI test coverage
Loan Officers	Tool usability, clarity of explanations	Local explanations, override workflow
Customers	Fair treatment, speed, transparency	Bias metrics, adverse-action reasons

Key Stakeholders

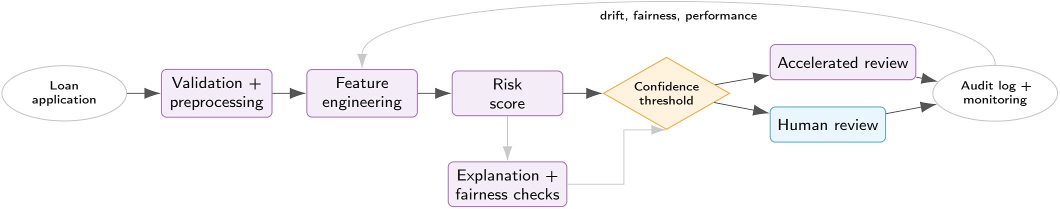
Stakeholder	Primary Concerns	Presentation Success Metrics
CEO / Board	Regulatory compliance, competitiveness	Time to market, audit, customer retention
Chief Risk Officer	Model accuracy, portfolio quality	AUC, recall, default-rate control
Retail Bank Head	Customer satisfaction, processing speed	Decision latency, churn reduction
Compliance Officer	Regulatory adherence, audit readiness	Auditable oversight, human review
IT Director	Reliability, integration, lifecycle	API uptime, monitoring, CI test coverage
Loan Officers	Tool usability, clarity of explanations	Local explanations, override workflow
Customers	Fair treatment, speed, transparency	Bias metrics, adverse-action reasons

Section 2

Data & Methodology

Section 2
Data & Methodology

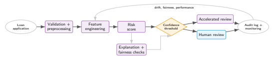
AI Augmented Process



2026-06-11

Loan Pirates
└Data & Methodology
└AI Augmented Process

AI Augmented Process



Data Assets

- Historical U.S.-based consumer lending data with financial indicators, repayment outcomes, temporal fields, categorical variables, and demographic attributes for audit.
- Binary target: borrower fully repaid the loan (non-default) versus default.

Data Group	Features from the project document	Modeling Role
Loan terms	Loan amount, term, installment burden	Affordability and repayment capacity
Borrower finances	Income, revolving balance, credit limits	Leverage, utilization, liquidity pressure
Credit profile	FICO score, account utilization aggregates	Creditworthiness, monitored for opacity
Categorical fields	Household, purpose, application type	Contextual factor after one-hot encoding
Protected attributes	Race indicators	Excluded from training, fairness audit

Data Assets

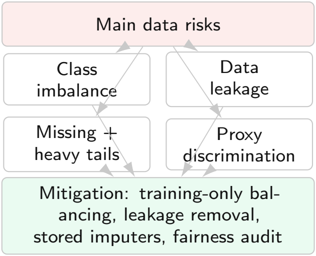
- Historical U.S.-based consumer lending data with financial indicators, repayment outcomes, temporal fields, categorical variables, and demographic attributes for audit.
- Binary target: borrower fully repaid the loan (non-default) versus default.

Data Group	Features from the project document	Modeling Role
Loan terms	Loan amount, term, installment burden	Affordability and repayment capacity
Borrower finances	Income, revolving balance, credit limits	Leverage, utilization, liquidity pressure
Credit profile	FICO score, account utilization aggregates	Creditworthiness, monitored for opacity
Categorical fields	Household, purpose, application type	Contextual factor after one-hot encoding
Protected attributes	Race indicators	Excluded from training, fairness audit

Data Characteristics Highlights and ETL

The raw dataset contains several characteristics that require special handling:

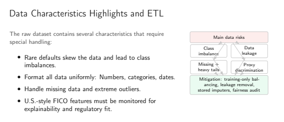
- Rare defaults skew the data and lead to class imbalances.
- Format all data uniformly: Numbers, categories, dates.
- Handle missing data and extreme outliers.
- U.S.-style FICO features must be monitored for explainability and regulatory fit.



2026-06-11

Loan Pirates
└Data & Methodology

└Data Characteristics Highlights and ETL



Feature Engineering Highlights

IAUR	BCUR	DTI
$bal_{IL}/limit_{IL}$	$(limit_{BC} - open_{BC})/limit_{BC}$	$revol\_bal/annual\_inc$
Installment credit utilization	Bank-card utilization pressure	Revolving debt versus income

PTI	Non-linear terms	Interaction flag
$12 \cdot payment/income$	$income^2$, DTI^2, $util^2$	High DTI $\wedge$ high utilization
Annualized installment burden	Curvature for linear models	Liquidity-stress indicator

Feature Engineering Highlights
IAUR $bal_{IL}/limit_{IL}$ Installment credit utilization
BCUR $(limit_{BC} - open_{BC})/limit_{BC}$ Bank-card utilization pressure
DTI $revol\_bal/annual\_inc$ Revolving debt versus income
PTI $12 \cdot payment/income$ Annualized installment burden
Non-linear terms $income^2$, DTI^2, $util^2$ Curvature for linear models
Interaction flag High DTI $\wedge$ high utilization Liquidity-stress indicator

Installment Account Utilization Ratio (IAUR) The Installment Account Utilization Ratio measures the proportion of outstanding installment balances relative to the total granted installment credit limit

Bank Card Utilization Ratio (BCUR) The Bank Card Utilization Ratio quantifies revolving credit pressure on bank-issued credit cards. In the implementation, utilization was derived using the open-to-buy variable:

Debt-to-Income Ratio (DTI) annual income: The Debt-to-Income ratio measures revolving debt relative to

Payment-to-Income Ratio (PTI) A proxy for installment burden was constructed by approximating monthly principal payments:

Non-Linear Transformations To allow linear models to capture curvature effects without sacrificing interpretability, selected quadratic terms were introduced:

Interaction Indicator This indicator captures scenarios in which high leverage and high credit utilization jointly occur, reflecting elevated liquidity stress.

AI/ML Algorithm Selection and Rationale

Model choice is not based on accuracy alone: fairness impact, explainability, computational efficiency, and regulated-deployment readiness are explicit selection dimensions.

Candidate	Why include it?	Selection Criteria
Logistic Regression	Transparent benchmark, useful for governance and sanity checks	Interpretability and baseline performance
Random Forest	Robust non-linear tabular model with stable performance	Recall, robustness, operational simplicity
XGBoost	**Strong gradient-boosting candidate for structured tabular credit-risk data**	**Highest AUC, efficient training, deployability**
MLP (sklearn)	Tests feed-forward neural alternative in tabular setting	Potential non-linear lift versus longer training time
MLP (PyTorch)	Engineering flexibility and hardware-acceleration path	Runtime flexibility versus predictive quality

This slide explains why we evaluated several model families instead of selecting a model based on accuracy alone. In a regulated credit-risk setting, we also need to consider explainability, fairness impact, computational efficiency, operational simplicity, and deployment readiness.

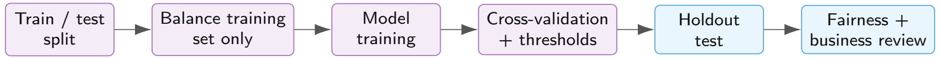
Logistic Regression was included as a transparent benchmark. It is useful for governance discussions and for checking whether more complex models provide a meaningful improvement over a simple baseline.

Random Forest was included because it is a robust non-linear model for tabular data. It is easier to operate than boosting in some cases, and it provides stable performance, but in our results it did not outperform XGBoost.

XGBoost is highlighted because it became the preferred model. It is well suited for structured tabular data, handles non-linear relationships effectively, and achieved the strongest predictive result, especially in terms of AUC. It also offers efficient training and a practical deployment path.

The MLP models were included to test whether a neural-network approach could provide additional non-linear lift. However, XGBoost still outperformed the MLP alternatives in our comparison.

Training Process & Validation



- Validation/test distributions remain unchanged so metrics reflect real-world class imbalance.
- Metrics combine AUC, accuracy, precision, recall, F1, training time, default-rate control, and fairness indicators.
- Thresholds are tuned with business costs in mind: missed defaults versus unnecessary rejection of good customers.

Training Process & Validation

Train / test split → Balance training set only → Model training → Cross-validation + thresholds → Holdout test → Fairness + business review

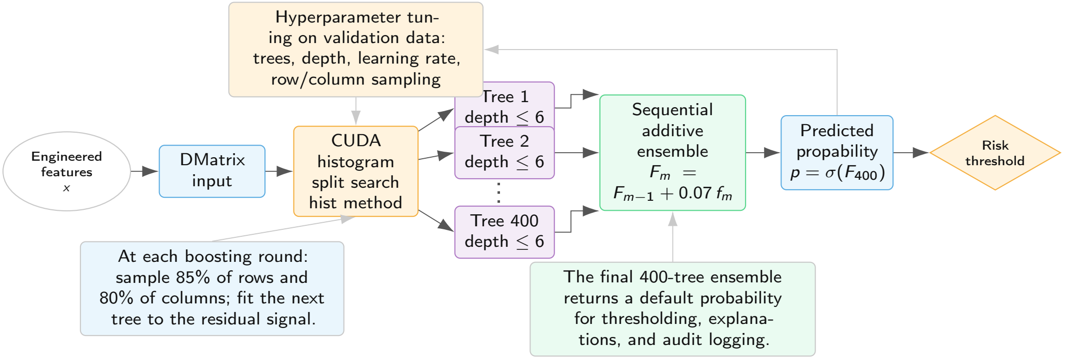
- Validation/test distributions remain unchanged so metrics reflect real-world class imbalance.
- Metrics combine AUC, accuracy, precision, recall, F1, training time, default-rate control, and fairness indicators.
- Thresholds are tuned with business costs in mind: missed defaults versus unnecessary rejection of good customers.

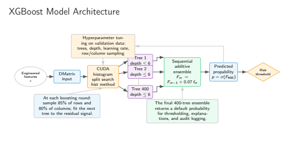
Section 3

Implementation & Technical Architecture

Section 3
Implementation &
Technical Architecture

XGBoost Model Architecture





This slide shows the XGBoost model up to the probability and risk-threshold step. The business decision after the threshold is covered later.

The input is the engineered feature set, converted into XGBoost's DMatrix format for efficient training.

The model uses 400 trees, a learning rate of 0.07, and a maximum tree depth of 6. The trees are built sequentially: each new tree corrects the remaining errors from the previous trees. The learning rate limits how much each tree contributes, and the depth limit controls model complexity.

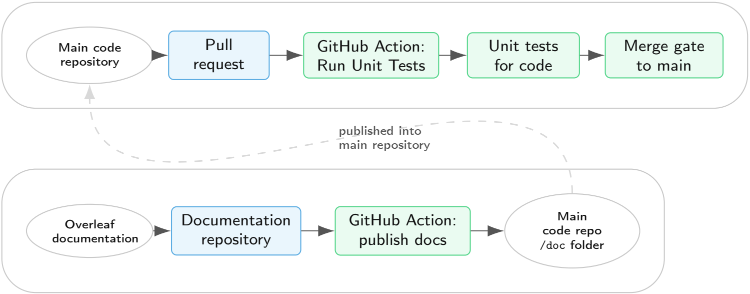
At each boosting round, the model samples 85% of rows and 80% of columns. This adds randomness and helps reduce overfitting. The random state is fixed at 42 for reproducibility.

Training uses the histogram tree method with CUDA acceleration. The model is evaluated with logloss because the output is a probability.

Compared with a Random Forest, the key difference is that Random Forest trees are independent and trained in parallel, while XGBoost trees are trained one after another, each improving the current ensemble. This usually gives stronger performance, but requires more careful tuning.

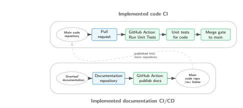
CI/CD Pipeline Integration

Implemented code CI



Implemented documentation CI/CD

2026-06-11



Unite testing for testing the code in /src directory for all functions. Documentation is done in overleaf and also pushed to /doc folder in main branch from overleaf The testing framework ensures robustness and alignment with regulatory lifecycle expectations (e.g., FINAME governance requirements.)

Key Technical Challenges & How they were addressed

Leakage control

Removed post-origination variables

Train only on decision-time information

Imbalance

Oversample minority class in training only

Preserve real validation/test mix

Proxy bias

Protected race attributes excluded

Retained solely for fairness audit

FICO opacity

Retain but monitor contribution

Balance realism and explainability

Deployment gap

U.S. dataset, Swiss/EU target context

Governance, validation, monitoring

Model trade-off

Tree models beat generic MLPs

Prioritize tabular performance + operations

2026-06-11
Loan Pirates
└Implementation & Technical Architecture
└Key Technical Challenges & How they were addressed

Section 4

Results & Evaluation

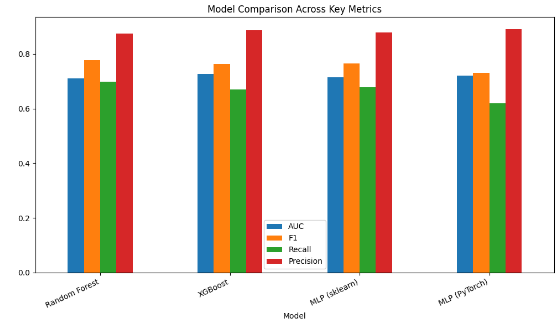
Section 4
Results & Evaluation

Performance metrics and benchmarks

Metric Family	Metrics	Why it matters for credit-risk decisions
Discrimination	AUC-ROC, precision-recall	Separates likely defaults from likely non-defaults across thresholds
Classification quality	Accuracy, recall, precision, F1, confusion matrix	Makes false-positive and false-negative trade-offs explicit
Business impact	Decision time, expected portfolio risk, default-rate control	Connects model choice to the under-30-minute and below-20% default constraints
Governance	Explainability, auditability, human review, fairness monitoring	Required for high-risk regulated credit decisions
Operations	Training time, API reliability, drift monitoring	Ensures scalable, maintainable deployment

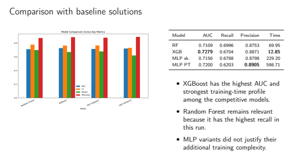
Model quality is necessary, but also requires fairness, explainability, and lifecycle control.

Comparison with baseline solutions



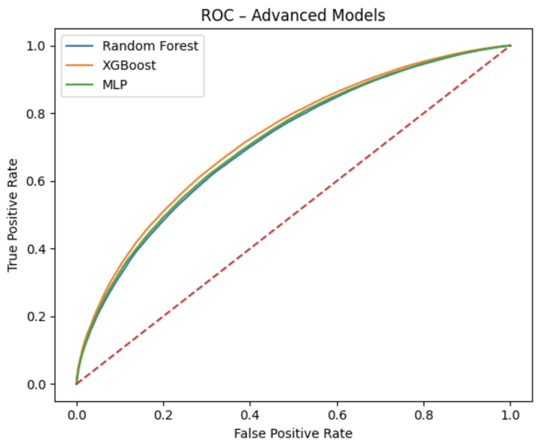
Model	AUC	Recall	Precision	Time
RF	0.7109	0.6996	0.8753	69.95
XGB	**0.7279**	0.6704	0.8871	**12.85**
MLP sk.	0.7156	0.6788	0.8798	229.20
MLP PT	0.7200	0.6203	**0.8905**	598.71

- XGBoost has the highest AUC and strongest training-time profile among the competitive models.
- Random Forest remains relevant because it has the highest recall in this run.
- MLP variants did not justify their additional training complexity.



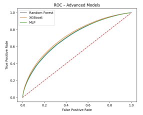
AUC: Area Under the Curve measures the model's ability to discriminate between classes across all classification thresholds. 0.5: no discriminative power (random classifier) 1.0 perfect discrimination **Recall:** measures the proption of actual defaults that are correctly identified **Precision:** measures the proportion of predicted defaults that are actually defaults.

AUC-ROC Curve



- The ROC curve visualizes the true Positive Rate (Recall) vs the False Positive Rate across all thresholds.
- TPR (True Positive Rate) = Proportion of actual defaults correctly identified.
- FPR (False Positive Rate) = Proportion of good borrowers incorrectly flagged as "risky".
- A score of 0.728 indicates strong separation capabiity for structured financial data.

AUC-ROC Curve



- The ROC curve visualizes the true Positive Rate (Recall) vs the False Positive Rate across all thresholds.
- TPR (True Positive Rate) = Proportion of actual defaults correctly identified.
- FPR (False Positive Rate) = Proportion of good borrowers incorrectly flagged as "risky".
- A score of 0.728 indicates strong separation capabiity for structured financial data.

Error analysis and model limitations

High precision

Predicted defaults are generally reliable

Lower recall

Some actual defaults may be missed

- **False negatives:** Defaulting borrowers classified as non-default create direct portfolio-loss risk.
- **False positives:** Good customers rejected or sent to review increase churn and fairness concerns.
- **Known limits:** Class imbalance, FICO opacity, heavy-tailed financial data, and U.S.-to-Swiss/EU transferability.
- **Operational implication:** Use thresholding and human review for uncertain, risky, or legally sensitive cases.

Error analysis and model limitations

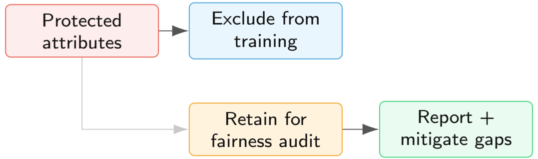
High precision
Predicted defaults are generally reliable
Lower recall
Some actual defaults may be missed

- **False negatives:** Defaulting borrowers classified as non-default create direct portfolio-loss risk.
- **False positives:** Good customers rejected or sent to review increase churn and fairness concerns.
- **Known limits:** Class imbalance, FICO opacity, heavy-tailed financial data, and U.S.-to-Swiss/EU transferability.
- **Operational implication:** Use thresholding and human review for uncertain, risky, or legally sensitive cases.

Bias assessment and mitigation strategies

Fairness design

- Race indicators are excluded from training to prevent direct discriminatory effects.
- The same attributes are retained only for post-training fairness audit.
- Evaluation checks demographic parity, equalized odds, and disparate impact.



Fairness metric	Purpose
Demographic parity	Acceptance rates across different demographic groups
Equalized odds	Accuracy of the decision (same FP/TP across groups)
Disparate impact	Benchmark for fair lending: ratio of approval rates are legally compliant

Operational Fairness Lifecycle

Measure subgroup performance → Diagnose proxy drivers → Reweighing or threshold tuning → Human review guardrail → Document model change

Before launch

Bias and error-rate tests

Evidence collection for approval

During operation

Drift & fairness monitoring

Detect distribution shifts in applicant pools

After change

Versioned documentation

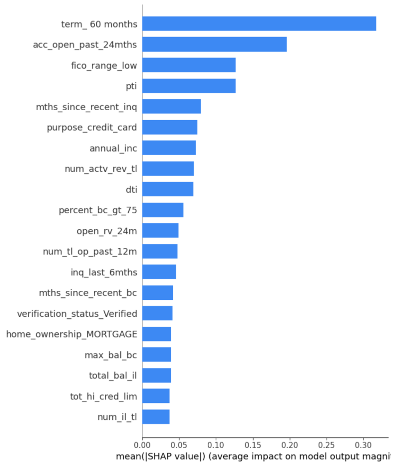
Audit-ready lifecycle logging

Operational Fairness Lifecycle
Measure subgroup performance → Diagnose proxy drivers → Reweighing or threshold tuning → Human review guardrail → Document model change
Before launch
Bias and error-rate tests
Evidence collection for approval
During operation
Drift & fairness monitoring
Detect distribution shifts in applicant pools
After change
Versioned documentation
Audit-ready lifecycle logging

Global interpretability and explainability

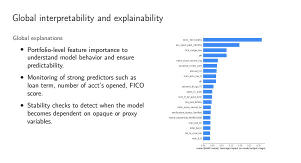
Global explanations

- Portfolio-level feature importance to understand model behavior and ensure predictability.
- Monitoring of strong predictors such as loan term, number of acct's opened, FICO score.
- Stability checks to detect when the model becomes dependent on opaque or proxy variables.

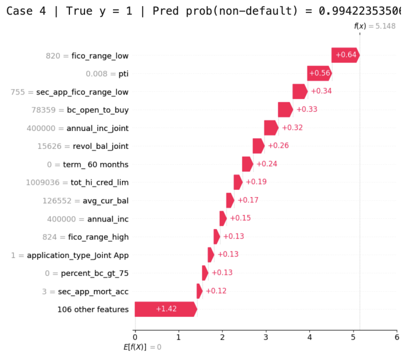


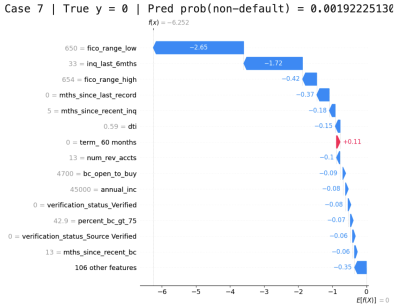
2026-06-11

Loan Pirates
└Results & Evaluation
└Global interpretability and explainability



Local explainability measures



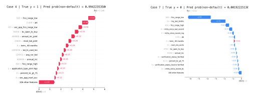


- For every reject or review decision a local explanation report is generated.
- Loan officer can see what features pushed the risk score over the threshold.

2026-06-11

Loan Pirates
└Results & Evaluation
└Local explainability measures

Local explainability measures



- For every reject or review decision a local explanation report is generated.
- Loan officer can see what features pushed the risk score over the threshold.

Why XGBoost is the Deployment Candidate

- **Strong benchmark position:** XGBoost delivers the highest AUC while keeping training time materially lower than the neural-network variants.
- **Fit for structured credit-risk data:** Gradient-boosted trees capture non-linear relationships and feature interactions across income, utilization, repayment, and credit-profile signals.
- **Efficient model lifecycle:** The histogram tree method, GPU support, and parallel execution support practical retraining, validation, and threshold tuning.
- **Responsible deployment fit:** The model can be wrapped with explanations, fairness checks, confidence thresholds, audit logging, and human review for uncertain or high-impact cases.

XGBoost is selected not as an autonomous decision engine, but as a governed risk-scoring component within the credit-decision process.

Why XGBoost is the Deployment Candidate

- **Strong benchmark position:** XGBoost delivers the highest AUC while keeping training time materially lower than the neural-network variants.
- **Fit for structured credit-risk data:** Gradient-boosted trees capture non-linear relationships and feature interactions across income, utilization, repayment, and credit-profile signals.
- **Efficient model lifecycle:** The histogram tree method, GPU support, and parallel execution support practical retraining, validation, and threshold tuning.
- **Responsible deployment fit:** The model can be wrapped with explanations, fairness checks, confidence thresholds, audit logging, and human review for uncertain or high-impact cases.

XGBoost is selected not as an autonomous decision engine, but as a governed risk-scoring component within the credit-decision process.

LLM-Enhanced Explainable AI & Inference Pipeline

Pipeline Interpretation Layer

1. ML Pipeline Output
Scaled Features & Probabilities

↓

2. Random Sampling

↓

3. LLM Audit Layer
Mapping Weights to Narratives

↓

4. Human-Readable Report
Transparent Credit Justifications

- Used a Large Language Model to randomly select three distinct loan applications directly from the inference pipeline output.
- Converted raw, scaled model outputs and complex feature weights into plain English narratives optimized for credit analysts.
- Connected directly to the end of our machine learning pipeline, acting as an automated, post-hoc interpretation layer.
- Demonstrates a practical framework for translating complex mathematical decisions into transparent, auditable loan justifications.

LLM-Enhanced Explainable AI & Inference Pipeline
Pipeline Interpretation Layer
1. ML Pipeline Output
Scaled Features & Probabilities
2. Random Sampling
3. LLM Audit Layer
Mapping Weights to Narratives
4. Human-Readable Report
Transparent Credit Justifications
- Used a Large Language Model to randomly select three distinct loan applications directly from the inference pipeline output.
- Converted raw, scaled model outputs and complex feature weights into plain English narratives optimized for credit analysts.
- Connected directly to the end of our machine learning pipeline, acting as an automated, post-hoc interpretation layer.
- Demonstrates a practical framework for translating complex mathematical decisions into transparent, auditable loan justifications.

Customer Explanations: Approved Case

==

SCENARIO: Random Case 1 (Prob: 0.7218)

==

Congratulations! I'm pleased to share that your loan application has been approved.

- **Positive Drivers:** Selection of a 60-month repayment term; positive history of managing recently opened accounts.
- **Risk Offset:** Higher stability counterbalanced a lower-than-average credit score.

"While your credit score is currently on the lower side, your positive history of managing recently opened accounts demonstrates financial stability. Additionally, the 60-month repayment term you selected aligns well with our requirements, helping us move forward with your request. To help you secure even better rates in the future, I recommend focusing on consistent, on-time payments to help boost your credit score further. We're excited to support your financial goals—please let us know if you have any questions!"

Customer Explanations: Rejected Case

==

SCENARIO: Random Case 2 (Prob: 0.4647)

==

Dear Customer,
We've carefully reviewed your application. While the 60-month term was a positive factor, we are unable to approve your loan at this time.

- **Negative Drivers:** High debt relative to income (DTI); seven new credit accounts opened in the last two years.

"The decision was primarily influenced by your current debt relative to your income and the seven new accounts opened in the last two years, which suggests a high level of recent credit activity. To improve your standing for future applications, we recommend focusing on paying down existing balances and pausing new credit inquiries. We value your business and look forward to assisting you down the road."

Customer Explanations: Rejected Case
SCENARIO: Random Case 2 (Prob: 0.4647)
Dear Customer,
We've carefully reviewed your application. While the 60-month term was a positive factor, we are unable to approve your loan at this time.
- **Negative Drivers:** High debt relative to income (DTI); seven new credit accounts opened in the last two years.
"The decision was primarily influenced by your current debt relative to your income and the seven new accounts opened in the last two years, which suggests a high level of recent credit activity. To improve your standing for future applications, we recommend focusing on paying down existing balances and pausing new credit inquiries. We value your business and look forward to assisting you down the road."

Section 5

Key Takeaways

Section 5
Key Takeaways

Our Takeaways

- Structured credit-risk data favors strong tree-based methods over generic neural networks in this experiment.
- XGBoost provides the best practical balance of AUC, training efficiency, interpretability path, and deployability.
- Accuracy alone is insufficient in regulated finance: fairness, explainability, auditability, and human review must be designed in from the start.
- The proposed system should be introduced as governed decision support, not unchecked fully automated credit approval.
- Next steps: validate on Swiss/EU-representative data, formalize API schemas, define thresholds with risk/compliance, and implement monitoring dashboards.

Our Takeaways

- Structured credit-risk data favors strong tree-based methods over generic neural networks in this experiment.
- XGBoost provides the best practical balance of AUC, training efficiency, interpretability path, and deployability.
- Accuracy alone is insufficient in regulated finance: fairness, explainability, auditability, and human review must be designed in from the start.
- The proposed system should be introduced as governed decision support, not unchecked fully automated credit approval.
- Next steps: validate on Swiss/EU-representative data, formalize API schemas, define thresholds with risk/compliance, and implement monitoring dashboards.