Auditor Evidence Pack - Sample

AI system: loan-approval-sample · 21 Sep 2026 to 5 Nov 2026

Sample Lending Co. (synthetic) · Generated 5 Nov 2026, 21:38 UTC

SAMPLE - SYNTHETIC ORGANIZATION, COMPUTED FIGURES

Report	Auditor Evidence Pack
Organization	Sample Lending Co. (synthetic)
Scope	AI system: loan-approval-sample
Reporting period	21 Sep 2026 to 5 Nov 2026
Generated	5 Nov 2026, 21:38 UTC
Frameworks referenced	ISO/IEC 42001, EU AI Act, NIST AI RMF, GDPR, CCPA
Sections	Trust Score, Risk Register, Compliance, Model Performance, Drift Analysis, Bias Audit, Explainability
Source of figures	SAMPLE. A synthetic organization and a synthetic dataset, published so the format can be inspected. Every figure was still computed by the production code paths named in the accompanying manifest.json; none was written by hand.
Forensic collection	Decoupled SDK and API telemetry recorded outside the model execution path.
Cryptographic verification	Audit events are chained with SHA-256 hashes. This file's SHA-256 hash, the sign-offs pinned to it and the audit chain are published in manifest.json; verify_sample.py recomputes all three offline.
Methodological rigor	Standard statistical drift tests (Kolmogorov-Smirnov, PSI, Chi-squared) and peer-reviewed explainability (SHAP, LIME).
Regulatory scope & isolation	Mapped to ISO/IEC 42001, GDPR, CCPA, EU AI Act, and NIST AI RMF under isolated databases and RBAC access logging.
How to read this report	Each section opens with a plain-language finding and its rating, then the figures it rests on. Every figure is computed from records held in WhiteBoxXAI; missing data is shown as missing, never as zero.

2	2	2
High findings	Medium findings	Areas with no issues

SAMPLE

Contents

01 At a Glance 2
02 Trust Score 3
03 Risk Register 5
04 Compliance 6
05 Model Performance 8
06 Drift Analysis 9
07 Bias Audit 11
08 Explainability 13

01 At a Glance

Area	Finding	Rating
Trust Score	Trust score: portfolio exception	MEDIUM
Risk Register	Risk register: high-rated exceptions open	HIGH
Compliance	Audit trail: activity recorded, no critical exceptions	NO ISSUES FOUND
Model Performance	Figures only; no automated finding	For review
Drift Analysis	Drift monitoring: material exception on loan-approval-sample	HIGH
Bias Audit	Bias audit: no exceptions for loan-approval-sample	NO ISSUES FOUND
Explainability	Explainability: gap on loan-approval-sample	MEDIUM

SAMPLE

02 Trust Score

WHAT THIS SECTION SHOWS AN AUDITOR

Whether each AI system is monitored well enough to stand behind: one score per model combining fairness, drift stability and explainability. It shows which systems carry the weakest evidence, and that systems without monitoring data are disclosed rather than hidden.

ISO/IEC 42001 | EU AI Act | NIST AI RMF | GDPR

Evidence relevant to:

- ISO/IEC 42001:2023, clause 9.1: Monitoring, measurement, analysis and evaluation
- EU AI Act, Article 72: Post-market monitoring by providers
- NIST AI RMF 1.0, MEASURE 3: Mechanisms for tracking identified AI risks over time are in place
- NIST AI RMF 1.0, MANAGE 4: Risk treatments, including response and recovery, and communication plans are documented and monitored regularly
- GDPR, Article 5: Principles relating to processing of personal data

Trust score	A score from 0 to 100 for one model. 80 and above is strong, 60 to 80 is moderate, and below 60 is weak. It is the weighted mean of the components that have data; the weights are listed with the scores.
Fairness, drift stability, explainability	The three components: the latest bias-audit fairness score, how stable the model's input data has been, and the share of its predictions that have an explanation.
Insufficient data	The model has no fairness, drift or explainability results yet. It is listed, and left out of averages rather than scored as zero.

MEDIUM

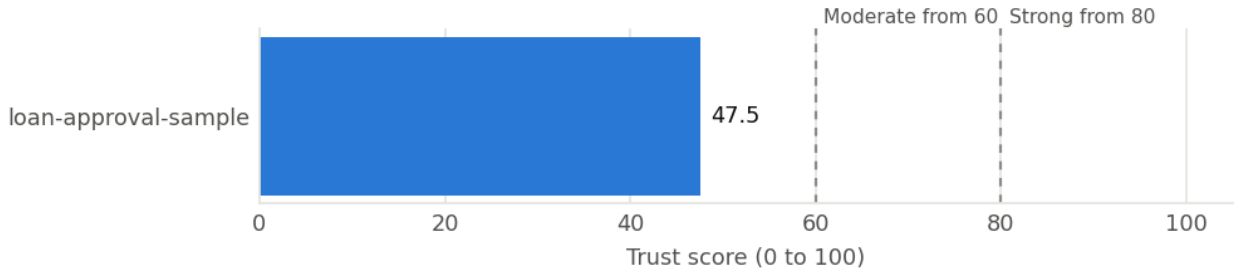
Trust score: portfolio exception

Across 1 scored model, the average trust score is 47.5 out of 100.

How this was determined: Mean of the per-model trust scores, over the models with enough monitoring data to be scored. A model with no fairness, drift or explainability data is left out of the average and counted separately; it is never scored as zero.

47.5 / 100	1	0
Average trust score	Models scored	Models without enough data

SAMPLE



Trust score by model

Model	Trust score	Fairness	Drift stability	Explainability
loan-approval-sample	47.5	100	25	0

Component	Weight in the score
Fairness	0.4
Drift	0.3
Explainability	0.3

How the trust score is calculated

Trust Score = weighted average of fairness (bias audit), drift severity, and explainability coverage, renormalized over whichever inputs are available for a model. Models with none of the three inputs are excluded from the score and counted separately.

SAMPLE

03 Risk Register

WHAT THIS SECTION SHOWS AN AUDITOR

The AI risks the organization has identified, how severe each one is, and whether the serious ones have a mitigation plan and a target date. An open high or critical risk with no plan is the first thing to ask about.

ISO/IEC 42001 | EU AI Act | NIST AI RMF | GDPR | CCPA

Evidence relevant to:

- ISO/IEC 42001:2023, clause 6.1: Actions to address risks and opportunities
- EU AI Act, Article 9: Risk management system
- NIST AI RMF 1.0, MAP 5: Impacts to individuals, groups, communities, organizations and society are characterized
- NIST AI RMF 1.0, MANAGE 1: AI risks based on MAP and MEASURE output are prioritized, responded to and managed
- GDPR, Article 24: Responsibility of the controller
- GDPR, Article 35: Data protection impact assessment
- CCPA, 11 CCR 7150 (CPPA regulations): When a business must conduct a risk assessment
- CCPA, 11 CCR 7152 (CPPA regulations): Risk assessment requirements

Risk score	Likelihood (1 to 5) multiplied by impact (1 to 5), so 1 to 25.
Severity	Set from the risk score. By default: critical at 20 or more, high from 12, medium from 6, low below 6. An organization can set its own bands in its risk settings.
Open	Any status other than closed.

SAMPLE

HIGH

Risk register: high-rated exceptions open

Sample Lending Co. (synthetic) has 1 open risk in its AI risk register, of which 1 is rated high. No open entry is rated critical.

How this was determined: Count of entries in the organization's AI risk register whose status is not closed, grouped by the severity recorded on each entry, at the time this report was generated.

1	1	0
Open risks	Open high or critical	Open critical

Risk	Category	Severity	Score (likelihood x impact)	Status	Mitigation plan	Target date
High-severity drift detected — loan-approval-sample	Drift	HIGH	12	Identified	None recorded	Not set

Risk	Detection source	ISO/IEC 42001 clauses
High-severity drift detected — loan-approval-sample	Drift report	ISO42001-8.2, ISO42001-9.1

04 Compliance

WHAT THIS SECTION SHOWS AN AUDITOR

Evidence that the AI systems were operated and overseen during the period: what was done on the platform, by whom, and with what outcome. An empty audit trail means oversight in the period is not evidenced.

ISO/IEC 42001 | NIST AI RMF | GDPR

Evidence relevant to:

- ISO/IEC 42001:2023, clause 9.2: Internal audit
- NIST AI RMF 1.0, GOVERN 1: Policies, processes, procedures and practices for mapping, measuring and managing AI risks are in place, transparent and implemented effectively
- NIST AI RMF 1.0, MEASURE 4: Feedback about the efficacy of measurement is gathered and assessed
- GDPR, Article 5: Principles relating to processing of personal data
- GDPR, Article 24: Responsibility of the controller

Audit event	One recorded action on the platform, such as a model change, a data access or a configuration change, with who did it and the outcome.
GDPR-tagged event	An audit event the platform marked as relevant to GDPR, such as a data-subject access. It counts events, not requests.
Tamper evidence	Each audit event stores a SHA-256 hash of itself and of the event before it, so a changed or deleted entry breaks the chain.

NO ISSUES FOUND

Audit trail: activity recorded, no critical exceptions

Between September 21, 2026 and November 5, 2026, the audit trail recorded 9 events for Sample Lending Co. (synthetic). No open risk-register entry is rated critical.

How this was determined: Count of audit-trail events recorded for the organization between the start and end of the reporting period, the subset tagged GDPR, and the number of open risk-register entries rated critical.

9
Audit events in period

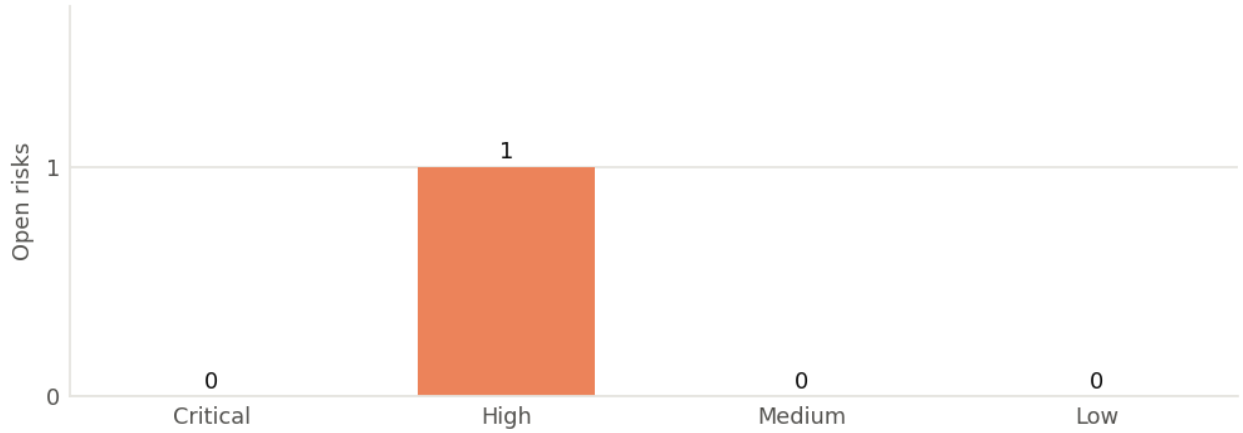
0
GDPR-tagged events in period

1
Open risks

0
Open critical risks

Verified (100%)
Audit chain integrity

SAMPLE



Open risks by severity

Cryptographic audit-chain verification

Every audit log entry is linked by a SHA-256 hash chain. Integrity verification checked 9 event(s) in the period with a 100.0% valid rate (9 valid, 0 invalid, 0 broken links).

Event category	Events
Compliance	5
Model operation	3
Report	1

Event severity	Events
INFO	9

Event outcome	Events
Success	9

Most active users	Events
riley@sample-lending.example	5
sam@sample-lending.example	2
morgan@sample-lending.example	1

Open risks by severity	Open risks
HIGH	1

SAMPLE

05 Model Performance

WHAT THIS SECTION SHOWS AN AUDITOR

Whether the model still performs as it did when it was registered, from the metrics recorded in the period and the baseline recorded with the model.

ISO/IEC 42001 | EU AI Act | NIST AI RMF | GDPR

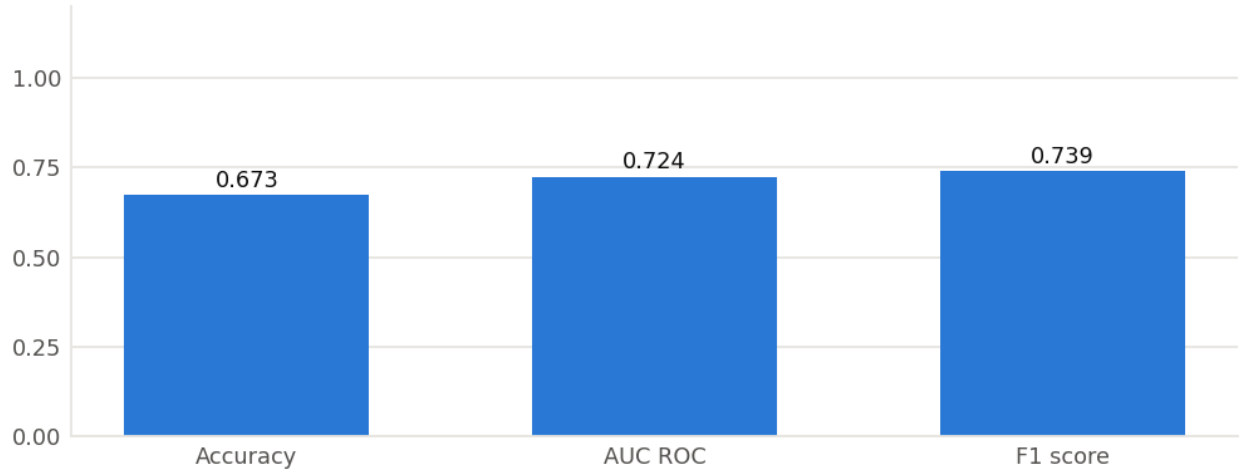
Evidence relevant to:

- ISO/IEC 42001:2023, clause 9.1: Monitoring, measurement, analysis and evaluation
- EU AI Act, Article 72: Post-market monitoring by providers
- NIST AI RMF 1.0, MEASURE 3: Mechanisms for tracking identified AI risks over time are in place
- NIST AI RMF 1.0, MANAGE 4: Risk treatments, including response and recovery, and communication plans are documented and monitored regularly
- GDPR, Article 5: Principles relating to processing of personal data

Latest	The most recent reading in the period.
Baseline	The value recorded for the model when it was registered.
Readings	How many measurements a figure rests on. Few readings mean weak evidence, whatever the value.

3 Metrics tracked

3 Readings in period



Latest performance against baseline (0 to 1 metrics)

Metric	Latest	Average in period	Readings
Accuracy	0.673	0.673	1
AUC ROC	0.724	0.724	1
F1 score	0.739	0.739	1

SAMPLE

06 Drift Analysis

WHAT THIS SECTION SHOWS AN AUDITOR

Whether the data the model sees in production has moved away from the data it was built on, and whether its predictions have become less reliable. Unnoticed drift is a common way a validated model stops being valid.

ISO/IEC 42001 | EU AI Act | NIST AI RMF | GDPR

Evidence relevant to:

- ISO/IEC 42001:2023, clause 9.1: Monitoring, measurement, analysis and evaluation
- EU AI Act, Article 72: Post-market monitoring by providers
- NIST AI RMF 1.0, MEASURE 3: Mechanisms for tracking identified AI risks over time are in place
- NIST AI RMF 1.0, MANAGE 4: Risk treatments, including response and recovery, and communication plans are documented and monitored regularly
- GDPR, Article 5: Principles relating to processing of personal data

Drift score (PSI)	Population Stability Index: how far recent input data has moved from the data the model was built on. 0 means no shift; larger means more. The alert threshold is set per feature (default 0.1).
Data drift and concept drift	Data drift is a change in what the model receives. Concept drift is a change in how well its predictions hold up, which can happen even when the inputs look the same.
Severity and trend	Each check is rated low, medium, high or critical. The trend says whether drift scores rose, fell or held steady across the period.

HIGH

Drift monitoring: material exception on loan-approval-sample

The drift report dated November 5, 2026 recorded high-severity drift over 1,200 samples. Evidence of investigation and remediation should be retained.

How this was determined: Worst severity across the model's latest data-drift report and latest concept-drift report. A report that detected no drift counts as no drift. PSI is the Population Stability Index of the recent window against the baseline.

1
Data-drift checks

Stable
Data-drift trend

0
Concept-drift checks

Checked	Drift score (PSI)	Features drifted	Severity
5 Nov 2026, 21:38 UTC	0.798	2	HIGH

SAMPLE

2 of 6 features show a distribution shift in the latest check

annual_income and credit_utilization moved away from the baseline. The remaining features did not cross any test threshold. Review whether the shift reflects a real change in the population before relying on the model's outputs.

How this was determined: A feature is flagged when any of its tests crosses its threshold. Defaults: Kolmogorov-Smirnov or Chi-squared p-value below 0.05, Jensen-Shannon divergence of 0.1 or more, PSI of 0.2 or more. Samples analyzed: 1,200.

Feature	Test	Statistic	p-value	Result
loan_amount	Kolmogorov-Smirnov	0.0333	0.8932	No shift
loan_amount	Jensen-Shannon	0.0563	No p-value (distance measure)	No shift
loan_amount	PSI	0.0489	No p-value (distance measure)	No shift
annual_income	Kolmogorov-Smirnov	0.1800	< 0.0001	Shift detected
annual_income	Jensen-Shannon	0.1256	No p-value (distance measure)	Shift detected
annual_income	PSI	0.1964	No p-value (distance measure)	No shift
debt_to_income	Kolmogorov-Smirnov	0.0467	0.5311	No shift
debt_to_income	Jensen-Shannon	0.0755	No p-value (distance measure)	No shift
debt_to_income	PSI	0.0472	No p-value (distance measure)	No shift
months_employed	Kolmogorov-Smirnov	0.0517	0.4001	No shift
months_employed	Jensen-Shannon	0.0571	No p-value (distance measure)	No shift
months_employed	PSI	0.0266	No p-value (distance measure)	No shift
recent_inquiries	Chi-squared	8.0048	0.1560	No shift
recent_inquiries	PSI	0.0151	No p-value (distance measure)	No shift
credit_utilization	Kolmogorov-Smirnov	0.2483	< 0.0001	Shift detected
credit_utilization	Jensen-Shannon	0.2608	No p-value (distance measure)	Shift detected

SAMPLE

Feature	Test	Statistic	p-value	Result
credit_utilization	PSI	0.8844	No p-value (distance measure)	Shift detected

Statistical drift methodology

Drift is evaluated using standard, non-proprietary statistical methodologies to enable independent auditor replication: Population Stability Index (PSI) and two-sample Kolmogorov-Smirnov (KS) tests for continuous feature distributions, and Chi-squared tests for categorical distributions. Severity thresholds follow standard evidentiary bands (< 0.10 stable, 0.10–0.25 moderate drift, >= 0.25 significant drift).

07 Bias Audit

WHAT THIS SECTION SHOWS AN AUDITOR

Whether the model's outcomes differ across groups, such as by age or gender, beyond the threshold set for each fairness metric, and whether audits were run in the period at all.

ISO/IEC 42001 | EU AI Act | NIST AI RMF | GDPR | CCPA

Evidence relevant to:

- ISO/IEC 42001:2023, clause 8.3: Data management for AI systems
- ISO/IEC 42001:2023, clause 8.4: AI system lifecycle management
- EU AI Act, Article 10: Data and data governance
- EU AI Act, Annex IV: Annex IV technical documentation content
- NIST AI RMF 1.0, MEASURE 2: AI systems are evaluated for trustworthy characteristics
- NIST AI RMF 1.0, MANAGE 2: Strategies to maximize AI benefits and minimize negative impacts are planned, implemented and documented
- GDPR, Article 22: Automated individual decision-making, including profiling
- GDPR, Article 35: Data protection impact assessment
- CCPA, 11 CCR 7152 (CPPA regulations): Risk assessment requirements

Fairness score	A summary from 0 to 1 across the metrics checked in one audit; higher is fairer.
Disparate impact	The ratio of favorable-outcome rates between two groups. 1 means both groups receive favorable outcomes at the same rate.
Demographic parity	The difference in favorable-outcome rates between groups. 0 means no difference.
Metrics failing	Metrics outside the threshold recorded with that audit. Thresholds are set per audit and shown with each metric in the platform.

SAMPLE

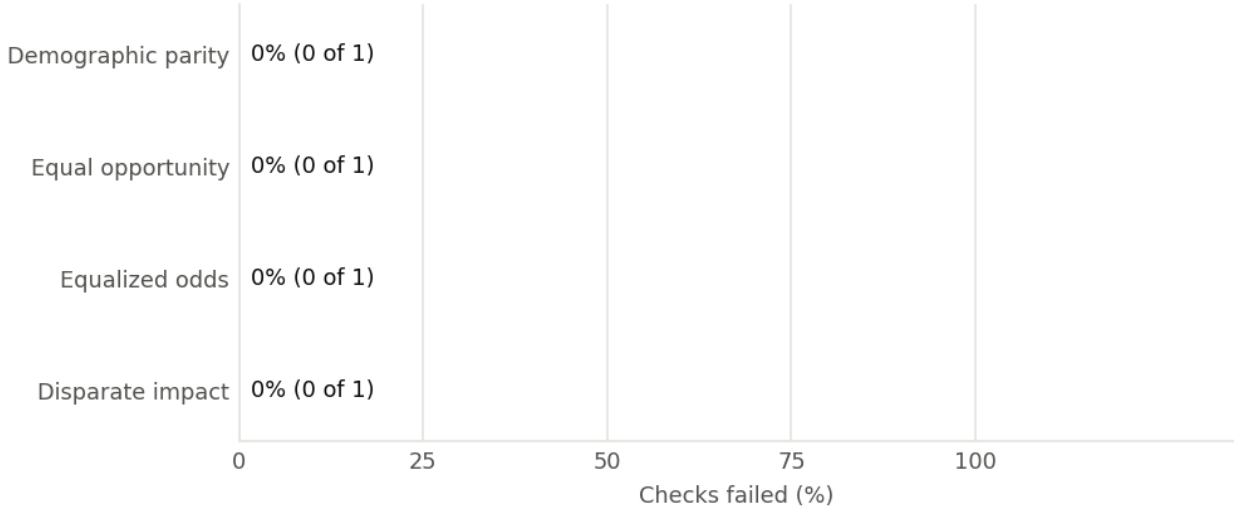
NO ISSUES FOUND

Bias audit: no exceptions for loan-approval-sample

The bias audit dated November 5, 2026 tested 4 fairness metrics; none exceeded its configured threshold. Overall fairness score: 100.0 out of 100.

How this was determined: The model's latest bias audit: each fairness metric compared with its configured threshold across the protected attributes tested. The overall score is the audit's own, scaled to 100.

1
Bias audits in period



Fairness checks failed in the period, by metric

Audit date	Fairness score	Bias detected	Worst severity	Metrics failing
5 Nov 2026, 21:38 UTC	1	No	None	0 of 4

Fairness metric	Checks	Failed	Failure rate
Demographic parity	1	0	0.0%
Equal opportunity	1	0	0.0%
Equalized odds	1	0	0.0%
Disparate impact	1	0	0.0%

08 Explainability

WHAT THIS SECTION SHOWS AN AUDITOR

Whether individual decisions can be explained after the fact: the basis for telling a person why a decision about them was made.

ISO/IEC 42001 | EU AI Act | NIST AI RMF | GDPR | CCPA

Evidence relevant to:

- ISO/IEC 42001:2023, clause 7.5: Documented information
- EU AI Act, Article 13: Transparency and provision of information to deployers
- EU AI Act, Annex IV: Annex IV technical documentation content
- NIST AI RMF 1.0, MAP 1: Context is established and understood
- NIST AI RMF 1.0, MAP 3: AI capabilities, targeted usage, goals and expected benefits and costs are understood against appropriate benchmarks
- GDPR, Article 13: Information to be provided (data collected from the data subject)
- GDPR, Article 22: Automated individual decision-making, including profiling
- CCPA, Cal. Civ. Code 1798.100: General duties of businesses that collect personal information

SHAP	Attributes one prediction to the input features that drove it, and by how much.
LIME	Explains one prediction by fitting a simple, readable model to the model's behavior around that prediction.
Explanations in period	Explanations generated in the reporting period. Compare with the number of decisions the model made to judge coverage.

MEDIUM

Explainability: gap on loan-approval-sample

Only 0 of 1,200 recorded predictions (0.0%) carry a completed explanation. A documented rationale or remediation plan should be retained.

How this was determined: Share of the model's recorded predictions that have a completed or cached SHAP or LIME explanation.

200
Explanations in period

13.011 ms
Average time to explain

Method	Explanations
SHAP tree	200

Status	Explanations
Completed	200

SAMPLE

annual_income, credit_utilization and debt_to_income drive this model's decisions

Across 200 computed explanations, the top 3 carry 63% of the total attribution across 6 features. Check that each is a legitimate basis for the decision and is not a proxy for a protected characteristic.

How this was determined: Mean absolute SHAP value per feature over the model's completed SHAP explanations in the period. The sign column shows the average direction: positive pushes the score up.

Rank	Feature	Mean |SHAP|	Mean SHAP (direction)
1	annual_income	0.4612	-0.0129
2	credit_utilization	0.2988	-0.0372
3	debt_to_income	0.2886	0.0204
4	recent_inquiries	0.2347	0.1231
5	months_employed	0.2308	-0.0032
6	loan_amount	0.1508	-0.0420

Explainability methodology

Local and global feature attributions are computed using peer-reviewed mathematical methodologies (SHAP Shapley Additive Explanations and LIME Local Interpretable Model-agnostic Explanations), providing mathematically grounded attribution vectors without proprietary black-box scoring.

SAMPLE