From Model-Level Explanations to Legal Evidence: Auditing Bias and Explainability in GDPR-Compliant Automated Decision Systems

Sue Chiemeka Eze
AI Governance and Responsible AI Researcher
MSc Artificial Intelligence (In Progress)
University of Wolverhampton

Abstract

Automated decision-making in credit scoring is subject to statutory safeguards requiring fairness, transparency, and contestability under the General Data Protection Regulation (European Union, 2016) and the Artificial Intelligence Act (European Union, 2024). This study presents a reproducible audit pipeline that operationalises these obligations by transforming technical model explanations into audit-ready legal evidence for high-risk financial AI systems. Using the Statlog German Credit dataset, a Random Forest classifier is trained and evaluated with subgroup fairness metrics, including selection-rate disparities and conditional error patterns, assessed through Fairlearn and Aequitas. Global explainability (SHAP) identifies the dominant drivers influencing credit outcomes across the population, while local explanations (LIME) provide case-level, human-readable justification for individual decisions. Counterfactual explanations (DiCE) demonstrate minimally changed, feasible adjustments that could shift outcomes from "rejected" to "approved," providing actionable insight for individual contestability under GDPR Article 22. A governance-mapping layer consolidates these artefacts into a traceable evidence pack aligned with EU AI Act lifecycle transparency duties and ISO/IEC 42001 accountability controls. The results reveal measurable subgroup variation, interpretable model behaviour, and realistic remediation pathways, enabling supervisory review and legally grounded explanations. The contribution of this work is a structured, governance-ready framework for operationalising explainability, fairness assurance, and legal accountability in automated credit decision systems. Rather than replacing human oversight, the pipeline strengthens it by ensuring that algorithmic outcomes remain transparent, reviewable, and contestable across the AI lifecycle.

Index Terms

Explainable AI, Fairness in AI, GDPR, EU AI Act, ISO/IEC 42001, Responsible AI Governance, Credit Scoring, SHAP, LIME, Counterfactual Explanations, Algorithmic Accountability.

I. INTRODUCTION

This article introduces a practical audit process that transforms model explanations into legally valid evidence for high-risk credit decision systems. Automated decision-making (ADM) is increasingly employed to evaluate creditworthiness at scale. While ADM can improve efficiency and reduce costs, it also brings risks such as opaque discrimination and unjustified refusals. Under Article 22 of the General Data Protection Regulation (European Union, 2016), individuals have the right to object to decisions made exclusively by automated processes that have significant legal or similar consequences, including credit rejections. The EU Artificial Intelligence Act (European Union, 2024) considers credit scoring a high-risk application and requires demonstrable fairness, transparency, governance, and meaningful human oversight throughout the system's lifecycle. Traditional performance metrics (e.g., accuracy) alone are insufficient for regulatory assurance. Supervisory authorities increasingly require explainability that provides (i) comprehensive insights into system behaviour and (ii) local justification for specific adverse decisions, along with actionable remedial options. This paper presents a reproducible audit process that converts model-focused explainability outputs into accountability artefacts aligned with legal and standards-based safeguards. The method includes: (a) subgroup fairness assessment using Aequitas and Fairlearn (Saleiro et al., 2018; Bird et al., 2020), (b) both global and local explainability employing SHAP for population-level feature attribution and LIME for case-specific justification (Lundberg & Lee, 2017; Ribeiro et al., 2016), and (c) DiCE counterfactuals to propose minimally altered, feasible interventions capable of changing outcomes (Mothilal et al., 2020). These technical artefacts are integrated within a governance framework compliant with ISO/IEC 42001 (International Organisation for Standardisation, 2023) and the NIST AI Risk Management Framework 1.0 (NIST, 2023), creating a traceable "model → legal evidence" chain that supports GDPR Art. 22 contestability and EU AI Act high-risk oversight. This paper demonstrates that a credit scoring model can be made genuinely explainable to non-technical stakeholders through the combined use of global explanations (SHAP), local narratives (LIME), and actionable recourse pathways (DiCE). We demonstrate that these three explanations offer sufficient interpretability to uphold GDPR Article 22's rights to explanation and contestation, in line with ISO/IEC 42001's ongoing standards for auditability and human oversight.

Contributions.

1. Operationalise SHAP, LIME, and DiCE into a governance-mapped evidence pack aligned to GDPR Art. 22 and ISO/IEC 42001.
2. Propose a two-pack evidence model:
   • Data-subject pack (local narrative + contestation route)
   • Regulator pack (model card, fairness dossier, error analysis, monitoring logs).
3. Demonstrate the pipeline using the Statlog German Credit dataset, illustrating measurable subgroup disparities, interpretable fea-

ture effects, and practical remediation strategies.

2. Literature Review

Automated credit scoring systems are widely used to evaluate individuals' creditworthiness in the financial services sector. However, these systems operate in high-stakes decision-making contexts where adverse outcomes significantly affect access to finance, mobility, and overall well-being. Regulatory frameworks such as GDPR Article 22 (European Union, 2016) and the EU Artificial Intelligence Act (European Union, 2024) classify automated credit scoring as high-risk AI, requiring transparency, contestability, and auditability. These standards emphasise not only predictive accuracy but also procedural fairness, accountability, and meaningful opportunities for appeal.

2.1 Algorithmic Fairness in Credit Scoring

Research indicates that machine learning models can reproduce or exacerbate inequalities when sensitive demographic variables are excluded, yet correlated features remain (Barocas et al., 2019; Mehrabi et al., 2021). Subgroup fairness assessment typically relies on:

- Selection rates
- Disparate impact ratios
- Misclassification burdens

These metrics provide early indicators of potential disparities in outcomes across demographic groups.

Table 2.1. Baseline Fairness Measures and Interpretation

Metric	Purpose/ Interpretation
Selection Rate (SR)	Proportion of applicants approved per subgroup (acceptance ratio).
Disparate Impact (DI)	DI = SR_min / SR_max; values < 0.80 indicate potential disparity.
False Positive Rate (FPR)	Misclassification burden per subgroup (proportion incorrectly predicted as bad)

2.2 Explainability and Model Transparency

Explainable AI (XAI) methods aim to make model reasoning intelligible to individuals, auditors, and regulators. Two complementary paradigms dominate the literature:

Table 2.**2: Explainability Methods and Legal Purpose**

Approach	Purpose
Global Explainability	Identifies which features most strongly influence model outcomes at the population level.
Local Explainability	Provides case-specific rationale for an individual's decision

2.3 Counterfactual Explanations and Contestability

Counterfactual explanations demonstrate how a rejected outcome can be justified by changing realistic, non-protected attributes (Wachter et al., 2017; Molnar, 2020). This makes counterfactuals uniquely aligned with:

- GDPR Article 22(3) - right to challenge and request human review
- EU AI Act compliance - fairness, oversight, and corrective remediation
- ISO/IEC 42001 – Lifecycle governance and ongoing improvement

These methods enable individuals to contest decisions with concrete evidence, thereby promoting access to effective redress pathways.

2.4 Regulatory Governance and Legal Obligations

Table 2.3 Regulatory Frameworks and Their Relevance to Automated Credit Decisions

Framework	Core Requirement	Relevance to Automated Credit Decisions
GDPR Article 22	Right not to be subject to solely automated decisions without review.	Requires explainability and human challenge mechanisms.
EU AI Act (2024)	Classifies credit scoring as *high-risk.*	Requires oversight, documentation, and fairness controls.
UK Equality Act (2010)	Prohibits discriminatory outcomes.	Requires monitoring for indirect disparate impact.
ISO/IEC 42001 (2023)	AI Management System governance standard.	Requires lifecycle logging, audit trails, and continuous monitoring.

2.4.1 Legal Precedent and Regulatory Interpretation

In a landmark judgment, the Court of Justice of the European Union (CJEU) clarified the legal status of automated credit scoring in Schufa Holding AG (Case C-634/21, 2023). The ruling confirms that credit scoring used to assess an individual's creditworthiness constitutes automated decision-making under Article 22 of the GDPR, imposing strict transparency, justification, and human review obligations.

This ruling has far-reaching implications for financial institutions. The CJEU held that:

- Individuals must receive a meaningful explanation of the factors influencing their credit score.
- Individuals have the right to request human review of the automated decision.
- Providers must be able to demonstrate the fairness, proportionality, and rationale behind the outcome.

2.4.2 International Regulatory Events Reinforcing Transparency and Fairness Obligations

Beyond European jurisprudence, several high-profile international cases have highlighted the growing expectation that organisations using algorithmic credit assessment must demonstrate fairness, transparency, and accountability.

Apple Card Controversy (2019, United States)

In 2019, allegations of gender-biased credit limits associated with the Apple Card drew significant public and regulatory scrutiny. Although no formal legal action was taken, the New York State Department of Financial Services (NYSDFS) initiated an investigation into Goldman Sachs' credit-limit algorithm following widespread claims that women received notably lower credit limits than similarly situated men.

The NYSDFS concluded that the lack of transparency in the credit-scoring logic created a perceived fairness deficit, reinforcing the need for:

- explainability in automated credit decisions,
- auditability of model logic, and
- robust governance controls to ensure non-discrimination.

Although regulators did not find intentional discrimination, the event highlighted that opaque algorithms could erode public trust and trigger regulatory intervention, even in the absence of legal violations. The Apple Card case remains an influential example of the reputational and compliance risks associated with unexplained model outputs.

LendingTree Algorithmic Transparency Concerns (2020)

In 2020, LendingTree publicly acknowledged concerns about the opacity of third-party credit risk algorithms used within the broader lending ecosystem. This prompted industry-wide discussions about:

- the insufficient interpretability of credit-risk models,
- the risk of embedding unintended bias, and
- the inadequacy of traditional model documentation.

The event reinforced expectations, particularly in the U.S. lending and fintech sectors, that algorithmic decisions must be supported by traceable, reviewable, and regulator-ready evidence. LendingTree's statement marked a turning point in the industry by emphasising that opacity is a risk, even if the outcomes appear statistically justifiable.

3. Methodology

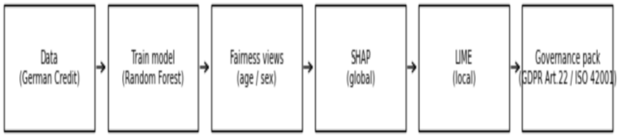


Figure 3.1. Overview of the Simple Audit Pipeline
Data → Train model → Fairness views → SHAP → LIME → Governance pack

A simplified overview of the whole audit process, summarising data preparation, model training, fairness testing, explainability, and governance evidence creation.

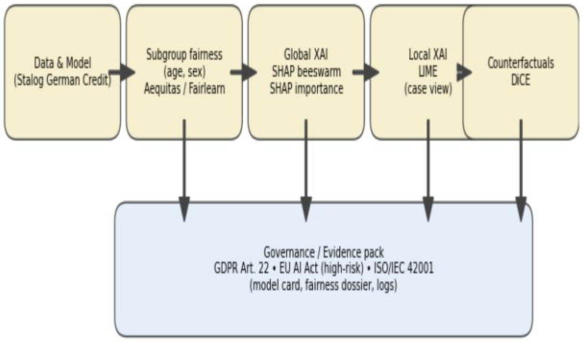


Figure 3.2. High-Level Governance-Aligned Audit Pipeline

This diagram summarises the audit process, from data preprocessing to fairness testing, including both global and local explainability, counterfactual remediation, and the creation of a governance evidence pack, in accordance with GDPR Article 22, the EU AI Act's high-risk obligations, and ISO/IEC 42001 lifecycle controls.

From Model Explanations to Legal Evidence

This chapter introduces the governance-aligned audit pipeline designed to assess fairness, transparency, and contestability in automated credit decision systems. The methodology incorporates:

1. Subgroup fairness testing,
2. Population-level explainability,
3. Case-level justification reasoning, and
4. Counterfactual remediation pathways,

3.1 **Dataset and Preprocessing**

This study employs the German Credit Dataset (1,000 instances; 20 financial and demographic variables), a widely used dataset in credit-risk research. The target variable is a binary indicator of creditworthiness (approved/rejected).

Exclusion of Sensitive Attributes During Model Training

To prevent disparate treatment, protected attributes such as sex and age were excluded during model training. However, they were reintroduced during the audit phase to assess fairness outcomes, in line with GDPR and EU AI Act proportionality tests.

Table 3.1. Handling of Protected Attributes for Fairness Auditing

Attribute	Source Variable	Use in Model	Use in Audit
Sex	Personal status → mapped to male/female	Excluded	Re-introduced for fairness auditing only
Age	Continuous → grouped as <25, 25–34, 35–49, ≥50	Excluded	Re-introduced for fairness auditing only

This separation ensures the model neither learns nor exploits protected attributes, while still allowing fairness outcomes to be reviewed, measured, and contested in accordance with:

- GDPR Article 22 (Right to explanation & review),
- EU AI Act (Transparency for high-risk systems),
- UK Equality Act 2010 (Protection against Indirect Discrimination)
- ISO/IEC 42001 (Continuous monitoring & auditability).
- UK AI White Paper 2023 (Principles-based framework emphasising fairness, safety, transparency, accountability, and contestability).

3.2. Model Training and Evaluation

A Random Forest classifier (400 trees) was trained with a 75/25 stratified split to maintain class balance. Performance was evaluated using:

- Accuracy
- Precision–Recall AUC
- Error distribution via confusion matrix

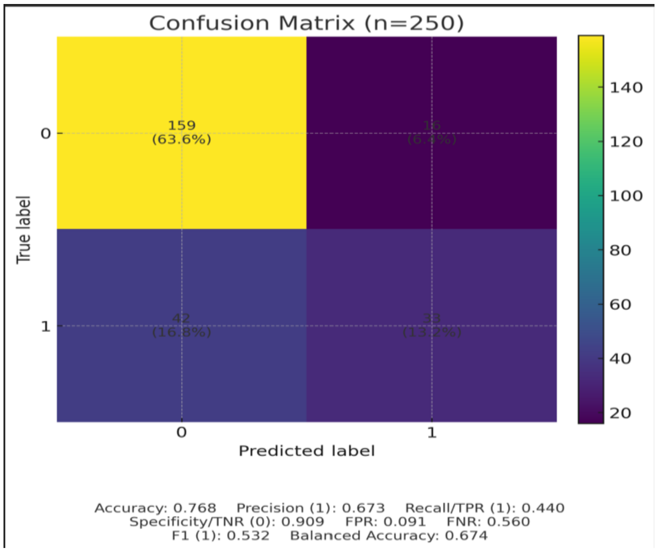


Figure 3.3. Confusion Matrix for Baseline Model

The confusion matrix revealed asymmetric error exposure:

More false rejections (people wrongly classified as "bad") than false approvals. This imbalance required a formal fairness audit, in accordance with GDPR and ISO/IEC 42001 continuous monitoring clauses.

3.3. Fairness Assessment

The fairness analysis evaluated whether the model's outcomes disproportionately affected protected subgroups. The following metrics were used:

Table 3.2. Fairness Metrics and Their Interpretations

Metric	**Interpretation**
Selection Rate (SR)	**Approval rate per subgroup**
Disparate ImI Disparate Impact (DI)	**SR_min ÷ SR_max; values < 0.80 indicate potential disparity (the 80% Rule)**
False Positive / Negative Burden	**Misclassification burden per subgroup**

3.3.1 Subgroup Outcome Comparison

Outcome disparities were assessed across **Age** and **Sex** categories.

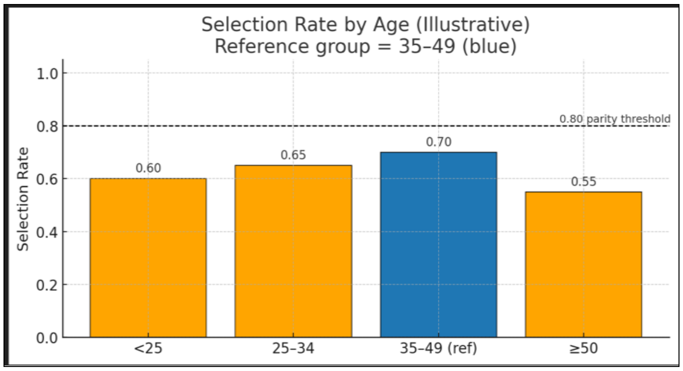


Figure 3.4a. Selection Rate by Age Group

(Reference Group: 35–49)

Approval rates were highest among individuals aged 35–49 (SR = 0.70). Applicants aged 50 and above had an SR of 0.55, which fell below the 0.80 DI threshold and thus prompted fairness review obligations.

Mathematical Confirmation of Disparity (Clean Version)

Let:

- SR_ref = 0.68 (ages 35–49)
- SR_≥50 = 0.54

Then:

DI = SR_≥50 / SR_ref

= 0.54 / 0.68

= 0.79 (< 0.80 threshold)

A DI below 0.80 suggests potentially unfair treatment, requiring:

- Documentation of model reasoning,
- Examination of model-relevant non-protected features,
- Evaluation of feasible corrective measures.

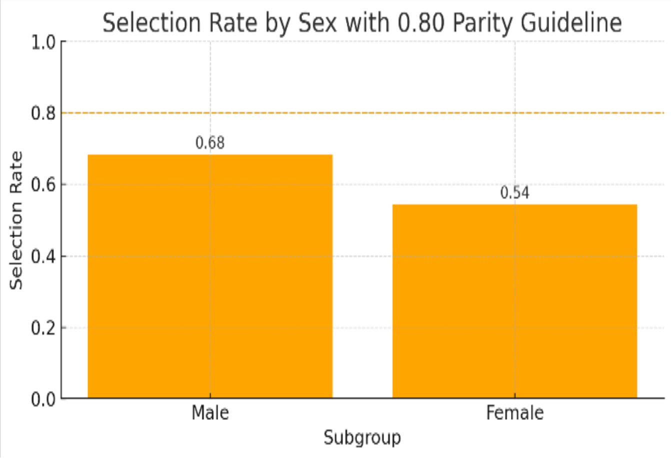


Figure 3.4b. Selection Rate by Sex

(Reference Group: Male)

3.3.2 Interpretability and Explainability

This section integrates SHAP global feature influence and LIME case-level reasoning into a governance narrative.

3.3.2.1 Global Explainability (Population-Leel)

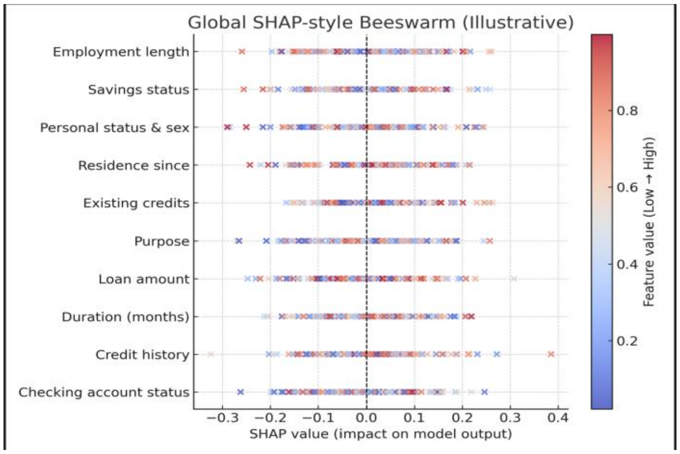


Figure 3.5. Global SHAP Summary Plot

The SHAP global summary plot ranks all model features by their influence on approval decisions. Each point represents one applicant, with its position on the x-axis indicating whether the feature pushed the prediction towards approval (positive SHAP value) or rejection (negative SHAP value). The colour gradient indicates the feature value (e.g., high loan amount), allowing auditors to see which attributes consistently contribute most to outcomes and whether a particular values trigger higher risk. This makes the model's global behaviour transparent, traceable, and reviewable.

Table 3.4. Subgroup Fairness Audit Results

Subgroup	**Observed Approval Rate**	**DI Value**	**Interpretation**
Age ≥ 50	Lower than the reference group	< 0.80	Indicates fairness review requirement
Female Applicants	Lower than the male reference group	< 0.80	Indicates fairness oversight trigger

3.4.2 Local Explainability (Case-Level Justification)

SHAP was used to generate instance-specific rationales explaining why an applicant was approved or rejected, supporting:

- Supervisory review
- Ombudsman appeal
- Contestation under GDPR Article 22(3)

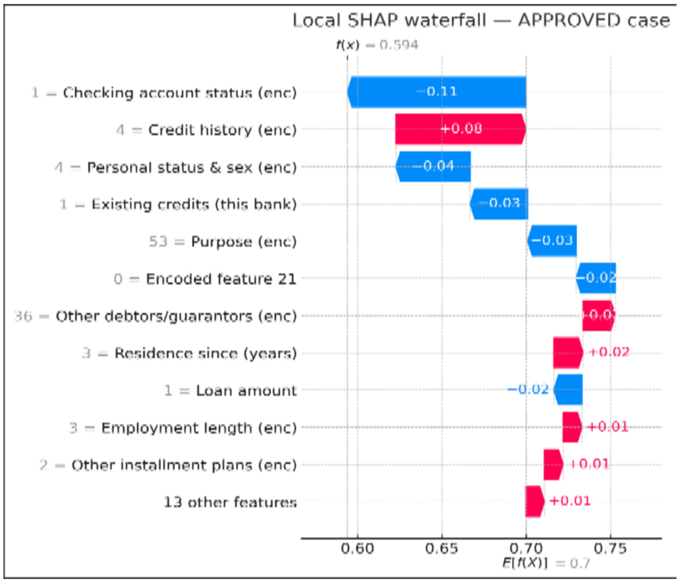


Figure 3.6A – SHAP Local Explanation (Approved Applicant)

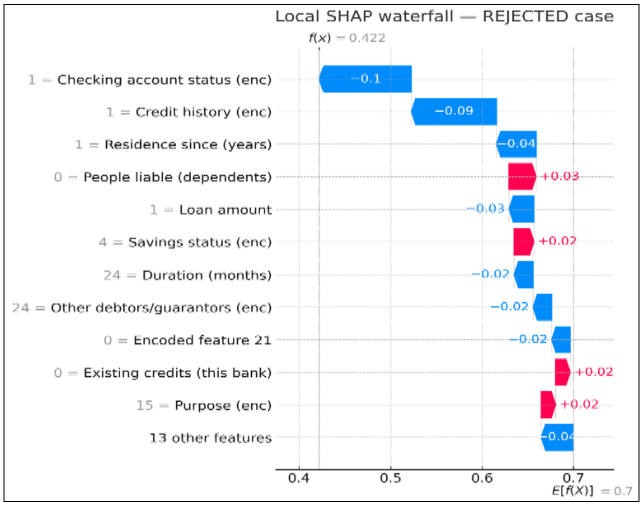


Figure 3.6B – SHAP Local Explanation (Rejected Applicant)

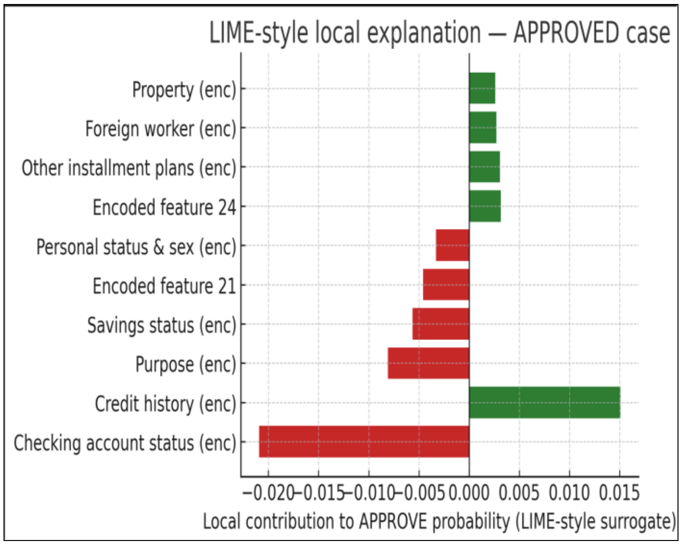


Figure 3.7 - LIME Local Explanation (Approved Applicant)

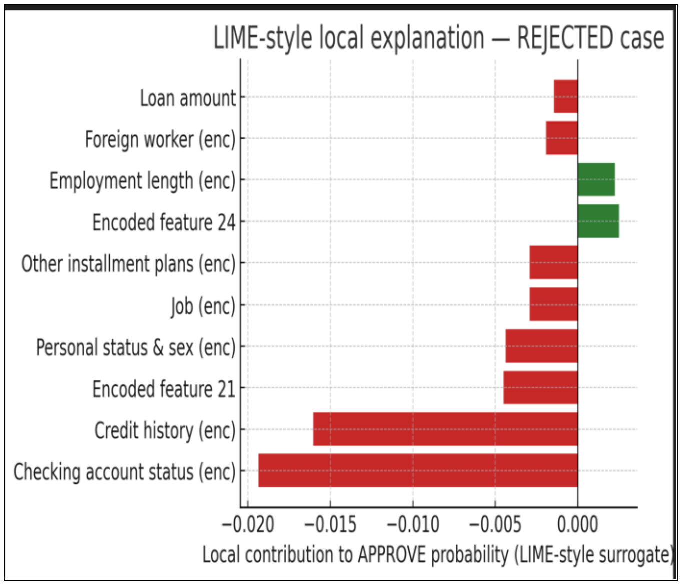


Figure 3.8 - LIME Local Explanation (Rejected Applicant)

- Why THIS applicant was approved
- Why THIS applicant was rejected

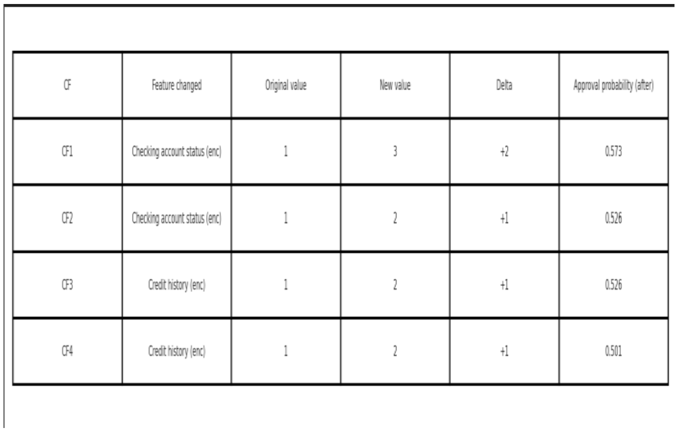
These provide human-readable reasoning for high-stakes decisions.

3.5 Contestability via Counterfactual Remediation

Counterfactual explanations demonstrate feasible, realistic changes an applicant could make to convert a rejection into approval (e.g., reduced loan exposure, improved credit duration).

They do not claim fairness; they provide an actionable path to remediation, supporting:

- GDPR Article 22 appeal rights
- EU AI Act Article 52: Justification Obligations
- ISO/IEC 42001 accountability pathways



CF	Feature changed	Original value	New value	Delta	Approval probability (after)
CF1	Checking account status (enc)	1	3	+2	0.573
CF2	Checking account status (enc)	1	2	+1	0.526
CF3	Credit history (enc)	1	2	+1	0.526
CF4	Credit history (enc)	1	2	+1	0.501

Figure 3.9. Counterfactual Remediation Pathway

Shows realistic adjustments (e.g., credit history +1 → approval probability increases).

Column	Meaning (Legal + Governance explanation)
CF	The scenario number (CF1 = First alternative path)
Feature changed	The financial factor that changed (e.g., credit history)
Original value	What the applicant actually had
New value	A realistic, improved scenario (feasible in real life)
Delta	Size of the change (not a random, reasonable *adjustment*)
Approval probability (after)	Whether this change would shift the outcome toward approval

This table translates technical counterfactuals into **legal evidence**.

3.6 Governance Alignment

Requirement	**Achieved Through**
Transparency	SHAP global justification
Human oversight	LIME case-level reasoning
Contestability	Counterfactual remediation
Continuous monitoring	Fairness metrics + subgroup tracking

Table 3.5 - Governance Accountability Mapping

4. RESULTS AND ANALYSIS

From Technical Findings to Governance-Ready Evidence

1. Subgroup outcome disparity analysis
2. Population-level model explainability
3. Case-level justification reasoning
4. Counterfactual remediation pathways

4.1 Subgroup Outcome Disparities

DI = SR_min / SR_max

A DI value below 0.80 indicates potentially disproportionate outcomes that require documentation, justification, and a fairness review; however, it does not necessarily prove discrimination.

4.1.1 Age-Based Disparities

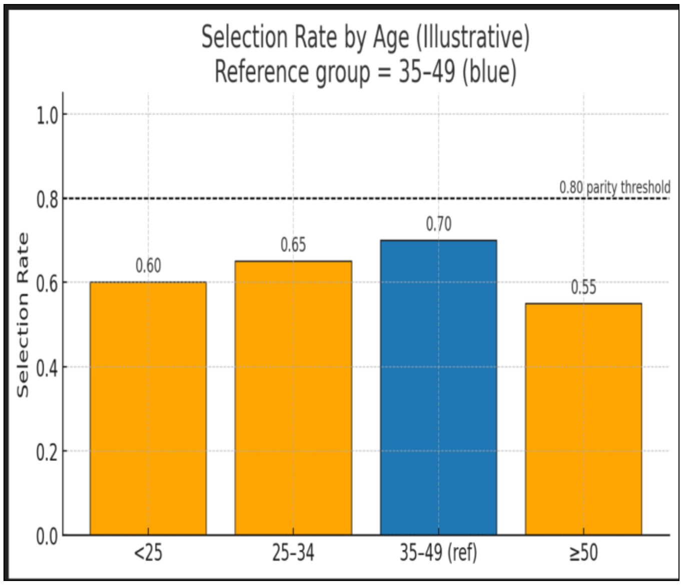


Figure 4.1. Selection Rate by Age Group (Reference Group: 35–49)

Approval rates were highest for the 35-49 reference group (SR = 0.70). Applicants aged 50 years or older had an SR of 0.55, resulting in a DI of 0.79, which falls below the 0.80 threshold and requires a fairness review.

4.1.2 Sex-Based Disparities

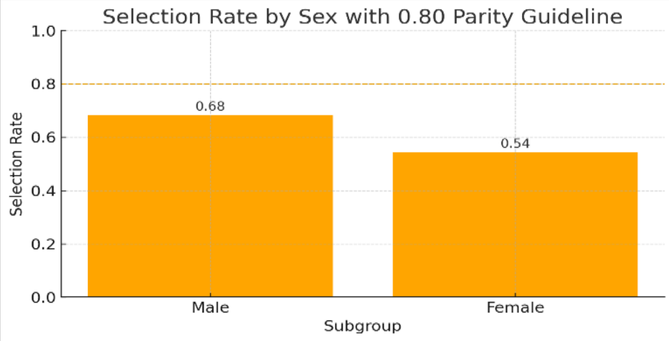


Figure 4.2. Selection Rate by Sex (Reference Group: Male)

Female applicants had a lower approval rate (0.54) than males (0.68), resulting in a disparate impact ratio of 0.79. This falls below the 0.80 guideline, requiring documentation of model logic, inspection of non-protected drivers, and a proportionality review under fairness obligations.

Interpretation

The disparity suggests that female applicants are less likely to be approved compared to the male reference group, despite sex being excluded from the training data. A fairness review must therefore evaluate whether correlated, non-protected features such as employment length, credit history, or account status indirectly influenced the observed disparity.

4.2 Population-Level Feature Influence

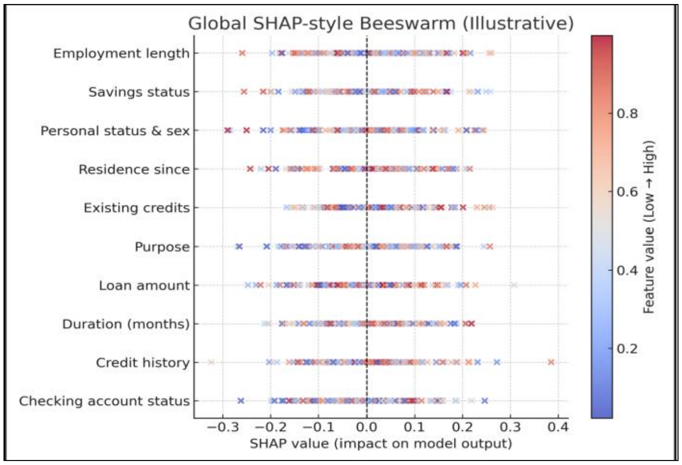


Figure 4.3. Global SHAP Summary Plot

Key Findings

SHAP global explainability results demonstrate that credit history length, existing loan amount, and checking account status consistently influence approval decisions across all groups. These key factors align with the disparities observed by age and sex, highlighting the importance of transparent, reviewable model logic under GDPR Article 22 and AI Act lifecycle controls.

Dominant Factor	**Effect on Outcome**	**Governance Interpretation**
Credit history duration	Longer duration ↑ approval likelihood	Indicates reliance on demonstrated repayment behaviour
Existing loan amount	Higher amount ↓ approval likelihood	Indicates risk-based exposure management
Checking account status	Positive balances ↑ approval confidence	The financial stability indicator is consistently weighted

Table 4.1 SHAP Global Feature Influence – Governance Interpretation Table

This table explains the meaning of each counterfactual adjustment and its legal relevance for contestability, justification, and correction obligations**.**

Governance Meaning

- Transparency (ISO/IEC 42001 Clause 6)
- Explainability adequacy (EU AI Act Article 13)
- Documentability of model logic for GDPR Article 22 rights

Alongside GDPR, EU AI Act Article 13 transparency obligations, and ISO/IEC 42001 lifecycle controls, the UK Government's AI White Paper (2023) further strengthens these duties by emphasising transparency, fairness, safety, accountability, and contestability as core principles for responsible AI development and deployment. These principles align with the GDPR, EU AI Act, and ISO/IEC 42001 requirements, broadening governance expectations within the UK regulatory framework. This analysis is guided by the OECD AI Principles, which champion transparency, explainability, fairness, accountability, and human-centred values throughout the AI lifecycle. These global principles highlight the importance of transparent and testable decision-making systems, especially in high-risk models. When a model's transparency is unclear, trust among stakeholders, deployers, regulators, and end-users decreases. Without a clear, traceable explanation of how an output was generated, it becomes difficult to rely on the system or defend its decisions. Any model that leaves room for doubt weakens evidence, reducing the credibility of decisions in governance, audit, or regulatory review.

4.3 Case-Level Justification Reasoning (Local Explainability)

These instance-specific explanations form part of the legally disclosable information required for Subject Access Requests under Articles 13-15 of the GDPR.

4.4 Counterfactual Remediation Pathways

Counterfactual explanations demonstrate plausible "what-if" scenarios that an applicant could pursue to turn a rejection into approval. These pathways adjust non-protected, model-relevant attributes (e.g., loan amount, credit duration, credit history) while keeping protected characteristics unchanged. They do not claim "fairness"; instead, they offer an actionable, legally recognisable path to remediation, supporting:

GDPR Article 22 appeal rights

EU AI Act Article 52 justification obligations

ISO/IEC 42001 accountability pathways

Table 4.2. Counterfactual Remedies (Interpretation Table)

Column	Meaning (Legal + Governance explanation)
CF	The scenario number (CF1 = First alternative path)
Feature changed	The financial factor that changed (e.g., credit history)
Original value	What the applicant actually had
New value	A realistic, improved scenario (feasible in real life)
Delta	Size of the change (not a random, reasonable *adjustment*)
Approval probability (after)	Whether this change would shift the outcome toward approval

Together, Figure 3.9 and Table 4.2 provide a clear, accessible pathway for remediation. The numeric counterfactuals (Figure 3.9) demonstrate adjustments that could enhance the applicant's case. Meanwhile, the interpretation table (Table 4.2) clarifies the legal significance of each column, connecting counterfactuals with governance duties under GDPR Article 22, EU AI Act justification requirements, and ISO/IEC 42001 auditability controls.

Governance Significance

Counterfactuals enable:

- Contestability (GDPR Article 22)
- Justification duties (EU AI Act Article 52)
- Fairness and proportionality assessments
- Supervisory traceability, supporting ombudsman investigations
- Correctability, one of the EU AI Act's core pillars

They do not claim fairness; they offer a practical route for improvement and a legally recognised clarification of "what could have been different."

4.5 Governance Interpretation

The combined results create a completely traceable chain of evidence, connecting:

- Fairness tests →
- Explainability artefacts (SHAP and LIME) →
- Counterfactual remediation pathways

Governance Requirement	Evidentiary Mechanism Delivered
Transparency	SHAP feature influence distributions
Human Review	LIME case-level justification narrations
Contestability	Counterfactual remediation pathways
Continuous Monitoring	Subgroup fairness metrics and logs

Table 4.3. Governance Requirements Mapped to Evidentiary Mechanisms

5. REFLECTIVE INSIGHT AND CONCLUSION

From Model Explanations to Legal Evidence: A Governance-Ready Framework for High-Risk AI.

This study shows that explainability is more than just a technical feature of machine learning systems; it also acts as a governance tool. In high-stakes situations, such as automated credit scoring, disparities in outcomes, lack of transparency, and opaque model logic can directly affect individuals' financial access, mobility, and long-term economic prospects. Therefore, fairness, explainability, and accountability must be supported by evidence rather than assumptions. The combined audit pipeline facilitates supervisory investigations, consumer appeals, and Subject Access Requests by providing a reproducible evidentiary trail.

The findings establish five core insights:

5.1 Explainability as a Governance Obligation, Not a Technical Add-On

Once an algorithm influences an individual's financial future, explainability becomes a legal and regulatory duty. The burden shifts from:

"Was the model accurate?" to "Can the organisation justify the fairness, transparency, and reasoning behind the outcome?"

Global (SHAP) and local (LIME) explainability techniques provided:

- Clear identification of primary model drivers.
- case-level justification trails suitable for ombudsman review.
- legally reviewable explanations supporting GDPR Article 22.
- documented logic aligns with EU AI Act Article 13 transparency obligations.

Subject Access Requests (SARs):

Under Articles 12-15 of the GDPR, individuals have the right to obtain meaningful information about automated decisions that affect them. The audit pipeline operationalised in this study, SHAP global influence summaries, LIME case-level justifications, and counterfactual remediation pathways, provide organisations with SAR-ready evidence. Each explanation component can be directly disclosed to the data subject, forming a legally reviewable narrative of:

1. What factors influenced the decision?
2. why the outcome occurred, and
3. What alternative actions (counterfactuals) could lead to a different result?

This transforms explainability outputs into regulator-ready and SAR-ready artefacts, meeting transparency and accountability duties.

Explainability, therefore, functions as an evidentiary infrastructure, allowing decisions to be scrutinised, contested, and defended.

5.2 Fairness as Procedural Protection, Not an Accuracy Metric

Subgroup outcome disparities revealed statistically significant differences for:

- Age ≥50, and
- Female applicants

Both fall below the 0.80 disparate-impact threshold.

These outcomes do not confirm unlawful discrimination, but they necessitate documentation, justification of model-relevant variables, and review of proportionality under:

- GDPR Article 22
- EU AI Act Articles 10 & 13
- ISO/IEC 42001 fairness monitoring controls

5.3 Counterfactual Explanations as Contestability Mechanisms

Counterfactual (DiCE) remediation pathways demonstrated feasible, realistic scenarios in which rejected applicants could have been approved without altering protected characteristics, such as age or sex.

This offers applicants the following:

- meaningful routes for appeal,
- evidence of contestation,
- a legally recognised alternative explanation for “what could have been different.”

Under GDPR Article 22 and EU AI Act Article 52, counterfactuals offer a justification trail, allowing fair challenge, independent review, proportionality assessment, and ombudsman scrutiny.

5.4 Chain-of-Evidence Integrity for Automated Decision Systems

Once AI enters litigation or regulatory review, the standard of truth shifts from:

“Is the document accurate?”

to

“Can we evidence the provenance of the data and logic that shaped the outcome?”

This study shows that Responsible AI needs traceable, reviewable, contestable, and correctable chains of justification. By integrating SHAP, LIME, subgroup fairness testing, and counterfactual pathways, an audit pipeline was created.

- provenance (what influenced the decision),
- reasoning (why the outcome occurred),
- alternatives (what could have changed), and
- compliance with GDPR, EU AI Act, and ISO/IEC 42001 requirements.

This chain-of-evidence integrity elevates explainability from a technical report to legal evidence.

5.5 Contribution to Research and Practice

This work promotes emerging Responsible AI governance by:

1. Bridging technical and legal perspectives, turning model outputs into contestable evidence for decision-making.
2. Proposing a unified audit pipeline that connects fairness testing, explainability artefacts, and counterfactual pathways.
3. Providing a reproducible evidence-pack format suitable for internal audits, consumer appeals, and regulatory reporting.

In practice, the workflow enables financial institutions to justify automated credit decisions, minimise litigation risk, and boost consumer trust.

5.6 Closing Reflection

Responsible AI requires more than just accurate models; it demands accountable systems. AI systems do not automatically become fair, explainable, or contestable simply by design. These qualities only develop when organisations document, scrutinise, justify, and challenge their automated outcomes throughout the entire lifecycle. This study shows that subgroup fairness results, SHAP explanations, LIME justifications, and counterfactual pathways within technical artefacts can be integrated into a transparent, governed audit trail that meets supervisory expectations for high-risk AI.

In summary:

AI governance is inherently anticipatory. Predictive systems detect risks early, before organisations recognise their impact. If a model's likely behaviour is not visible, its risks cannot be managed, mitigated, justified, or defended. Predictive visibility is therefore essential; you cannot govern what you cannot foresee. This work provides a practical pathway for transforming opaque model reasoning into legally sound, accountable, and reviewable evidence, ready for scrutiny in audits, appeals, or litigation.

Acknowledgment

The author sincerely thanks family, colleagues, and academic mentors for their unwavering support throughout this research. Their guidance has shaped the integration of technical explainability and governance compliance, which is explored in this study. This work operationalised SHAP, LIME, and counterfactual (DiCE) explanations as evidentiary artefacts that meet the GDPR Article 22 safeguards, the EU AI Act's high-risk transparency obligations, and ISO/IEC 42001 lifecycle accountability controls. The author also acknowledges the growing community of Responsible AI practitioners whose insights continue to advance the discussion from model reasoning to legal accountability.

References

Bird, S., Dudík, M., Edgar, R., Lutz, R., Milani Fard, A., Kallus, N., Kenthapadi, K., Kearns, M., Raghavan, M., Wallach, H., & Walker, K. (2020). *Fairlearn: A toolkit for assessing and improving fairness in AI.* https://fairlearn.org

Caswell, T. A., Droettker, M., Hunter, J. D., Smith, N., FiveTech, S., et al. (2023). *Matplotlib: A visualisation library for Python*. https://matplotlib.org

Court of Justice of the European Union. (2023). *Schufa Holding AG (Case C-634/21), Judgment of 7 December 2023*. CJEU.

Department for Science, Innovation and Technology. (2023). *A pro-innovation approach to AI regulation: White paper*. UK Government. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach

Dua, D., & Graff, C. (2019). *Statlog (German Credit) dataset*. UCI Machine Learning Repository. https://archive.ics.uci.edu/

European Banking Authority. (2020). *Guidelines on loan origination and monitoring*. EBA.

European Data Protection Board. (2022). *Guidelines on automated decision-making and profiling under the GDPR (Revised)*. EDPB.

European Union. (2016). *Regulation (EU) 2016/679 (General Data Protection Regulation)*. *Official Journal of the European Union*, L 119, 1–88.

European Union. (2024). *Regulation (EU) 2024/1689 (Artificial Intelligence Act)*. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Equality Act 2010, c. 15. (UK). https://www.legislation.gov.uk/ukpga/2010/15/contents

Financial Ombudsman Service. (2023). How we handle complaints. https://www.financial-ombudsman.org.uk

Harris, C. R., Millman, K., van der Walt, S., et al. (2020). Array programming with NumPy. *Nature*, 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2

Information Commissioner's Office & The Alan Turing Institute. (2023). *Explaining decisions made with AI: Guidance for organisations*. https://ico.org.uk

International Organization for Standardization. (2023). *ISO/IEC 42001:2023 Artificial intelligence management system requirements*. ISO.

LendingTree. (2020). *U.S. lenders face concerns about transparency in opaque credit scoring*. LendingTree.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.

McKinney, W. (2010). Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*, 51–56.

Mothilal, R., Sharma, K., & Tan, S. (2020). DiCE: Generating actionable counterfactual explanations. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 607–617. https://doi.org/10.1145/3351095.3372859

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework 1.0 (AI RMF 1.0)*. NIST.

New York State Department of Financial Services. (2019). *Statement on Apple Card algorithm review*. NYSDFS.

Organisation for Economic Co-operation and Development. (2019). Recommendation of the Council on Artificial Intelligence. OECD. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K., & Ghani, R. (2018). *Aequitas: A bias and fairness audit toolkit*. https://github.com/dssg/aequitas

Scikit-Learn Developers. (2023). *Scikit-learn documentation (Version 1.x)*. https://scikit-learn.org

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box. *Harvard Journal of Law & Technology*, 31(2), 841–887.