h_da hochschule darmstadt
fachbereich informatik

member of
EUT+
EUROPEAN UNIVERSITY OF TECHNOLOGY

Hochschule Darmstadt

– Fachbereich Informatik –

Explainable AI zur Analyse von Verzerrungen in ML-basierten Kreditrisikomodellen: Ein Vergleich von SHAP, LIME und PFI im Kontext gruppenbezogener Bias-Indikatoren

Abschlussarbeit zur Erlangung des akademischen Grades

Bachelor of Science (B. Sc.)

vorgelegt von

Denis Cvach

Matrikelnummer: 1115725

Referent: Prof. Dr. Stefan Zander
Korreferent: Prof. Dr. Jens Heuschkel

EIGENSTÄNDIGKEITSERKLÄRUNG

Ich versichere hiermit, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die im Literaturverzeichnis angegebenen Quellen benutzt habe.

Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder noch nicht veröffentlichten Quellen entnommen sind, sind als solche kenntlich gemacht.

Die Zeichnungen oder Abbildungen in dieser Arbeit sind von mir selbst erstellt worden oder mit einem entsprechenden Quellennachweis versehen.

Diese Arbeit ist in gleicher oder ähnlicher Form noch bei keiner anderen Prüfungsbehörde eingereicht worden.

KI-gestützte Werkzeuge habe ich ausschließlich im zulässigen Rahmen eingesetzt.

Darmstadt, 5. August 2026

Denis Cvach

ABSTRACT

Credit risk assessment increasingly relies on black-box models such as gradient-boosted tree ensembles, while the EU AI Act establishes requirements for transparency, human oversight and, under certain conditions, explanations for affected persons. SHAP, LIME and Permutation Feature Importance (PFI) have been widely studied as explanation methods, but have less frequently been compared with respect to their methodological properties and their technical utility for analysing group-related bias indicators.

This thesis compares the three post-hoc XAI methods using an XGBoost model trained on the *South German Credit* dataset (Grömping 2019). A binary sex variable derived from the confounded `personal_status_sex` attribute, together with the `foreign_worker` attribute, defines two sensitive dimensions. The global SHAP and PFI analyses are conducted at the level of the original variables: SHAP values are aggregated across encoded columns, while PFI is computed using grouped permutations. For the local explanations, LIME operates directly on the original variables. Ranking stability across repeated training runs is assessed for SHAP and PFI, while SHAP is additionally used in an ablation analysis of redundant sex-related representations. The technical capabilities of the three methods are subsequently assessed with reference to Articles 13, 14 and 86 of the EU AI Act.

At the variable level, SHAP and grouped PFI yield highly stable rankings for a fixed train/test split, with mean pairwise Spearman correlations of 0.99 and 0.94, respectively. LIME is excluded from this stability comparison because it does not provide an established global ranking and is therefore evaluated only locally. The outcome-based fairness metrics Equal Opportunity and Disparate Impact are reported separately as independent model diagnostics. The ablation analysis further shows that neither of the redundant sex-related representations becomes prominent after the other is removed.

Within the selected technical criteria, SHAP provides the broadest coverage of the examined explanation capabilities. Its measured stability applies to global rankings rather than individual explanations. PFI provides a complementary performance-based global perspective but no instance-specific explanations. LIME offers an alternative local surrogate explanation, although its incremental practical value over SHAP was not directly evaluated.

ZUSAMMENFASSUNG

Die Kreditrisikobewertung stützt sich zunehmend auf Black-Box-Modelle wie baumbasierte Gradient-Boosting-Ensembles. Zugleich stellt der EU AI Act Anforderungen an Transparenz, menschliche Aufsicht und unter bestimmten Voraussetzungen an Erläuterungen für betroffene Personen. SHAP, LIME und Permutation Feature Importance (PFI) wurden als Erklärungsverfahren vielfach untersucht, jedoch seltener hinsichtlich ihrer methodischen Eigenschaften und ihres technischen Unterstützungspotenzials bei der Analyse gruppenbezogener Bias-Indikatoren miteinander verglichen.

Die vorliegende Arbeit vergleicht die drei Post-hoc-XAI-Methoden anhand eines XGBoost-Modells, das auf dem *South German Credit*-Datensatz von Grömping (2019) trainiert wurde. Eine aus dem konfundierten Attribut `personal_status_sex` abgeleitete binäre Geschlechtsvariable und das Attribut `foreign_worker` bilden zwei sensible Dimensionen. Die globalen Auswertungen von SHAP und PFI erfolgen auf Ebene der Ursprungsvariablen: SHAP-Werte werden über die kodierten Spalten aggregiert und PFI wird mittels gruppierter Permutationen berechnet. LIME arbeitet für die lokalen Erklärungen direkt mit den Ursprungsvariablen. Für SHAP und PFI wird die Stabilität der Rankings über wiederholte Trainingsläufe untersucht. Zusätzlich wird mit SHAP eine Ablationsanalyse redundanter geschlechtsbezogener Repräsentationen durchgeführt. Anschließend werden die technischen Fähigkeiten der drei Methoden mit Bezug auf die Artikel 13, 14 und 86 des EU AI Act eingeordnet.

Auf Variablenebene erzeugen SHAP und die gruppierte PFI bei fester Datenaufteilung sehr stabile Rankings mit mittleren paarweisen Spearman-Korrelationen von 0,99 beziehungsweise 0,94. LIME wird von dieser Stabilitätsanalyse ausgenommen, da es kein etabliertes globales Ranking bereitstellt, und ausschließlich lokal ausgewertet. Die outcome-basierten Fairnessmetriken Equal Opportunity und Disparate Impact werden getrennt davon als unabhängige Modelldiagnostik berichtet. Die Ablationsanalyse zeigt zudem, dass keine der beiden redundanten geschlechtsbezogenen Repräsentationen nach Entfernung der jeweils anderen prominent wird.

Innerhalb der ausgewählten technischen Kriterien bietet SHAP die breiteste Abdeckung der betrachteten Erklärungsmöglichkeiten. Die gemessene Stabilität bezieht sich ausschließlich auf globale Rankings und nicht auf einzelne Erläuterungen. PFI ergänzt eine performancebasierte globale Perspektive, bietet jedoch keine Einzelfallerklärungen. LIME stellt eine alternative lokale Surrogaterklärung bereit, deren zusätzlicher praktischer Nutzen gegenüber SHAP nicht direkt untersucht wurde.

INHALTSVERZEICHNIS

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

ABKÜRZUNGSVERZEICHNIS

AUC	Area Under the Curve
EUAIA	EU Artificial Intelligence Act
LIME	Local Interpretable Model-agnostic Explanations
ML	Machine Learning
NBER	National Bureau of Economic Research
PFI	Permutation Feature Importance
ROC	Receiver Operating Characteristic
SHAP	Shapley Additive Explanations
TPR	True-Positive-Rate
XAI	Explainable Artificial Intelligence
XGBoost	Extreme Gradient Boosting

1 EINLEITUNG

1.1 MOTIVATION

Kreditscoring bewertet das Kreditrisiko einer Person auf Basis finanzieller, beruflicher und soziodemografischer Merkmale und überführt das geschätzte Ausfallrisiko typischerweise in einen Score oder eine Risikoklasse [2], die Entscheidungen über den Zugang zu Krediten, Hypotheken und weiteren Finanzdienstleistungen beeinflussen kann [7, 31]. Mit der Verfügbarkeit größerer Datenmengen werden neben regelbasierten und klassischen statistischen Modellen zunehmend komplexe Verfahren des maschinellen Lernens (ML) eingesetzt [2]. Diese können gegenüber klassischen Verfahren eine höhere Vorhersagequalität erreichen [8, 40], weisen jedoch häufig eine Black-Box-Charakteristik auf, deren Entscheidungslogik nicht unmittelbar nachvollziehbar ist [50].

Diese Intransparenz ist im Kreditkontext kein rein technisches Problem. Modelle können Verzerrungen aus historischen Daten übernehmen und Benachteiligungen entlang sensibler Attribute wie Geschlecht oder Herkunft reproduzieren [7]. Bei komplexen Black-Box-Modellen sind solche Zusammenhänge jedoch besonders schwer zu erkennen [50]. Empirische Studien dokumentieren geschlechtsbezogene Diskriminierung im Konsumkreditmarkt [45], ethnizitäts- und geschlechtsbezogene Unterschiede in Kredit-Scores [27] sowie gruppenbezogene Auswirkungen ML-gestützter Kreditvergabe auch ohne direkte Nutzung geschützter Merkmale [21]. Eine juristische Analyse beschreibt darüber hinaus diskriminierende Wirkungen des Kreditscorings auf Communities of Color [49]. Vor diesem Hintergrund stuft der EU Artificial Intelligence Act (EUAIA) KI-Systeme, die bestimmungsgemäß zur Kreditwürdigkeitsprüfung oder Bonitätsbewertung natürlicher Personen eingesetzt werden, grundsätzlich als Hochrisiko-KI-Systeme ein. Er stellt Anforderungen an Transparenz und menschliche Aufsicht und sieht unter bestimmten Voraussetzungen ein Recht betroffener Personen auf Erläuterung vor [17]. XAI-Methoden sollen zur Nachvollziehbarkeit solcher Modelle beitragen und Hinweise auf potenzielle Verzerrungen liefern [4], wobei mit SHAP, LIME und Permutation Feature Importance (PFI) mehrere etablierte Verfahren zur Verfügung stehen, die sich in ihren Erklärungen jedoch unterscheiden können [44, 51].

1.2 PROBLEMSTELLUNG

Das domänenspezifische Problem besteht in der Spannung zwischen potenziell leistungsfähigen, aber schwer nachvollziehbaren ML-basierten Kreditrisikomodellen und regulatorischen Anforderungen an Transparenz und Er-

klärbarkeit. Untersuchungsgegenstand ist dabei kein vollständiges Kreditvergabesystem, sondern ein ML-basiertes Kreditrisikomodell, das den späteren Verlauf bereits vergebener Kredite prognostiziert. An XAI-Methoden richtet sich die Erwartung, das Verhalten solcher Modelle nachvollziehbar zu machen und dabei auch Hinweise auf potenzielle Benachteiligungen sichtbar werden zu lassen. In welchem Umfang sie diese Erwartung erfüllen können, ist jedoch offen. SHAP, LIME und PFI beruhen auf unterschiedlichen methodischen Konzepten und erzeugen unterschiedliche Erklärungsobjekte. Damit ist unklar, welche methodischen Eigenschaften und Grenzen die Verfahren bei der Analyse potenzieller Bias-Indikatoren aufweisen und wie weit ihre Erklärungen als Grundlage einer solchen Analyse überhaupt tragen.

Daraus ergeben sich drei Teilprobleme, die die Forschungsfragen in Abschnitt 1.3 aufgreifen. Ein *Stabilitätsproblem* entsteht, weil sich die Rangfolge der Merkmale zwischen Trainingsläufen allein aufgrund stochastischer Komponenten des Modelltrainings verändern kann. Ohne Kenntnis dieser Stabilität lässt sich die Aussagekraft eines einzelnen beobachteten Rankings schwer beurteilen (F1).

Ein *Redundanzproblem* entsteht, wenn sensible Information nicht isoliert in einer Variable vorliegt, sondern über mehrere Modelleingaben verteilt ist, weil Merkmale einander teilweise enthalten oder gemeinsam variieren. Die vorliegende Arbeit untersucht diesen Fall nicht in seiner ganzen Breite, sondern an einer gezielt hergestellten Konstellation zweier deterministisch verbundener Geschlechtsrepräsentationen, die Abschnitt 4.1 im Einzelnen beschreibt. Attributionen können sich in solchen Fällen auf mehrere Repräsentationen verteilen, sodass der Rang einer einzelnen Variable ihre Rolle im Modell nicht vollständig widerspiegelt. Ob eine niedrige Attribution einen geringen eigenständigen Beitrag oder lediglich die Verteilung der Attribution über mehrere Repräsentationen anzeigt, geht aus der Erklärung allein nicht hervor (F2).

Regulatorisch stellt der EUAIA Anforderungen an Transparenz und menschliche Aufsicht und sieht unter bestimmten Voraussetzungen eine fallspezifische Erläuterung vor, benennt jedoch keine konkreten XAI-Verfahren zu deren technischer Unterstützung. Offen bleibt daher, in welchem Umfang die von XAI-Methoden erzeugten Erklärungen die Anforderungen technisch unterstützen (F3).

Die Untersuchung bleibt in allen drei Teilproblemen auf der Ebene deskriptiver Befunde; kausale oder rechtliche Diskriminierungsaussagen werden daraus nicht abgeleitet. Die zugehörige Forschungslücke leitet Abschnitt 3.3 aus den verwandten Arbeiten her.

1.3 FORSCHUNGSFRAGEN

Aus der Problemstellung leiten sich drei Forschungsfragen ab, die die empirische Untersuchung anhand operationalisierter Kriterien strukturieren:

F1 (STABILITÄT). Wie stark verändern sich die globalen Variablen-Rankings von SHAP und PFI, wenn das XGBoost-Modell mit unterschiedlichen Zufalls-Seeds erneut trainiert wird?

F2 (REDUNDANZ- UND ABLATIONSEFFEKTE). Wie verändern sich die SHAP-Ränge der redundanten Repräsentationen `sex_binary` und `personal_status_sex`, wenn jeweils eine der beiden Variablen aus den Modelleingaben entfernt und das Modell anschließend erneut trainiert wird?

F3 (REGULATORISCHES UNTERSTÜTZUNGSPOTENZIAL). In welchem Umfang eignen sich die von SHAP, LIME und PFI erzeugten Erklärungen als technische Unterstützung für Transparenz, menschliche Aufsicht und fallspezifische Erläuterungen nach den Artikeln 13, 14 und 86 des EUAIA?

Die methodische Operationalisierung der Forschungsfragen F1 und F2 sowie der ergänzend berichteten Fairnessmetriken erfolgt in Abschnitt 4.4. Die Methodik der regulatorischen Einordnung von F3 wird in Abschnitt 4.5 erläutert.

1.4 AUFBAU DER ARBEIT

Kapitel 2 führt Kreditrisikomodelle, die XAI-Methoden SHAP, LIME und PFI, die Fairnessmetriken sowie die relevanten Anforderungen des EUAIA ein, Kapitel 3 ordnet die Untersuchung in den Forschungsstand ein. Kapitel 4 beschreibt den empirischen Untersuchungsaufbau, Kapitel 5 berichtet die Modell-, XAI- und Fairnessergebnisse. Kapitel 6 diskutiert die Befunde entlang der Forschungsfragen und ordnet ihre Aussagekraft unter Berücksichtigung der methodischen Limitationen ein. Kapitel 7 fasst die Ergebnisse zusammen, beantwortet die Forschungsfragen und zeigt Ansatzpunkte für weiterführende Untersuchungen auf.

2 GRUNDLAGEN

Dieses Kapitel führt die für die Untersuchung relevanten Begriffe und Verfahren ein: den Anwendungskontext von Kreditrisikomodellen und Kreditscoring, die XAI-Methoden SHAP, LIME und PFI, die outcome-basierten Fairnessmetriken sowie die relevanten Bestimmungen des EUAIA.

2.1 KREDITRISIKOMODELLE UND KREDITSCORING

Ein Kreditrisikomodell schätzt aus den Merkmalen einer kreditnehmenden Person die Wahrscheinlichkeit eines zukünftigen Zahlungsausfalls beziehungsweise eines guten oder schlechten Kreditverlaufs [2]. Die im Kreditscoring daraus abgeleitete Risikoklasse kann Finanzinstitute bei Entscheidungen über Kreditvergabe, Zinssätze und Kreditlinien unterstützen. Im hier verwendeten Datensatz ist die Zielvariable der beobachtete Rückzahlungsverlauf bereits vergebener Kredite (vgl. 4.1).

Einen frühen Ausgangspunkt der statistischen Kreditrisikobewertung bildet die Studie von Durand für das National Bureau of Economic Research (NBER), die als eine der ersten systematisch untersuchte, anhand welcher Merkmale sich gute von schlechten Ratenkrediten unterscheiden lassen [2, 14]. Ab den 1990er-Jahren finden sich erste Anwendungen maschineller Lernverfahren [33], deren breite Untersuchung im Kreditscoring der Benchmark-Vergleich von Lessmann et al. [40] dokumentiert. Zu den im Kreditscoring häufig untersuchten Verfahren zählen logistische Regressionen sowie baumbasierte Ensembleverfahren wie Gradient-Boosting-Modelle [8, 40]. Bei Modellen wie XGBoost lässt sich die Entscheidungslogik jedoch, anders als bei logistischen Regressionen oder klassischen Scorecards, nicht unmittelbar aus den Modellparametern und der Gesamtheit der trainierten Entscheidungsbäume ablesen (vgl. 4.2).

Bereits in der frühen statistischen Kreditrisikobewertung galten gruppenbezogene Effekte als erklärungsbedürftig. Durand stellte fest, dass Frauen in den untersuchten Kreditdaten als bessere Risiken erschienen, und berichtete die von Kreditpraktikern geäußerte Vermutung, dieser Zusammenhang könne auf dem indirekten Einfluss anderer, mit dem Geschlecht korrelierter Faktoren beruhen. Eine einfache Kreuzklassifikation nach Beruf bestätigte diese Erklärung in seinen Daten nicht [14, S. 74]. Die heutige Proxy-Diskussion setzt an derselben Grundproblematik korrelierter Merkmale an, betrachtet jedoch eine andere Wirkungsrichtung. Während das Geschlecht bei Durand möglicherweise lediglich als Indikator für andere risikorelevante Faktoren erschien, können umgekehrt andere Merkmale geschlechtsbezogene Information im Modell verfügbar halten. Als potenzielles Proxy-Merkmal gilt dabei ein Merkmal, das selbst nicht geschützt ist, aber so mit einem geschütz-

ten Merkmal zusammenhängt, dass dessen Information indirekt in die Modellvorhersage einfließen kann [7]. Von einer Proxy-Nutzung im engeren Sinne lässt sich sprechen, wenn das Modell diese korrelierte Information tatsächlich für seine Vorhersage verwendet [13]. Der Ausschluss geschützter Merkmale aus den Modelleingaben beseitigt eine mögliche Benachteiligung daher nicht zwangsläufig, da solche Proxies wirksam bleiben [13, 31, 56].

2.2 EXPLAINABLE ARTIFICIAL INTELLIGENCE

Unter *Explainable Artificial Intelligence* werden Methoden zusammengefasst, die Entscheidungen maschineller Lernmodelle nachvollziehbar machen sollen [4]. Sie lassen sich entlang dreier Achsen klassifizieren. Lokale Erklärungen beschreiben einzelne Vorhersagen, während globale Erklärungen das Modellverhalten über eine Menge von Instanzen zusammenfassen [44]. Modellspezifische Verfahren nutzen die interne Struktur einer bestimmten Modellklasse, modellagnostische Verfahren greifen dagegen ausschließlich auf die Ein- und Ausgaben des Modells zu [3]. Ante-hoc-Ansätze setzen auf intrinsisch interpretierbare Modelle, wohingegen Post-hoc-Verfahren nachträglich auf ein bereits trainiertes Modell angewendet werden [25].

2.2.1 *SHAP*

SHAP überträgt das spieltheoretische Konzept der Shapley-Werte [52] auf die Bestimmung des Beitrags einzelner Features zu einer Modellvorhersage [43]. Der Shapley-Wert eines Features entspricht dessen gewichtetem durchschnittlichem Grenzbeitrag über mögliche Koalitionen der übrigen Features. Unter gegebener Wertfunktion wird diese Zuordnung durch die vier Axiome *Efficiency*, *Symmetry*, *Dummy* und *Additivity* eindeutig festgelegt [43, 52]. Für diese Arbeit sind zwei davon unmittelbar relevant. *Efficiency* verlangt, dass sich Basiswert und Attributionen einer Instanz exakt zur Modellausgabe summieren, sodass eine lokale Erklärung die Vorhersage vollständig auf die Features aufteilt. *Additivity* überträgt diese Zuordnung auf zusammengesetzte Modelle. Setzt sich die Modellausgabe als Summe mehrerer Teilmodelle zusammen, so ist die Attribution eines Features die Summe seiner Attributionen in diesen Teilmodellen. Da ein XGBoost-Modell die Ausgaben seiner einzelnen Bäume addiert, lassen sich die Attributionen dadurch baumweise berechnen und anschließend aufsummieren [42].

Da die allgemeine Berechnung von Shapley-Werten exponentiell aufwendig ist, wird für baumbasierte Modelle mit TreeSHAP in dieser Arbeit eine modellspezifische Variante eingesetzt, die die instanzweisen Attributionen in polynomieller Zeit berechnet [42, 44]. SHAP ermöglicht sowohl die lokale Erklärung einer Einzelvorhersage als auch die Aggregation lokaler Attributionen zu einer modellweiten Übersicht, beispielsweise über den Mittelwert ihrer absoluten Werte je Feature [41, 51]. Ein solches globales SHAP-Ranking beschreibt, wie groß die einem Feature zugeschriebenen Beiträge zu den betrachteten Modellvorhersagen im Durchschnitt ausfallen. Es misst nicht,

wie stark die Vorhersageleistung des Modells von diesem Feature abhängt, und ist daher von einer performancebasierten Wichtigkeit wie der PFI zu unterscheiden [20, 41], die Abschnitt 2.2.3 einführt. Kumar et al. [39] argumentieren dabei allgemeiner, dass sich das Problem der Feature-Wichtigkeit nicht allein durch die Übertragung eines spieltheoretischen Rahmens lösen lässt.

2.2.2 *LIME*

LIME [48] erklärt eine einzelne Vorhersage, indem in einer durch Perturbation der Eingabe erzeugten und um die zu erklärende Instanz gewichteten Nachbarschaft ein lineares Surrogatmodell an das Verhalten des komplexen Modells angepasst wird. Da dafür ausschließlich die Vorhersagefunktion des erklärten Modells benötigt wird, ist das Verfahren modellagnostisch anwendbar. Die Koeffizienten des Surrogats geben an, wie die ausgewählten lokalen Bedingungen innerhalb der erzeugten Nachbarschaft mit der Surrogatvorhersage zusammenhängen, und stellen anders als SHAP-Werte keine additive Zerlegung des Modelloutputs dar. LIME erzeugt ausschließlich lokale Surrogaterklärungen und stellt damit kein etabliertes globales Wichtigkeitsmaß bereit. Es wird in dieser Arbeit daher nicht in die Stabilitätsanalyse globaler Rankings einbezogen, sondern anhand einer lokalen Erklärung dargestellt und hinsichtlich seines technischen Unterstützungspotenzials für fallspezifische Erläuterungen eingeordnet (vgl. 4.3 und 4.5).

Zwei literaturgestützte Einschränkungen sind dafür unmittelbar relevant. Erstens beruht die Erklärung auf einer stochastisch erzeugten Perturbationsstichprobe und hängt zusätzlich von Nachbarschaftsdefinition, Kernelgewichtung, Diskretisierung und Feature-Auswahl ab; wiederholt berechnete Erklärungen derselben Instanz können daher voneinander abweichen [5, 48]. Die TreeSHAP-Berechnung ist demgegenüber bei festgelegtem Baummodell, fester Implementierung und identischen Eingabedaten deterministisch [42]. Die vorliegende Arbeit quantifiziert die Variabilität wiederholter LIME-Erklärungen nicht (vgl. 6.5). Zweitens kann ein lineares Surrogat die lokal nichtlinearen Entscheidungsstrukturen baumbasierter Ensembles nur näherungsweise abbilden, was die lokale Treue zum Originalmodell begrenzt [44].

2.2.3 *Permutation Feature Importance*

Die Permutationswichtigkeit geht auf Breiman [10] zurück. In ihrer modellagnostischen Verallgemeinerung nach Fisher et al. [20] misst PFI die globale, performancebasierte Wichtigkeit eines Features, indem dessen Werte über die Instanzen eines Evaluationsdatensatzes zufällig permutiert werden. Die dadurch verursachte Verschlechterung der Modellgüte dient als Importance-Maß. Die in dieser Arbeit verwendete Umsetzung folgt der Darstellung von Molnar [44, Kap. 23] und liefert kein lokales Attributionsmaß.

Eine wesentliche Limitation betrifft statistisch oder strukturell abhängige Features. Werden deren Werte unabhängig voneinander permutiert, können unrealistische Merkmalskombinationen entstehen und die geschätzte Importance verzerrt werden [30]. Da die One-Hot-kodierten Indikatorspalten derselben Ursprungsvariable strukturell voneinander abhängig sind, würde ihre getrennte Permutation die Kategorienstruktur verletzen. Die vorliegende Untersuchung permutiert die Indikatorspalten einer Ursprungsvariable daher gemeinsam und wertet PFI auf Variablenebene aus (vgl. 4.3). Die gruppierte Permutation adressiert damit ausschließlich die strukturelle Abhängigkeit innerhalb einer Ursprungsvariable. Statistische oder deterministische Abhängigkeiten zwischen verschiedenen Ursprungsvariablen bleiben unberücksichtigt, da diese weiterhin getrennt permutiert werden (vgl. 6.5).

2.3 FAIRNESSMETRIKEN IM MASCHINELLEN LERNEN

Fairnessmetriken quantifizieren gruppenbezogene Disparitäten auf der *Outcome-Ebene*, also anhand der Modellklassifikationen und ihres Verhältnisses zu den tatsächlichen Klassen, etwa über Unterschiede in positiven Klassifikationsraten oder in gruppenspezifischen Fehlerraten. Damit setzen sie an einer anderen Ebene an als die XAI-Verfahren, deren Erklärungen sich auf Features und deren Attributionen beziehen. Eine allgemeingültige Fairnessmetrik existiert nicht, da unterschiedliche Definitionen jeweils einen anderen Fairnessbegriff formalisieren [54]. Zentrale Kriterien wie Kalibrierung und gruppenweise gleiche Fehlerraten lassen sich zudem bei unterschiedlichen gruppenspezifischen Basisraten und nicht perfekten Vorhersagen nur in Spezialfällen gleichzeitig erfüllen [36]. Die Wahl der Metriken ist daher eine kontextabhängige inhaltliche Festlegung. Diese Arbeit verwendet zwei Fairnessmetriken, die einen unmittelbar interpretierbaren Vergleich zweier Gruppen ermöglichen.

Equal Opportunity [26] fordert gleiche True-Positive-Raten zwischen sensiblen Subgruppen. Im Kreditkontext bedeutet dies, dass Personen mit tatsächlich vertragsgemäß erfülltem Kredit unabhängig von ihrer Gruppenzugehörigkeit mit gleicher Wahrscheinlichkeit korrekt als gutes Kreditrisiko eingestuft werden sollten. Die als Betrag berichtete Differenz der True-Positive-Raten (TPR), der Equal-Opportunity-Gap, dient als Kennwert für die jeweilige sensible Achse (vgl. 4.1).

Disparate Impact [19] misst dagegen das Verhältnis der positiven Klassifikationsrate der betrachteten Gruppe zur entsprechenden Rate der Referenzgruppe. Werte unter eins zeigen an, dass die betrachtete Gruppe seltener als gutes Kreditrisiko eingestuft wird, was einer geringeren Zustimmungsrate nur unter der Annahme entspricht, dass die Modellklassifikation eine Vergabeentscheidung steuert. Nach der Four-Fifths-Rule gilt ein Wert unter 0,8 als deskriptiver Hinweis auf eine mögliche gruppenbezogene Benachteiligung [19]. Die Schwelle stammt als Aufgreifkriterium für eine genauere Prüfung aus der US-amerikanischen Verwaltungspraxis zu Einstellungsverfahren [16], ist im europäischen Kreditkontext rechtlich nicht bindend

und dient hier ausschließlich als konventioneller Orientierungswert, nicht als Nachweis einer Diskriminierung. Da Equal Opportunity nur Personen mit tatsächlich gutem Kreditverlauf betrachtet, Disparate Impact dagegen alle Klassifikationen einbezieht, können beide Metriken zu unterschiedlichen Befunden führen.

Beide Metriken zeigen damit ausschließlich gruppenbezogene Unterschiede auf der Outcome-Ebene; sie geben nicht an, welche Variablen zu diesen Unterschieden beitragen (vgl. 4.4).

2.4 EU ARTIFICIAL INTELLIGENCE ACT

Der EUAIA [17] gilt grundsätzlich ab dem 2. August 2026. KI-Systeme zur Kreditwürdigkeits- oder Bonitätsbewertung natürlicher Personen gelten nach Art. 6 Abs. 2 i. V. m. Anhang III Nr. 5 lit. b als Hochrisiko-KI-Systeme. Die im Folgenden herangezogenen Art. 13 und 14 stehen in Kapitel III Abschnitt 2 und gelten aufgrund der Verordnung (EU) 2026/1744 [18] für Systeme nach Anhang III jedoch erst ab dem 2. Dezember 2027. Von dieser Verschiebung nicht erfasst ist der in Kapitel IX stehende Art. 86, der ab dem 2. August 2026 gilt. Das in dieser Arbeit betrachtete Recht auf Erläuterung greift damit zeitlich vor den hier betrachteten Transparenz- und Aufsichtspflichten, steht mit diesen jedoch sachlich in engem Zusammenhang.

Das in dieser Arbeit trainierte XGBoost-Modell (vgl. 4.2) dient als experimentelles Kreditrisikomodell zur methodischen Untersuchung und zum Vergleich der XAI-Verfahren. Der EUAIA dient daher nicht der rechtlichen Bewertung des Forschungsmodells, sondern als regulatorischer Referenzrahmen für den hypothetischen Einsatz eines vergleichbaren Modells in einem realen Kreditscoring-System.

Für die Untersuchung sind insbesondere die Artikel 13, 14 und 86 relevant. Artikel 13 verlangt eine hinreichende Transparenz, damit Betreiber die Systemausgaben interpretieren und angemessen verwenden können. Artikel 14 fordert eine wirksame menschliche Aufsicht, einschließlich des Verständnisses der Fähigkeiten und Grenzen des Systems sowie der Möglichkeit, dessen Ausgaben zu übergehen. Artikel 86 sieht unter den dort genannten Voraussetzungen ein Recht betroffener Personen auf eine klare und aussagekräftige Erläuterung zur Rolle des KI-Systems und zu den wichtigsten Elementen der betreffenden Entscheidung vor.

Weitere Bestimmungen des EUAIA sind für ein solches System ebenfalls einschlägig, etwa zur Daten-Governance, zur Protokollierung sowie zu Genauigkeit und Robustheit. Sie richten sich jedoch an die Datengrundlage und den Systembetrieb und nicht an die Erklärungseigenschaften der hier verglichenen Verfahren. Die Einordnung in Abschnitt 4.5 beschränkt sich daher auf die drei genannten Artikel.

3 VERWANDTE ARBEITEN

Dieses Kapitel ordnet die Untersuchung in den Forschungsstand ein und leitet daraus die Forschungslücke ab. Dazu werden methodische Vergleiche von XAI-Verfahren sowie empirische Untersuchungen zu Verzerrungen in Kreditrisikomodellen und beim Einsatz solcher Modelle im Kreditscoring betrachtet. Der EUAIA dient dabei als regulatorischer Referenzrahmen für F3 (vgl. 2.4 und 4.5).

3.1 XAI-METHODEN IM VERGLEICH

Aufbauend auf den Klassifikationsachsen aus Abschnitt 2.2 zeigt die Überblicksarbeit von Ali et al. [4], dass SHAP, LIME und PFI unterschiedliche Erklärungsebenen und methodische Konzepte von Feature-Wichtigkeit verwenden. Salih et al. [51] stellen SHAP und LIME für tabellarische Daten gegenüber und zeigen, dass die Ergebnisse beider Verfahren von der verwendeten Modellklasse und von Abhängigkeiten zwischen den Merkmalen beeinflusst werden können. PFI beziehen sie nicht ein, sodass der Vergleich mit einer globalen, performancebasierten Feature-Wichtigkeit offenbleibt. Kumar et al. [39] zeigen zudem, dass aggregierte lokale Shapley-Attributionen nicht ohne Weiteres als globale Feature-Wichtigkeit interpretiert werden können. Die drei Verfahren erzeugen unterschiedliche und daher nur eingeschränkt vergleichbare Erklärungsobjekte. SHAP und PFI ermöglichen globale Rangfolgen auf Grundlage unterschiedlicher Wichtigkeitskonzepte, während LIME lokale Surrogaterklärungen ohne etabliertes globales Wichtigkeitsmaß liefert. Aas et al. [1] weisen nach, dass sich Shapley-Attributionen bei korrelierten Features je nach Verteilungsannahme unterschiedlich auf die beteiligten Variablen verteilen. Dies ist für die vorliegende Untersuchung relevant, da die One-Hot-kodierten Ausprägungen von `personal_status_sex` strukturell abhängig sind und `sex_binary` zusätzlich deterministisch aus dieser Ursprungsvariable abgeleitet wird (vgl. 4.1).

Zugleich sind Grenzen post-hoc erzeugter Erklärungen dokumentiert. Rudin [50] bewertet sie bei risikoreichen Entscheidungen nicht als vollständigen Ersatz für intrinsisch interpretierbare Modelle. Slack et al. [53] zeigen darüber hinaus, dass adversarial konstruierte Modelle die perturbationsbasierten Verfahren LIME und KernelSHAP gezielt plausible, aber irreführende Erklärungen erzeugen lassen. Der Angriff nutzt aus, dass beide Verfahren das Modell auf künstlich erzeugten Instanzen auswerten; ob er sich auf die hier eingesetzte pfadabhängige TreeSHAP-Berechnung übertragen lässt, ist damit nicht gesagt. Derartige Manipulationen liegen ohnehin außerhalb des hier untersuchten Modellkontexts.

In der hier berücksichtigten Literatur dienen XAI-Methoden im Kreditkontext überwiegend der allgemeinen Erklärung von Modellvorhersagen und nicht der systematischen Bias-Prüfung. Gramegna und Giudici [22] vergleichen SHAP und LIME bei der Ausfallprognose kleiner und mittlerer Unternehmen und kommen zu dem Ergebnis, dass sich die Erklärungsmuster von Ausfall- und Nichtausfallfällen anhand von SHAP deutlicher voneinander abgrenzen lassen als anhand von LIME. Bitetto et al. [9] identifizieren mit PFI und Shapley-Werten die vorhersagerelevanten Merkmale von Kreditrisikomodellen, betrachten dabei jedoch keine gruppenbezogenen Verzerrungen. Hlongwane et al. [28] berichten, dass aus Shapley-Werten abgeleitete Kredit-Scorecards eine mit logistischer Regression vergleichbare Transparenz erreichen, untersuchen jedoch ebenfalls keine gruppenbezogenen Verzerrungen.

Methodisch am nächsten stehen der vorliegenden Untersuchung Kłosok und Chlebus [37] sowie Chen et al. [12]. Kłosok und Chlebus stellen auf einem Kreditausfall-Datensatz ein breites Spektrum modellagnostischer Verfahren einschließlich PFI, LIME und Shapley-Werten gegenüber, verknüpfen die Ergebnisse jedoch weder mit outcome-basierten Fairnessmetriken noch mit regulatorischen Anforderungen. Chen et al. untersuchen auf mehreren Kreditdatensätzen, darunter dem *South German Credit*-Datensatz [23] (vgl. 4.1), die Stabilität von aus SHAP- und LIME-Erklärungen abgeleiteten Ranglisten und berichten eine abnehmende Stabilität bei zunehmendem Klassenungleichgewicht. Sie betrachten Stabilität jedoch als allgemeines Qualitätsmerkmal und nicht als Bestandteil einer fairnessorientierten Bias-Prüfung. Für die gewählte Kombination aus Modell und Datensatz bleibt damit offen, wie stabil das globale SHAP-Ranking und die gruppierte PFI unter erneutem Training mit unterschiedlichen Zufalls-Seeds sind (F1; zur Operationalisierung vgl. 4.4).

3.2 BIAS UND FAIRNESS IN KREDITRISIKOMODELLEN UND KREDITENTSCHEIDUNGEN

Empirische Untersuchungen zeigen, dass Kreditvergabe und Kreditscoring Benachteiligungen entlang sensibler Merkmale reproduzieren können. Henderson et al. [27] finden ethnizitäts- und geschlechtsbezogene Unterschiede in den Kredit-Scores US-amerikanischer Unternehmensgründungen und Montoya et al. [45] weisen geschlechtsbezogene Diskriminierung im chilenischen Konsumkreditmarkt experimentell nach. Methodisch beruhen diese Untersuchungen auf der statistischen Auswertung beobachteter Kreditergebnisse beziehungsweise auf experimentell variierten Kreditanträgen. XAI-Methoden zur Untersuchung der zugrunde liegenden Modelle kommen nicht zum Einsatz. Fuster et al. [21] zeigen ergänzend, dass leistungsfähigere ML-Verfahren gruppenbezogene Auswirkungen auch ohne direkte Nutzung geschützter Merkmale erzeugen können, wobei sie als mögliche Mechanismen größere funktionale Flexibilität und die Rekonstruktion ausgeschlossener Gruppeninformation aus anderen Variablen diskutieren. Diese Arbeiten unterstreichen die praktische Relevanz gruppenbezoge-

ner Benachteiligungen im Kreditkontext, sind jedoch keine unmittelbaren methodischen Vergleichsstudien für die vorliegende Analyse eines Kreditrisikomodells.

Auf der Outcome-Ebene untersuchen Kozodoi et al. [38] verschiedene gruppenbezogene Fairnesskriterien und Fairnessinterventionen systematisch für das Kreditscoring und verknüpfen die in Abschnitt 2.3 eingeführten outcome-basierten Fairnessmetriken [19, 26] unmittelbar mit der Bias-Bewertung von Kreditmodellen. Hurlin et al. [32] entwickeln für Kreditrisikomodelle ein statistisches Verfahren zur Prüfung gruppenbezogener Fairness und zur Identifikation der zu einer Fairnessverletzung beitragenden Variablen und nutzen diese anschließend zur Verbesserung des Fairness-Performance-Trade-offs. Ihre Untersuchung verdeutlicht, dass Vorhersageleistung und gruppenbezogene Fairness getrennt bewertet werden müssen. Die Trennung zwischen Vorhersageleistung und gruppenbezogener Fairness wird in dieser Arbeit durch die ergänzend berichteten outcome-basierten Fairnessmetriken berücksichtigt.

Der methodischen Fragestellung näher steht die Studie von Nwafor et al. [47], die sensible Merkmale nicht auf Outcome-Ebene, sondern über XAI-Erklärungen betrachtet. Nwafor et al. erklären ein hybrides Kreditmodell mit SHAP und zeigen, dass die Entfernung von Alter und Geschlecht die Klassifikationsgüte kaum verändert. Der Befund belegt zunächst lediglich, dass diese Merkmale keinen wesentlichen zusätzlichen Beitrag zur gemessenen Vorhersageleistung leisten. Da die gruppenbezogene Fairness des reduzierten Modells und die mögliche Verfügbarkeit sensibler Information über korrelierte Merkmale nicht umfassend geprüft werden, erlaubt er keine abschließende Aussage über Biasfreiheit. Daraus ergibt sich die Notwendigkeit, nicht nur die Entfernung eines sensiblen Merkmals, sondern auch das Verhalten verbleibender redundanter oder korrelierter Repräsentationen zu untersuchen. F2 greift dies für einen einzelnen kontrollierten, deterministischen Redundanzfall auf und leistet keine allgemeine Proxy-Detektion (vgl. 4.4).

3.3 FORSCHUNGSLÜCKE

Die genannten Arbeiten behandeln jeweils Teilaspekte der Methodencharakterisierung oder der empirischen Bias-Identifikation, ohne sie mit einer regulatorischen Einordnung zu verbinden. Zwei aktuelle Übersichtsarbeiten ergänzen dieses Bild um eine methodische und eine dimensionsübergreifende Perspektive. Khan et al. [35] ordnen SHAP und LIME als die prominentesten modellagnostischen Erklärungsverfahren im Finanzbereich ein, benennen instabile lokale Erklärungen und eine unzureichende Berücksichtigung von Fairness jedoch als wiederkehrende Schwächen. Bahlool et al. [6] betrachten für das Kreditscoring Vorhersageleistung, Fairness und Erklärbarkeit gemeinsam und stellen fest, dass diese drei Dimensionen in der Literatur überwiegend getrennt behandelt werden.

Aus den benannten offenen Punkten ergeben sich die drei Forschungsfragen dieser Arbeit. F1 greift die in Abschnitt 3.1 identifizierte Frage auf, wie stabil die globalen Rangfolgen von SHAP und gruppierter PFI bei wiederholtem Training desselben Modells auf demselben Datensatz sind. F2 untersucht den in Abschnitt 3.2 hergeleiteten kontrollierten Redundanzfall zweier deterministisch verbundener geschlechtsbezogener Repräsentationen. Für die übergreifende Einordnung beider Befunde sind die in Abschnitt 3.1 herausgearbeiteten Unterschiede zwischen den Erklärungsobjekten der Verfahren maßgeblich; ihre Ausgaben sind keine austauschbaren globalen Wichtigkeitsmaße.

F3 betrifft die regulatorische Einordnung der Verfahren. In der berücksichtigten Literatur werden die technischen Möglichkeiten und Grenzen von SHAP, LIME und PFI nicht gemeinsam auf die Anforderungen des EUAIA an Transparenz, menschliche Aufsicht und fallspezifische Erläuterung nach den Artikeln 13, 14 und 86 bezogen (vgl. 4.5). Der EUAIA bildet dabei keine zusätzliche empirische Untersuchungsebene, sondern den Referenzrahmen, anhand dessen die Eigenschaften und Ergebnisse der drei Verfahren eingeordnet werden.

Die empirischen Fragestellungen F1 und F2 werden anhand eines XGBoost-Modells auf dem von Grömping [24] fehlerbereinigten *South German Credit*-Datensatz untersucht (vgl. 4.1). Dieser wurde in der berücksichtigten Literatur bislang nicht in einem Untersuchungsaufbau ausgewertet, der die Stabilität globaler Rankings und den deterministischen Redundanzfall geschlechtsbezogener Repräsentationen zur Analyse potenzieller gruppenbezogener Bias-Indikatoren zusammenführt. Ergänzend berichtet die Arbeit die outcome-basierten Fairnessmetriken des Modells (vgl. 4.4). Sie liefern eine von den Attributionen unabhängige Ergebnisperspektive auf dasselbe Modell und dienen als Bezugspunkt für die Einordnung der XAI-Befunde, sind jedoch keiner der drei Forschungsfragen unmittelbar zugeordnet.

4 METHODIK

Dieses Kapitel beschreibt die empirische Vorgehensweise zur Beantwortung der in Abschnitt 1.3 formulierten Forschungsfragen: Datensatz, Klassifikationsmodell und XAI-Pipeline, die Operationalisierung der Bewertungskriterien und der regulatorischen Einordnung sowie die Reproduzierbarkeit der Analyse.

4.1 DATENSATZ

Als empirische Grundlage dient der *South German Credit*-Datensatz [23, 24], eine fehlerbereinigte Fassung des klassischen *Statlog German Credit*-Datensatzes [29], dessen verbreitete UCI-Fassung mehrere Kodierungsfehler enthält. Der zugrunde liegende *German Credit*-Datensatz ist ein häufig verwendeter Referenzdatensatz [40], dessen korrigierte Fassung auch Chen et al. [12] einsetzen. Der Datensatz umfasst 1 000 tatsächlich vergebene Kredite einer deutschen Bank mit 20 finanziellen, beruflichen und soziodemografischen Merkmalen; Variablen und Ausprägungen tragen durchgängig die Bezeichner des Codebooks. Die binäre Zielvariable `credit_risk` bildet den späteren Kreditverlauf ab und unterscheidet zwischen vertragsgemäß erfüllten Krediten (*good*, 70 %) und nicht vertragsgemäß erfüllten Krediten (*bad*, 30 %). Abgelehnte Anträge sind nicht enthalten, der Kreditverlauf ist also nur für die damals positiv beschiedenen Anträge beobachtbar. Die Analyse bezieht sich daher durchgehend auf die finanzierte Stichprobe. Die Konsequenzen für die Fairnessmetriken werden in Abschnitt 6.5 aufgegriffen. Der Datensatz ist zudem eine geschichtete Stichprobe. Die Fälle wurden getrennt nach Kreditverlauf gezogen, wobei Fälle der Klasse *bad* häufiger aufgenommen wurden [24].

Für die Fairnessanalyse sind zwei sensible Merkmale zentral: die binäre Geschlechtsvariable `sex_binary`, abgeleitet aus dem Attribut `personal_status_sex`, und das Attribut `foreign_worker`. Beide Merkmale bleiben bewusst Modelleingaben, um ihre Nutzung durch das Kreditrisikomodell und ihre Sichtbarkeit in den XAI-Erklärungen untersuchen zu können. Da `sex_binary` zusätzlich zum unverändert beibehaltenen Attribut `personal_status_sex` in das Modell eingeht, umfasst der Merkmalsraum insgesamt 21 Modelleingangsvariablen. Dies ist eine methodische Entscheidung für die Bias-Analyse und soll keine Empfehlung für den produktiven Einsatz darstellen.

Das Attribut `personal_status_sex` vermengt Geschlecht und Familienstand. Eine seiner vier Ausprägungen fasst nicht alleinstehende Frauen und alleinstehende Männer zusammen (*female: non-single or male: single*) [24]. Für die betroffenen 310 Fälle (31 %) ist das Geschlecht daher individuell nicht

eindeutig zuordenbar. Da das Codebook keine Zuordnung als korrekt auszeichnet, werden zwei pauschale Zuordnungsszenarien betrachtet, die den Bereich der möglichen Zuordnungen einrahmen, ihn aber nicht formal begrenzen. Szenario A ordnet diese Kategorie vollständig der weiblich kodierten Gruppe zu, Szenario B vollständig der männlich kodierten Gruppe. Keines der beiden wird als wahre Geschlechtszuordnung interpretiert. Ein Ausschluss der betroffenen Fälle wurde verworfen, da dadurch fast ein Drittel des kleinen Datensatzes entfiele.

Die geschlechtsbezogenen Fairnessergebnisse werden für beide Szenarien berichtet (vgl. 5.8, Anhang A.6) und sind unter der in Abschnitt 6.5 diskutierten Einschränkung zu interpretieren. `sex_binary` und `personal_status_sex` bilden gemeinsam den in Abschnitt 4.4 operationalisierten kontrollierten deterministischen Redundanzfall für Forschungsfrage F2, dessen Robustheit gegenüber beiden Szenarien in Anhang A.6 geprüft wird.

Das Attribut `foreign_worker` gibt binär an, ob eine kreditnehmende Person als ausländischer Arbeitnehmer geführt wird. Es darf jedoch nicht mit Staatsangehörigkeit, ethnischer Herkunft oder Migrationserfahrung gleichgesetzt werden. In der korrigierten Datensatzfassung stehen 37 Fälle mit der Ausprägung *yes* 963 Fällen mit *no* gegenüber. Die Variable dient als sekundäre sensible Achse, um das Verhalten der outcome-basierten Fairnesskennzahlen bei stark ungleich großen Subgruppen zu untersuchen. Im nach der Zielvariable stratifizierten 80/20-Split (vgl. 4.2) verbleiben nur fünf als ausländische Arbeitnehmer kodierte Personen in der Testmenge. Diese Fallzahl begrenzt die Aussagekraft der Fairnesskennzahlen erheblich; die Bootstrap-Konfidenzintervalle werden daher nur ergänzend berichtet und erlauben keine belastbare Inferenz für diese Subgruppe (vgl. 4.4).

4.2 KLASSIFIKATIONSMODELL

Als Klassifikationsmodell wird XGBoost [11] eingesetzt. Das baumbasierte Gradient-Boosting-Verfahren eignet sich für strukturierte tabellarische Daten [8] und wird im Kreditscoring regelmäßig eingesetzt [9, 12]. Seine Modellstruktur ermöglicht zudem die effiziente und modellspezifische Berechnung von SHAP-Werten mittels TreeSHAP [42]. Die Arbeit untersucht dabei nicht, ob XGBoost gegenüber intrinsisch interpretierbaren Modellen die geeignetste Modellklasse für diesen Datensatz darstellt. Vielmehr dient das Modell als kontrollierter Untersuchungsgegenstand für den Vergleich der Post-hoc-Erklärungsverfahren und die Analyse potenzieller Bias-Indikatoren.

Der Datensatz wird mit Seed 42 stratifiziert im konventionellen Verhältnis 80/20 in Trainings- und Testdaten aufgeteilt. Diese Partition wird für sämtliche Analysen beibehalten, damit Unterschiede zwischen den XAI-Verfahren nicht mit Unterschieden zwischen Datenaufteilungen vermengt werden [34]. Die Bindung sämtlicher Befunde an diese eine Aufteilung wird in Abschnitt 6.5 als Limitation eingeordnet. Standardisierung und One-Hot-Kodierung werden ausschließlich anhand der Trainingsdaten angepasst und anschließend unverändert auf die Testdaten übertragen. Die Standardisie-

rung der numerischen Merkmale dient der einheitlichen Vorverarbeitung und ist für das Modell ohne Wirkung, da eine monotone Skalierung die Aufteilungen eines baumbasierten Verfahrens nicht verändert [44]. Aus den acht numerischen Merkmalen, von denen vier geordnete Klassencodes und keine Messwerte sind, und den One-Hot-kodierten kategorialen Merkmalen resultieren die 62 Modellfeatures, deren vollständige Zuordnung zu den Ursprungsvariablen dem Begleitmaterial beiliegt.

Die dem ursprünglichen Datensatz beigefügte Kostenmatrix bewertet die fälschliche Einstufung eines Falls der Klasse *bad* als *good* fünfmal so hoch wie den umgekehrten Fehler [29]. Diese historische Kostenannahme wird nicht als Trainingsgewicht von 5:1 übernommen, da sie weder als aktuelles empirisches Kostenmodell noch als geeignete Grundlage für den methodischen Vergleich interpretiert wird. Stattdessen werden aus den Klassenhäufigkeiten der Trainingsmenge inverse Klassengewichte berechnet, um beide Zielklassen trotz ihrer unterschiedlichen Häufigkeiten gleichgewichtig in der Modellanpassung zu berücksichtigen. Für die Minderheitsklasse *bad* ergibt sich gegenüber der Klasse *good* ein Gewichtungsverhältnis von rund $2{,}33 : 1$. Die Herleitung ist in Anhang A.1 dokumentiert. Die Gewichtung gleicht ausschließlich die Klassenhäufigkeiten im Training aus und stellt keine Korrektur auf die historische Klassenprävalenz dar.

Da nicht die Vorhersageleistung, sondern der kontrollierte Vergleich der XAI-Verfahren im Mittelpunkt steht, erfolgt keine Hyperparameteroptimierung. Die vorab festgelegten Parameter wie Baumtiefe, Lernrate, Zeilen- und Spalten-Subsampling sind in Anhang A.2, Tabelle A.1, dokumentiert. Die Testdaten werden weder zur Auswahl der Hyperparameter noch zur Festlegung der Klassifikationsschwelle verwendet. Für alle schwellenwertabhängigen Auswertungen wird die Klassifikationsschwelle einheitlich auf 0,5 festgelegt. Aufgrund der Klassengewichtung und der fehlenden Kalibrierungsanalyse ist sie weder als optimaler noch als kalibrierter Entscheidungspunkt zu interpretieren. Die Abhängigkeit des Disparate Impact von dieser Festlegung wird durch eine Variation des Betriebspunkts geprüft (vgl. 4.4).

Da 70 % der Fälle zur Klasse *good* gehören, entspricht bereits deren konstante Vorhersage einer Accuracy von 0,70. Die Modellgüte wird daher anhand mehrerer komplementärer Metriken bewertet: für die als Klasse 1 kodierte Klasse *good* Precision, Recall und F_1-Score, ergänzend der Recall der Klasse *bad*, also der Anteil korrekt erkannter *bad*-Fälle. Die Area Under the Curve (AUC) unter der Receiver Operating Characteristic (ROC)-Kurve misst darüber hinaus, wie gut das Modell beide Klassen über unterschiedliche Klassifikationsschwellen hinweg trennt [40]. Eine fünffache stratifizierte Kreuzvalidierung auf den Trainingsdaten ergänzt die Bewertung auf der festen Testmenge um Ergebnisse über mehrere Datenpartitionen. Standardisierung und One-Hot-Kodierung werden dabei in jeder Iteration ausschließlich auf dem Trainingsteil angepasst und nur auf den zugehörigen Validierungsteil angewendet, sodass kein Informationsabfluss entsteht. Die Kreuzvalidierung dient ausschließlich der Einordnung der Modellleistung und nicht der Auswahl oder Optimierung der Hyperparameter.

4.3 XAI-PIPELINE

Untersucht werden mit SHAP, LIME und PFI drei Post-hoc-Verfahren, die unterschiedliche Erklärungsansätze vertreten (vgl. 2.2). Die Auswahl folgt der Verbreitung dieser Verfahren in der Kreditrisikoliteratur (vgl. Kapitel 3). Die drei XAI-Methoden werden auf dasselbe trainierte XGBoost-Modell und dieselbe Testmenge angewendet. Da die Wichtigkeitswerte von SHAP und PFI auf unterschiedlichen Skalen beruhen, werden ausschließlich die resultierenden Variablen-Rangfolgen und nicht die absoluten Werte miteinander verglichen. Ihre Übereinstimmung wird als Spearman-Rangkorrelation zwischen beiden Rankings über die 21 Variablen des Referenzlaufs berichtet. Sie vergleicht zwei Verfahren auf einem Modell und nicht, wie die Stabilitätskennzahl in Abschnitt 4.4, ein Verfahren über mehrere Modelle. Die quantitativen globalen Auswertungen erfolgen durchgängig auf der Ebene der 21 Ursprungsvariablen und nicht der 62 kodierten Spalten.

Für SHAP wird der `TreeExplainer` der `shap`-Bibliothek mit der Einstellung `tree_path_dependent` verwendet. Dieser berechnet die Attributionen für das Baummodell effizient und deterministisch [42]. Die pfadabhängige Variante stützt sich auf die im Modell gespeicherten Häufigkeiten der Trainingsdaten an den Baumknoten und benötigt keine externen Hintergrunddaten, deren Auswahl die Attributionen zusätzlich beeinflussen würde. Bei korrelierten oder deterministisch verbundenen Variablen verteilt sie die Attribution jedoch entlang der tatsächlich verwendeten Baumpfade, während eine interventionelle Berechnung die Abhängigkeit zwischen berücksichtigten und entfernten Merkmalen unter Verwendung einer Hintergrundverteilung aufbricht. Je nach Variante entfallen auf redundante Repräsentationen daher unterschiedliche Anteile. Für die in Forschungsfrage F2 betrachteten Repräsentationen ist diese Wahl damit eine Randbedingung (vgl. 4.4). Für jede Testinstanz wird ein Attributionsvektor relativ zum erwarteten Modelloutput, dem Basiswert, bestimmt. Als Modelloutput dient der Rohwert der Klasse *good*, bei binärer Klassifikation also die Log-Odds. Auf dieser Skala addieren sich Basiswert und Attributionen gemäß dem Efficiency-Axiom exakt zum Modelloutput; alle SHAP-Auswertungen beziehen sich daher auf die Log-Odds-Skala.

Das globale SHAP-Ranking wird anhand des Mittelwerts der absoluten Attributionswerte je Feature gebildet [41]. Für den Methodenvergleich werden diese mittleren absoluten Werte über die One-Hot-Spalten jeder Ursprungsvariable aufsummiert. Diese Aggregation bildet den Betrag vor der Summe und kann dadurch Variablen mit vielen One-Hot-Spalten begünstigen. Die Alternative, je Instanz zuerst vorzeichenbehaftet zu summieren und erst danach den Betrag zu bilden, fällt nie größer aus; ihre vorzeichenbehaftete Zwischensumme erhält die lokale Additivität auf Variablenebene (Herleitung in Anhang A.3). Die berichteten SHAP-Ränge sind daher an die pfadabhängige Berechnung wie an diese Aggregationsentscheidung gebunden. Wie stark diese Bindung ausfällt, prüft eine Gegenrechnung des globalen Rankings unter der alternativen Aggregation (Anhang A.3). Ergänzend wer-

den für drei mit festem Seed gezogene Testinstanzen Wasserfall-Diagramme erstellt. Diese zeigen abweichend von den globalen Auswertungen die transformierten Modellfeatures, also die einzelnen One-Hot-Spalten, und dienen ausschließlich der Veranschaulichung (vgl. 5.7).

LIME übernimmt im Untersuchungsaufbau eine andere Rolle als SHAP und PFI. Da es kein etabliertes globales Wichtigkeitsmaß bereitstellt, nimmt es weder an der Stabilitätsanalyse (F1) noch an der Redundanz- und Ablationsanalyse (F2) teil und dient ausschließlich als zweites, methodisch andersartiges Verfahren für fallspezifische Erklärungen (F3). Die Gegenüberstellung von LIME und den lokalen SHAP-Attributionen an denselben Instanzen macht sichtbar, ob und worin sich die Erläuterung einer einzelnen Entscheidung zwischen den Verfahren unterscheidet. Dieser einzelfallbezogene Vergleich ist für die Einordnung zu Artikel 86 EUAIA relevant, der unter den dort genannten Voraussetzungen ein Recht betroffener Personen auf eine klare und aussagekräftige Erläuterung der betreffenden Entscheidung vorsieht (vgl. 2.4 und 4.5). Herangezogen wird allein der in Abschnitt 4.5 definierte technische Indikator, ob ein Verfahren eine fallspezifische Erklärung überhaupt bereitstellt, und darüber hinaus die Beobachtung, welchen Inhalt diese Erklärung hat und wie stark er zwischen den Verfahren variiert.

LIME wird mithilfe der Klasse `LimeTabularExplainer` [48] angewendet. Als Eingaberaum erhält es die 21 Ursprungsvariablen, sodass die Perturbation zur Erzeugung der lokalen Nachbarschaft auf dieser Ebene und nicht auf der der 62 kodierten Spalten stattfindet. Über den Parameter `categorical_features` werden die 13 kategorialen Ursprungsvariablen sowie die beiden binären Variablen `sex_binary` und `people_liable` als kategorial gekennzeichnet. Je so gekennzeichneter Variable wird eine vollständige Ausprägung gezogen; andernfalls würden sie wie numerische Merkmale quartilsbasiert diskretisiert (vgl. Anhang A.4). Die übrigen numerischen Merkmale werden anhand ihrer Quartile in Intervalle eingeteilt, sodass sich die ausgegebenen lokalen Koeffizienten auf Intervallbedingungen in den ursprünglichen Maßeinheiten beziehen (z. B. `duration <= 12.00`). Die Transformation in die vom XGBoost-Modell verwendeten standardisierten und One-Hot-kodierten Spalten erfolgt erst innerhalb der Vorhersagefunktion. Die von LIME erzeugten Instanzen der lokalen Nachbarschaft werden dort durch die ausschließlich auf den Trainingsdaten angepasste Standardisierungs- und One-Hot-Pipeline in die 62 Modellfeatures transformiert und anschließend dem unveränderten XGBoost-Modell zur Wahrscheinlichkeitsvorhersage übergeben. Die vollständige Konfiguration ist in Anhang A.4 dokumentiert. Die nicht explizit gesetzten Parameter bleiben auf den Standardwerten der Bibliothek, sodass die Erklärungen der voreingestellten Anwendung des Verfahrens entsprechen.

Für dieselben drei Testinstanzen wird je ein lokal gewichtetes lineares Surrogatmodell über alle 21 Variablen angepasst. Die Gewichte beziehen sich auf die vorhergesagte Wahrscheinlichkeit der Klasse *good*. Ergänzend werden je Instanz zwei Maße der lokalen Treue des Surrogats berichtet. Das Bestimmtheitsmaß der gewichteten Regression in der erzeugten Nachbar-

schaft sowie die Abweichung zwischen der Surrogatvorhersage und der Wahrscheinlichkeit des XGBoost-Modells an derselben Instanz. Ohne diese Angaben bleibt offen, wie gut das jeweilige Surrogat das Modell in der Umgebung der erklärten Instanz überhaupt approximiert (vgl. 2.2.2 und 5.7). Die berichteten lokalen Gewichte sind Beobachtungen an drei Einzelinstanzen und keine über den Datensatz verallgemeinerbaren Kennzahlen.

PFI wird als gruppierte Permutation auf Variablenebene berechnet. Alle One-Hot-Spalten einer Ursprungsvariable werden gemeinsam mit derselben Zeilenpermutation permutiert, sodass gültige One-Hot-Kombinationen erhalten bleiben (vgl. 2.2.3). Numerische Merkmale bilden Gruppen der Größe eins. In jedem Lauf werden je Variable zwanzig Permutationswiederholungen verwendet, sowohl im Referenzmodell als auch in allen Läufen mit erneutem Training und bei festem Modell. Als Importance-Wert dient die mittlere Abnahme der AUC auf der permutierten Testmenge. PFI quantifiziert damit die leistungsbezogene Abhängigkeit des trainierten Modells von einer Variable und nicht deren Beitrag im kausalen Sinne.

4.4 BEWERTUNGSKRITERIEN

Der Vergleich erfolgt anhand zweier methodischer Kernkriterien, der Ranking-Stabilität (F1) und der Redundanz- und Ablationsanalyse (F2). Ergänzend werden zwei outcome-basierte Fairnessmetriken berichtet.

Die Redundanz- und Ablationsanalyse stützt sich ausschließlich auf die SHAP-Attributionen. LIME bleibt aus den in Abschnitt 4.3 genannten Gründen ausgenommen, und die gruppierte PFI permutiert die beiden deterministisch verbundenen Repräsentationen getrennt voneinander (vgl. 6.5).

RANKING-STABILITÄT (F1). Bei unveränderter Datenaufteilung werden elf XGBoost-Modelle mit unterschiedlichen Zufalls-Seeds trainiert. Die Modelle unterscheiden sich ausschließlich im Seed der Modellanpassung, nicht in Datenaufteilung oder Hyperparametern, sodass Unterschiede zwischen den trainierten Modellen allein auf diesen Seed zurückgehen. Für SHAP und PFI wird jeweils das vollständige Ranking aller 21 Variablen bestimmt und die Stabilität als paarweise Spearman-Rangkorrelation über alle 55 Modellpaare berechnet. Der paarweise Vergleich vermeidet die Bevorzugung eines einzelnen Referenzlaufs. Die Korrelation gewichtet alle Rangpositionen gleich und misst damit die Stabilität des Gesamtrankings. Die rangbasierte Kennzahl ist einem wertbasierten Maß vorzuziehen, da für die Stabilitätsbewertung die Reihenfolge der Variablen und nicht die absolute Höhe der methodenspezifischen Wichtigkeitswerte entscheidend ist [46]. Berichtet werden Mittelwert und Standardabweichung der paarweisen Korrelationswerte. Ergänzend wird dieselbe Kennzahl auf die zehn und fünf höchstplatzierten Variablen des Referenzlaufs beschränkt, um die Stabilität der vorderen Ränge getrennt auszuweisen. Diese Variante ist an den Referenzlauf gebunden und misst nur, ob dessen Reihenfolge erhalten bleibt, nicht ob dieselben Variablen in den übrigen Läufen weiterhin zur Spitzengruppe gehören. Die

Mengenzugehörigkeit wird daher getrennt geprüft. Anstelle mengenbasierter Ähnlichkeitsmaße wie der Jaccard-Ähnlichkeit der Top-k-Mengen [34] wird bei elf Läufen und den Werten $k = 5$ und $k = 10$ unmittelbar geprüft, ob die Top-k-Menge über alle Läufe identisch bleibt (vgl. 5.5).

Die beiden Stabilitätswerte erfassen dabei nicht dasselbe. Der `TreeExplainer` ist bei festem Modell deterministisch, sodass die SHAP-Variabilität auf die unterschiedlichen Trainingsläufe zurückgeht. Die beobachtete PFI-Variabilität enthält darüber hinaus eine permutationsbedingte Schätzkomponente. Um diese klein zu halten, verwenden alle Trainingsläufe denselben Permutations-Seed und damit identische Zeilenpermutationen, sodass die paarweisen Unterschiede überwiegend aus den Modellen und nicht aus den Ziehungen stammen. Die nachfolgend beschriebene Robustheitsprüfung bei festem Modell weist die Empfindlichkeit der PFI-Rankings gegenüber der Permutationsziehung getrennt aus; eine Zerlegung der in F1 beobachteten Variabilität in unabhängige Komponenten leistet sie nicht.

RANGPOSITIONEN SENSIBLER VARIABLEN. Ergänzend werden die Rangpositionen der drei sensiblen Modellvariablen `sex_binary`, `personal_status_sex` und `foreign_worker` (vgl. 4.1) in beiden globalen Rankings berichtet. Sie geben allein an, wie prominent ein sensibles Merkmal in der jeweiligen globalen Wichtigkeitsdarstellung erscheint, und treffen keine Aussage über gruppenbezogene Fairness (vgl. 2.3). Verfahren, die eine gemessene Disparität auf einzelne Variablen zurückführen [32] oder entlang kausaler Pfade zerlegen [57], beantworten die davon verschiedene Frage, welche Variablen zur Disparität beitragen, und werden hier nicht eingesetzt.

REDUNDANZ- UND ABLATIONSANALYSE GESCHLECHTSBEZOGENER REPRÄSENTATIONEN (F2). Forschungsfrage F2 wird über die SHAP-Ränge der beiden redundanten Repräsentationen `sex_binary` und `personal_status_sex` operationalisiert. Beide sind deterministisch verbunden und verbleiben im Hauptmodell bewusst parallel (vgl. 4.1). Die Redundanz betrifft dabei ausschließlich die enthaltene Geschlechtsinformation; informationsäquivalent sind die beiden Variablen nicht. Untersucht wird damit nicht eine Umverteilung der SHAP-Attribution innerhalb eines festen Modells, sondern die Rangänderung zwischen zwei getrennt angepassten Modellen.

Verglichen werden zwei Ablationsmodelle: eines ohne `sex_binary` und eines ohne `personal_status_sex`. Beide werden mit unveränderter Vorverarbeitungs-, Trainings- und Hyperparameterkonfiguration auf demselben Trainingssplit neu angepasst; verändert wird ausschließlich der Eingabemerkmalsraum. Betrachtet wird der SHAP-Rang der jeweils verbleibenden Repräsentation im vollständigen Modell und im Ablationsmodell, auf Basis der rohen Variablenränge. Die Bezugsgröße sinkt dabei von 21 auf 20 Variablen. Eine Variable, die im vollständigen Modell hinter der entfernten liegt, gewinnt daher allein durch deren Entfernung eine Rangpo-

sition, ohne dass sich ihre Attribution geändert hätte. Dieser mechanische Anteil entspricht der Zahl der entfernten Variablen, die vor der betrachteten platziert waren, bei je einer entfernten Variable also einer oder keiner Position. Er wird von der beobachteten Rangänderung abgezogen; nur der verbleibende Anteil wird als Befund gewertet. Die beiden Richtungen sind dabei nicht symmetrisch. `sex_binary` ist eine deterministische Funktion von `personal_status_sex`, die Ursprungsvariable trägt jedoch zusätzlich Familienstandsinformation. Ihre Entfernung nimmt dem Modell daher mehr Information als die Entfernung der abgeleiteten Variable, sodass die beiden Rangänderungen nicht als spiegelbildliche Gegenstücke zu lesen sind. Änderungen der Test-AUC werden zurückhaltend berichtet und Differenzen auf einer einzelnen Datenaufteilung nicht als Leistungsgewinn interpretiert.

ROBUSTHEITSPRÜFUNGEN. Drei zusätzliche Kontrollanalysen prüfen die Belastbarkeit zentraler Ergebnisse. Erstens wird die gruppierte PFI bei festem Modell unter zehn vorab festgelegten Permutations-Seeds mit je zwanzig Wiederholungen wiederholt. Die Anzahl entspricht der Zahl der zusätzlichen Trainingsläufe in F1. Die Übereinstimmung wird wie in F1 als paarweise Spearman-Rangkorrelation über alle 45 Paare gebildet, sodass beide Kennzahlen denselben Schätzer verwenden. Zweitens wird der Disparate Impact für Klassifikationsschwellen zwischen 0,40 und 0,60 in Schritten von 0,05 berechnet, also symmetrisch um die festgelegte Schwelle (vgl. 4.2), um die Abhängigkeit des Ergebnisses vom gewählten Betriebspunkt sichtbar zu machen. Drittens werden die geschlechtsbezogenen Fairness- und Ablationsanalysen unter beiden Szenarien A und B durchgeführt, um zu prüfen, ob die qualitativen Schlussfolgerungen von der Zuordnung der mehrdeutigen `personal_status_sex`-Kategorie abhängen (vgl. 4.1, Anhang A.6).

OUTCOME-BASIERTE FAIRNESSMETRIKEN. Als von den XAI-Erklärungen unabhängige Modelldiagnostik werden für beide sensiblen Achsen der Equal-Opportunity-Gap [26] und der Disparate Impact [19] berechnet. Die Auswahl deckt mit dem Equal-Opportunity-Gap ein fehlerratenbasiertes und mit dem Disparate Impact ein klassifikationsratenbasiertes Kriterium ab (vgl. 2.3). Kalibrierungsbasierte Kriterien werden nicht berücksichtigt, da keine Kalibrierungsanalyse der Modellwahrscheinlichkeiten durchgeführt wird (vgl. 4.2).

Als günstiges Ergebnis gilt die Einstufung in die Klasse *good*, also die als Klasse 1 kodierte positive Klasse. Bezugsgruppe des Disparate Impact ist die männlich kodierte beziehungsweise die nicht als ausländische Arbeitnehmer geführte Gruppe. Werte unter eins bezeichnen somit eine niedrigere positive Klassifikationsrate der weiblich kodierten beziehungsweise der als ausländische Arbeitnehmer geführten Gruppe. Der Equal-Opportunity-Gap wird als Betrag $|\Delta\mathrm{TPR}|$ berichtet, da die Richtung der geschlechtsbezogenen Differenz zwischen den beiden Szenarien wechselt.

Die Schätzunsicherheit der gruppenspezifischen True-Positive-Raten und ihrer Differenzen wird mithilfe eines nichtparametrischen Perzentil-

Bootstrap-Verfahrens mit 1 000 Wiederholungen und einem Konfidenzniveau von 95 % quantifiziert [15]. Die Intervalle beziehen sich auf die vorzeichenbehafteten TPR-Differenzen und nicht auf deren Betrag, damit Richtung und Unsicherheit erhalten bleiben. Sie quantifizieren allein die Resampling-Variabilität innerhalb der beobachteten Teststichprobe und können bei extrem kleinen oder homogenen Subgruppen degenerieren (vgl. Tabelle A.6). Die Wahl des Perzentilverfahrens, die je Metrik unterschiedliche Ziehungslogik und der Umgang mit undefinierten Replikaten sind in Anhang A.5 dokumentiert.

4.5 REGULATORISCHE EINORDNUNGSMETHODIK

Die regulatorische Einordnung orientiert sich an den in Abschnitt 2.4 eingeführten Artikeln 13, 14 und 86 des EUAIA. Sie stellt keine formale Konformitätsbewertung dar, sondern untersucht qualitativ, in welchem Umfang die technischen Eigenschaften der drei XAI-Methoden die jeweiligen Anforderungen unterstützen können.

Aus den Anforderungen in Abschnitt 2.4 und den technischen Eigenschaften der Verfahren aus Abschnitt 4.3 werden je Artikel Einordnungsindikatoren abgeleitet. Sie sind eine Operationalisierung dieser Arbeit und keine im Verordnungstext genannten Anforderungen. Für Artikel 13 sind dies die Verfügbarkeit globaler und lokaler Erklärungen sowie die Dokumentierbarkeit der Ergebnisse, für Artikel 14 die Visualisierbarkeit der Erklärungen und ihre prinzipielle Integrierbarkeit in einen menschlichen Prüfprozess, für Artikel 86 die Bereitstellung fallspezifischer Erklärungen. Die globale Ranking-Stabilität wird ergänzend als methodisches Qualitätskriterium berücksichtigt, das im EUAIA nicht verankert ist und die Konsistenz einer individuellen Erläuterung nicht misst. Das resultierende Mapping wird in Abschnitt 6.3 mit den empirischen Ergebnissen verknüpft und seine Grenzen in Abschnitt 6.5 diskutiert.

4.6 REPRODUZIERBARKEIT

Zur Reproduzierbarkeit werden die stochastischen Komponenten des Hauptlaufs durch den Zufalls-Seed 42 kontrolliert. Dies betrifft die stratifizierte Aufteilung in Trainings- und Testdaten, die stochastischen Komponenten der Modellanpassung, die Erzeugung der LIME-Nachbarschaften, die gruppierten PFI-Permutationen und das Bootstrap-Verfahren. Die Stabilitäts- und Robustheitsprüfungen werden mit vorab festgelegten Seeds durchgeführt. Anzahl und jeweiliger Einsatzzweck der Wiederholungen sind in Abschnitt 4.4 beschrieben. Die konkreten Seed-Werte sind im Analysecode dokumentiert.

Die Implementierung erfolgt in Python 3.9 mit `scikit-learn` 1.6.1, XGBoost 2.1.4, `shap` 0.49.1, `lime` 0.2.0.1, NumPy 2.0.2, pandas 2.3.3 und SciPy 1.13.1. Der vollständige Analysecode und die verwendete korrigierte Datensatzfassung liegen der Arbeit bei.

5 ERGEBNISSE

Dieses Kapitel berichtet die Ergebnisse der in Kapitel 4 beschriebenen Untersuchung: zunächst die Modell-Performance und die beiden globalen Variablen-Rankings mit ihrem Vergleich, anschließend entlang der Forschungsfragen die Ranking-Stabilität (F1), die Redundanz- und Ablationsanalyse (F2) und die lokalen Erklärungsbeispiele (F3), abschließend die ergänzend erhobenen outcome-basierten Fairnessmetriken.

5.1 MODELL-PERFORMANCE

Das auf 800 Instanzen trainierte XGBoost-Modell erreicht auf der Testmenge von 200 Instanzen eine AUC von 0,820. Die Accuracy liegt mit 0,755 nur knapp über den 0,70 der konstanten Vorhersage der Klasse *good* (vgl. 4.2) und ist wegen des Klassenungleichgewichts nur eingeschränkt interpretierbar (vgl. 6.5). Die klassenbezogenen Kennzahlen sind hier aussagekräftiger. Diese fallen laut Tabelle 5.1 ungleich aus. Die Klasse *good* wird mit hoher Precision und hohem Recall erkannt, während knapp ein Drittel der tatsächlichen *bad*-Fälle als *good* eingestuft wird. Der nach der historischen Kostenmatrix teure Fehlerfall bleibt damit trotz Klassengewichtung häufig (vgl. 4.2). In der Kreuzvalidierung fällt der Recall der Klasse *bad* mit $0{,}671 \pm 0{,}023$ ähnlich aus wie der Testwert von 0,683. Die übrigen Maße liegen in der Kreuzvalidierung durchgängig niedriger als auf der Testmenge, bei kleiner Streuung über die Folds. Am deutlichsten fällt der Abstand bei der AUC aus (vgl. Tabelle 5.1 und 6.5).

Tabelle 5.1: Performance von XGBoost auf der Testmenge ($n = 200$) und in der fünffachen stratifizierten Kreuzvalidierung auf den Trainingsdaten ($n = 800$), dort als Mittelwert $\pm$ Standardabweichung über die fünf Folds (vgl. 4.2).

Metrik	Testmenge	Kreuzvalidierung
Accuracy	0,755	$0{,}741 \pm 0{,}039$
ROC-AUC	0,820	$0{,}774 \pm 0{,}024$
Precision (*good*)	0,853	$0{,}845 \pm 0{,}015$
Recall (*good*)	0,786	$0{,}771 \pm 0{,}051$
F_1 (*good*)	0,818	$0{,}806 \pm 0{,}034$
Recall (*bad*)	0,683	$0{,}671 \pm 0{,}023$

5.2 GLOBALES SHAP-RANKING

Die über die One-Hot-Spalten jeder Ursprungsvariable aufsummierten mittleren absoluten SHAP-Beiträge der 200 Testinstanzen ergeben das in Tabelle A.7 in Anhang A.7 vollständig berichtete globale Ranking über alle 21

Variablen, dessen Werte auf der Log-Odds-Skala stehen (vgl. 4.3). Die höchsten Werte entfallen auf `status`, `purpose` und `duration`, wobei `status` mit einem mittleren absoluten Beitrag von 0,806 gegenüber 0,445 für `purpose` deutlich dominiert. Der zweite Rang von `purpose` ist dabei zurückhaltend zu lesen. Die Variable besitzt mit zehn Ausprägungen die meisten One-Hot-Spalten, und die Aggregation bildet den Betrag vor der Summe, was Variablen mit vielen Spalten begünstigen kann (vgl. 4.3). Die gruppierte PFI führt dieselbe Variable auf Rang 5 (vgl. 5.3). Da beide Verfahren unterschiedliche Wichtigkeitskonzepte messen, bestätigt diese Abweichung den Aggregationseffekt jedoch nicht (vgl. 6.1). Unter den sensiblen Variablen erscheint `personal_status_sex` auf Rang 10 von 21, während `foreign_worker` auf Rang 19 und die abgeleitete Variable `sex_binary` mit einem mittleren absoluten SHAP-Wert von 0,004 auf dem letzten Rang 21 liegt. Tabelle 5.2 stellt diese Positionen denen der gruppierten PFI gegenüber.

5.3 PERMUTATION FEATURE IMPORTANCE

Die gruppierte PFI wird nach Abschnitt 4.3 mit zwanzig Permutationswiederholungen je Variable und ROC-AUC als Bewertungsmaß berechnet. Tabelle A.7 in Anhang A.7 berichtet beide Rankings vollständig, Abbildung A.1 zeigt ihre jeweils zehn höchstplatzierten Variablen. Auf `status` folgen `duration`, `amount`, `credit_history` und `purpose`. Der Abstand an der Spitze ist groß. Die Permutation dieser Variable verschlechtert die AUC im Mittel um 0,090, die von `duration` um 0,034 und die des fünftplatzierten `purpose` um 0,016. Alle drei sensiblen Variablen werden dagegen niedrig gerankt. `foreign_worker` auf Rang 16, `sex_binary` auf Rang 18 und `personal_status_sex` auf dem letzten Rang 21 von 21. Die Werte der hinteren Ränge liegen nahe null oder leicht darunter; für `sex_binary` und `personal_status_sex` sind die geschätzten Werte geringfügig negativ, die mittlere AUC der permutierten Testmenge liegt dort also minimal über dem Ausgangswert (vgl. Tabelle A.7).

5.4 VERGLEICH DER GLOBALEN RANKINGS

Beide Verfahren führen `status` mit deutlichem Abstand auf Rang 1 und teilen acht der jeweils zehn höchstplatzierten Variablen. Unterschiedlich besetzt sind allein die jeweils zwei nicht geteilten Positionen. SHAP weist dort `housing` und `personal_status_sex` aus, die gruppierte PFI dagegen `other_debtors` und `property`. Diese Besetzung bezieht sich auf den Referenzlauf; welche Variablen die Spitzengruppe bilden, ist über die Trainingsläufe hinweg nicht konstant (vgl. 5.5). Abbildung A.1 in Anhang A.7 stellt die jeweils zehn höchstplatzierten Variablen beider Rankings gegenüber. Die Spearman-Rangkorrelation beider Rankings auf Variablenebene beträgt im Referenzlauf $\rho_{\text{SHAP,PFI}} = 0{,}72$ und zeigt eine positive, aber nicht vollständige Übereinstimmung der Rangfolgen (vgl. 4.3; zur Einordnung vgl. 6.1).

Tabelle 5.2 berichtet die Rangpositionen der sensiblen Variablen in beiden globalen Rankings. `sex_binary` und `foreign_worker` führen beide Verfahren übereinstimmend im hinteren Drittel. Bei `personal_status_sex` weichen sie dagegen deutlich voneinander ab. SHAP platziert die Variable im Mittelfeld, die gruppierte PFI auf dem letzten Rang (zur eingeschränkten Interpretierbarkeit dieser PFI-Ränge vgl. 6.5).

Tabelle 5.2: Rangpositionen der sensiblen Variablen in den beiden globalen Rankings unter $n = 21$ Variablen (Rang 1 = höchste Wichtigkeit). LIME ist mangels globalen Wichtigkeitsmaßes nicht aufgeführt.

Variable	SHAP-Rang	PFI-Rang
`sex_binary`	21	18
`personal_status_sex`	10	21
`foreign_worker`	19	16

5.5 RANKING-STABILITÄT (F1)

Über die 55 Paare der elf Trainingsläufe beträgt die mittlere paarweise Spearman-Rangkorrelation 0,989 für SHAP und 0,941 für die gruppierte PFI. Tabelle 5.3 berichtet diese Werte samt Streuung sowie die auf die vorderen Ränge beschränkten Varianten, Abbildung A.2 in Anhang A.7 die zugehörigen Verteilungen. Beide Verfahren erzeugen auf Variablenebene damit stabile Rankings; über das Gesamtranking fällt die mittlere Korrelation für SHAP um 0,048 höher aus als für die gruppierte PFI. Beschränkt auf die vorderen Ränge gilt dieses Verhältnis jedoch nicht durchgängig. Über die zehn höchstplatzierten Variablen des Referenzlaufs bleibt SHAP vor der gruppierten PFI, über die fünf höchstplatzierten kehrt es sich um. Die fünf im Referenzlauf höchstplatzierten PFI-Variablen behalten in allen elf Trainingsläufen ihre relative Reihenfolge, während SHAP zwischen benachbarten Spitzenrängen wechselt.

Da die Rangkorrelation nur die Reihenfolge und nicht die Mengenzugehörigkeit prüft (vgl. 4.4), wurde ergänzend die Top-k-Menge selbst verglichen. Bei der gruppierten PFI bilden in allen elf Läufen dieselben fünf Variablen die Spitzengruppe, bei SHAP tauschen in zwei Läufen `amount` und `credit_history` die Zugehörigkeit. Über die zehn höchstplatzierten Variablen ist die Mengenzugehörigkeit bei keinem der beiden Verfahren durchgängig konstant. Die geringere Stabilität des vollständigen PFI-Rankings gegenüber seinen vorderen Rängen geht damit überwiegend auf Rangpositionen außerhalb der Spitzengruppe zurück. Bei festem Modell erreicht die gruppierte PFI über die 45 Paare von zehn Permutations-Seeds eine mittlere paarweise Spearman-Rangkorrelation von $0{,}978 \pm 0{,}009$ (vgl. 4.4).

Tabelle 5.3: Paarweise Spearman-Rangkorrelation der globalen Variablen-Rankings über die 55 Paare der elf Trainingsläufe, jeweils über das vollständige Ranking und beschränkt auf die zehn beziehungsweise fünf höchstplatzierten Variablen des Referenzlaufs.

	alle 21		Top-10		Top-5	
	$\bar{\rho}$	σ_ρ	$\bar{\rho}$	σ_ρ	$\bar{\rho}$	σ_ρ
SHAP	0,989	0,004	0,984	0,010	0,956	0,050
PFI	0,941	0,028	0,964	0,027	1,000	0,000

Auch die Rangpositionen der sensiblen Variablen sind keine Eigenheit des Referenzlaufs. Über die elf Trainingsläufe liegt `personal_status_sex` bei SHAP durchgängig im Mittelfeld (Ränge 10 bis 12) und bei der gruppierten PFI auf einem der drei letzten Ränge. Die SHAP-Ränge von `sex_binary` (20 bis 21) und `foreign_worker` (18 bis 19) bleiben ebenfalls eng begrenzt, während die PFI-Ränge beider Variablen mit 16 bis 20 beziehungsweise 12 bis 18 breiter streuen.

5.6 REDUNDANZ UND ABLATION (F2)

Tabelle 5.4 berichtet die SHAP-Ränge beider Repräsentationen je Modell. Wird eine der beiden Variablen entfernt, verändert sich der Rang der jeweils anderen nur geringfügig. `sex_binary` steigt vom letzten Rang auf Rang 19 von 20, `personal_status_sex` fällt von Rang 10 auf Rang 11. Bei `sex_binary` ist davon gemäß Abschnitt 4.4 eine Position mechanisch, sodass ein zusätzlicher Aufstieg um eine Rangposition verbleibt; bei `personal_status_sex` entfällt dieser Anteil. Auch nach Entfernung der jeweils anderen geschlechtsbezogenen Repräsentation erreicht damit keine der beiden die vorderen SHAP-Ränge; dasselbe gilt unter Szenario B (Anhang A.6). Keine Ablation senkt die Test-AUC. Diese Unterschiede werden gemäß Abschnitt 4.4 nicht als Leistungsgewinn interpretiert. Die Einordnung dieser Rangänderungen erfolgt in Abschnitt 6.2.

Tabelle 5.4: SHAP-Variablenränge der beiden redundanten geschlechtsbezogenen Repräsentationen im Hauptmodell und in den beiden Ablationsmodellen. Rang 1 bezeichnet die höchste Wichtigkeit. Die Bezugsgröße sinkt in den Ablationsmodellen von 21 auf 20 Variablen. Ein Strich kennzeichnet die jeweils entfernte Variable. Die Test-AUC ist zusätzlich ausgewiesen.

Modell	`sex_binary`	`personal_status_sex`	**Test-AUC**
Hauptmodell (21 Variablen)	21 von 21	10 von 21	0,820
ohne `personal_status_sex`	19 von 20	—	0,844
ohne `sex_binary`	—	11 von 20	0,829

5.7 LOKALE ERKLÄRUNGSBEISPIELE IM VERGLEICH (F3)

Anhang B zeigt die lokalen SHAP- und LIME-Erklärungen für drei mit festem Seed gezogene Testinstanzen mit den Indizes 130, 17 und 154 innerhalb der 200 Testinstanzen (vgl. 4.3; Abbildung B.1 für Index 130, Abbildungen B.2 und B.3 für Index 17 sowie Abbildungen B.4 und B.5 für Index 154). Innerhalb der jeweiligen Erklärung sind die Merkmale nach dem Betrag ihrer

Beiträge sortiert; das Vorzeichen gibt die Richtung des jeweiligen Beitrags an. Da beide Verfahren auf unterschiedlichen Skalen berichten (vgl. 2.2.2 und 4.3), bleibt der Vergleich darauf beschränkt, welche Merkmale in der jeweiligen Erklärung prominent erscheinen.

Tabelle B.1 in Anhang B stellt die jeweils drei höchstgewichteten Beiträge beider Verfahren gegenüber. In den unmittelbar ausgegebenen Darstellungen teilen beide Verfahren das jeweils prominenteste Merkmal nur bei Index 130; dahinter weichen die Reihenfolgen durchgängig voneinander ab. Aufgrund der unterschiedlichen Repräsentationsebenen ist dieser unmittelbare Vergleich jedoch nur eingeschränkt interpretierbar. Eine Abweichung wiederholt sich dabei in allen drei Instanzen. Die Bedingung `foreign_worker=2`, also die Zugehörigkeit zur nicht als ausländische Arbeitnehmer geführten Gruppe (vgl. 4.1), gehört jeweils zu den drei höchstgewichteten LIME-Bedingungen, während dieselbe Variable in keinem der drei SHAP-Diagramme (Abbildungen B.1a, B.2 und B.4) unter den elf einzeln ausgewiesenen Features liegt.

Dieser Kontrast könnte allein aus den unterschiedlichen Repräsentationsebenen folgen, da die Wasserfall-Diagramme nur die elf betragsmäßig größten kodierten Spalten einzeln ausweisen, während LIME je Ursprungsvariable genau einen Koeffizienten ausgibt. Er bleibt jedoch bestehen, wenn die lokalen SHAP-Beiträge zum Vergleich je Instanz vorzeichenbehaftet über die One-Hot-Spalten jeder Ursprungsvariable summiert werden, sodass beide Erklärungen 21 Ursprungsvariablen gegenüberstellen. Diese Aggregation entspricht der in Abschnitt 4.3 genannten Alternative, die die lokale Additivität auf Variablenebene erhält. Tabelle B.2 stellt beide Sichtweisen gegenüber. Der aggregierte lokale SHAP-Beitrag von `foreign_worker` bleibt betragsmäßig klein und belegt die Ränge 15 bis 18 von 21, während LIME die Variable auf den Rängen 1 bis 3 führt. Bei diesen drei Instanzen geht die Abweichung also nicht auf die Aggregationsebene oder die abgeschnittene Darstellung zurück. Die vorderen lokalen LIME-Ränge stehen zudem im Kontrast zu den globalen Rankings, die `foreign_worker` jeweils im hinteren Drittel führen (vgl. 5.2 und 5.3). Dieselbe Aggregation zeigt bei Index 17 außerdem eine Abweichung zwischen lokaler und globaler SHAP-Prominenz. `personal_status_sex` erreicht den fünften Rang der lokalen SHAP-Beiträge, während die Variable global auf Rang 10 liegt. Die Einordnung dieser Beobachtungen erfolgt in Abschnitt 6.3.

Die lokale Treue der drei LIME-Surrogate fällt dabei gering aus und schwankt zwischen den Instanzen erheblich (Tabelle B.3). Das Bestimmtheitsmaß der gewichteten Ridge-Regression in der erzeugten Nachbarschaft liegt zwischen 0,135 und 0,460, und die Surrogatvorhersage weicht um 0,043 bis 0,175 von der Modellwahrscheinlichkeit ab. Bei Index 154 liegen Modellwahrscheinlichkeit und Surrogatvorhersage mit 0,486 und 0,610 auf verschiedenen Seiten der Klassifikationsschwelle 0,5. Die Surrogate approximieren das Modell in der Umgebung dieser Instanzen damit nur eingeschränkt und unterschiedlich genau (vgl. 6.3).

5.8 FAIRNESSMETRIKEN

Tabelle 5.5 berichtet die gruppenbasierten Fairnessmetriken für beide sensiblen Achsen.

Tabelle 5.5: Outcome-basierte Fairnessmetriken auf der Testmenge ($n = 200$). Die geschlechtsbezogenen Werte gelten für Szenario A, für Szenario B siehe Tabelle A.3. n ist die Gruppengröße und damit der Nenner der positiven Klassifikationsrate, n_+ die Zahl der Fälle der Gruppe mit tatsächlicher Klasse *good* und damit der Nenner der TPR. Disparate Impact und Equal-Opportunity-Gap (Betrag der TPR-Differenz, aus den ungerundeten Raten berechnet) beziehen sich auf die gesamte sensible Achse und stehen daher nur in der Zeile der erstgenannten Gruppe *female* beziehungsweise *foreign* (vgl. 4.4).

Gruppe	n	n_+	TPR	Pos.-Klass.-Rate	DI	EO-Gap
female	81	49	0,776	0,556	0,787	0,016
male	119	91	0,791	0,706	—	—
foreign	5	3	1,000	0,600	0,929	0,219
domestic	195	137	0,781	0,646	—	—

Auf der geschlechtsbezogenen Achse liegen die Punktschätzungen der True-Positive-Raten unter Szenario A nahe beieinander, während der Disparate Impact von 0,787 die in Abschnitt 2.3 eingeführte Four-Fifths-Schwelle knapp unterschreitet. Die als weiblich kodierte Gruppe wird nur mit rund 79 % der Rate der männlichen Gruppe der Klasse *good* zugeordnet. Unter Szenario B beträgt der Disparate Impact dagegen 1,050 und kehrt damit die Richtung der Ratendifferenz um (vgl. Tabelle A.3 in Anhang A.6). Innerhalb von Szenario A bleibt die Unterschreitung über den gesamten geprüften Schwellenbereich von 0,40 bis 0,60 bestehen (vgl. Tabelle A.5 in Anhang A.7); da für den Disparate Impact kein Konfidenzintervall berichtet wird, ist der geringe Abstand zur Schwelle nicht gegen die Stichprobenvariabilität abgesichert (vgl. 6.5). Das Bootstrap-Intervall der TPR-Differenz reicht von $-0{,}17$ bis 0,14 und schließt null ein (vgl. Tabelle A.6 in Anhang A.7).

Auf der `foreign_worker`-Achse weichen die True-Positive-Raten deutlich voneinander ab, während der Disparate Impact von 0,929 klar oberhalb der Schwelle liegt. Beide Kennzahlen ergeben damit ein uneinheitliches Bild, das angesichts der Subgruppengröße von nur fünf Testfällen nicht belastbar interpretiert werden kann. Die True-Positive-Rate beruht auf drei Fällen mit tatsächlicher Klasse *good*, und ihr Bootstrap-Intervall ist mit $[1{,}00,\ 1{,}00]$ degeneriert (vgl. Tabelle A.6 in Anhang A.7 und 6.5).

6 DISKUSSION

Dieses Kapitel beantwortet die drei Forschungsfragen unter Bezug auf die Ergebnisse aus Kapitel 5: die Ranking-Stabilität (F1), die Redundanz- und Ablationseffekte (F2) und das technische Unterstützungspotenzial (F3), anschließend die ergänzend berichteten outcome-basierten Fairnessmetriken und abschließend die methodischen Grenzen der Untersuchung.

6.1 STABILITÄT DER GLOBALEN RANKINGS (F1)

Die Stabilitätsanalyse (vgl. 5.5) zeigt auf Variablenebene für beide Verfahren sehr stabile globale Rankings, mit einem geringen Vorsprung von SHAP ($\bar{\rho} = 0{,}989$) gegenüber der gruppierten PFI ($\bar{\rho} = 0{,}941$). Naheliegend ist, diesen Vorsprung teilweise darauf zurückzuführen, dass TreeSHAP bei festem Modell deterministisch arbeitet, während die PFI zusätzlich von den gezogenen Permutationen abhängt. Wie groß der Beitrag dieser zusätzlichen Zufallsvariation ist, lässt sich im vorliegenden Untersuchungsaufbau jedoch nicht isolieren, da SHAP und PFI zugleich unterschiedlich empfindlich auf Änderungen der neu trainierten Modelle reagieren können. Bei festem Modell und variierenden Permutations-Seeds erreicht die gruppierte PFI eine mittlere Rangkorrelation von $\bar{\rho} = 0{,}978$. Dieser Wert beschreibt ausschließlich die Empfindlichkeit gegenüber der Permutationsziehung und ist daher nicht unmittelbar mit der über unterschiedliche Trainingsläufe bestimmten SHAP-Stabilität von $\bar{\rho} = 0{,}989$ vergleichbar. Die Stabilitätsmessung über die Trainingsläufe hält umgekehrt die Permutationsziehung fest (vgl. 4.4). Beide Analysen variieren jeweils einen Faktor und erlauben keine Zerlegung der PFI-Variabilität in unabhängige Komponenten. Der Stabilitätsvorsprung von SHAP lässt sich daher nicht allein der Modellsensitivität der gruppierten PFI zuschreiben. Ob die gruppierte Permutation zur hohen PFI-Stabilität beiträgt, kann mangels einer ungruppierten Kontrollbedingung nicht beurteilt werden (vgl. 2.2.3).

Der Stabilitätsvorsprung von SHAP gilt zudem nicht durchgängig. Über die zehn höchstplatzierten Variablen des Referenzlaufs bleibt SHAP vorn, über die fünf höchstplatzierten weist dagegen die gruppierte PFI das stabilere Ranking auf (vgl. 5.5). Für Anwendungen, die nur eine kleine Spitzengruppe von Variablen berichten, kann die relative Stabilitätsbewertung der Verfahren damit anders ausfallen als für das vollständige Ranking.

Eine Garantie für stabile Rankings unter erneutem Training folgt aus der deterministischen SHAP-Berechnung nicht. Die hier gemessene Ranking-Stabilität beschreibt die Konsistenz der globalen Rankings innerhalb des untersuchten Modellierungs- und Auswertungsaufbaus und nicht die Konsistenz einer individuellen Erläuterung.

Chen et al. [12] berichten eine mit zunehmendem Klassenungleichgewicht abnehmende Stabilität von Ranglisten, die aus SHAP- und LIME-Erklärungen abgeleitet werden. Der vorliegende Befund steht dem nicht entgegen, ist aufgrund der unterschiedlichen Untersuchungs- und Stabilitätskonzepte jedoch nicht unmittelbar vergleichbar (vgl. 3.1).

Von der Stabilität innerhalb eines Verfahrens ist die Übereinstimmung der beiden Verfahren untereinander zu unterscheiden. Mit $\rho_{\text{SHAP,PFI}} = 0{,}72$ fällt sie niedriger aus als die mittlere Übereinstimmung der Rankings desselben Verfahrens über die Trainingsläufe (vgl. 5.4). Beide Verfahren liefern somit stabile, aber nicht identische Rankings, was mit ihren unterschiedlichen Wichtigkeitskonzepten vereinbar ist (vgl. 2.2.1 und 2.2.3). Weder ist die Übereinstimmung eine unabhängige Validierung noch eine Abweichung ein Hinweis auf einen Fehler in einem der Verfahren. Die geringere zwischenmethodische Übereinstimmung verdeutlicht, dass die Verfahrenswahl die berichtete Rangfolge sichtbar beeinflusst. Eine quantitative Zerlegung des Einflusses von Verfahrenswahl und erneutem Training ist mit dem Untersuchungsaufbau nicht möglich.

6.2 REDUNDANZ- UND ABLATIONSEFFEKTE GESCHLECHTSBEZOGENER REPRÄSENTATIONEN (F2)

Die Ablation (vgl. 5.6) verändert die SHAP-Ränge beider Repräsentationen nur geringfügig. Keine von ihnen wird nach Entfernung der jeweils anderen prominent. Aufgrund des einzelnen Ablationsmodells je Richtung lässt sich daraus weder eine belastbare systematische Umverteilung der globalen SHAP-Wichtigkeit auf die jeweils verbleibende Repräsentation ableiten noch der Einfluss der Modellstochastik von den beobachteten Rangänderungen trennen. Insbesondere liegt die Verschiebung von `personal_status_sex` auf Rang 11 innerhalb des in der Stabilitätsanalyse über die elf Trainingsläufe beobachteten Rangbereichs von 10 bis 12 (vgl. 5.5).

Der Befund ist zudem an die gewählte Operationalisierung gebunden. Verglichen werden rohe Variablenränge, die erst dann reagieren, wenn eine Attributionsänderung groß genug ist, um eine benachbarte Position zu überholen. Eine Umverteilung unterhalb dieser Auflösung bliebe unsichtbar. Hinzu kommt die in Abschnitt 4.3 als Randbedingung ausgewiesene pfadabhängige SHAP-Berechnung, die die Attribution entlang der tatsächlich verwendeten Baumpfade verteilt, während eine interventionelle Berechnung die Anteile zwischen deterministisch verbundenen Repräsentationen anders zuweisen könnte. Ob zwischen `sex_binary` und `personal_status_sex` überhaupt eine Umverteilung stattfindet, lässt sich mit dem gewählten Aufbau daher nicht entscheiden. Die geringfügige Rangänderung ist mit beiden Möglichkeiten vereinbar.

Die beiden Ablationsrichtungen sind ferner aufgrund der zusätzlichen Familienstandsinformation in `personal_status_sex` nicht unmittelbar miteinander vergleichbar (vgl. 4.4). Und selbst ein eindeutiges Ausbleiben der Umverteilung besagte nicht, dass geschlechtsbezogene Information dem Modell

nicht zur Verfügung steht, da sie über korrelierte oder als Proxy wirkende Merkmale verfügbar bleiben kann (vgl. 6.5).

6.3 TECHNISCHES UNTERSTÜTZUNGSPOTENZIAL (F3)

Die Einordnung entlang der in Abschnitt 4.5 abgeleiteten Indikatoren fällt je Artikel unterschiedlich aus. Tabelle A.8 in Anhang A.7 fasst die technischen Fähigkeiten der drei Methoden zusammen. Für den Indikator zu Artikel 13 ergibt sich ein dreiteiliges Bild: SHAP liefert sowohl globale als auch fallspezifische Erklärungen, LIME fallspezifische Erklärungen ohne etabliertes globales Maß und PFI ausschließlich globale, performancebasierte Importance ohne Einzelfallbezug. Innerhalb der ausgewählten Indikatoren bietet SHAP damit die breiteste Abdeckung, während die gruppierte PFI eine globale performancebasierte Perspektive und LIME eine methodisch andersartige lokale Surrogaterklärung ergänzt. Diese Abdeckung folgt unmittelbar aus den binär operationalisierten Indikatoren und den Verfahrenseigenschaften und ist damit eine strukturierte technische Zuordnung, keine empirische Aussage über die Eignung der Verfahren im praktischen Prüfbetrieb. Für den Indikator zu Artikel 14 differenziert die Einordnung dagegen nicht zwischen den Verfahren. Alle drei erzeugen dokumentierbare und visualisierbare Artefakte, die sich prinzipiell in einen menschlichen Prüfprozess einbinden lassen. Der technische Indikator zu Artikel 86, die Bereitstellung einer fallspezifischen Erklärung, liegt für SHAP und LIME vor, jedoch nur auf der Ebene der unaufbereiteten technischen Ausgabe. Die lokalen Erklärungen berichten auf der Log-Odds-Skala beziehungsweise über diskretisierte Bedingungen und setzen damit methodisches Vorwissen voraus, wie Abbildung B.1 zeigt. Aus der technischen Verfügbarkeit folgt daher weder eine adressatengerechte Erläuterung nach Artikel 86 noch eine wirksame menschliche Aufsicht nach Artikel 14. Beide setzen eine nachgelagerte Aufbereitung und Einbettung in einen geeigneten Prüfprozess voraus.

Die Gegenüberstellung der lokalen SHAP- und LIME-Erklärungen zeigt zunächst, dass der Inhalt einer fallspezifischen Erläuterung von der Verfahrenswahl abhängen kann. Der in allen drei Instanzen beobachtete Kontrast bei `foreign_worker` bleibt auch nach einer Aggregation der lokalen SHAP-Beiträge auf die Ebene der 21 Ursprungsvariablen bestehen (vgl. 5.7). Die Verfahrenswahl ist daher eine dokumentationsbedürftige methodische Entscheidung und nicht lediglich eine Frage der Darstellung. Aufgrund der illustrativen Auswahl von drei Instanzen lässt sich dieser Befund jedoch nicht verallgemeinern.

Die Beispiele zeigen außerdem, dass ein sensibles Merkmal in der LIME-Erklärung lokal prominent erscheinen kann, obwohl es sowohl in den globalen Rankings als auch in den lokalen SHAP-Beiträgen derselben Instanzen nur eine geringe Prominenz besitzt (vgl. Tabelle B.2). Erklärungsebene und Verfahrenswahl können damit gemeinsam beeinflussen, welches Merkmal prominent erscheint. Globale und lokale Erklärungen beantworten unterschiedliche Fragen und sollten bei der Analyse potenzieller Bias-

Indikatoren nicht gegeneinander ersetzt werden. Ob die auffälligen lokalen LIME-Gewichte von `foreign_worker` mit der extrem ungleichen Verteilung dieser Variable zusammenhängen, wurde nicht geprüft.

Der technische Nutzen der hier erzeugten LIME-Erklärungen wird zusätzlich durch ihre geringe und zwischen den Instanzen schwankende lokale Treue eingeschränkt (vgl. 5.7). Die ausgegebenen Gewichte beschreiben daher zunächst das jeweilige Surrogat und können nur insoweit als Erläuterung des XGBoost-Modells dienen, wie dieses lokal hinreichend genau approximiert wird. Ein über die betrachteten Beispiele hinausgehender Zusatznutzen von LIME gegenüber SHAP wurde nicht untersucht.

6.4 OUTCOME-BASIERTE FAIRNESSMETRIKEN

Die ergänzend erhobenen outcome-basierten Fairnessmetriken verdeutlichen, dass XAI-Attributionen eine gruppenbezogene Prüfung auf Outcome-Ebene nicht ersetzen. Im untersuchten Modell liegt `sex_binary` in beiden globalen Rankings im hinteren Drittel, während `personal_status_sex` im SHAP-Ranking eine mittlere Position einnimmt. Gleichzeitig unterschreitet der geschlechtsbezogene Disparate Impact unter Szenario A die Four-Fifths-Schwelle (vgl. 5.4 und 5.8). Eine geringe globale Prominenz sensibler beziehungsweise geschlechtsbezogener Variablen schließt eine Ergebnisdisparität somit nicht aus. Umgekehrt erlaubt das Vorliegen einer solchen Disparität keine Bewertung der verwendeten Erklärungsverfahren.

Auch die outcome-basierte Bewertung selbst ergibt keinen einheitlichen geschlechtsbezogenen Befund. Der Disparate Impact liegt unter Szenario A unter 0,8, unter Szenario B dagegen über eins. Der Equal-Opportunity-Gap fällt in beiden Szenarien klein aus (0,016 und 0,034), und das Bootstrap-Konfidenzintervall der zugehörigen TPR-Differenz schließt unter Szenario A die Null ein (vgl. 5.8). Die Bewertung hängt damit sowohl von der gewählten Fairnessmetrik als auch von der unsicheren Zuordnung der mehrdeutigen Geschlechtskategorie ab. Eine eindeutige geschlechtsbezogene Bewertung des Modells lässt sich daraus nicht ableiten. Für die `foreign_worker`-Achse wird die Aussagekraft zusätzlich durch die sehr kleine Subgruppe begrenzt (vgl. 6.5).

6.5 LIMITATIONEN

DATENGRUNDLAGE UND SENSIBLE GRUPPEN. Der Datensatz bildet historische Kreditverläufe der 1970er-Jahre innerhalb einer bereits durch damalige Vergabeentscheidungen ausgewählten Stichprobe ab. Sein Alter begrenzt die inhaltliche Übertragbarkeit und historische Merkmals- und Gruppenstrukturen können auch die methodischen Befunde beeinflussen.

Die geschlechtsbezogenen Auswertungen sind durch die mehrdeutige Kodierung von 310 Fällen eingeschränkt. Da die Szenarien A und B pauschale plausible Zuordnungen und keine formalen Identifikationsgrenzen darstellen, lässt sich der tatsächliche Wertebereich der geschlechtsbezogenen Fair-

nessmetriken nicht bestimmen (vgl. 4.1 und Anhang A.6). Der Szenarienvergleich verändert zudem nicht nur die Gruppenzuordnung, sondern über die abweichende Kodierung von `sex_binary` auch die Eingabedaten des jeweils neu trainierten Modells; Gruppen- und Modelleffekt lassen sich daher nicht voneinander trennen.

Die `foreign_worker`-Subgruppe umfasst nur $n = 5$ Testfälle, erlaubt daher keine belastbare Fairnessaussage und wird ausschließlich explorativ als Stresstest für extrem unbalancierte Subgruppen interpretiert (vgl. 5.8 und 6.4).

Der Datensatz enthält zudem ausschließlich vergebene Kredite, sodass die Fairnessmetriken nur Disparitäten der Modellklassifikation innerhalb der bereits finanzierten Stichprobe beschreiben und nicht die historischen Vergabeentscheidungen der Bank. Eine Übertragung auf die gesamte Antragstellerpopulation würde eine mit den verfügbaren Daten nicht prüfbare gruppenneutrale Vorauswahl voraussetzen (vgl. 4.1).

OUTCOME-STRATIFIZIERTE STICHPROBE. Die 1 000 Fälle wurden getrennt nach Kreditverlauf gezogen und Fälle der Klasse *bad* wurden überrepräsentiert. Der Anteil von 30 % *bad*-Fällen beschreibt daher das Stichprobendesign und nicht die historische Grundgesamtheit.

Prävalenzabhängige Kennzahlen wie Accuracy, Precision, F1, die positiven Klassifikationsraten und der daraus gebildete Disparate Impact gelten nur für die untersuchte Stichprobe. Ohne Kenntnis der historischen Auswahlwahrscheinlichkeiten sind sie nicht auf die ursprünglich finanzierte Population übertragbar. Klassenbedingte Kennzahlen wie die True-Positive-Rate sind weniger unmittelbar betroffen, sofern die Auswahl ausschließlich vom Kreditverlauf und nicht zusätzlich von den Merkmalen abhing.

Für den internen Vergleich der Erklärungsverfahren bildet die Stichprobenzusammensetzung eine gemeinsame Randbedingung. Auswirkungen auf das trainierte Modell und damit auf die resultierenden Attributionen und Rankings sind dennoch nicht ausgeschlossen.

MODELL- UND AUFTEILUNGSABHÄNGIGKEIT. Die Untersuchung beschränkt sich auf ein XGBoost-Modell und erlaubt daher keine Aussage darüber, ob die Befunde auf andere Modellklassen übertragbar sind. Ebenso wird nicht untersucht, ob auf diesem Datensatz ein intrinsisch interpretierbares Modell einem Black-Box-Modell mit nachgelagerter XAI vorzuziehen wäre (vgl. 7.3).

Sämtliche berichteten Rankings, Fairnesskennzahlen und Testwerte beruhen zudem auf einer einzigen Trainings-Test-Aufteilung. Die feste Aufteilung hält den Methodenvergleich frei von Unterschieden zwischen Partitionen, lässt aber offen, wie stark die Ergebnisse bei einer erneuten Datenaufteilung variieren würden. Die Kreuzvalidierung der Modellperformance ersetzt eine solche Wiederholung der vollständigen Analyse nicht, da die berichteten Erklärungen und Fairnessmetriken an die feste Testmenge gebunden bleiben.

AGGREGATION DER SHAP-WERTE. Das globale SHAP-Ranking ist an die gewählte Aggregation der One-Hot-Spalten auf Variablenebene gebunden (vgl. 4.3). Die Gegenrechnung unter der alternativen Aggregation in Anhang A.3 zeigt zwar eine hohe Rangübereinstimmung und unveränderte Positionen der drei sensiblen Variablen, die vorderen Ränge nicht-sensibler Variablen verschieben sich jedoch um bis zu drei Positionen. Einzelne Rangangaben des globalen SHAP-Rankings sind daher nur zusammen mit der zugrunde liegenden Aggregationsvariante zu interpretieren.

LIME UND UNSICHERHEITSQUANTIFIZIERUNG. Für LIME wurde keine Stabilitätsanalyse durchgeführt. Über die Variabilität seiner lokalen Erklärungen wird daher keine Aussage getroffen. Die lokalen Erklärungen sind zudem an die Behandlung der ordinal kodierten Merkmale `installment_rate`, `present_residence` und `number_credits` gebunden, die in der LIME-Konfiguration nicht als kategorial gekennzeichnet und daher wie numerische Merkmale quartilsbasiert diskretisiert werden (vgl. 4.3). Die erzeugte Nachbarschaft und die daraus geschätzten Gewichte hängen damit von dieser Modellierungsentscheidung ab. Konfidenzintervalle werden nur für die gruppenbezogenen True-Positive-Raten und ihre Differenzen berichtet, nicht jedoch für den Disparate Impact oder die Rangpositionen. Da die Unsicherheit nicht über die gesamte Auswertungskette fortgepflanzt wird, bleiben diese Größen Punktschätzungen.

GRUPPIERTE PFI UND GESCHLECHTSBEZOGENE REPRÄSENTATIONEN. Die gruppierte PFI erhält die Kategorienstruktur innerhalb einer One-Hot-kodierten Ursprungsvariable. Sie erhält jedoch nicht die deterministische Beziehung zwischen `sex_binary` und `personal_status_sex`, da beide Variablen getrennt permutiert werden. Dadurch können Kombinationen entstehen, die im ursprünglichen Datensatz nicht vorkommen. Die individuellen PFI-Ränge dieser beiden Repräsentationen sind daher besonders zurückhaltend zu interpretieren (vgl. 4.4).

ASSOZIATIVER CHARAKTER UND REGULATORISCHE REICHWEITE. Die eingesetzten Erklärungsverfahren beschreiben statistische Zusammenhänge im trainierten Modell und erlauben keine kausale Aussage über Diskriminierung oder die Wirkung einzelner Merkmale. Insbesondere schließen niedrige Ränge sensibler Variablen und geringe Rangänderungen in der Ablation nicht aus, dass gruppenbezogene Information über korrelierte oder als Proxy wirkende Merkmale verfügbar bleibt.

Die regulatorische Einordnung untersucht ausschließlich das technische Unterstützungspotenzial der erzeugten Erklärungen anhand ausgewählter Indikatoren zu den Artikeln 13, 14 und 86 (vgl. 4.5 und 6.3). Sie stellt keine rechtliche Konformitätsprüfung dar.

7 FAZIT UND AUSBLICK

7.1 KERNERGEBNIS

Die vorliegende Arbeit hat ausgewählte methodische Eigenschaften, Grenzen und technische Unterstützungsmöglichkeiten von SHAP, LIME und PFI bei der Analyse potenzieller Bias-Indikatoren in einem XGBoost-Kreditrisikomodell auf dem korrigierten *South German Credit*-Datensatz untersucht und im Kontext des EUAIA eingeordnet. Sie verbindet damit drei Aspekte, die in der berücksichtigten Literatur bislang getrennt behandelt werden: den in Kapitel 3 als offen benannten Stabilitätsvergleich beider globaler Rankings auf demselben Modell und Datensatz, einen kontrollierten deterministischen Redundanzfall und eine regulatorische Einordnung.

Für Forschungsfrage F1 zeigt sich, dass SHAP und die gruppierte PFI über wiederholte Trainingsläufe bei unveränderter Datenaufteilung sehr stabile globale Variablen-Rankings erzeugen. Über das vollständige Ranking ist SHAP geringfügig stabiler, während die gruppierte PFI im Top-5-Ausschnitt des Referenzlaufs die höhere Stabilität erreicht (vgl. 5.5 und 6.1). Die zwischenmethodische Rangkorrelation von $\rho_{\text{SHAP,PFI}} = 0{,}72$ im Referenzlauf zeigt zugleich, dass beide Verfahren trotz ihrer jeweils hohen Stabilität keine identischen Rangfolgen liefern. Eine hohe Stabilität innerhalb eines Verfahrens belegt daher keine verfahrensunabhängige globale Rangordnung. Die Stabilitätsmessung betrifft ausschließlich globale Rankings und erlaubt keine Aussage über die Konsistenz einzelner fallspezifischer Erläuterungen.

Für Forschungsfrage F2 ergibt die Ablation nur geringfügige Rangänderungen. Keine der deterministisch verbundenen Repräsentationen `sex_binary` und `personal_status_sex` wird nach Entfernung der jeweils anderen prominent (vgl. 5.6 und 6.2). Aufgrund des einzelnen Ablationsmodells je Richtung lässt sich daraus kein belastbarer systematischer Übergang globaler SHAP-Wichtigkeit auf die jeweils verbleibende Repräsentation ableiten, und die geringen Ränge belegen nicht, dass geschlechtsbezogene Information für das Modell unerheblich ist.

Für Forschungsfrage F3 deckt SHAP die ausgewählten technischen Einordnungsindikatoren am breitesten ab, da es als einziges der drei Verfahren globale und fallspezifische Erklärungen bereitstellt, während die gruppierte PFI eine globale performancebasierte Perspektive und LIME eine methodisch andersartige lokale Surrogaterklärung beisteuert (vgl. 6.3). Aus dieser technischen Verfügbarkeit folgt jedoch weder eine wirksame menschliche Aufsicht noch eine klare und aussagekräftige Erläuterung für betroffene Personen. Beides erfordert eine adressatengerechte Aufbereitung und eine geeignete Einbettung in den jeweiligen Prüfprozess.

Die Gegenüberstellung der lokalen SHAP- und LIME-Erklärungen zeigt an drei illustrativ ausgewählten Instanzen, dass der Inhalt einer fallspezifischen Erläuterung vom gewählten Verfahren abhängen kann und dass ein sensibles Merkmal in der LIME-Erklärung lokal prominent erscheinen kann, obwohl es sowohl in den globalen Rankings als auch in den lokalen SHAP-Beiträgen derselben Instanzen nur eine geringe Prominenz besitzt (vgl. 5.7). Aufgrund der kleinen Auswahl und der geringen lokalen Treue der untersuchten LIME-Surrogate folgt daraus jedoch keine allgemeine Überlegenheit oder Ergänzungswirkung eines Verfahrens.

Insgesamt erzeugen die drei Verfahren unterschiedliche und nicht austauschbare Erklärungsobjekte, deren technischer Nutzen von der jeweiligen Fragestellung, der Erklärungsebene, den methodischen Randbedingungen und der Qualität der verfügbaren Daten abhängt. Keines von ihnen erlaubt für sich allein eine rechtliche Bewertung möglicher Diskriminierung.

7.2 PRAXISRELEVANZ

Für die Praxis folgt zunächst, dass XAI-Methoden eine eigenständige outcome-basierte Fairnessprüfung nicht ersetzen. Eine technische Prüfung potenzieller Bias-Indikatoren gewinnt an Aussagekraft, wenn sie globale und fallspezifische Erklärungen mit gruppenbezogenen Fairnessmetriken, den jeweiligen Subgruppengrößen und geeigneten Unsicherheitsangaben verbindet. Ergänzend bieten sich Sensitivitätsanalysen für Klassifikationsschwellen und mehrdeutige Gruppenzuordnungen sowie eine Prüfung der Daten- und Feature-Konstruktion an.

Auch die Verfahrenswahl sollte dokumentiert und begründet werden. Dabei sind die jeweilige Erklärungsebene, das zugrunde liegende Wichtigkeitskonzept, die verwendete Aggregation und die empirisch geprüfte Treue oder Stabilität zu berücksichtigen. Auf der fallspezifischen Ebene können alternative Formate wie kontrafaktische Erläuterungen [55] die Darstellung ergänzen.

7.3 AUSBLICK

Aus den methodischen Grenzen ergeben sich mehrere zentrale Anschlussrichtungen.

Erstens sollte die vollständige Analyse über mehrere geeignete Modellklassen und wiederholt gezogene Trainings-Test-Aufteilungen hinweg durchgeführt werden. Dadurch ließen sich die Modell- und Partitionsabhängigkeit der Rankings, Fairnessmetriken und lokalen Erklärungen getrennt untersuchen. Ergänzend sollte die Stabilität wiederholt berechneter LIME-Erklärungen quantifiziert und die gruppierte PFI einer ungruppierten Kontrollbedingung gegenübergestellt werden.

Zweitens könnte eine kontrollierte synthetische Bias-Injektion unterschiedliche Verzerrungsmechanismen gezielt variieren. Dazu gehören die direkte Nutzung eines sensiblen Merkmals, dessen indirekte Nutzung

über Proxy-Merkmale, gruppenspezifische Fehlerraten sowie Verzerrungen in den Trainingsdaten oder der Zielvariable. Auf dieser Grundlage ließe sich systematisch vergleichen, wie die verschiedenen XAI-Ausgaben und outcome-basierten Fairnessmetriken auf die jeweils eingespielte Verzerrung reagieren.

Drittens ist eine Wiederholung auf einem zeitgemäßeren Datensatz mit getrennt erfassten und hinreichend großen sensiblen Gruppen erforderlich. Dadurch würden sowohl die mehrdeutige Geschlechtskodierung als auch die statistisch nicht belastbare `foreign_worker`-Subgruppe vermieden. Falls eine eindeutige Zuordnung weiterhin nicht möglich ist, könnten die geschlechtsbezogenen Fairnessmetriken als Wertebereich über alle mit dem Codebook vereinbaren Zuordnungen statt als einzelne Punktschätzer berichtet werden.

Viertens ist eine nutzerzentrierte Evaluation erforderlich, um das in F3 untersuchte technische Unterstützungspotenzial praktisch zu prüfen. Dabei sollte untersucht werden, ob unterschiedliche Nutzergruppen die aufbereiteten Erklärungen verstehen, Modellfehler erkennen und fundiertere Prüf- oder Entscheidungsprozesse durchführen können. Ein solcher Vergleich könnte außerdem zeigen, ob ein kombinierter Einsatz mehrerer Erklärungsverfahren tatsächlich einen messbaren Zusatznutzen bietet.

Darüber hinaus könnte die korrelative XAI-Perspektive durch kausale Pfadzerlegungen ergänzt werden, die unter expliziten Annahmen über den kausalen Graphen untersuchen, ob und über welche Pfade ein sensibles Merkmal die Modellvorhersage beeinflusst [57]. Im Sinne von Rudin [50] bleibt zudem im Einzelfall zu prüfen, ob ein Black-Box-Modell mit nachgelagerter XAI gegenüber einem intrinsisch interpretierbaren Modell überhaupt einen nachvollziehbar belegbaren Zusatznutzen bietet.

A

ERGÄNZENDE TABELLEN UND HERLEITUNGEN

Dieser Anhang dokumentiert ergänzende Herleitungen und Auswertungen zu Kapitel 4 und 5: die Herleitung der im Training verwendeten Klassengewichte, die Hyperparameter des XGBoost-Modells, die Aggregation der SHAP-Werte auf Variablenebene, die Konfiguration von LIME, eine Robustheitsprüfung der beiden Szenarien für die mehrdeutige Geschlechtskategorie sowie ergänzende Ergebnisdarstellungen.

A.1 HERLEITUNG DER KLASSENGEWICHTE

Die inversen Klassengewichte werden aus den Klassenhäufigkeiten der Trainingsmenge abgeleitet. Der stratifizierte 80/20-Split ergibt eine Trainingsmenge von $n = 800$ Instanzen mit $n_{\text{good}} = 560$ Krediten der Klasse *good* und $n_{\text{bad}} = 240$ der Klasse *bad*. Das Gewicht der Klasse c folgt der balancierten Heuristik $w_c = n/(2\,n_c)$, woraus $w_{\text{good}} \approx 0{,}714$ und $w_{\text{bad}} \approx 1{,}667$ folgen. Ihr Verhältnis $w_{\text{bad}}/w_{\text{good}} = 560/240 \approx 2{,}33$ entspricht dem inversen Klassenverhältnis der Trainingsmenge und stimmt aufgrund der stratifizierten Aufteilung mit dem Klassenverhältnis des Gesamtdatensatzes (700:300) überein. Die Gewichte gehen als Instanzgewichte in das Training des XGBoost-Modells ein. In der Kreuzvalidierung (vgl. 4.2) werden dieselben Gewichte auf die Instanzen des jeweiligen Trainingsteils angewendet.

A.2 HYPERPARAMETER DES XGBOOST-MODELLS

Tabelle A.1 listet die in Abschnitt 4.2 begründete Konfiguration des XGBoost-Modells.

Tabelle A.1: Hyperparameter des XGBoost-Modells.

Parameter	Wert
Anzahl Bäume (`n_estimators`)	200
Maximale Baumtiefe (`max_depth`)	4
Lernrate (η)	0,05
Zeilen-Subsampling (`subsample`)	0,80
Spalten-Subsampling (`colsample_bytree`)	0,80

A.3 AGGREGATION DER SHAP-WERTE AUF VARIABLENEBENE

Ergänzend zu Abschnitt 4.3 dokumentiert dieser Abschnitt den Zusammenhang der beiden Aggregationsvarianten. Für die Attributionen ϕ_{ic} der ko-

dierten Spalten c einer Ursprungsvariable v gilt je Testinstanz i nach der Dreiecksungleichung

$$\left|\sum_{c \in v} \phi_{ic}\right| \leq \sum_{c \in v} |\phi_{ic}|,$$

mit Gleichheit nur bei durchgängig gleichem Vorzeichen. Die für das globale Ranking verwendete Variante bildet den Betrag vor der Summe und entspricht der rechten Seite, die Alternative summiert je Instanz zuerst vorzeichenbehaftet und bildet erst danach den Betrag. Sie fällt daher nie größer aus. Der Abstand beider Varianten kann insbesondere bei Variablen mit vielen kodierten Spalten und gegenläufigen Beiträgen größer ausfallen. Ihre vorzeichenbehaftete Zwischengröße $\sum_{c \in v} \phi_{ic}$ erhält die Additivität auf Variablenebene, summiert sich also weiterhin exakt zum Modelloutput, und wird in Abschnitt 5.7 für die lokale Gegenüberstellung mit LIME verwendet.

Da die berichteten SHAP-Ränge an diese Entscheidung gebunden sind, wurde das globale Ranking des Hauptmodells zusätzlich unter der alternativen Aggregation berechnet. Die Spearman-Rangkorrelation beider Rankings über alle 21 Variablen beträgt 0,987; keine Variable verschiebt sich um mehr als drei Positionen. Die größte Verschiebung betrifft mit `purpose` erwartungsgemäß die Variable mit den meisten kodierten Spalten (zehn), die von Rang 2 auf Rang 5 fällt. Die übrigen Verschiebungen um ein bis zwei Positionen entstehen als Folge davon in den benachbarten Rängen. Die Ränge der drei sensiblen Variablen bleiben unter beiden Aggregationen identisch: `personal_status_sex` auf Rang 10, `foreign_worker` auf Rang 19 und `sex_binary` auf Rang 21. Auch die zwischenmethodische Übereinstimmung bleibt erhalten. Die Spearman-Rangkorrelation zur gruppierten PFI beträgt unter der alternativen Aggregation 0,718 gegenüber 0,721 unter der berichteten und damit auf zwei Nachkommastellen in beiden Fällen 0,72 (vgl. 5.4). Die in Kapitel 5 berichteten Aussagen zu den sensiblen Variablen und zum Vergleich mit der gruppierten PFI hängen damit nicht an der gewählten Aggregationsvariante. Die Begünstigung spaltenreicher Variablen wirkt sich im vorliegenden Modell nur auf die vorderen Ränge nicht-sensibler Variablen aus. Aggregationsabhängig ist allein der in Abschnitt 5.2 berichtete Rangunterschied von `purpose` zwischen beiden Verfahren. Unter der alternativen Aggregation nimmt die Variable denselben Rang 5 ein wie in der gruppierten PFI.

A.4 LIME-KONFIGURATION

Tabelle A.2 listet die in Abschnitt 4.3 verwendete Konfiguration von `LimeTabularExplainer`. Explizit gesetzt werden `discretize_continuous`, `categorical_features`, `categorical_names`, `random_state` und `num_features`. Die übrigen Parameter entsprechen, sofern nicht anders ausgewiesen, den Standardwerten der verwendeten `lime`-Version (0.2). Die beiden zweiwertigen Variablen `sex_binary` und `people_liable` werden ebenfalls als kategorial gekennzeichnet, damit LIME ihre Ausprägungen als

diskrete Kategorien und nicht als numerische Größen mit quartilsbasierter Diskretisierung behandelt (vgl. 4.3). Bei `people_liable` fallen ohne diese Auszeichnung sämtliche Quartilsgrenzen der Trainingsmenge zusammen, sodass als einzige Bedingung `people_liable <= 2.00` entsteht, die jede Instanz erfüllt und daher ein Gewicht von exakt null erhält.

Tabelle A.2: Konfiguration von `LimeTabularExplainer`. Mit * markierte Werte sind Bibliotheksstandards und wurden nicht explizit gesetzt.

Parameter	Wert
Nachbarschaftsstichprobe (`num_samples`)*	5 000
Ausgegebene Variablen (`num_features`)	21 (alle Variablen)
Distanzmetrik*	euklidisch
Kernel*	exponentiell, Breite $\sqrt{21} \cdot 0{,}75 \approx 3{,}44$
Diskretisierung (`discretize_continuous`)	aktiviert (Quartile)
Kategoriale Merkmale (`categorical_features`)	15 Variablen: die 13 kategorialen Modellvariablen sowie `sex_binary` und `people_liable`
Kategorienbezeichner (`categorical_names`)	je kategorialer Variable gesetzt
Feature-Selektion (`feature_selection`)*	automatisch (`highest_weights`, Ridge $\alpha = 0{,}01$); bei Ausgabe aller Variablen ohne Auswahlwirkung
Lokales Surrogatmodell*	Ridge-Regression ($\alpha = 1$)
Instanzzentrierte Stichprobe (`sample_around_instance`)*	deaktiviert
Zufallsstartwert (`random_state`)	42

A.5 BOOTSTRAP-KONFIDENZINTERVALLE DER TPR-KENNZAHLEN

Ergänzend zu Abschnitt 4.4 dokumentiert dieser Abschnitt die Konfiguration des dort eingesetzten Perzentil-Bootstrap-Verfahrens. Auf die bei Efron und Tibshirani [15] beschriebene bias- und beschleunigungskorrigierte BCa-Variante wird verzichtet. Diese Abwägung ist eine eigene methodische Einschätzung. Bei einer Subgruppe von nur fünf Instanzen und teilweise degenerierten Kennzahlen lässt sich der Beschleunigungsparameter nicht verlässlich schätzen.

Die Ziehungslogik unterscheidet sich je Metrik. Für die gruppenweisen True-Positive-Raten wird innerhalb der jeweiligen Gruppe neu gezogen, für ihre Differenz dagegen je Replikat die gesamte Testmenge. Damit gehen in die Intervalle der Differenz die Unsicherheiten beider Gruppen gemeinsam ein, und bei kleinen Subgruppen wird zusätzlich die Variabilität der Gruppengröße selbst erfasst. Eine auch für die Differenz gruppenweise geschichtete Ziehung wäre die konsistentere Alternative, hielte die Gruppengrößen jedoch künstlich fest.

Replikate, in denen die Metrik mangels positiver Fälle undefiniert ist, werden verworfen. Betroffen sind systematisch jene Replikate, die keinen positiven Fall der kleineren Gruppe enthalten, sodass das Intervall die Unsicherheit dieser Gruppe eher unterschätzt.

A.6 ROBUSTHEITSPRÜFUNG DER BEIDEN SZENARIEN FÜR DIE MEHRDEUTIGE GESCHLECHTSKATEGORIE

Die beiden in Abschnitt 4.1 eingeführten Szenarien für die mehrdeutige Kategorie 2 von `personal_status_sex` unterscheiden sich in der Kodierung von `sex_binary`, die unter Szenario B nur für Kategorie 4 den Wert 1 annimmt. Um zu prüfen, ob die geschlechtsbezogenen Befunde von dieser Zuordnung abhängen, wird das gesamte Modell unter beiden Szenarien mit identischem Seed, identischer Datenaufteilung und identischen Hyperparametern trainiert. Tabelle A.3 vergleicht die zentralen geschlechtsbezogenen Kennzahlen. Der Vergleich stellt zwei plausible pauschale Zuordnungen der 310 Fälle gegenüber, jedoch keine formalen Identifikationsgrenzen.

Tabelle A.3: Vergleich der zentralen geschlechtsbezogenen Kennzahlen unter Szenario A und Szenario B der mehrdeutigen Kategorie 2 von `personal_status_sex`. Szenario A ordnet diese Kategorie der weiblich, Szenario B der männlich kodierten Gruppe zu.

Kennzahl	**Szenario A**	**Szenario B**
n weiblich (Testmenge)	81	19
Modellgüte (AUC)	0,820	0,823
SHAP-Variablenrang `sex_binary` (von 21)	21	20
Equal-Opportunity-Gap (sex)	0,016	0,034
Disparate Impact (sex)	0,787	1,050

Die Modellgüte und der SHAP-basierte Kernbefund erweisen sich als robust. Unter beiden Szenarien bleibt `sex_binary` auf einem der letzten SHAP-Variablenränge. Auch unter Szenario B führt die Ablation zu keiner deutlichen Rangverschiebung: `sex_binary` erreicht ohne `personal_status_sex` Rang 19 von 20 (Test-AUC 0,833), ausgehend von Rang 20 im Hauptmodell, sodass der Aufstieg vollständig der mechanischen Rangverschiebung entspricht (vgl. 4.4); `personal_status_sex` verbleibt ohne `sex_binary` auf Rang 11 von 20 (Test-AUC 0,829). Eine deutliche Rangverschiebung auf die jeweils verbleibende Repräsentation ist damit unter keinem der beiden Szenarien zu beobachten.

Die outcome-basierten Fairnessmetriken der geschlechtsbezogenen Achse sind dagegen nicht robust. Der Equal-Opportunity-Gap verdoppelt sich näherungsweise von Szenario A zu Szenario B, und der Disparate Impact wechselt von einem Wert unterhalb der Four-Fifths-Schwelle zu einem Wert oberhalb von eins, kehrt also die Richtung der scheinbaren Disparität um. Maßgeblich hierfür ist, dass Szenario B die weiblich kodierte Testgruppe von $n = 81$ auf $n = 19$ reduziert. Der Szenarienvergleich verändert allerdings zwei Dinge gleichzeitig: die Gruppenzuordnung der 310 mehrdeutigen Fälle und, über die veränderte Kodierung von `sex_binary` als Modelleingabe, das neu angepasste Modell selbst. Der isolierte Effekt der Gruppenzuordnung lässt sich davon nicht trennen, sodass die nahezu unveränderte Modellgüte (AUC 0,820 gegenüber 0,823) und die um höchstens eine Position verschobenen SHAP-Ränge lediglich deskriptiv berichtet werden. Die Zuordnung der mehrdeutigen Kategorie ist damit keine neutrale technische Entscheidung,

sondern bestimmt die Gruppengröße und darüber die statistische Belastbarkeit der geschlechtsbezogenen Fairnessbefunde (vgl. 6.5).

Ergänzend wurde geprüft, ob die in Abschnitt 5.4 berichteten Rangpositionen der sensiblen Variablen unter Szenario B zum selben qualitativen Bild führen. Dazu wurden SHAP und die gruppierte PFI auch für das unter Szenario B trainierte Modell auf Variablenebene ausgewertet (Tabelle A.4). Das Muster aus Szenario A bleibt erhalten. `sex_binary` und `foreign_worker` bleiben in beiden Rankings niedrig platziert (Ränge 14 bis 20 von 21), während `personal_status_sex` auch unter Szenario B bei SHAP im Mittelfeld und bei der gruppierten PFI auf einem der letzten Ränge erscheint. Die verfahrensabhängige Einordnung dieser Variable ist damit kein Artefakt der gewählten Geschlechtszuordnung.

Tabelle A.4: Variablenränge der sensiblen Variablen unter beiden Szenarien (von 21 Variablen, Rang 1 = höchste Wichtigkeit).

	Szenario A		Szenario B	
Variable	SHAP	PFI	SHAP	PFI
`sex_binary`	21	18	20	17
`foreign_worker`	19	16	19	14
`personal_status_sex`	10	21	11	20

A.7 ERGÄNZENDE ERGEBNISDARSTELLUNGEN

Dieser Abschnitt enthält die im Ergebniskapitel referenzierten ergänzenden Tabellen und Abbildungen.

Tabelle A.5: Geschlechtsbezogener Disparate Impact unter Szenario A für Klassifikationsschwellen von 0,40 bis 0,60 (vgl. 4.4). Angegeben sind zusätzlich die positiven Klassifikationsraten beider Gruppen. Der Betriebspunkt der übrigen Auswertungen ist 0,50.

Schwelle	Rate weiblich	Rate männlich	Disparate Impact
0,40	0,617	0,798	0,773
0,45	0,593	0,765	0,775
0,50	0,556	0,706	0,787
0,55	0,519	0,672	0,771
0,60	0,420	0,630	0,666

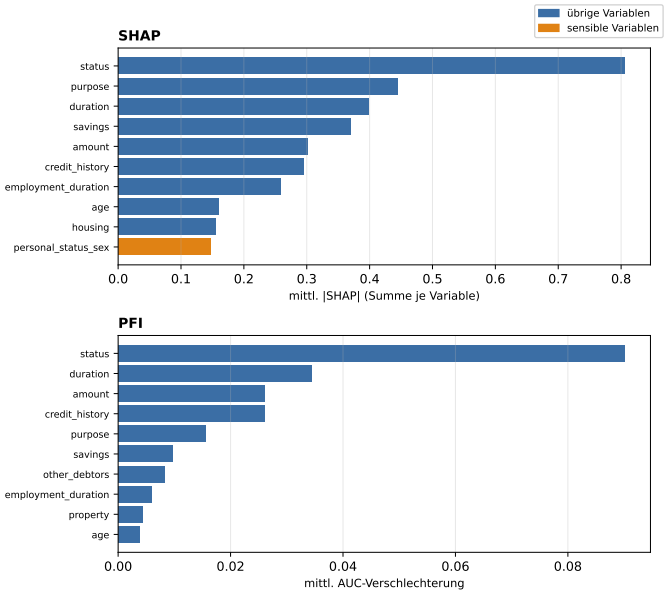


Abbildung A.1: Die jeweils zehn höchstplatzierten Variablen der beiden globalen Rankings (vgl. 5.4; vollständig in Tabelle A.7): globales SHAP-Ranking (aufsummierte mittlere |SHAP|) und gruppiertes PFI-Ranking (mittlere AUC-Verschlechterung). Sensible Variablen sind orange hervorgehoben. Beide Rankings stimmen in den dominanten Variablen weitgehend überein.

Tabelle A.6: 95 %-Bootstrap-Konfidenzintervalle über 1 000 Replikate (vgl. 5.8). n ist die Zahl der Testfälle, aus denen die Kennzahl gezogen wird, n_+ die Teilmenge der darin tatsächlich der Klasse *good* zugehörigen Fälle. Bei den gruppenweisen Raten ist n_+ der Nenner der TPR. Bei den Differenzen ist es die Gesamtzahl positiver Fälle der Testmenge, aus der sich je Replikat die beiden gruppenspezifischen Nenner ergeben. Die Zahl der gültigen Replikate liegt unter 1 000, wenn eine Subgruppe in einem Replikat keine tatsächlich positiven Fälle enthält und die True-Positive-Rate daher undefiniert bleibt. Die TPR-Differenzen sind vorzeichenbehaftet (vgl. 4.4).

Kennzahl	n	n_+	**95 %-Intervall**	**gültige Repl.**
TPR *foreign*	5	3	[1,00, 1,00]	989
TPR *domestic*	195	137	[0,71, 0,85]	1 000
TPR-Differenz weiblich − männlich	200	140	[−0,17, 0,14]	1 000
TPR-Differenz *foreign* − *domestic*	200	140	[0,15, 0,29]	961

Das Intervall der `foreign_worker`-TPR-Differenz schließt null zwar aus, folgt darin aber mechanisch daraus, dass die drei Fälle mit tatsächlicher Klasse *good* sämtlich korrekt klassifiziert werden und die `foreign_worker`-TPR damit in jedem gültigen Replikat eins beträgt; der Ausschluss der Null trägt daher keine Aussagekraft (vgl. 6.5).

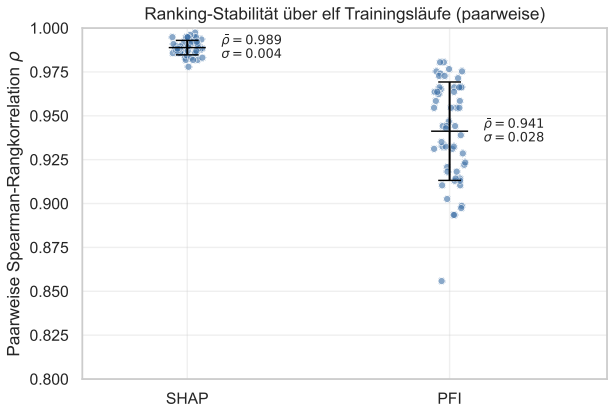


Abbildung A.2: Paarweise Spearman-Rangkorrelation der globalen Variablen-Rankings über die 55 Paare der elf Trainingsläufe. Jeder Punkt entspricht einem Modellpaar; die schwarzen Marker zeigen Mittelwert und Standardabweichung. Beide Rankings bleiben über die elf Trainingsläufe sehr stabil, SHAP über das Gesamtranking geringfügig stabiler als die gruppierte PFI.

Tabelle A.7: Globale Rankings beider Verfahren über alle 21 Modelleingangsvariablen, sortiert nach dem SHAP-Rang (vgl. 5.2, 5.3 und 5.4). Der SHAP-Wert ist der über die One-Hot-Spalten aufsummierte mittlere |SHAP| über 200 Testinstanzen auf der Log-Odds-Skala, der PFI-Wert die mittlere AUC-Verschlechterung der gruppierten Permutation. Negative PFI-Werte bedeuten, dass die mittlere AUC der permutierten Testdaten geringfügig über dem Ausgangswert lag.

Rang	Variable	mittl. |SHAP|	PFI	Rang
1	status	0,806	0,0902	1
2	purpose	0,445	0,0156	5
3	duration	0,399	0,0344	2
4	savings	0,369	0,0096	6
5	amount	0,301	0,0261	3
6	credit_history	0,296	0,0260	4
7	employment_duration	0,259	0,0060	8
8	age	0,160	0,0038	10
9	housing	0,155	−0,0029	19
10	personal_status_sex	0,147	−0,0052	21
11	property	0,145	0,0044	9
12	present_residence	0,110	0,0014	15
13	installment_rate	0,107	0,0019	12
14	other_installment_plans	0,106	0,0021	11
15	job	0,083	−0,0008	17
16	other_debtors	0,079	0,0082	7
17	telephone	0,053	0,0016	13
18	number_credits	0,048	−0,0031	20
19	foreign_worker	0,040	0,0012	16
20	people_liable	0,013	0,0014	14
21	sex_binary	0,004	−0,0010	18

Tabelle A.8: Qualitative technische Fähigkeitsübersicht der XAI-Methoden entlang der in Abschnitt 4.5 abgeleiteten Einordnungsdimensionen. In Klammern steht der Artikel, dessen Indikator die jeweilige Zeile operationalisiert. *Die Ranking-Stabilität ist die mittlere paarweise Spearman-Rangkorrelation über die elf Trainingsläufe (vgl. Tabelle 5.3). Sie ist ein methodisches Qualitätskriterium und im EUAIA nicht verankert. Die tatsächliche Nutzbarkeit durch aufsichtführende Personen und die Wirksamkeit menschlicher Aufsicht wurden nicht untersucht.

Eigenschaft	**SHAP**	**LIME**	**PFI**
Globale Erklärung beziehungsweise Wichtigkeitsdarstellung (Art. 13)	ja (aggregierte Attribution)	nein (kein etabliertes Maß)	ja (performancebasiert)
Fallspezifische Erklärung (Art. 13, 86)	ja	ja	nein
Dokumentierbares und visualisierbares Ergebnisartefakt (Art. 13, 14)	ja	ja	ja
Prinzipiell in einen menschlichen Prüfprozess integrierbar (Art. 14)	ja	ja	ja
Performancebasierte Wichtigkeit	nein	nein	ja
Globale Ranking-Stabilität*	0,99	nicht anwendbar	0,94
Wesentliche technische Grenze	Attribution statt Performance; Log-Odds-Skala	stochastisches Surrogat; diskretisierte Bedingungen; hier gemessen geringe lokale Treue	nur global; keine Einzelfallerklärung
Weiterer Aufbereitungs- beziehungsweise Einbettungsbedarf	adressatengerechte Darstellung (Art. 86)	adressatengerechte Darstellung (Art. 86)	Einbettung in globalen Aufsichtsprozess (Art. 14)

B

LOKALE ERKLÄRUNGSBEISPIELE

Dieser Anhang enthält die in Abschnitt 5.7 referenzierten lokalen Erklärungsbeispiele: die tabellarische Gegenüberstellung der jeweils drei höchstgewichteten Beiträge (Tabelle B.1), die lokalen Beiträge von `foreign_worker` in beiden Verfahren (Tabelle B.2), die lokale Treue der drei LIME-Surrogate (Tabelle B.3) sowie die SHAP- und LIME-Erklärungen derselben drei gezogenen Testinstanzen (Index 130, 17 und 154). Sie dienen ausschließlich der Veranschaulichung und als empirische Grundlage der Einordnung zu Forschungsfrage F3 (vgl. 6.3), gehen nicht in die Auswertung der Forschungsfragen F1 und F2 ein und werden nicht über die drei Einzelinstanzen hinaus verallgemeinert.

Tabelle B.1: Die jeweils drei höchstgewichteten Beiträge von SHAP und LIME für dieselben drei Testinstanzen (vgl. 5.7). SHAP weist die kodierten Modellfeatures auf der Log-Odds-Skala aus, LIME Bedingungen über die Ursprungsvariablen auf der Wahrscheinlichkeitsskala. Die Werte beider Spalten sind daher nicht ineinander umrechenbar und werden mit unterschiedlicher Stellenzahl ausgewiesen. Aufgrund der unterschiedlichen Repräsentationsebenen ist auch die Reihenfolge der Beiträge nicht unmittelbar methodenübergreifend vergleichbar. Tabelle B.2 stellt beide Verfahren dazu auf der Ebene der 21 Ursprungsvariablen gegenüber. Sortiert ist jeweils nach dem Betrag. Bei Index 154 liegen der zweite und dritte LIME-Beitrag nahezu gleichauf. Hervorgehoben ist die in allen drei Instanzen unter den LIME-Bedingungen vertretene Bedingung `foreign_worker=2`.

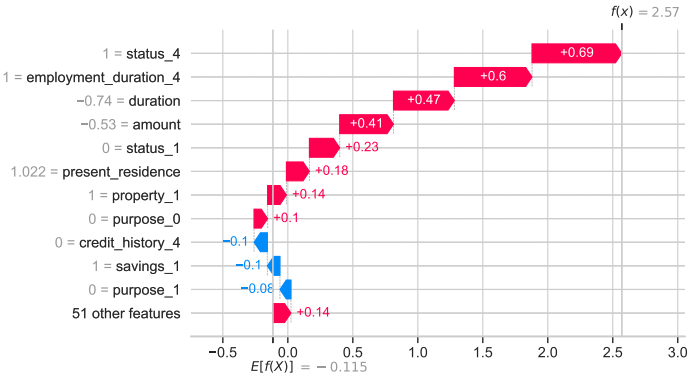
Instanz	SHAP-Feature	Beitrag	LIME-Bedingung	Gewicht
130	`status_4`	+0,69	`status=4`	+0,2161
	`employment_duration_4`	+0,60	`duration <= 12.00`	+0,1927
	`duration`	+0,47	**`foreign_worker=2`**	−0,1192
17	`amount`	+0,67	`status=1`	−0,2025
	`status_1`	−0,42	**`foreign_worker=2`**	−0,1428
	`savings_5`	+0,35	`savings=5`	+0,0980
154	`savings_5`	+0,48	**`foreign_worker=2`**	−0,1358
	`amount`	+0,40	`savings=5`	+0,0978
	`savings_1`	+0,29	`2281.50 < amount <= 3907.50`	+0,0975

Tabelle B.2: Lokale Beiträge von `foreign_worker` je Instanz (vgl. 5.7). Der SHAP-Beitrag ist je Instanz vorzeichenbehaftet über die One-Hot-Spalten der Variable summiert. Die Ränge beziehen sich in beiden Verfahren auf den Betrag über alle 21 Ursprungsvariablen.

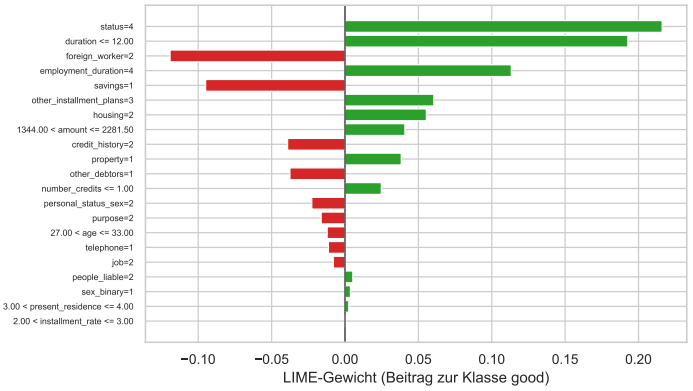
Instanz	SHAP-Beitrag	Rang	LIME-Gewicht	Rang
130	−0,018	18	−0,119	3
17	−0,022	15	−0,143	2
154	−0,023	17	−0,136	1

Tabelle B.3: Lokale Treue der drei LIME-Surrogatmodelle (vgl. 5.7). R^2 ist das Bestimmtheitsmaß der lokal gewichteten Ridge-Regression in der von LIME erzeugten Nachbarschaft. Die beiden Wahrscheinlichkeiten sind die Vorhersage des Surrogats und die des XGBoost-Modells für die Klasse *good* an der erklärten Instanz. Die Differenz ist als Surrogat minus XGBoost gebildet. Bei Index 154 liegen beide auf verschiedenen Seiten der Klassifikationsschwelle 0,5.

Instanz	R^2	Surrogat	XGBoost	Differenz
130	0,460	0,972	0,929	+0,043
17	0,259	0,585	0,760	−0,175
154	0,135	0,610	0,486	+0,125



(a) SHAP-Wasserfall-Diagramm auf der Log-Odds-Skala



(b) LIME-Erklärung auf der Wahrscheinlichkeitsskala

Abbildung B.1: Lokale SHAP- und LIME-Erklärung derselben zufällig gezogenen Testinstanz (Index 130). In (a) summieren sich die Beiträge der 62 kodierten Modellfeatures ausgehend vom Erwartungswert $E[f(X)]$ zur Modellausgabe $f(x)$; beide liegen auf der Margin-Skala des Modells, weshalb $f(x)$ Werte über 1 annehmen kann. Einzeln ausgewiesen sind die elf betragsmäßig größten Beiträge, die übrigen 51 fasst die Darstellung zu einer Restzeile zusammen. Rote Beiträge erhöhen, blaue senken die Log-Odds der Klasse *good*. Die links neben den Featurenamen ausgewiesenen Merkmalswerte sind die dem Modell übergebenen Werte, für numerische Merkmale also die standardisierten und nicht die ursprünglichen Größen (vgl. 4.2). In (b) sind die Gewichte aller 21 Variablen des lokalen linearen Surrogatmodells nach Betrag absteigend sortiert; grüne Gewichte wirken im lokalen Surrogat in Richtung einer höheren, rote in Richtung einer niedrigeren vorhergesagten Wahrscheinlichkeit der Klasse *good*. Kategoriale Bedingungen erscheinen als Gleichheiten, numerische als Intervalle über die ursprünglichen Werte.

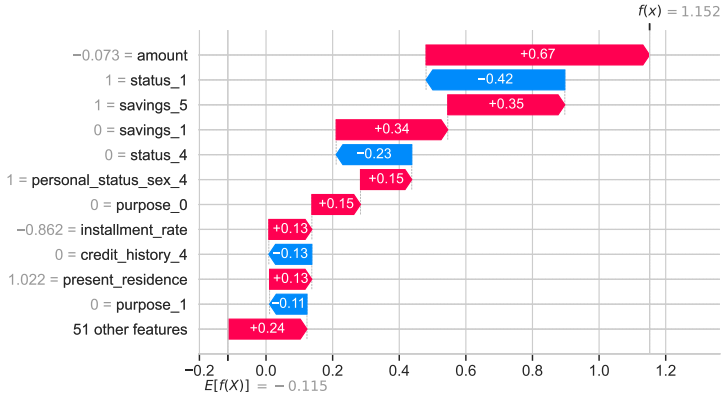


Abbildung B.2: Lokales SHAP-Wasserfall-Diagramm für die zweite gezogene Testinstanz (Index 17). Darstellung wie in Abbildung B.1a.

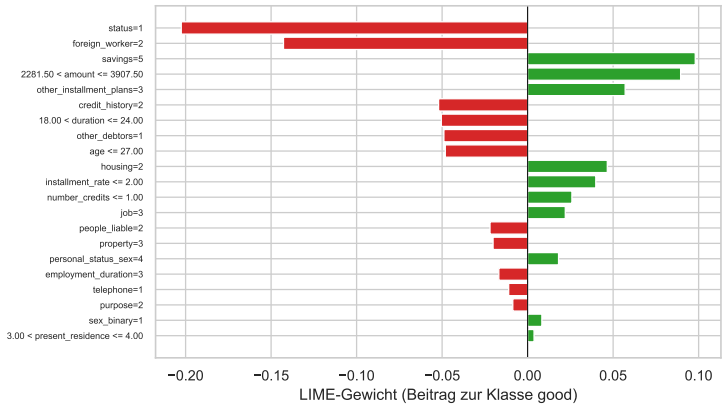


Abbildung B.3: Lokale LIME-Erklärung für dieselbe Testinstanz (Index 17) wie in Abbildung B.2. Darstellung wie in Abbildung B.1b.

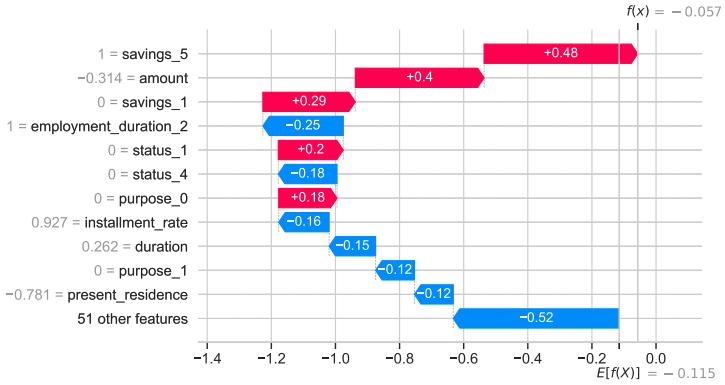


Abbildung B.4: Lokales SHAP-Wasserfall-Diagramm für die dritte gezogene Testinstanz (Index 154). Darstellung wie in Abbildung B.1a.

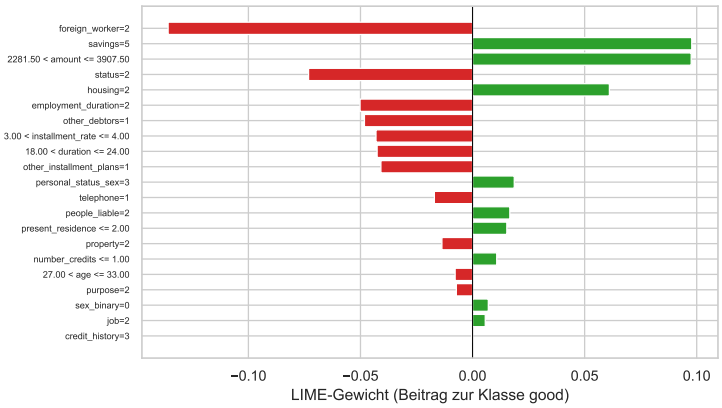


Abbildung B.5: Lokale LIME-Erklärung für dieselbe Testinstanz (Index 154) wie in Abbildung B.4. Darstellung wie in Abbildung B.1b.

LITERATURVERZEICHNIS

[1] Aas, K., Jullum, M., Løland, A.: Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. Artificial Intelligence **298**, 103502 (2021). https://doi.org/10.1016/j.artint.2021.103502

[2] Abdou, H.A., Pointon, J.: Credit scoring, statistical techniques and evaluation criteria: A review of the literature. Intelligent Systems in Accounting, Finance and Management **18**(2–3), 59–88 (2011). https://doi.org/10.1002/isaf.325

[3] Adadi, A., Berrada, M.: Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access **6**, 52138–52160 (2018). https://doi.org/10.1109/ACCESS.2018.2870052

[4] Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J.M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., Herrera, F.: Explainable Artificial Intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. Information Fusion **99**, 101805 (2023). https://doi.org/10.1016/j.inffus.2023.101805

[5] Alvarez-Melis, D., Jaakkola, T.S.: On the Robustness of Interpretability Methods. In: Proceedings of the 2018 ICML Workshop on Human Interpretability in Machine Learning (WHI) (2018). https://doi.org/10.48550/arXiv.1806.08049

[6] Bahlool, R., Hewahi, N., Elmedany, W.: Performance, Fairness, and Explainability in AI-Based Credit Scoring: A Systematic Literature Review. Journal of Risk and Financial Management **19**(2), 104 (2026). https://doi.org/10.3390/jrfm19020104

[7] Barocas, S., Selbst, A.D.: Big Data's Disparate Impact. California Law Review **104**(3), 671–732 (2016). https://doi.org/10.15779/Z38BG31

[8] Bentéjac, C., Csörgő, A., Martínez-Muñoz, G.: A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review **54**(3), 1937–1967 (2021). https://doi.org/10.1007/s10462-020-09896-5

[9] Bitetto, A., Cerchiello, P., Filomeni, S., Tanda, A., Tarantino, B.: Can we trust machine learning to predict the credit risk of small businesses? Review of Quantitative Finance and Accounting **63**(3), 925–954 (2024). https://doi.org/10.1007/s11156-024-01278-0

[10] Breiman, L.: Random Forests. Machine Learning **45**(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[11] Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794 (2016). https://doi.org/10.1145/2939672.2939785

[12] Chen, Y., Calabrese, R., Martin-Barragan, B.: Interpretable machine learning for imbalanced credit scoring datasets. European Journal of Operational Research **312**(1), 357–372 (2024). https://doi.org/10.1016/j.ejor.2023.06.036

[13] Datta, A., Fredrikson, M., Ko, G., Mardziel, P., Sen, S.: Proxy non-discrimination in data-driven systems. arXiv preprint arXiv:1707.08120 (2017). https://doi.org/10.48550/arXiv.1707.08120

[14] Durand, D.: Risk Elements in Consumer Instalment Financing, Technical Edition. National Bureau of Economic Research, New York, NY (1941), https://www.nber.org/books-and-chapters/risk-elements-consumer-instalment-financing-technical-edition

[15] Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap, Monographs on Statistics and Applied Probability, vol. 57. Chapman & Hall, New York (1993). https://doi.org/10.1201/9780429246593

[16] Equal Employment Opportunity Commission and Civil Service Commission and Department of Labor and Department of Justice: Uniform Guidelines on Employee Selection Procedures. 29 C.F.R. Part 1607, 43 Fed. Reg. 38290–38315 (1978), https://www.ecfr.gov/current/title-29/part-1607

[17] European Parliament and Council of the European Union: Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, OJ L, 2024/1689, 12.7.2024 (2024), https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[18] European Parliament and Council of the European Union: Regulation (EU) 2026/1744 of the European Parliament and of the Council of 8 July 2026 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI). Official Journal of the European Union, OJ L, 2026/1744, 24.7.2026 (2026), https://eur-lex.europa.eu/eli/reg/2026/1744/oj

[19] Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and Removing Disparate Impact. In: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 259–268 (2015). https://doi.org/10.1145/2783258.2783311

[20] Fisher, A., Rudin, C., Dominici, F.: All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. Journal of Machine Learning Research **20**(177), 1–81 (2019), https://jmlr.org/papers/v20/18-760.html

[21] Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., Walther, A.: Predictably unequal? The effects of machine learning on credit markets. The Journal of Finance **77**(1), 5–47 (2022). https://doi.org/10.1111/jofi.13090

[22] Gramegna, A., Giudici, P.: SHAP and LIME: An evaluation of discriminative power in credit risk. Frontiers in Artificial Intelligence **4**, 752558 (2021). https://doi.org/10.3389/frai.2021.752558

[23] Grömping, U.: South German Credit. Datensatz, UCI Machine Learning Repository (2019). https://doi.org/10.24432/C5X89F

[24] Grömping, U.: South German Credit Data: Correcting a Widely Used Data Set. Reports in Mathematics, Physics and Chemistry 4/2019, Beuth University of Applied Sciences Berlin (2019), https://www1.bht-berlin.de/FB_II/reports/Report-2019-004.pdf

[25] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys **51**(5), 93:1–93:42 (2018). https://doi.org/10.1145/3236009

[26] Hardt, M., Price, E., Srebro, N.: Equality of Opportunity in Supervised Learning. In: Advances in Neural Information Processing Systems. vol. 29, pp. 3315–3323 (2016), https://papers.nips.cc/paper_files/paper/2016/hash/6a9659feb1216f14f7384ba499518b38-Abstract.html

[27] Henderson, L., Herring, C., Horton, H.D., Thomas, M.: Credit Where Credit is Due?: Race, Gender, and Discrimination in the Credit Scores of Business Startups. The Review of Black Political Economy **42**(4), 459–479 (2015). https://doi.org/10.1007/s12114-015-9215-4

[28] Hlongwane, R., Ramaboa, K., Mongwe, W.: A novel framework for enhancing transparency in credit scoring: Leveraging Shapley values for interpretable credit scorecards. PLOS ONE **19**(8), e0308718 (2024). https://doi.org/10.1371/journal.pone.0308718

[29] Hofmann, H.: Statlog (German Credit Data). Datensatz, UCI Machine Learning Repository (1994). https://doi.org/10.24432/C5NC77

[30] Hooker, G., Mentch, L., Zhou, S.: Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. Statistics and Computing **31**(6), 82 (2021). https://doi.org/10.1007/s11222-021-10057-z

[31] Hurley, M., Adebayo, J.: Credit Scoring in the Era of Big Data. Yale Journal of Law and Technology **18**(1), 148–216 (2016), https://yjolt.org/credit-scoring-era-big-data

[32] Hurlin, C., Pérignon, C., Saurin, S.: The Fairness of Credit Scoring Models. Management Science **72**(1), 406–425 (2026). https://doi.org/10.1287/mnsc.2022.03888

[33] Jensen, H.L.: Using Neural Networks for Credit Scoring. Managerial Finance **18**(6), 15–26 (1992). https://doi.org/10.1108/eb013696

[34] Kalousis, A., Prados, J., Hilario, M.: Stability of feature selection algorithms: a study on high-dimensional spaces. Knowledge and Information Systems **12**(1), 95–116 (2007). https://doi.org/10.1007/s10115-006-0040-8

[35] Khan, F.S., Mazhar, S.S., Mazhar, K., AlSaleh, D.A., Mazhar, A.: Model-agnostic explainable artificial intelligence methods in finance: a systematic review, recent developments, limitations, challenges and future directions. Artificial Intelligence Review **58**(8), 232 (2025). https://doi.org/10.1007/s10462-025-11215-9

[36] Kleinberg, J., Mullainathan, S., Raghavan, M.: Inherent Trade-Offs in the Fair Determination of Risk Scores. In: Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS). Leibniz International Proceedings in Informatics (LIPIcs), vol. 67, pp. 43:1–43:23 (2017). https://doi.org/10.4230/LIPIcs.ITCS.2017.43

[37] Kłosok, M., Chlebus, M.: Towards better understanding of complex machine learning models using Explainable Artificial Intelligence (XAI) – case of Credit Scoring modelling. Working Paper 2020-18, University of Warsaw, Faculty of Economic Sciences (2020), https://ideas.repec.org/p/war/wpaper/2020-18.html

[38] Kozodoi, N., Jacob, J., Lessmann, S.: Fairness in credit scoring: Assessment, implementation and profit implications. European Journal of Operational Research **297**(3), 1083–1094 (2022). https://doi.org/10.1016/j.ejor.2021.06.023

[39] Kumar, I.E., Venkatasubramanian, S., Scheidegger, C., Friedler, S.: Problems with Shapley-value-based explanations as feature importance measures. In: Proceedings of the 37th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 119, pp. 5491–5500 (2020), https://proceedings.mlr.press/v119/kumar20e.html

[40] Lessmann, S., Baesens, B., Seow, H.V., Thomas, L.C.: Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research **247**(1), 124–136 (2015). https://doi.org/10.1016/j.ejor.2015.05.030

[41] Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I.: From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence **2**(1), 56–67 (2020). https://doi.org/10.1038/s42256-019-0138-9

[42] Lundberg, S.M., Erion, G.G., Lee, S.I.: Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888 (2018). https://doi.org/10.48550/arXiv.1802.03888

[43] Lundberg, S.M., Lee, S.I.: A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems. vol. 30, pp. 4765–4774 (2017), https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[44] Molnar, C.: Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Independently published, 3rd edn. (2025), https://christophm.github.io/interpretable-ml-book

[45] Montoya, A.M., Parrado, E., Solís, A., Undurraga, R.: Discriminación de género en el mercado de créditos de consumo en Chile. Serie de Políticas Públicas y Transformación Productiva 34, CAF – Banco de Desarrollo de América Latina, Caracas (2020), https://scioteca.caf.com/handle/123456789/1533

[46] Nogueira, S., Sechidis, K., Brown, G.: On the use of Spearman's rho to measure the stability of feature rankings. In: Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA). pp. 381–391. Springer (2017). https://doi.org/10.1007/978-3-319-58838-4_42

[47] Nwafor, C.N., Nwafor, O., Brahma, S.: Enhancing transparency and fairness in automated credit decisions: an explainable novel hybrid machine learning approach. Scientific Reports **14**, 25174 (2024). https://doi.org/10.1038/s41598-024-75026-8

[48] Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144 (2016). https://doi.org/10.1145/2939672.2939778

[49] Rice, L., Swesnik, D.: Discriminatory Effects of Credit Scoring on Communities of Color. Suffolk University Law Review **46**(3), 935–966 (2013), https://sites.suffolk.edu/lawreview/2013/12/19/discriminatory-effects-of-credit-scoring/

[50] Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x

[51] Salih, A.M., Raisi-Estabragh, Z., Galazzo, I.B., Radeva, P., Petersen, S.E., Lekadir, K., Menegaz, G.: A perspective on explainable artificial intelligence methods: SHAP and LIME. Advanced Intelligent Systems **7**(1), 2400304 (2025). https://doi.org/10.1002/aisy.202400304

[52] Shapley, L.S.: A Value for n-Person Games. In: Kuhn, H.W., Tucker, A.W. (eds.) Contributions to the Theory of Games, Volume II, pp. 307–318. Princeton University Press, Princeton (2016). https://doi.org/10.1515/9781400881970-018

[53] Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H.: Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES). pp. 180–186 (2020). https://doi.org/10.1145/3375627.3375830

[54] Verma, S., Rubin, J.: Fairness Definitions Explained. In: Proceedings of the International Workshop on Software Fairness (FairWare). pp. 1–7 (2018). https://doi.org/10.1145/3194770.3194776

[55] Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology **31**(2), 841–887 (2018), https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf

[56] Weerts, H., Kelly-Lyth, A., Binns, R., Adams-Prassl, J.: Unlawful Proxy Discrimination: A Framework for Challenging Inherently Discriminatory Algorithms. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT). pp. 1850–1860 (2024). https://doi.org/10.1145/3630106.3659010

[57] Zhang, J., Bareinboim, E.: Fairness in Decision-Making – The Causal Explanation Formula. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32, pp. 2037–2045 (2018). https://doi.org/10.1609/aaai.v32i1.11564