AStA Wirtschafts- und Sozialstatistisches Archiv
https://doi.org/10.1007/s11943-024-00346-0

ORIGINALVERÖFFENTLICHUNG

Fairness als Qualitätskriterium im Maschinellen Lernen – Rekonstruktion des philosophischen Konzepts und Implikationen für die Nutzung außergesetzlicher Merkmale bei qualifizierten Mietspiegeln

Ludwig Bothmann · Kristina Peters

Eingegangen: 13. April 2023 / Angenommen: 23. Juli 2024

Zusammenfassung Mit der verstärkten Nutzung von Modellen des Maschinellen Lernens (ML) innerhalb von Systemen der automatisierten Entscheidungsfindung wachsen die Anforderungen an die Qualität von ML-Modellen. Die reine Prognosegüte ist nicht länger das alleinige Qualitätskriterium; insbesondere wird vermehrt gefordert, dass Fairnessaspekte berücksichtigt werden. Dieser Beitrag verfolgt zwei Ziele. Zum einen werden die aktuelle Fairnessdiskussion im Bereich ML (fairML) zusammengefasst und die aktuellsten Entwicklungen, insbesondere in Bezug auf die philosophischen Grundlagen des Fairnessbegriffs innerhalb ML, beschrieben. Zum anderen wird die Frage behandelt, inwiefern sogenannte „außergesetzliche" Merkmale bei der Erstellung qualifizierter Mietspiegel genutzt werden dürfen. Ein aktueller Vorschlag von Kauermann und Windmann (AStA Wirtschafts- und Sozialstatistisches Archiv, Band 17, 2023) zur Nutzung außergesetzlicher Merkmale in qualifizierten Mietspiegeln beinhaltet eine modellbasierte Imputationsmethode, welche wir den gesetzlichen Anforderungen gegenüberstellen. Schließlich zeigen wir auf, welche Alternativen aus dem Bereich fairML genutzt werden könnten und legen dar, welche unterschiedlichen philosophischen Grundannahmen hinter den verschiedenen Verfahren stehen.

Schlüsselwörter Amtliche Statistik · Automatisierte Entscheidungsfindung · Fairness · Maschinelles Lernen · Mietspiegel

✉ Ludwig Bothmann
Department of Statistics, LMU Munich, München, Deutschland
E-Mail: ludwig.bothmann@stat.uni-muenchen.de

Munich Center for Machine Learning (MCML), München, Deutschland

Kristina Peters
Faculty of Law, LMU Munich, München, Deutschland
E-Mail: kristina.peters@jura.uni-muenchen.de

Published online: 23 August 2024

Fairness as a quality criterion in machine learning – Reconstruction of the philosophical concept and implications for the use of extra-legal features in qualified rent indices

Abstract With the increased use of machine learning (ML) models within automated decision-making systems, the demands on the quality of ML models are growing. Pure prediction quality is no longer the sole quality criterion; in particular, there is an increasing demand to consider fairness aspects. This paper pursues two goals. First, it summarizes the current fairness discussion in the field of ML (fairML) and describes the most recent developments, especially with respect to the philosophical foundations of the concept of fairness within ML. On the other hand, the question is addressed to what extent so-called 'extra-legal' characteristics may be used in the compilation of qualified rent indices. A recent proposal by Kauermann and Windmann (AStA Wirtschafts- und Sozialstatistisches Archiv, Volume 17, 2023) on using extra-legal features in qualified rent indices includes a model-based imputation method, which we contrast with the legal requirements. Finally, we show which alternatives from the field of fairML could be used and outline the different basic philosophical assumptions behind the various methods.

Keywords Official Statistics · Automated Decision Making · Fairness · Machine Learning · Rent Indices

1 Fairness im Maschinellen Lernen

Systeme der automatisierten Entscheidungsfindung (automated decision making – ADM) werden immer populärer; sie werden beispielsweise zur Bewertung von Kreditrisiken, zur vorausschauenden Gefahrenabwehr (Predictive Policing) oder zur Zulassung von Bewerber:innen zu einem Universitätsstudium eingesetzt.[1] Diese ADM-Systeme können direkt individuelle Lebensumstände beeinflussen. Eine naheliegende Forderung in Bezug auf die Qualität dieser Systeme ist daher, dass sie „fair" sein sollen. Da eine Kernkomponente eines ADM-Systems ein Modell Maschinellen Lernens (ML-Modell) sein kann, überträgt sich die Forderung nach Fairness auf das ML-Modell. In diesem Beitrag wollen wir aktuelle Entwicklungen in der Debatte um Fairness im Maschinellen Lernen (ML) systematisieren, die damit zusammenhängenden Überlegungen in die deutsche Diskussion einbringen und aufzeigen, welche Implikationen diese für die Nutzung außergesetzlicher Merkmale bei der Erstellung qualifizierter Mietspiegel hat.

Als Beispiel zur Darstellung der Debatte um Fairness in ML in Abschn. 2 nutzen wir COMPAS (Correctional Offender Management Profiling for Alternative Sanctions).[2] Dieses von Northpointe entwickelte System soll Richter:innen bei ver-

[1] siehe https://algorithmwatch.org/en/wp-content/uploads/2019/02/Automating_Society_Report_2019.pdf (Aufrufdatum 31.01.2024) für einen aktuellen Überblick über ADM-Systeme in einigen EU-Ländern.

[2] https://s3.documentcloud.org/documents/2840784/Practitioner-s-Guide-to-COMPAS-Core.pdf (Aufrufdatum 31.01.2024).

schiedenen gerichtlichen Entscheidungen unterstützen. Die Grundidee ist – vereinfacht dargestellt –, dass ein ML-Modell die Wahrscheinlichkeit $\pi^{(i)} = \mathbb{P}(y^{(i)} = 1|\mathcal{H}_i)$ vorhersagen soll, dass die angeklagte Person i in den kommenden zwei Jahren „rückfällig" wird ($y^{(i)} \in \{0,1\}$), also eine Ordnungswidrigkeit oder eine Straftat begehen wird. Hierbei beschreibt $\mathcal{H}_i$ die Menge aller Charakteristika und der Vorgeschichte des Individuums i – welche in einem empirischen Anwendungsfall durch einen Kovariablenvektor $\mathbf{x}^{(i)}$ lediglich angenähert werden kann. Die Behandlung $t^{(i)}$ der Person (also die gerichtliche Entscheidung) basiert dann auf der vorhergesagten Wahrscheinlichkeit $\widehat{\pi}(\mathbf{x}^{(i)})$.

COMPAS wurde vorgeworfen, es sei unfair in dem Sinne, dass das System unter anderem einen Bias bezüglich Schwarzer Menschen hat (Angwin et al. 2016); diesem Vorwurf trat Northpointe entgegen und legte dar, wieso dies aus ihrer Sicht nicht der Fall sei; einen guten Einblick in die Debatte gibt (Corbett-Davies et al. 2016). Der Dissens in dieser Frage rührt daher, dass der Begriff „Fairness" unterschiedlich verstanden wird, und dass deshalb unterschiedliche Metriken zur Messung von Fairness zugrunde gelegt werden.

FairML. Der Bereich des Fairness-bewussten ML (fairness-aware ML – fairML) setzt sich genau mit der Frage auseinander, wie man die ML-bezogene Unfairness in einem ADM-System messen und abmildern kann. Die Relevanz des Themas spiegelt sich in einer Fülle von aktuellen Beiträgen zu den großen ML-Konferenzen; auch hat sich etwa mit der „Conference on Fairness, Accountability, and Transparency" (FAccT) eine Konferenz etabliert, die sich speziell mit den Themen Fairness, Verantwortlichkeit und Transparenz beschäftigt.[3]

Typischerweise haben Beiträge in diesem Bereich zwei Ziele, nämlich (a) eine Metrik vorzuschlagen, mit der die (Un-)Fairness eines ML-Modells gemessen wird, und (b) eine Methode vorzuschlagen, mit der sichergestellt werden soll, dass das resultierende ML-Modell einen guten Wert bezüglich dieser Metrik erreicht.

Dabei haben sich zwei Gruppen von Metriken herauskristallisiert. Auf der einen Seite stehen die **gruppenbasierten Fairnessmetriken**: Hier werden oft bedingte Wahrscheinlichkeiten in bestimmten Subgruppen verglichen. Wenn A die *geschützten Merkmale* (protected attributes – PA) bezeichnet (also die in besonderer Weise geschützten Merkmale wie im COMPAS-Beispiel die als binär angenommene Ethnie [b: Black / w: White]), X andere Merkmale, C vorhergesagte Wahrscheinlichkeitsscores und $\widehat{y}$ die vorhergesagte Zielgröße, können einige Fairnessmetriken wie folgt definiert werden (siehe Verma und Rubin 2018, für einen ausführlichen Überblick):

- Statistische Parität (Statistical parity – SP): $\mathbb{P}(\widehat{y} = 1|A = w) = \mathbb{P}(\widehat{y} = 1|A = b)$.
- Bedingte SP (Conditional SP): $\mathbb{P}(\widehat{y} = 1|A = w, X = x) = \mathbb{P}(\widehat{y} = 1|A = b, X = x)$.
- Wohlkalibrierung (Well-calibration): $\mathbb{P}(y = 1|A = w, C = c) = \mathbb{P}(y = 1|A = b, C = c) = c$.

[3] https://facctconference.org/2024/ (Aufrufdatum 31.01.2024).

Auf der anderen Seite stehen die **individuellen Fairnessmetriken**, von denen wir zwei besonders erwähnen möchten. *Fairness durch Aufmerksamkeit* (Fairness through awareness, Dwork et al. 2012) verfolgt die Idee, dass ähnliche Individuen ähnlich behandelt werden sollten („similar individuals should be treated similarly"). *Kontrafaktische Fairness* (Counterfactual fairness, Kusner et al. 2017) führt den Begriff der Kausalität in die Diskussion ein und definiert Behandlungen als fair „if it is the same in (a) the actual world and (b) a counterfactual world where the individual belonged to a different demographic group". Das Konzept der Kausalität ist seither stark in die Fairnessdebatte eingebunden, siehe Makhlouf et al. (2022) für einen Überblick.

Wenig überraschend ist, dass es Kompatibilitätsprobleme gibt, also dass nicht bezüglich aller Metriken gleichzeitig Fairness erreicht werden kann. Ebensowenig überrascht, dass es in Bezug auf die gängigsten Fairnessmetriken einen Tradeoff zwischen Fairness und Vorhersagegüte gibt. Vielmehr war letzteres zu erwarten, da das ML-Modell durch Forderungen wie (bedingte) statistische Parität in seiner Flexibilität (stark) eingeschränkt wird.

Für den vorliegenden Beitrag relevanter ist jedoch der Umstand, dass zwar viele Fairnessmetriken vorgeschlagen werden – also Messinstrumente bezüglich des Konzeptes „Fairness"; die Frage, was Fairness konkret bedeutet – was also die philosophische Basis des Konzepts Fairness ist –, wurde im Bereich fairML bis vor kurzem aber kaum beleuchtet. Insbesondere aus diesem Grund stehen die vorgeschlagenen Metriken relativ unverbunden nebeneinander. Zudem ist oft nicht klar ersichtlich, auf welches Konzept die verschiedenen Metriken abzielen; wenn das zu messende Konzept aber nicht klar definiert ist, ist es schwierig zu evaluieren inwiefern die Metrik zur Messung dieses Konzepts überhaupt geeignet ist. Um einen tragfähigen Beitrag zur gesellschaftlichen Diskussion bezüglich des Einsatzes von ADM-Systemen leisten zu können, muss aus unserer Sicht diese Auseinandersetzung und die Anbindung der Fairnessmetriken an ein anschlussfähiges philosophisches Konzept stattfinden. Die Überführung der verschiedenen Ansätze in ein einheitliches Konzept hat auch zum Ziel, den Dialog zwischen denjenigen zu erleichtern, die sich um Fairness von ADM-Systemen bemühen. Ohne einen solchen Dialog besteht die Gefahr, dass sich die Diskussion im Bereich ML von der gesellschaftlichen Debatte entfernt und als „technische Fingerübung" keine Relevanz mehr für die real existierenden Fragestellungen hat.

Im nächsten Abschnitt werden wir das von Bothmann et al. (2024) vorgestelle Konzept der Fairness in ML und dessen Implikationen für ML-Modelle in fairen ADM-Systemen erläutern. Daran anschließend werden wir in Abschn. 3 intensiv auf die Nutzung außergesetzlicher Merkmale bei der Erstellung von Mietspiegeln eingehen. Hierbei werden wir zunächst die gesetzliche Grundlage beleuchten, dann aktuell vorgeschlagene Methoden diesbezüglich darstellen und die Bezüge zur Fairnessdebatte in ML aufzeigen. Schließlich legen wir dar, wie ein Mietspiegel erstellt werden müsste, der ebendiese Fairnessaspekte als Qualitätsanforderungen berücksichtigt.

2 Philosophische Grundlagen und Übertragung auf ML

In Bothmann et al. (2024) wurde der Versuch unternommen, die philosophischen Grundlagen des Konzepts Fairness zu formalisieren und daraus Implikationen für die weitere Beschäftigung mit dem Thema innerhalb der fairML abzuleiten. Im Folgenden wollen wir die Ergebnisse dieser Überlegungen kurz zusammenfassen, um diese für die Diskussion über den Umgang mit außergesetzlichen Merkmalen nutzen zu können. Für eine ausführliche Darstellung sei auf die Originalquelle (Bothmann et al. 2024) verwiesen.

2.1 Grundkonzept der Fairness

Die Idee von Fairness lässt sich bis zu Aristoteles zurückverfolgen (Aristoteles 1831). Demnach ist eine Handlung fair, wenn gilt: „Gleiche werden gleich behandelt, Ungleiche werden ungleich behandelt." Das allgemeine Verständnis von **Fairness** ist regelmäßig dadurch geprägt, dass es um eine Behandlung von Menschen durch Menschen geht. Hieraus folgt, dass ein ML-Modell per se nicht fair oder unfair sein kann; als fair oder unfair können nur die Handlungen bezeichnet werden, die basierend auf den ML-Prognosen ausgeführt werden.

Die Konkretisierung dessen, was mit „Fairness" bezeichnet wird, hängt davon ab, was den Bezugspunkt der Bewertung bilden soll: Soll (a) eine Sache oder ein sonstiges Gut verteilt werden, ohne dass es auf die betroffenen Personen ankommen soll? Oder soll hierbei (b) berücksichtigt werden, welche Individuen konkret betroffen sind? Während im ersten Fall die Fairness mittels einer einfachen Rechnung ermittelt werden kann – Aristoteles spricht hier von der **arithmetischen Proportionalität** –, hängt sie im zweiten Fall von der Beurteilung der Personen ab. In diesem zweiten Fall muss die von Aristoteles sogenannte „Würdigkeit" der Personen festgelegt werden (man kann dies als task-spezifische Gleichheit interpretieren), um sodann die sogenannte **geometrische Proportionalität** zu ermitteln. Die Behandlung $t^{(i)}$ einer Person i soll dann proportional zu ihrer Würdigkeit $w^{(i)}$ sein, also

$$t^{(i)} = k \cdot w^{(i)}. \tag{1}$$

Im Fall der arithmetischen Proportionalität wird davon ausgegangen, dass die natürliche Unterschiedlichkeit der betroffenen Person für ihre Behandlung keine Rolle spielt – beispielsweise bei Kaufverträgen, bei denen sich der Preis nur aus der Ware bestimmt und nicht zusätzlich aus den Merkmalen der Käuferin –, also

$$t^{(i)} = k \quad \forall i \in \{1, ..., n\}. \tag{2}$$

Das Konzept der geometrischen Proportionalität kann verallgemeinert werden, indem für die Behandlungsfunktion $s(\cdot)$ nicht nur lineare Funktionen der Würdigkeit zugelassen werden, sondern allgemein monotone Funktionen, so dass

$$t^{(i)} = s(w^{(i)}), \quad s(\cdot) \quad \text{monoton}. \tag{3}$$

Darüber hinaus könnte sogar strenge Monotonie gefordert werden, damit wirklich immer Ungleiche ungleich behandelt werden. In der Realität haben sich aber Konstellationen durchgesetzt, in denen Plateaupunkte als fair angesehen werden, wenn z. B. der Steuersatz mit dem Einkommen zunächst steigt, ab einer gewissen Grenze aber konstant ist. In diesem Zusammenhang sind zwei normative Entscheidungen zu treffen, nämlich (a) wie die Würdigkeit oder task-spezifische Gleichheit im vorliegenden Fall zu definieren ist und (b) wie die Behandlungsfunktion $s(w^{(i)})$ konkret ausgestaltet ist. Gemäß Bothmann et al. (2024) kann eine faire Behandlung also wie folgt definiert werden:

Definition 1 (Fair treatment) Die Behandlung $t^{(i)}$ eines Individuums i wird **fair** genannt, genau dann wenn sie als normative Funktion der Würdigkeit des Individuums $w^{(i)}$ abgeleitet ist, also, $\exists s(\cdot)$ sodass $t^{(i)} = s(w^{(i)})$, wobei $s(\cdot)$ eine (streng) monotone Funktion ist.

Die Aufgabe von ML oder der Statistik ist es nun, die Würdigkeit $w^{(i)}$ – als Entscheidungsgrundlage für eine Behandlung – möglichst gut zu schätzen; typischerweise wird die Würdigkeit eine (Funktion einer) Erfolgswahrscheinlichkeit $\pi^{(i)}$ oder ein(es) Erwartungswert(s) $\mu^{(i)}$ sein. Wie bereits erwähnt, kann ML-Modell also per se nicht unfair sein, kann aber – durch schlechte Schätzung – Unfairness induzieren, wie Bothmann et al. (2024) wie folgt begründen:

Durch die Reduktion der (potentiell unendlichdimensionalen) Information bezüglich eines Individuums auf eine endliche Menge von Variablen $\mathbf{x}^{(i)}$ einerseits und über die Schätzung des Zusammenhangs dieser Variablen mit der Würdigkeit, also der Schätzung von $\pi(\mathbf{x}^{(i)})$ andererseits, wird Unschärfe eingeführt. Dies kann dazu führen, dass die vorhergesagte Würdigkeit einer Person $\widehat{\pi}(\mathbf{x}^{(i)})$ nicht der wahren Würdigkeit $\pi^{(i)}$ entspricht, also $\pi^{(i)} \neq \widehat{\pi}(\mathbf{x}^{(i)})$, was bei streng monotonem $s(\cdot)$ notwendig zu einer unfairen Behandlung $t^{(i)} = s(\widehat{\pi}(\mathbf{x}^{(i)}))$ führt. Bothmann et al. (2024) definieren in diesem Zusammenhang den Begriff der **individuellen Wohlkalibrierung** (individual well-calibration), die eingehalten ist, wenn $\pi^{(i)} = \widehat{\pi}(\mathbf{x}^{(i)})$. Natürlich kann in der Praxis aufgrund der genannten Unschärfen strikte *individuelle Wohlkalibrierung* nie erreicht werden, weshalb im Grunde jede Behandlung, die auf einem ML-Modell basiert, als unfair angesehen werden muss. Die Definition einer gewissen, normativ als unproblematisch eingestuften Differenz zwischen $\pi^{(i)}$ und $\widehat{\pi}(\mathbf{x}^{(i)})$ kann hier Abhilfe schaffen.

2.2 Behandlung geschützter Merkmale

Einen besonderen Stellenwert nehmen in fairML die PAs ein, die sich beispielsweise aus Art. 3 GG ergeben können. Dies sind Merkmale, auf Basis derer Individuen nicht diskriminiert werden dürfen. Nun ist es natürlich inhärentes Ziel eines ML-Modells, unterschiedlichen Individuen unterschiedliche Scores $\widehat{\pi}(\mathbf{x}^{(i)})$ zuzuweisen – die möglichst nah an den wahren Scores $\pi^{(i)}$ sind –, basierend auf den vorhandenen Merkmalen, also in einem nicht wertenden Sinn zu diskriminieren. Die von Bothmann et al. (2024) gegebenen Definitionen der deskriptiv bzw. normativ unfairen Behandlung bringen Klarheit darüber, was den Kern der PAs formal

ausmacht (wie oben werden PAs mit A bezeichnet, während X andere, nicht geschützte Merkmale bezeichnet):

Definition 2 (Deskriptiv unfaire Behandlung) Nehmen wir ein Paar von Individuen i und j an, die sich nur in Bezug auf das Merkmal X unterscheiden. Angenommen, das Merkmal X ist kein kausaler Grund für einen Unterschied in den wahren Wahrscheinlichkeiten, d. h. $\pi^{(i)} = \pi^{(j)}$. Eine Behandlung wird **deskriptiv unfair in Bezug auf das Merkmal** X genannt, wenn diese Individuen in einem Prozess – aufgrund von unterschiedlichen geschätzten individuellen Wahrscheinlichkeiten $\widehat{\pi}^{(i)} \neq \widehat{\pi}^{(j)}$ – unterschiedlich behandelt werden, d. h. $t^{(i)}(= s(\widehat{\pi}^{(i)})) \neq t^{(j)}(= s(\widehat{\pi}^{(j)}))$.

Im Zusammenhang mit COMPAS kann das beispielsweise bedeuten: Zwei Individuen i und j unterscheiden sich ausschließlich bezüglich der Herkunft X. Sollte (a) die Herkunft nicht kausal sein für einen Unterschied in der Rückfallwahrscheinlichkeit, also $\pi^{(i)} = \pi^{(j)}$, dann wäre eine Behandlung aufgrund $\widehat{\pi}^{(i)} \neq \widehat{\pi}^{(j)}$ unfair. Sollte (b) die Herkunft allerdings kausal sein, also $\pi^{(i)} \neq \pi^{(j)}$, dann wäre eine Behandlung aufgrund $\widehat{\pi}^{(i)} \neq \widehat{\pi}^{(j)}$ nicht als unfair anzusehen – Ungleiche würden ungleich behandelt werden.

Den Kern der Idee geschützter Merkmale macht allerdings aus, dass auch obiger Fall (b) normativ als unfair angesehen werden kann. Beispielsweise soll der Umstand (so er denn wahr sein sollte), dass Herkunft einen kausalen Effekt auf die Rückfallwahrscheinlichkeit hat, nicht zulasten der betroffenen Individuen gehen; möglicherweise ist dieser Zusammenhang Folge von historischem Rassismus, für welchen die Mitglieder der benachteiligten Gruppe nicht die Verantwortung tragen sollen. Wie Bothmann et al. (2024) zeigen, passt allerdings auch diese Idee in das Framework, und zwar mittels Vorstellung einer fiktiven, normativ gewünschten Welt, in der diese ungewünschte Kausalität nicht existiert:

Definition 3 (Normativ unfaire Behandlung) Nehmen wir ein Paar von Individuen i und j an, die sich nur in Bezug auf das Merkmal A unterscheiden. Nehmen wir ferner an, dass Merkmal A ein kausaler Grund für einen Unterschied in den wahren Wahrscheinlichkeiten ist, d. h. $\pi^{(i)} \neq \pi^{(j)}$, und dass Merkmal A ein PA ist. Eine Behandlung wird **normativ unfair in Bezug auf das Merkmal** A genannt, wenn diese Individuen in einem Prozess aufgrund unterschiedlicher geschätzter individueller Wahrscheinlichkeiten $\widehat{\pi}^{(i)} \neq \widehat{\pi}^{(j)}$ unterschiedlich behandelt werden, d. h. $t^{(i)}(= s(\widehat{\pi}^{(i)})) \neq t^{(j)}(= s(\widehat{\pi}^{(j)}))$, da Merkmal A nicht herangezogen werden darf für die Bestimmung der Gleichheit, d. h. der Entscheidungsgrundlage für die Behandlung.

Im Beispiel mit COMPAS ist $\pi^{(i)}$ die Rückfallwahrscheinlichkeit in der wahren Welt, $\widetilde{\pi}^{(i)}$ die Rückfallwahrscheinlichkeit in der fiktiven Welt (in der die Herkunft nicht kausal auf die Rückfallwahrscheinlichkeit wirkt). Wir betrachten wieder Individuen i and j, die sich nur bezüglich der Herkunft A unterscheiden. Wenn diese sich bezüglich der Rückfallwahrscheinlichkeiten $\pi^{(i)} \neq \pi^{(j)}$ unterscheiden, wäre eine ungleiche Behandlung aus normativen Gründen dennoch unfair. Die wahre **korrigierte** Rückfallwahrscheinlichkeit $\widetilde{\pi}^{(i)} = \widetilde{\pi}^{(j)}$ wäre aber identisch (selbst

wenn die Herkunft in der wahren Welt kausal wäre). Eine Entscheidung sollte also auf Grundlage der Schätzungen der $\widetilde{\pi}^{(i)}$ getroffen werden.

Eine methodische Herausforderung bleibt die Frage, wie aus vorhandenem Datenmaterial die korrigierten Rückfallwahrscheinlichkeiten in der fiktiven Welt geschätzt werden sollen. Hierzu geben Bothmann et al. (2024) erste Hinweise auf konkrete Algorithmen, die in Bothmann et al. (2023) detaillierter ausgeführt sind und uns hier aber zunächst nicht weiter beschäftigen sollen.

Wenden wir uns nun der Thematik des Mietspiegels zu, bei der wir die außergesetzlichen Merkmale als PA definieren werden.

3 Implikationen für qualifizierte Mietspiegel

In Deutschland unterscheidet das Gesetz sogenannte „einfache“ und „qualifizierte“ Mietspiegel. Während an einen einfachen Mietspiegel kaum Anforderungen gestellt werden, ist ein qualifizierter Mietspiegel „nach anerkannten wissenschaftlichen Grundsätzen erstellt“[4]. Ferner wird unterschieden zwischen Tabellen- und Regressionsmietspiegeln, wobei wir uns im Folgenden ausschließlich auf Regressionsmietspiegel fokussieren (für einen methodischen Vergleich dieser Mietspiegelarten siehe etwa Kauermann und Windmann 2016). Eine besondere Rolle nehmen die sogenannten „außergesetzlichen Merkmale“ ein, die für die Feststellung der ortsüblichen Vergleichsmiete nicht genutzt werden dürfen (zur Begriffsdefinition siehe unten).

Wir fassen zunächst in Abschn. 3.1 die aktuell gültige Rechtslage zusammen, beschreiben in Abschn. 3.2 das in Kauermann und Windmann (2023) vorgeschlagene statistische Vorgehen zur Behandlung außergesetzlicher Merkmale zur Erstellung eines qualifizierten Mietspiegels, beleuchten in Abschn. 3.3 inwieweit und mit welchen Mitteln die gesetzlichen Anforderungen bei diesem Vorgehen umgesetzt werden und beschreiben schließlich in Abschn. 3.4, wie der aktuelle Forschungsstand im Bereich FairML für die Erstellung qualifizierter Regressionsmietspiegel nutzbar gemacht werden kann.

3.1 Gesetzliche Forderung

Den gesetzlichen Rahmen für die Bildung der ortsüblichen Vergleichsmiete bildet §558 Absatz 2 BGB (Bürgerliches Gesetzbuch):[5]

> „Die ortsübliche Vergleichsmiete wird gebildet aus den üblichen Entgelten, die in der Gemeinde oder einer vergleichbaren Gemeinde für Wohnraum vergleichbarer Art, Größe, Ausstattung, Beschaffenheit und Lage einschließlich der energetischen Ausstattung und Beschaffenheit in den letzten sechs Jahren vereinbart [...] worden sind.“

[4] BGB §558d, https://www.gesetze-im-internet.de/bgb/__558d.html (Aufrufdatum 31.01.2024).

[5] BGB §558, https://www.gesetze-im-internet.de/bgb/__558.html (Aufrufdatum 31.01.2024).

Ziel eines „einfachen“ und „qualifizierten“ Mietspiegels ist es, diese ortsübliche Vergleichsmiete zu bestimmen. Der Bundesregierung ermöglicht §558c Absatz 5, diese allgemein gehaltene Vorgabe zu konkretisieren:[6]

> „Die Bundesregierung wird ermächtigt, durch Rechtsverordnung mit Zustimmung des Bundesrates Vorschriften zu erlassen über den näheren Inhalt von Mietspiegeln und das Verfahren zu deren Erstellung und Anpassung einschließlich Dokumentation und Veröffentlichung.“

Von dieser Möglichkeit hat die Bundesregierung mit ihrer Verordnung vom 28.10.2021 mit Inkrafttreten zum 01.07.2022 Gebrauch gemacht, welche die Anforderungen an einen qualifizierten Mietspiegel konkretisiert.[7] Insbesondere legt die Verordnung in Form einer Positivliste fest, wozu die sogenannten „außergesetzlichen Merkmale“ genutzt werden dürfen. Außergesetzliche Merkmale sind dabei

> „Merkmale in Bezug auf die Wohnung oder das Mietverhältnis, die in §558 Absatz 2 Satz 1 des Bürgerlichen Gesetzbuchs nicht genannt sind, aber dennoch für die Mietpreisbildung relevant sind oder im Erstellungsstadium des Mietspiegels relevant sein können.“[8]

Demnach dürfen außergesetzliche Merkmale „insbesondere zur Wahl des Regressionsmodells [...] herangezogen werden“.[9] Wozu diese nicht benutzt werden dürfen, wird in der Verordnung selbst nicht explizit gemacht, kann aber aus der Begründung[10] geschlossen werden. Demnach ist die „direkte Berücksichtigung solcher Faktoren bei der Feststellung der Einzelvergleichsmiete“ nicht zulässig, es muss also „durch statistische Verfahren sichergestellt werden, dass außergesetzliche Merkmale bei der Feststellung der Einzelvergleichsmiete nicht direkt berücksichtigt werden und dadurch keine Verzerrung der Ergebnisse entsteht“.[11] Leider wird auch hier nicht näher darauf eingegangen, was genau als „direkte“ Berücksichtigung gilt und was nicht. Bei der Erstellung eines qualifizierten Mietspiegels muss diese Anforderung also interpretiert und in ein konkretes Modell überführt werden.

3.2 Außergesetzliche Merkmale als fehlende Werte

Kauermann und Windmann (2023) beschreiben Verfahren, mit denen die oben genannten gesetzlichen Anforderungen konkret umgesetzt werden können, und eva-

[6] BGB §558c, https://www.gesetze-im-internet.de/bgb/__558c.html (Aufrufdatum 31.01.2024).

[7] *Verordnung über den Inhalt und das Verfahren zur Erstellung und zur Anpassung von Mietspiegeln sowie zur Konkretisierung der Grundsätze für qualifizierte Mietspiegel*, http://www.bgbl.de/xaver/bgbl/start.xav?startbk=Bundesanzeiger_BGBl&jumpTo=bgbl121s4779.pdf (Aufrufdatum 31.01.2024).

[8] ebd. §2 Abs. 2.

[9] ebd. §14 Abs. 1.

[10] *Verordnung über den Inhalt und das Verfahren zur Erstellung und zur Anpassung von Mietspiegeln sowie zur Konkretisierung der Grundsätze für qualifizierte Mietspiegel*, Drucksache 766/20 vom 17.12.2020, https://dserver.bundestag.de/brd/2020/0766-20.pdf. Weitere Dokumente zu dem Gesetzgebungsverfahren können abgerufen werden unter: https://dip.bundestag.de/vorgang/verordnung-ueber-den-inhalt-und-das-verfahren-zur-erstellung-und/271557 (Aufrufdatum 31.01.2024).

[11] ebd., S. 35.

luieren die resultierenden Modelle empirisch. Wir fassen die dort beschriebenen Ansätze und Ergebnisse im Folgenden kurz zusammen und verweisen für eine detailliertere Beschreibung auf die Originalquelle.

Notation. Der zugrunde liegende Datensatz wird bezeichnet mit $\mathcal{D} = ((\mathbf{x}^{(1)}, \mathbf{z}^{(1)}, y^{(1)}), ..., (\mathbf{x}^{(n)}, \mathbf{z}^{(n)}, y^{(n)}))$, wobei $y^{(i)}$ die Nettomiete pro qm in EUR, $\mathbf{x}^{(i)}$ den Vektor der gesetzlichen Merkmale (bspw. Wohnfläche und Lage) und $\mathbf{z}^{(i)}$ den Vektor der außergesetzlichen Merkmale (bspw. Mietdauer) von Wohnung i bezeichnen. Die Vektoren $\mathbf{x}^{(i)}$ und $\mathbf{z}^{(i)}$ können auch Transformationen der ursprünglichen Merkmale beinhalten, so dass im Folgenden auch nichtlineare Effekte mit linearen Modellen modelliert werden können, z. B. via Splines.

Modelle. Es werden zwei lineare Modelle zur Modellierung der Nettomiete vorgeschlagen, (M0) berücksichtigt alle verfügbaren Merkmale, (M1) lediglich die gesetzlichen Merkmale:

$$\text{(M0):} \quad y^{(i)} = \beta_0 + \boldsymbol{\beta}_x \mathbf{x}^{(i)} + \boldsymbol{\beta}_z \mathbf{z}^{(i)} + \epsilon_0^{(i)} \tag{4}$$

$$\text{(M1):} \quad y^{(i)} = \theta_0 + \boldsymbol{\theta}_x \mathbf{x}^{(i)} + \epsilon_1^{(i)}, \tag{5}$$

wobei für die Fehlerterme angenommen wird, dass sie unabhängig und identisch normalverteilt sind mit Erwartungswert 0 und homoskedastischer Varianz, also $\epsilon_m^{(i)} \overset{iid}{\sim} N(0, \sigma^2) \forall i \in \{1, ..., n\}, m \in \{0,1\}$. Die Parameter $\beta_0, \boldsymbol{\beta}_x, \boldsymbol{\beta}_z, \theta_0, \boldsymbol{\theta}_x$ können per Kleinste-Quadrate-Schätzung oder Maximum-Likelihood-Schätzung basierend auf Trainingsdaten geschätzt werden, wobei die Schätzung in M1 unter einem *omitted variable bias* leiden wird, sofern die Merkmale $\mathbf{x}$ uns $\mathbf{z}$ korreliert sind (siehe auch die formalere Behandlung in Kauermann und Windmann 2023); die Schätzer werden bezeichnet mit $\widehat{\beta}_0$ etc.

Direkte Prognosen aus diesen Modellen ergeben sich für eine neue Beobachtung $(\mathbf{x}^{(*)}, \mathbf{z}^{(*)})$ zu:

$$\text{(P0):} \quad \widehat{y}^{(*)} = \widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}_x \mathbf{x}^{(*)} + \widehat{\boldsymbol{\beta}}_z \mathbf{z}^{(*)} \tag{6}$$

$$\text{(P1):} \quad \widehat{y}^{(*)} = \widehat{\theta}_0 + \widehat{\boldsymbol{\theta}}_x \mathbf{x}^{(*)} \tag{7}$$

Da bei Prognose (P0) der Vektor der außergesetzlichen Merkmale der Wohnung, $\mathbf{z}^{(*)}$, direkt benutzt wird, widerspricht dieses Vorgehen klar den gesetzlichen Vorgaben; es dient aber als Benchmark um andere Modelle hinsichtlich der Prognosegüte evaluieren zu können. Um die Information der außergesetzlichen Merkmale im Trainingsprozess gesetzeskonform zu nutzen – verbunden mit der Hoffnung eines prognostisch stärkeren Modells im Vergleich zu (M1)/(P1) –, schlagen Kauermann und Windmann (2023) vor, dennoch Modell (M0) zu schätzen, zum Prognosezeitpunkt allerdings nicht die tatsächlichen außergesetzlichen Merkmale $\mathbf{z}^{(*)}$ zu verwenden, sondern diese als fehlende Werte zu betrachten und Imputationen derselben in (6) einzusetzen. Unter Verwendung von Mittelwertimputation (P2, bereits früher zu fin-

den bei Malottki et al., 2018) bzw. modellbasierter Imputation (P3) leiten sie zwei weitere Prognosemodelle her:

$$\text{(P2):} \quad \widehat{y}^{(*)} = \widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}_x \mathbf{x}^{(*)} + \widehat{\boldsymbol{\beta}}_z \bar{\mathbf{z}}, \quad \text{mit} \quad \bar{\mathbf{z}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}^{(i)} \tag{8}$$

$$\text{(P3):} \quad \widehat{y}^{(*)} = \widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}_x \mathbf{x}^{(*)} + \widehat{\boldsymbol{\beta}}_z \widehat{\mathbf{z}}^{(*)}, \tag{9}$$

wobei $\widehat{\mathbf{z}}^{(*)}$ die modellbasierte Imputation von $\mathbf{z}^{(*)}$ ist: Nach Schätzung der Parameter $\alpha_{j,0}, \boldsymbol{\alpha}_j, \forall j \in \{1, ..., k\}$ in

$$\begin{aligned} &\text{(ZM)}_j: \quad z_j^{(i)} = \alpha_{j,0} + \boldsymbol{\alpha}_j \mathbf{x}^{(i)} + \epsilon_j^{(i)}, \epsilon_j^{(i)} \overset{iid}{\sim} N(0, \sigma^2) \\ &\forall i \in \{1, ..., n\}, j \in \{1, ..., k\} \end{aligned} \tag{10}$$

auf den Trainingsdaten, wird die Imputation $\widehat{\mathbf{z}}^{(*)} = (\widehat{z}_1^{(*)}, ..., \widehat{z}_k^{(*)})$ bestimmt durch

$$\text{(ZP)}_j: \quad \widehat{z}_j^{(*)} = \widehat{\alpha}_{j,0} + \widehat{\boldsymbol{\alpha}}_j \mathbf{x}^{(*)}, \qquad \forall j \in \{1, ..., k\}. \tag{11}$$

Hier sei darauf hingewiesen (wie auch schon in Kauermann und Windmann 2023, berücksichtigt), dass dies nur für stetige außergesetzliche Merkmale ratsam ist; bei anderen Skalentypen sollten passendere Regressionsmodelle für die modellbasierte Imputation genutzt werden, wie beispielsweise das Logit-Modell bei binären z_j. Da die grundlegende Idee sich hierdurch allerdings nicht verändert und die aufwändigere Formelarbeit von den zentralen Punkten ablenken könnte, gehen wir im Folgenden o. B. d. A. von stetigen außergesetzlichen Merkmalen aus (kategoriale Merkmale könnten beispielsweise in stetige Merkmale umkodiert werden).

Empirische Evaluation. Die empirische Evaluation in Kauermann und Windmann (2023) basierend auf den Daten des Mietspiegels für München zeigt unter anderem:

1. Erwartungstreue: Ein Vergleich der Schätzer der Effekte der gesetzlichen Merkmale $\mathbf{x}$, also ein Vergleich von $\widehat{\boldsymbol{\beta}}_x$ und $\widehat{\boldsymbol{\theta}}_x$, zeigt, dass die Schätzung des Bias $\widehat{\boldsymbol{\theta}}_x - \boldsymbol{\beta}_x$ besonders für diejenigen Merkmale (betragsmäßig) groß ist, die stark mit einem oder mehreren außergesetzlichen Merkmalen korreliert sind. Zum Beispiel wird der Effekt des Merkmals „neuer Boden“ – welches eine hohe Korrelation mit der Mietdauer aufweist – stark überschätzt. Dies ist damit zu erklären, dass in Abwesenheit der Variablen „Mietdauer“ die Bodenart als Proxy für die Mietdauer ausgenutzt wird, was aus Perspektive der reinen Prognosegüte vorteilhaft ist – allerdings auf Kosten der Erwartungstreue der Schätzer geht.

Tab. 1 Vergleich der Bestimmtheitsmaße der verschiedenen Modelle – Zahlen übernommen aus Kauermann und Windmann (2023)

	(P0)	(P1)	(P2)	(P3)
R^2	0,53	0,39	0,20	0,47
Anteil R^2	100%	74%	38%	89%

2. Bestimmtheitsmaß: Tab. 1 zeigt die in Kauermann und Windmann (2023) angegebenen Bestimmtheitsmaße für die vier verschiedenen Ansätze (P0) – (P3), sowie den Anteil am Benchmarkmodell (P0). Wie zu erwarten, schneiden (P1) – (P3) schlechter ab als das Benchmarkmodell (P0). Am wenigsten Erklärungskraft hat hierbei (P2): Im Trainingsprozess wird verhindert, dass gesetzliche Merkmale als Proxies für außergesetzliche Merkmale genommen werden, im Prognoseprozess wird diese Information ebensowenig zugelassen, was letztlich die Prognosegüte reduziert. Bei (P1) wird dem Modell jedoch gestattet, diese Proxyinformation zu berücksichtigen. Am besten schneidet (P3) ab: Hier wird die Information dadurch in die Prognose geholt, dass die außergesetzlichen Merkmale explizit durch die gesetzlichen Merkmale vorhergesagt werden – was hier anscheinend besser ist als die Information indirekt in die Schätzer $\widehat{\boldsymbol{\theta}}_x$ aufzunehmen (P1).

3.3 Verhältnis der Imputationsmethode zur gesetzlichen Forderung

Wegen der größeren Prognosegüte bei gleichzeitiger Unverzerrtheit der Schätzer empfehlen Kauermann und Windmann (2023) also das Vorgehen (P3) für einen qualifizierten Mietspiegel. Nun stellt sich die Frage, ob die gesetzliche Forderung, die außergesetzlichen Merkmale nicht „direkt" zur Feststellung der Einzelvergleichsmiete heranzuziehen, hiermit eingehalten ist. Bei genauerer Betrachtung stellt sich heraus, dass die Imputation letztlich nichts anderes bewirkt als eine Änderung der Regressionsparameter – die Struktur des Modells ohne außergesetzliche Merkmale (M1) bleibt erhalten. Bei Modell (P2) verschiebt sich durch die Mittelwertimputation lediglich der Intercept:

$$\text{(P2):}\quad \widehat{y}^{(*)} = \widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}_x \mathbf{x}^{(*)} + \widehat{\boldsymbol{\beta}}_z \bar{\mathbf{z}} \tag{12}$$

$$= \underbrace{\left(\widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}_z \bar{\mathbf{z}}\right)}_{\widehat{\beta}_0^*} + \widehat{\boldsymbol{\beta}}_x \mathbf{x}^{(*)} \tag{13}$$

$$= \widehat{\beta}_0^* + \widehat{\boldsymbol{\beta}}_x \mathbf{x}^{(*)} \tag{14}$$

Bei Modell (P3) ändert sich neben dem Intercept auch der Steigungsparameter:

$$\text{(P3):}\quad \widehat{y}^{(*)} = \widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}_x \mathbf{x}^{(*)} + \widehat{\boldsymbol{\beta}}_z \widehat{\mathbf{z}}^{(*)} \tag{15}$$

$$= \widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}_x \mathbf{x}^{(*)} + \sum_{j=1}^{k} \widehat{\beta}_{j,z} \left(\widehat{\alpha}_{j,0} + \widehat{\boldsymbol{\alpha}}_j \mathbf{x}^{(*)}\right) \tag{16}$$

$$= \underbrace{\left(\widehat{\beta}_0 + \sum_{j=1}^{k} \widehat{\beta}_{j,z} \widehat{\alpha}_{j,0}\right)}_{\widetilde{\widehat{\beta}}_0} + \underbrace{\left(\widehat{\boldsymbol{\beta}}_x + \sum_{j=1}^{k} \widehat{\beta}_{j,z} \widehat{\boldsymbol{\alpha}}_j\right)}_{\widetilde{\widehat{\boldsymbol{\beta}}}} \mathbf{x}^{(*)} \tag{17}$$

$$= \widetilde{\widetilde{\beta}}_0 + \widetilde{\widetilde{\boldsymbol{\beta}}}\mathbf{x}^{(*)} \tag{18}$$

In (18) ist nun klar ersichtlich, dass die konkrete Ausprägung $\mathbf{z}^{(*)}$ nicht direkt zur Feststellung der Einzelvergleichsmiete herangezogen wird.[12] Die Informationen bezüglich der außergesetzlichen Merkmale werden nur im Trainingsprozess genutzt, was durchaus mit den Vorgaben vereinbar ist, die dies explizit gestatten: „Zur Vermeidung von verzerrten geschätzten Parametern kann es auch sachgerecht sein, außergesetzliche Merkmale im finalen Regressionsmodell zu berücksichtigen."[13]

3.4 Anwendung FairML auf den Mietspiegel

Auch wenn die außergesetzlichen Merkmale also nicht „direkt" für die Feststellung der Einzelvergleichsmiete benutzt werden – die vorgeschlagene Methodik von Kauermann und Windmann (2023) aus unserer Sicht damit gesetzeskonform ist (und innerhalb dessen eine beträchtliche Verbesserung der Prognosegüte zu erlauben scheint) –, wird die Information, die in den außergesetzlichen Merkmalen enthalten ist, immerhin indirekt genutzt. Im Folgenden erläutern wir, was diese „indirekte" Berücksichtigung konkret impliziert, welche Parallelen es zur aktuellen Diskussion bezüglich der Fairness von ML-Modellen in der ML-Literatur gibt und wie eine Methode aussehen müsste, welche die außergesetzlichen Merkmale auch indirekt nicht berücksichtigt – insbesondere bezüglich der Zulässigkeit von (M1/P1) mag das folgende Ergebnis überraschend sein. Ob diese indirekte Berücksichtigung in einem qualifizierten Mietspiegel gewünscht ist oder nicht, muss letztendlich die Gesetzgebung entscheiden – unser Ziel ist es hier, konzeptionelle Klarheit zu schaffen, auf deren Grundlage die Gesetzgebung dann entscheiden kann, was zulässig sein soll und was nicht, welche Qualitätsanforderungen also konkret gestellt werden.

Fairness Through Unawareness (FTU). Wie eingangs bereits erwähnt, hat die ML-Community eine ganze Reihe von Fairnessmetriken entwickelt. Um sich einen Überblick über die Fülle und Diversität zu verschaffen, bietet sich der oben bereits erwähnte Review von Verma und Rubin (2018) an. Eine Definition möchten wir aber speziell herausgreifen, da sie direkt mit einem der oben von Kauermann und Windmann (2023) vorgeschlagenen Modelle korrespondiert, nämlich *Fairness durch Unaufmerksamkeit* (Fairness through unawareness – FTU). Demnach gilt:

> „An algorithm is fair so long as any protected attributes A are not explicitly used in the decision-making process." (Kusner et al. 2017)

Diese Definition, die auch im Zusammenhang mit dem Begriff der „process fairness" in der Literatur genannt wird (Grgic-Hlaca et al. 2016) und eine enge Verbindung zur Idee des „disparate treatment" (Zafar et al. 2017) aufweist, kann wohl als

[12] Hinweis: Wenn bei nichtstetigen außergesetzlichen Merkmalen andere Modelle in (10) zugelassen werden, gilt die Formel in der Form nicht mehr zwingend, es bleibt aber dabei, dass die konkrete Ausprägung $\mathbf{z}^{(*)}$ nicht direkt zur Feststellung der Einzelvergleichsmiete herangezogen wird – der gesetzlichen Forderung also auch hier Genüge geleistet würde.

[13] *Verordnung über den Inhalt und das Verfahren zur Erstellung und zur Anpassung von Mietspiegeln sowie zur Konkretisierung der Grundsätze für qualifizierte Mietspiegel*, Drucksache 766/20 vom 17.12.2020, S. 35, https://dserver.bundestag.de/brd/2020/0766-20.pdf (Aufrufdatum 31.01.2024).

äquivalent zur oben genannten gesetzlichen Anforderung an qualifizierte Mietspiegel in Deutschland angesehen werden – die außergesetzlichen Merkmale haben in dem Fall die Rolle der *protected attributes*. Obige gesetzliche Forderung stellt also keine Ausnahme oder gar eine neue Idee dar, was als Glücksfall betrachtet werden kann, da wir nun auf zahlreiche Ergebnisse der Untersuchung von FTU zurückgreifen können, um die Implikationen dieser Forderung besser zu verstehen.

Es kann als in der fairML-Community breit akzeptiertes Resultat angesehen werden, dass FTU kein angemessenes Konzept zur Absicherung der Fairness von ADM-Systemen basierend auf ML-Modellen ist (Kusner et al. 2017), aus zwei Gründen: Zum einen verhindert das Ignorieren eines PA nicht, dass das ML-Modell PA-bezogene Informationen im Training und bei der Prognose benutzt; solange es Features gibt, die stark mit dem PA zusammenhängen, kann das ML-Modell diese Features als Proxies für die PA verwenden (dies zeigt auch die bessere Prognosegüte von (P1) gegenüber (P2) in Tab. 1). Zum anderen gibt es Situationen, in denen die direkte Verwendung der PA im Modell eine faire Behandlung überhaupt erst möglich macht, siehe beispielsweise das „Red Car“ - Beispiel in Kusner et al. (2017). Die Intuition zu diesem Ergebnis ist, dass bei direkter Berücksichtigung der PA im Modell beim Trainingsprozess die Effekte von PA und anderen Features besser getrennt werden können und Proxyeffekte keine Rolle spielen.

Die Modellierungsansätze (M1)/(P1) verfolgen genau die Idee von FTU. Ergänzend zur von Kauermann und Windmann (2023) gezeigten relativ schlechten Vorhersagegüte kann also festgehalten werden, dass dieser Ansatz auch aus Fairnessgesichtspunkten nicht optimal ist.

Fairness der Imputationsmethode. Die obigen Imputationsmethoden können als Versuch angesehen werden, die Fairnessprobleme des FTU-Ansatzes dadurch in den Griff zu bekommen, dass die PA im Trainingsprozess explizit genutzt werden (um Proxyeffekte – verursacht durch den *omitted variable bias* – zu vermeiden), gleichzeitig der gesetzlichen Forderung Genüge zu leisten (die PA bei der Prognose nicht zu nutzen) und hierbei die Vorhersagegüte noch zu optimieren (modellbasierte Imputation statt Mittelwertimputation). Wie bereits erwähnt, kann dem Ansatz (M0)/(P3) hinsichtlich der Einhaltung der gesetzlichen Forderung und Vorhersagegüte Erfolg bescheinigt werden. Wie verhält es sich aber mit der Vermeidung von Fairnessproblemen?

Folgendes Beispiel soll hier Klarheit bringen: Nehmen wir an, zur Entscheidung bezüglich der Gewährung eines Kredits darf das Geschlecht der anstragstellenden Person nicht direkt verwendet werden. Eine Bank entscheidet sich daher, obiges Imputationsverfahren zu verwenden; das Geschlecht wird zwar zum Modelltraining benutzt, die konkrete Ausprägung wird bei der Vorhersage aber durch eine modellbasierte Imputation ersetzt. Statt das wahre Geschlecht zu benutzen, versucht die Bank also, das Geschlecht möglichst gut aus den verfügbaren Daten vorherzusagen; man könnte also sagen, eine geschlechtsbezogene Diskriminierung wird angestrebt, allerdings unter „erschwerten Bedingungen“, da das wahre Geschlecht zur Prognosezeit „vergessen“ wird. Nehmen wir weiter an, der Bank stünde als Feature die Information zur Verfügung, ob eine Person bereits eine Schwangerschaft hinter sich hat. Sollte dies bei einer anstragstellenden Person zutreffen, könnte damit das (biologische) Geschlecht eindeutig bestimmt werden – die Entscheidung würde also zwar

formal das Geschlecht nicht „direkt“ berücksichtigen, aber indirekt eben schon und damit genauso diskriminieren (für diese Person) wie wenn zur Prognose direkt das wahre Geschlecht verwendet worden wäre.

Dieses Beispiel mag sehr plakativ sein und in Fragen des Mietspiegels mag es eine solche direkte Verbindung von gesetzlichen und außergesetzlichen Merkmalen möglicherweise nicht geben. Dennoch können solche Proxyinformationen auch hier zum Prognosezeitpunkt einfließen, falls eine modellbasierte Imputation der außergesetzlichen durch die gesetzlichen Merkmale hinreichend gut möglich ist. Wenn dies nicht möglich ist, hat die modellbasierte Imputation gegenüber der Mittelwertimputation keine prognostischen Vorteile. Anders ausgedrückt: Wenn die modellbasierte Imputation gegenüber der Mittelwertimputation prognostische Vorteile hat, ist davon auszugehen, dass Proxyeffekte eine Rolle spielen.

Kausale Fairness. Wenn diese indirekte Nutzung der Informationen der PA nicht gewünscht sein sollte, könnte beipielsweise das oben beschriebene Framework der kausalen Fairness verwendet werden. Hierfür müssten folgende Schritte ausgeführt werden – die zweifelsohne bedeutend aufwändiger sind als die anderen bisher beschriebenen Methoden – siehe Bothmann et al. (2024) für eine detailliertere Erklärung:

1. Das kausale Modell der wahren Welt wird mittels Methoden der *causal discovery* und *causal inference* in Verbindung mit Fachwissen geschätzt.
2. Ein Mapping der Daten von der wahren Welt in die normativ gewünschte Welt – in der die PA keinen Einfluss auf die Zielvariable haben – wird basierend auf dem obigen kausalen Modell definiert.
3. Die Trainingsdaten werden mit diesem Mapping in die normativ gewünschte Welt „übersetzt“.
4. Das ML-Modell wird mit den gemappten Daten trainiert.
5. Zum Prognosezeitpunkt werden die Features zunächst in die normativ gewünschte Welt gemappt und dann mit dem dort trainierten Modell vorhergesagt.

So wird sichergestellt, dass es im finalen Modell auch keine indirekten Einflüsse der PA gibt.

Vergleich der Ansätze. Der Unterschied bei Verwendung der kausalen Fairness und der Imputationsmethode ist, dass bei ersterer die Information bezüglich der PA weder direkt noch indirekt benutzt wird, während bei letzterer nur die direkte, jedoch nicht die indirekte Nutzung ausgeschlossen wird. Es ist nun eine normative Entscheidung, welche Anforderungen ein qualifizierter Mietspiegel erfüllen soll. Diesem Beitrag sind zwei Dinge wichtig, nämlich (a) herauszuarbeiten, welcher Unterschied zwischen einer direkten und einer indirekten Nutzung der PA besteht, und (b) aufzuzeigen, dass es für beide Forderungen mögliche Lösungsansätze gibt (wenn auch das Entfernen der indirekten Effekte deutlich komplexer ist als das Entfernen der direkten Effekte). Hiermit soll die Gesetzgebung in die Lage versetzt werden, eine fundierte Entscheidung bezüglich der gesetzlichen Anforderungen an qualifizierte Mietspiegel zu treffen – im vollen Wissen um die unterschiedlichen Implikationen möglicher Gesetzesformulierungen.

4 Zusammenfassung und Ausblick

Die Erstellung qualifizierter Mietspiegel unterliegt strengen gesetzlichen Anforderungen. Ein besonderes Augenmerk liegt hierbei auf außergesetzlichen Merkmalen, die zwar mietpreisrelevant sein können, aber für die Feststellung der ortsüblichen Vergleichsmiete nicht herangezogen werden dürfen – wie zum Beispiel die Mietdauer.

In diesem Beitrag haben wir einen Vorschlag zur modellbasierten Imputation der außergesetzlichen Merkmale von Kauermann und Windmann (2023) genauer beleuchtet und den gesetzlichen Anforderungen gegenübergestellt. Es stellte sich heraus, dass die Idee der modellbasierten Imputation einerseits den gesetzlichen Rahmenbedingungen entspricht und andererseits qualitative Vorteile gegenüber anderen Ansätzen (Mittelwertimputation, Nichtberücksichtigung) hat.

Die Frage der Behandlung außergesetzlicher Merkmale weist starke Parallelen zur Fairnessdebatte in ML auf. Daher haben wir die Grundzüge dieser Debatte skizziert, eine aktuelle Neuentwicklung von Bothmann et al. (2024) vorgestellt und die Verbindungen zur Mietspiegelfrage aufgezeigt. Wir haben erläutert, inwiefern auch das Vorgehen der modellbasierten Imputation indirekt Informationen der außergesetzlichen Merkmale nutzt und wie mit der Methode der kausalen Fairness ein Mietspiegel erstellt werden könnte, der auch indirekt diese Informationen nicht nutzen würde – bei gleichzeitig ungleich höherem Modellierungsaufwand. Diese Darstellung soll konzeptionelle Klarheit bringen, welche es wiederum der Gesetzgebung ermöglichen kann, noch genauer festzulegen, welche Kriterien ein qualifizierter Mietspiegel erfüllen soll.

Hiervon ausgehend lässt sich auch die Praxis zur Umsetzung von Diskriminierungsfreiheit in anderen Bereichen evaluieren. Beispielsweise ist die Nichtberücksichtigung von geschützten Merkmalen wie zum Beispiel Geschlecht bei der Erstellung von Versicherungstarifen ebenso wenig dafür geeignet, geschlechtsbezogene Diskriminierung auszuschließen. Dieser und andere, ähnlich gelagerte Fälle, sind relevante Anwendungsfälle von fairML und interessante Ansatzpunkte für künftige Forschungsarbeiten bezüglich der Qualität von ADM-Systemen.

Funding Open Access funding enabled and organized by Projekt DEAL.

Interessenkonflikt Die Autor:innen haben keine Interessenkonflikte zu erklären.

Literatur

Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine bias: there's software used across the country to predict future criminals. and it's biased against blacks. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Zugegriffen: 31. Jan. 2024

Aristoteles (1831) Aristotelis Opera Bd. 2. de Gruyter, S Book V (I. Bekker, Hrsg)

Bothmann L, Dandl S, Schomaker M (2023) Causal Fair Machine Learning via Rank-Preserving Interventional Distributions. In: Proceedings of the 1st Workshop on Fairness and Bias in AI co-located with 26th European Conference on Artificial Intelligence (ECAI 2023). CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3523/

Bothmann L, Peters K, Bischl B (2024) What is fairness? On the role of protected attributes and fictitious worlds. arXiv. https://doi.org/10.48550/arXiv.2205.09622

Corbett-Davies S, Pierson E, Feller A, Goel S (2016) A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. Washington Post. https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithmbe-racist-our-analysis-is-more-cautious-than-propublicas/. Zugegriffen: 31. Jan. 2024

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012) Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, S 214–226 https://doi.org/10.1145/2090236.2090255

Grgic-Hlaca N, Zafar MB, Gummadi KP, Weller A (2016) The case for process fairness in learning: feature selection for fair decision making. In: Symposium on Machine Learning and the Law at the 29th Conference on Neural Information Processing Systems. https://api.semanticscholar.org/CorpusID:13633339

Kauermann G, Windmann M (2016) Mietspiegel heute. AStA Wirtsch Sozialstat Arch 10(4):205–223. https://doi.org/10.1007/s11943-016-0197-x

Kauermann G, Windmann M (2023) Die Berücksichtigung von außergesetzlichen Merkmalen bei der Mietspiegelerstellung – Kausalität versus Vorhersage. AStA Wirtsch Sozialstat Arch 17(2):145–160. https://doi.org/10.1007/s11943-023-00321-1

Kusner MJ, Loftus J, Russell C, Silva R (2017) Counterfactual fairness. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (Hrsg) Advances in neural information processing systems. Curran Associates, (https://proceedings.neurips.cc/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5Paper.pdf)

Makhlouf K, Zhioua S, Palamidessi C (2022) Survey on causal-based machine learning fairness notions. arXiv. https://doi.org/10.48550/arXiv.2010.09553

v. Malottki C, Krapp M-C, Vaché M (2018) Außergesetzliche Mietpreisdeterminanten im Mietspiegel – Auswirkungen und statistische Behandlung. Wohnungswirt Mietrecht 71(11):665–675

Verma S, Rubin J (2018) Fairness definitions explained. In: Proceedings of the International Workshop on Software Fairness https://doi.org/10.1145/3194770.3194776

Zafar MB, Valera I, Rogriguez MG, Gummadi KP (2017) Fairness constraints: mechanisms for fair classification. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 962–970. https://proceedings.mlr.press/v54/zafar17a.html