DAV/DGVFM
Herbsttagung
2024

Marvin Schneller, Munich Re
Max Bagus, viadico

Vermeidung von Bias und Diskriminierung

Eine Diskussion für Aktuarinnen und Aktuare

19. November 2024

Agenda

Kontext, Arbeitsgruppe und Auftrag	1
Begriffe und Rahmenbedingungen	2
Nichttechnische Handlungsempfehlungen	3
Technische Handlungsempfehlungen	4
Diskussion und weiterführende Fragen	5

Verantwortungsbewusster Umgang wird erwartet

Bias und Diskriminierung werden sowohl in branchenübergreifenden und versicherungs-spezifischen Leitlinien für eine vertrauenswürdige Künstliche Intelligenz (KI) als auch im European Artificial Intelligence Act (EU AI Act) adressiert.

DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Arbeitsgruppe und Auftrag

- Gleichnamige Arbeitsgruppe des Ausschusses ADS mit Kick-Off im Januar 2023
- Mit dem Ziel, Maßnahmen zum Umgang mit Bias und zur Vermeidung von Diskriminierung bei Einsatz von Data Science Verfahren zu identifizieren und zu diskutieren

Marvin Schneller

Max Bagus

Christian Bökenheide

Maximilian Hudlberger

Olga Jungnitsch

Dr. Axel Kaiser

Marcel Meuthen

Dr. Michael Preischl

Suyang Pu

Dr. Johannes Schupp

Leonid Zeldin

Abgrenzung von Bias, Diskriminierung und Fairness (1/2)

- **Bias** („schief“ franz., „Vorurteil“ engl.) meint im statistischen Kontext, dass das Ergebnis einer Analyse aufgrund eines systematischen Fehlers insgesamt oder bezüglich einzelner Merkmale verzerrt ist. Diese Verzerrung kann schwer zu identifizieren sowie zu quantifizieren sein und kann in den Daten, in der darauf aufbauenden Analyse sowie in der Anwendung der Ergebnisse auftreten.
- **Diskriminierung im rechtlichen Sinne** bezieht sich auf die Ungleichbehandlung von Personen basierend auf bestimmten geschützten Merkmalen, wie sie im AGG[1] definiert sind. Diese Art der Diskriminierung ist verboten und kann sowohl unmittelbar („direkt“) als auch mittelbar („indirekt“) sein.
- **Diskriminierung im statistischen Sinne** meint die Anwendung von Gruppenmerkmalen, um Entscheidungen für Individuen zu treffen. Diese Art der Diskriminierung ist wertneutral.

1 Allgemeines Gleichbehandlungsgesetz

Abgrenzung von Bias, Diskriminierung und Fairness (2/2)

- In der Versicherung kommen beide Arten von Diskriminierung zusammen. Bei der aktuariellen Herleitung und Kalkulation von Prämien, der sogenannten **Differenzierung**, gibt die Diskriminierung im statistischen Sinne den Weg vor, während die rechtliche Diskriminierung gleichzeitig einen festen Rahmen setzt.
- Die Idee hinter der **algorithmischen Fairness** ist, dass statistische Modelle und KI möglichst objektive Entscheidungen treffen oder unterstützen. Durch die Anwendung quantitativer Techniken, sogenannter **Fairness-Metriken**, soll sichergestellt werden, dass sich zwischen verschiedenen Personen oder Gruppen entweder die Verteilung von Vorhersagen und Entscheidungen oder die Modellierungsfehler nicht anhand sensibler Merkmale wie z.B. Geschlecht oder ethnischer Herkunft unterscheiden.

Rechtliche Rahmenbedingungen in Deutschland

Allgemeines Gleichbehandlungsgesetz

- Verbot von rassistischen Diskriminierungen und Diskriminierungen aufgrund der ethnischen Herkunft, des Geschlechts, der Religion, einer Behinderung, des Alters und der sexuellen Identität [1]
- Differenzierung wegen Religion, Behinderung, Alter oder sexueller Identität zulässig, wenn diese auf anerkannten Prinzipien risiko-adäquater Kalkulation beruht [2]

Europäische KI-Verordnung

Insb. Artikel 10 „Daten und Daten-Governance“ der Anforderungen an Hochrisiko-KI-Systeme

Datenschutz-Grundverordnung

Insb. Artikel 5, 13, 14, 15, 16, 17 und 22 fördernd für einen verantwortungsbewussten Umgang mit Bias

Versicherungsaufsichtsgesetz

Bei Tarifierung § 138 Absatz 2, Einfordern einer integren und fairen Geschäftstätigkeit nach Artikel 23 Absatz 1

[1] Umsetzung der Europäischen Richtlinien 2000/43/EG, 2000/78/EG, 2002/73/EG, 2004/113/EG, 08/2006 **[2]** AGG § 20 Absatz 2

Nicht-technische Handlungsempfehlungen

zur Vermeidung von Bias und Diskriminierung

- Effektive Governance und Integration in das bestehende Risikomanagementsystem
  z.B. klare Verantwortlichkeit, transparente Entscheidungsprozesse und Dokumentation, Feedbackmechanismen, neue Risikokategorien
- Ethische Leitlinien, verantwortungsvolle Nutzungskultur und menschliche Aufsicht
  z.B. Ethik-Gremium und klarer ethischer Rahmen bei Nutzung von Data Science Methoden
- Ausbildung und Sensibilisierung der Mitarbeiter
  z.B. Schulung von methodischen Grundkenntnissen, Risiken bei Nutzung und rechtlichen Rahmenbedingungen
- Vielfalt in Entwicklungsteams, eine besondere Rolle für Aktuarinnen und Aktuare sowie rege Stakeholder-Beteiligung
  z.B. Einbindung verschiedener Funktionen und Personen, insbesondere von Aktuarinnen und Aktuaren, Ethik-Gremium inkl. ext. Stakeholder
- Kontinuierliches Monitoring von Gesetzgebung und Best Practices
  z.B. Verfolgen weltweiter KI-Regulierungen, Hilfestellungen der Aufsichten sowie konstante Evaluierung von Data Science Best Practices
- Externes Stakeholder-Engagement, Kommunikation und Transparenz
  z.B. Mitarbeit in Branchen-Initiativen und Data Science Communities, externes Reporting

DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Technische Handlungsempfehlungen (1/2)

zur Vermeidung von Bias und Diskriminierung

Cross-Industry Standard Process for Data Mining (CRISP-DM)

1. Geschäftsverständnis
2. Datenverständnis *Wichtige Fragestellungen in Bezug zu Bias*
   - Wer hat die Daten erhoben? Zu welchem Zweck wurden die Daten erhoben?
   - Auf welche Personengruppe beziehen sich die Daten? Wie repräsentativ ist der Datensatz mit Blick auf diese Gruppe?
   - Auf welche Art und Weise wurden die Daten erhoben? Welches Dateiformat wurde dabei verwendet?
   - Wie umfangreich ist der Datensatz?
   - Sind vertrauliche / personenbezogene Daten enthalten?
   - Welche Merkmale im Datensatz sind vor dem Hintergrund der gesetzlichen Vorgaben zur Diskriminierung (z.B. AGG) sensibel? Sind Proxys enthalten? Welche Merkmale sind vor beiden Hintergründen unkritisch?
3. Datenvorbereitung
4. Modellierung
5. Evaluierung
6. Bereitstellung

DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Technische Handlungsempfehlungen (2/2)

zur Vermeidung von Bias und Diskriminierung

Praxisbeispiel: Schadeneintritt stationärer Aufenthalt

Wichtig: Abwägen von Performance und algorithmischer Fairness

	Ansatz	AUC	Demographic Parity	Equalized Odds
Logistische Regression	(1) Mit Geschlecht	0,798	0.272	0.434
	(2) Ohne Geschlecht	0,791	0.111	0.180
	(3) Nach Correlation Remover	0,779	0.059	0.076
XGBoost	(1) Mit Geschlecht	0,736	0.258	0.402
	(2) Ohne Geschlecht	0,726	0.085	0.163
	(3) Nach Correlation Remover	0,740	0.230	0.303
Random Forest	(1) Mit Geschlecht	0,760	0.268	0.352
	(2) Ohne Geschlecht	0,756	0.091	0.126
	(3) Nach Correlation Remover	0,754	0.282	0.388

Fazit: Correlation Remover nicht zwingend effektive Technik zur Bias-Reduktion

4. Praxisbeispiel aus der KFZ-Versicherung: Identifikation indirekter Diskriminierung und Diskussion des discrimination-free pricing-Ansatzes

Am Beispiel der bei Kompositversicherern üblichen Risikomodellierung der Schadenhäufigkeit wird gezeigt, wie indirekte Diskriminierung identifiziert und der „discrimination-free Pricing" Ansatz von (Lindholm M. , Richman, Tsanakas, & Wüthrich, 2023) angewandt werden kann. Die algorithmische Fairness der daraus

5. Praxisbeispiel aus der Krankenversicherung: Einfluss von Bias-Vermeidungsstrategien im Pre-Processing auf Performance und Fairness von Klassifikationsmodellen

Das folgende Praxisbeispiel untersucht den Einfluss von zwei Pre-Processing Techniken zur Bias-Verhinderung auf die Performance und algorithmische Fairness von Schadenhäufigkeitsmodellen für ambulante Krankenversicherungskosten. Die erste Technik ist die Entfernung des sensitiven Merkmals in der Modellierung, wäh-

6. Praxisbeispiel aus der Krankenversicherung: Einfluss von Bias-Vermeidungsstrategien im Pre-Processing auf Performance und Fairness von Regressionsmodellen

Dieses Praxisbeispiel führt das vorangegangene fort, indem die beiden Ansätze zur Bias-Verhinderung – Entfernung des sensitiven Merkmals und Correlation Remover – auf ein Regressionsproblem angewandt werden. Dabei werden erneut die Einflüsse auf Performance und Fairness diskutiert.

Im Vordergrund steht, den Aktuarinnen und Aktuaren einen praktischen, ersten Einblick in den Zusammenhang von Modellperformance und algorithmischer Fairness am Beispiel der Regression zu geben und nicht die dahinterliegende aktuarielle Fragestellung vollumfänglich zu beantworten. Es wurden deshalb bewusst einfache Modelle verwendet und keine Hyperparameter-Optimierung durchgeführt. Die vollständige Analyse und der zugrundeliegenden Python Code finden sich im Notebook XYZ auf dem Github-Account der Deutschen Aktuarvereinigung.

AUC = Area-under-the-ROC-curve. Demographic Parity misst, ob ein Modell die gleiche Prognose unabhängig des geschützten Merkmals liefert.
Equalized Odds messen die Unterschiede eines Modells in den False Positive Rates und True Positive Rates zwischen den Geschlechtern.

Diskussion und weiterführende Fragen

Wichtig zu beachten, dass die vollständige Vermeidung von Bias und Diskriminierung eine kontinuierliche Herausforderung bleibt, die ständige Wachsamkeit und Anpassung erfordert. Warum?

- Technologie und Methoden werden komplexer, Best Practices im Bereich Data Science ändern sich laufend.
- Gesellschaftliche Wahrnehmung und regulatorische Anforderungen können sich (schnell) ändern.

In welchem Verhältnis stehen Performance, algorithmische Fairness und Erklärbarkeit?

Wie ist mit Bias in dem externen, großen Sprachmodellen umzugehen?

Wie werden Fairness und Diskriminierung von Gesellschaft und Politik zukünftig bewertet?

…

DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

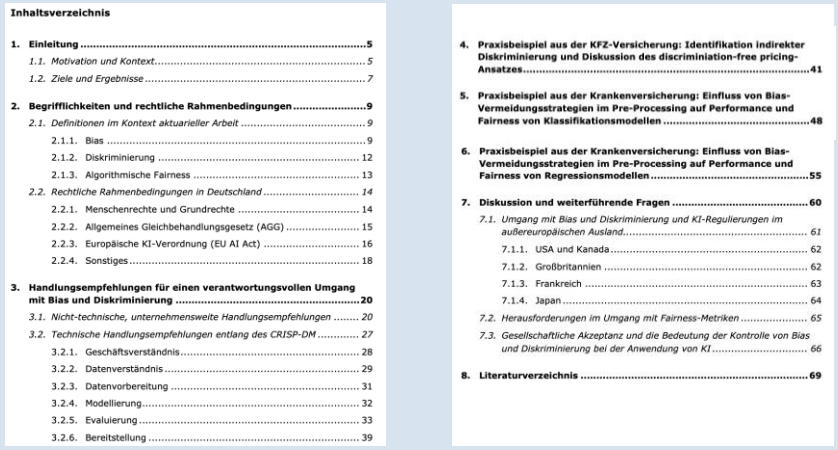
Ergebnisbericht und Notebooks

Liegt dem Ausschuss ADS zur Verabschiedung vor

Behandlung weiterer Fragen im bald veröffentlichten Ergebnisbericht

- Welche Arten von Bias gibt es?
- Welche Fragen sollten im jeweiligen CRISP-DM-Schritt beantwortet sein?
- Welche Herausforderungen gibt es bei der Anwendung von Fairness-Metriken?
- Wie werden Bias und Diskriminierung in anderen Ländern bewertet?
- …

Ergebnisbericht



Inhaltsverzeichnis

1. Einleitung 5
   1.1. Motivation und Kontext 5
   1.2. Ziele und Ergebnisse 7
2. Begrifflichkeiten und rechtliche Rahmenbedingungen 9
   2.1. Definitionen im Kontext aktuarieller Arbeit 9
      2.1.1. Bias 9
      2.1.2. Diskriminierung 12
      2.1.3. Algorithmische Fairness 13
   2.2. Rechtliche Rahmenbedingungen in Deutschland 14
      2.2.1. Menschenrechte und Grundrechte 14
      2.2.2. Allgemeines Gleichbehandlungsgesetz (AGG) 15
      2.2.3. Europäische KI-Verordnung (EU AI Act) 16
      2.2.4. Sonstiges 18
3. Handlungsempfehlungen für einen verantwortungsvollen Umgang mit Bias und Diskriminierung 20
   3.1. Nicht-technische, unternehmensweite Handlungsempfehlungen 20
   3.2. Technische Handlungsempfehlungen entlang des CRISP-DM 27
      3.2.1. Geschäftsverständnis 28
      3.2.2. Datenverständnis 29
      3.2.3. Datenvorbereitung 31
      3.2.4. Modellierung 32
      3.2.5. Evaluierung 33
      3.2.6. Bereitstellung 39
4. Praxisbeispiel aus der KFZ-Versicherung: Identifikation indirekter Diskriminierung und Diskussion des discriminiation-free pricing-Ansatzes 41
5. Praxisbeispiel aus der Krankenversicherung: Einfluss von Bias-Vermeidungsstrategien im Pre-Processing auf Performance und Fairness von Klassifikationsmodellen 48
6. Praxisbeispiel aus der Krankenversicherung: Einfluss von Bias-Vermeidungsstrategien im Pre-Processing auf Performance und Fairness von Regressionsmodellen 55
7. Diskussion und weiterführende Fragen 60
   7.1. Umgang mit Bias und Diskriminierung und KI-Regulierungen im außereuropäischen Ausland 61
      7.1.1. USA und Kanada 62
      7.1.2. Großbritannien 62
      7.1.3. Frankreich 63
      7.1.4. Japan 64
   7.2. Herausforderungen im Umgang mit Fairness-Metriken 65
   7.3. Gesellschaftliche Akzeptanz und die Bedeutung der Kontrolle von Bias und Diskriminierung bei der Anwendung von KI 66
8. Literaturverzeichnis 69

+ 3 Notebooks für die praktische Anwendung

DAV/DGVFM
Herbsttagung
2024

Vielen Dank
für Ihre Aufmerksamkeit.

Marvin Schneller, Munich Re
Max Bagus, viadico

DAV/DGVFM
Herbsttagung
2024

Besuchen Sie unsere Webseite

www.aktuar.de