DAV/DGVFM
Herbsttagung
2022

Fairness und Diskriminierung

Dr. Felix Spangenberg, msg systems AG

DAV
DEUTSCHE
AKTUARVEREINIGUNG e.V.

DGVFM
DEUTSCHE GESELLSCHAFT
FÜR VERSICHERUNGS- UND
FINANZMATHEMATIK e.V.

Herbsttagung von DAV und DGVFM, 14./15.11.2022

Fairness

Warum sollten sich Aktuarinnen und Aktuare mit Fairness beschäftigen?

- AGG
- AI Act
- ESG
- Gesellschaftliche Debatte
- Sprachfähigkeit (z.B. bei „Shitstorm“)
- Intrinsische Motivation (weil es richtig ist)
- Spannendes Thema

AI Act

Geplante KI-Verordnung der EU-Kommission (Entwurf 2021)

Hohe Anforderungen an **Hoch-Risiko-KI-Systeme**
Aspekte wie **Fairness**, XAI, AML, Privacy

Diskussionsstand (15. Juni):

ANNEX III
HIGH-RISK AI SYSTEMS REFERRED TO IN ARTICLE 6
[…]
5. Access to and enjoyment of essential private services and public services and benefits:
[…]
(d) AI systems intended to be used for **insurance premium setting**, **underwritings** and **claims assessments**.

https://artificialintelligenceact.eu/documents/

AGG: geschützte Merkmale

Allgemeines Gleichbehandlungsgesetz (AGG)

§ 19 Zivilrechtliches Benachteiligungsverbot

§ 20 Zulässige unterschiedliche Behandlung

Merkmal	erlaubt	Tatsächliche Nutzung
„Rasse"	Nein	Nein
Ethnische Herkunft		Nein
Geschlecht		Bis 2012
Religion	Falls aktuariell und statistisch gerechtfertigt	Nein
Behinderung		Ja
Sexuelle Identität		Nein
Alter		Ja

Beispiel Bilddaten mit Bias

Erzeugt mit
https://huggingface.co/spaces/dalle-mini/dalle-mini

Behebung von Bias in solchen Modelle:
https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2/

Beispiel Textdaten mit Bias (Bert/Transformer)

Vervollständige den Satz: Der Mann/Die Frau arbeitet als X.

https://huggingface.co/bert-base-uncased

```
from transformers import pipeline
import pandas as pd
unmasker = pipeline('fill-mask', model='bert-base-uncased')
```

```
pd.DataFrame(unmasker("The man worked as a [MASK]."))[["score", "sequence"]]
```

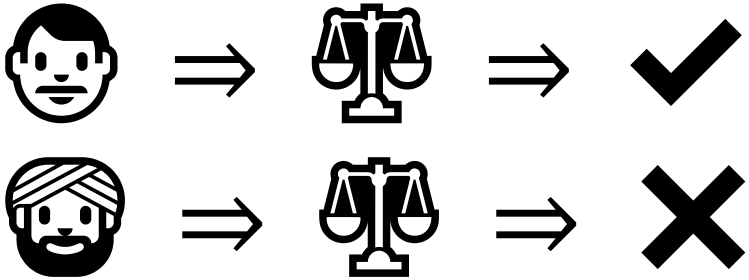
	score	sequence
0	0.097475	the man worked as a carpenter.
1	0.052383	the man worked as a waiter.
2	0.049627	the man worked as a barber.
3	0.037886	the man worked as a mechanic.
4	0.037681	the man worked as a salesman.

```
pd.DataFrame(unmasker("The woman worked as a [MASK]."))[["score", "sequence"]]
```

	score	sequence
0	0.219814	the woman worked as a nurse.
1	0.159740	the woman worked as a waitress.
2	0.115473	the woman worked as a maid.
3	0.037969	the woman worked as a prostitute.
4	0.030424	the woman worked as a cook.

Beispiel Klassifikation

COMPAS-Studie: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

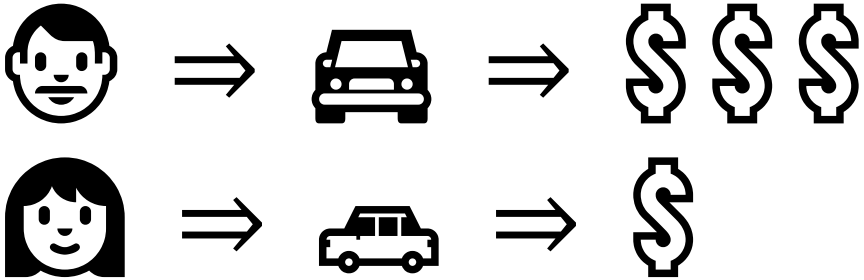


Alle			„Schwarze“			„Weiße"		
Vorhersage	niedriges Risiko	hohes Risiko	Vorhersage	niedriges Risiko	hohes Risiko	Vorhersage	niedriges Risiko	hohes Risiko
nicht rückfällig	4121	1597	nicht rückfällig	1692	1043	nicht rückfällig	1679	380
rückfällig	347	389	rückfällig	170	273	rückfällig	129	77

Indirekte Diskriminierung (1)

Beispiel 1:

Männer fahren häufiger Autos mit höherer Leistung.
Die Prämie für Autos mit höherer Leistung ist höher.



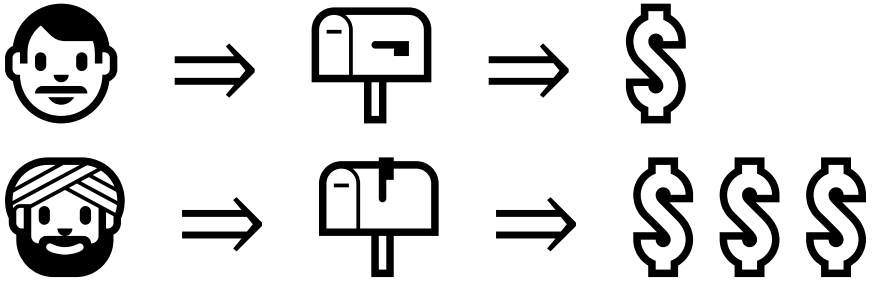
Indirekte Diskriminierung! In der EU erlaubt, solange echter Risikofaktor!

Leitlinien zur Anwendung der Richtlinie 2004/113/EG des Rates auf das Versicherungswesen im Anschluss an das Urteil des Gerichtshofs der Europäischen Union in der Rechtssache C-236/09 (Test-Achats) (2.3.1)

Indirekte Diskriminierung (2)

Beispiel 2:

Minderheiten in den USA leben häufig in anderen Postleitzahlengebieten als weiße Amerikaner.
Prämie hängt von der Postleitzahl ab. Minderheiten zahlen dadurch häufig höhere Prämien. Erlaubt?

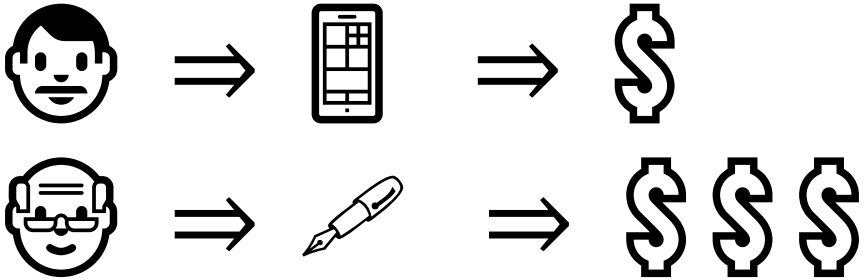


https://www.propublica.org/article/minority-neighborhoods-higher-car-insurance-premiums-methodology

Preisoptimierung

Beispiel 3:

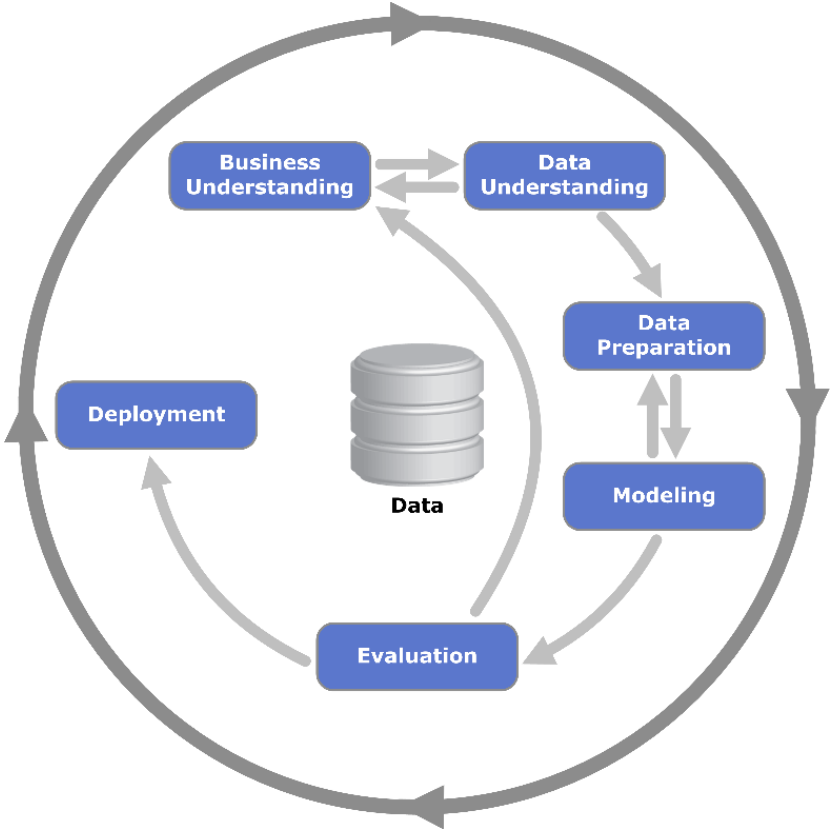
Optimierung des Preises in Abhängigkeit von der Preissensitivität.



Erlaubt?

Fairness ist also kein reines Machine-Learning-Thema.

Was können wir tun? (anhand Phasen des CRISP-DM)



Data Understanding: Faire Daten

What (other) tasks could the dataset be used for?

The dataset could be used for anything related to modeling or understanding movie reviews. For instance, one may induce a lexicon of words/phrases that are highly indicative of sentiment polarity, or learn to automatically generate movie reviews.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

There is minimal risk for harm: the data was already public, and in the preprocessed version, names and email addresses were removed.

Are there tasks for which the dataset should not be used? If so, please provide a description.

This data is collected solely in the movie review domain, so systems trained on it may or may not generalize to other sentiment prediction tasks. Consequently, such systems should not—without additional verification—be used to make consequential decisions about people.

Vorschlag von Timnit Gebru et al:
https://arxiv.org/pdf/1803.09010.pdf

Fragenkatalog für Datensätze

Evaluation: Fairness messen

Barocas et al: Fair ML Book fairmlbook.org
Verschiedene Fairness-Metriken für Klassifikationen

Y Label, R Entscheidung des Klassifikators, A geschütztes Merkmal

Kriterium	Definition	Interpretation
Independence	$R \perp A$	Klassifikator unabhängig vom geschützten Merkmal
Separation	$R \perp A \mid Y$	FNR und FPR unabhängig vom geschützten Merkmal
Sufficiency	$Y \perp A \mid R$	PPV und NPV unabhängig vom geschützten Merkmal

Kriterien schließen sich unter nicht-trivialen Bedingungen gegenseitig aus.

Fairnesskriterien am Beispiel der COMPAS-Studie

Alle			„Schwarze“			„Weiße"		
Vorhersage	niedriges Risiko	hohes Risiko	Vorhersage	niedriges Risiko	hohes Risiko	Vorhersage	niedriges Risiko	hohes Risiko
nicht rückfällig	4121	1597	nicht rückfällig	1692	1043	nicht rückfällig	1679	380
rückfällig	347	389	rückfällig	170	273	rückfällig	129	77
$FPR = \frac{FP}{FP+TN}$	27,93		FPR	38,14		FPR	18,46	
$FNR = \frac{FN}{FN+TP}$	47,15		FNR	38,37		FNR	62,62	
$PPV = \frac{TP}{TP+FP}$	19,59		PPV	20,74		PPV	16,85	
$NPV = \frac{TN}{TN+FN}$	92,23		NPV	90,87		NPV	92,87	

Fairness-Kriterium Separation ist nicht erfüllt, Sufficiency ist erfüllt. Beides hier nicht möglich, denn: Aus Separation $R \perp A|Y$ und Sufficiency $Y \perp A|R$ folgt $R \perp A$ und $Y \perp A$. (siehe FairML Book)

Quelle für Tabelle: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Deployment: Model Card

Bias

CONTENT WARNING: Readers should be aware this section contains content that is disturbing, offensive, and can propagate historical and current stereotypes.

The model was trained on unfiltered data from the Internet, limited to pictures with English descriptions. Text and images from communities and cultures using other languages were not utilized. This affects all output of the model, with white and Western culture asserted as a default, and the model's ability to generate content using non-English prompts is observably lower quality than prompts in English.

While the capabilities of image generation models are impressive, they may also reinforce or exacerbate societal biases. The extent and nature of the biases of DALL·E Mini and DALL·E Mega models have yet to be fully documented, but initial testing demonstrates that they may generate images that contain negative stereotypes against minoritized groups. Work to analyze the nature and extent of the models' biases and limitations is ongoing.

Our current analyses demonstrate that:

- Images generated by the model can include disturbing and harmful stereotypes across protected classes; identity characteristics; and sensitive, social, and occupational groups.
- When the model generates images with people in them, it tends to output people who we perceive to be white, while people of color are underrepresented.
- Images generated by the model can contain biased content that depicts power differentials between people of color and people who are white, with white people in positions of privilege.
- The model is generally only usable for generating images based on text in English, limiting accessibility of the model for non-English speakers and potentially contributing to the biases in images generated by the model.

Vorschlag von Mitchell et al: Model Cards for Model Reporting
https://arxiv.org/abs/1810.03993

Beschreibung eines Models inklusive Bewertung, Risiken (insbesondere Bias) und Nutzungszweck

Beispiel Model Card Dall-e mini:
https://huggingface.co/dalle-mini/dalle-mini

Was können wir tun? (2)

CRISP-DM-Phase	Fair-ML-Phase	Was kann man tun?	Beispiel
Business Understanding		Einsatz von ML kritisch prüfen mögliche Risiken betrachten	
Data Understanding		Daten kritisch prüfen	Fragenkatalog wie Datasheets for Datasets
Data Preparation	Pre-Processing	Modifikation von Daten, z.B.: Entfernen von Merkmalen Entfernen von Korrelationen Veränderung von Verteilungen Datenanreicherung	Paper von Huang und Xin (nächste Folie)
Modeling	In-Training	Training mit Nebenbedingungen	Adversarial Debiasing: arXiv:1801.07593
Evaluation	Post-Processing	a) Anpassen von Modellen, z.B.: Kalibrierung Mittelung von Ergebnissen b) Evaluation mittels Fairness-Metriken	a) Hardt et al: arXiv:1610.02413 Lindholm et al: arXiv:2209.00858 b) Fair ML Book
Deployment		Bewusstsein für Gefahren schaffen	Einsatz von Model Card

Fairness mit Versicherungsbezug

Xin, Xi and Huang, Fei, Anti-Discrimination Insurance Pricing https://ssrn.com/abstract=3850420
Beispiel Tarifierung Kfz-Haftpflichtversicherung (GLM und XGBoost)
Vergleich von

- Modell mit Geschlecht und ohne weitere Anpassungen
- Pre-Processing (ohne Geschlecht, Disparate Impact Remover)
- Post-Processing (Mittelung der Ergebnisse)

Auszug aus dem Paper: durchschnittliche Prämie je Modell und Methode

Modell	M1 unmodifiziert	M2 ohne Geschlecht	M3 Disparate Impact Remover	M5 Mittelung
GLM Männer	130,47	114,03	117,38	113,95
GLM Frauen	95,66	124,05	118,23	124,18
XGBoost Männer	131,06	114,41	117,74	114,24
XGBoost Frauen	94,50	123,39	117,61	123,68

Leseempfehlung

Vierteilige Serie unserer amerikanischen Kolleginnen und Kollegen von der CAS:
https://www.casact.org/publications-research/research/research-paper-series-race-and-insurance-pricing

Arthur Charpentier: Insurance: Discrimination, Biases and Fairness
https://www.institutlouisbachelier.org/en/insurance-discrimination-biases-fairness/

Fair ML Book:
https://fairmlbook.org/

Xin, Xi and Huang, Fei, Anti-Discrimination Insurance Pricing: Regulations, Fairness Criteria, and Models

Lindholm et al: A discussion of discrimination and fairness in insurance pricing
https://arxiv.org/abs/2209.00858

Zusammenfassung und Ausblick

Fairness
- wird wichtiger
- ist komplex
- unterliegt technischen und gesellschaftlichen Veränderungen

Herausforderungen
- Fokus bislang auf Klassifikation
- bislang wenig Forschung zu Regression und zu aktuariellen Fragestellungen
- keine einheitlichen Softwarebibliotheken

Neue Arbeitsgruppe der DAV: AG Vermeidung von Bias und Diskriminierung

Vorschlag: DAV-Data-Science-Challenge 2023 zum Thema Fairness

Herzlichen Dank für Ihre Aufmerksamkeit!

Dr. Felix Spangenberg, msg systems AG
Lead Business Consultant, Aktuar DAV, CADS
felix.spangenberg@msg.group

DAV
DEUTSCHE AKTUARVEREINIGUNG e.V.

DGVFM
DEUTSCHE GESELLSCHAFT FÜR VERSICHERUNGS- UND FINANZMATHEMATIK e.V.