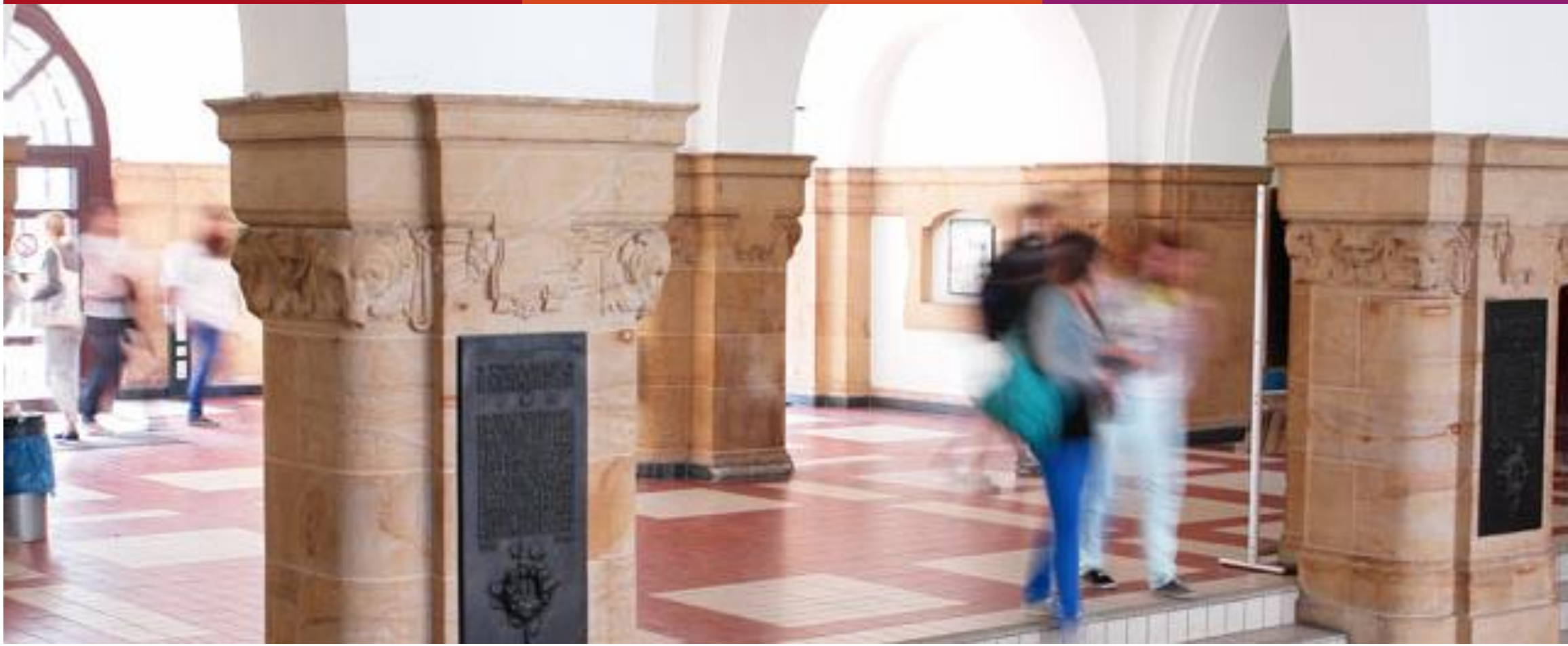


Wer besitzt eine private Krankenzusatzversicherung?

Eine Analyse mit Machine-Learning-Methoden

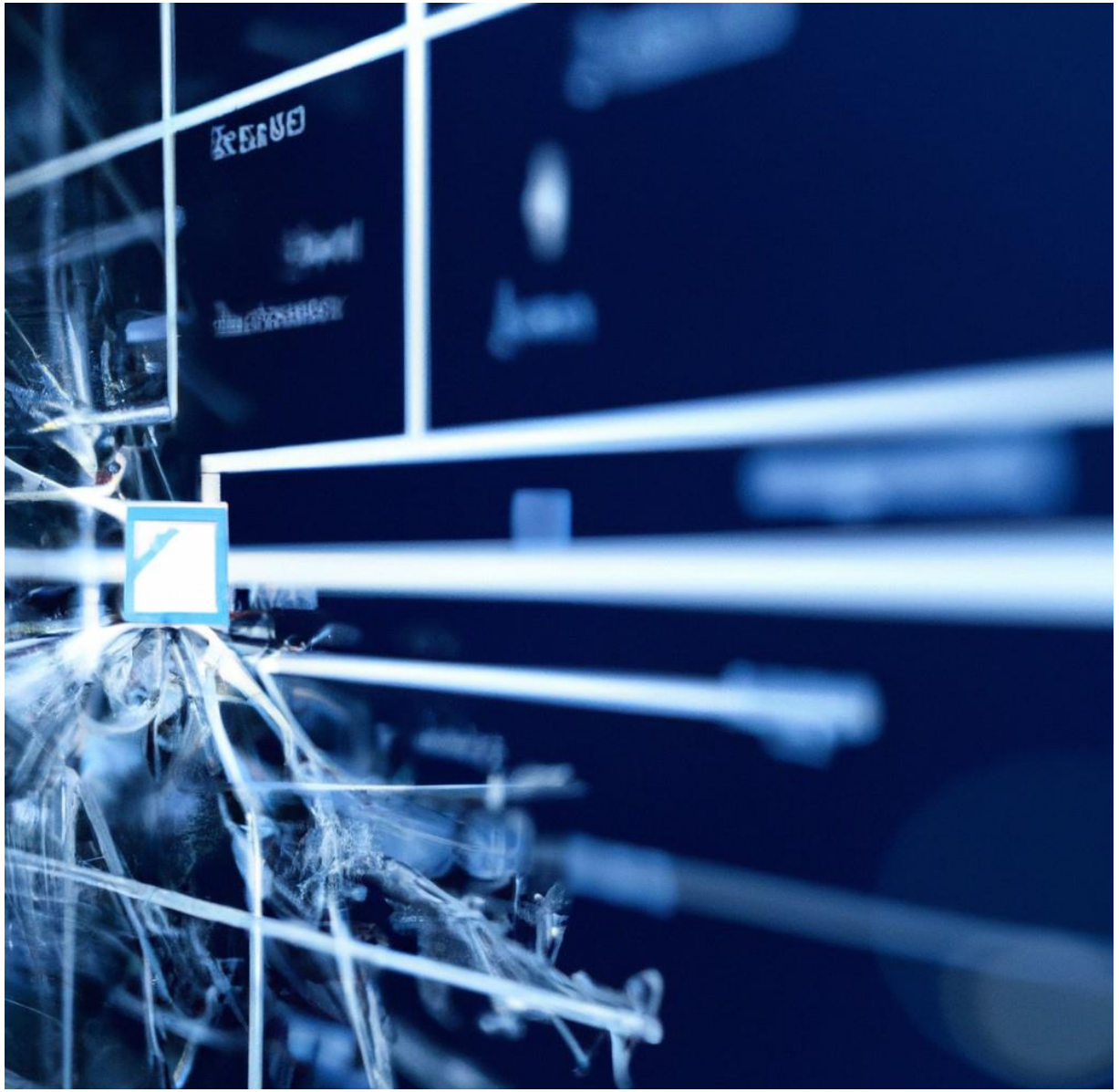
07.11.2023

Prof. Dr. Benedikt Funke, Dr. Simon Hatzesberger, Dr. Lars Kunze

qx-Club Köln/Bonn/Düsseldorf

ivwKöln
Institut für
Versicherungswesen

Technology
Arts Sciences
TH Köln

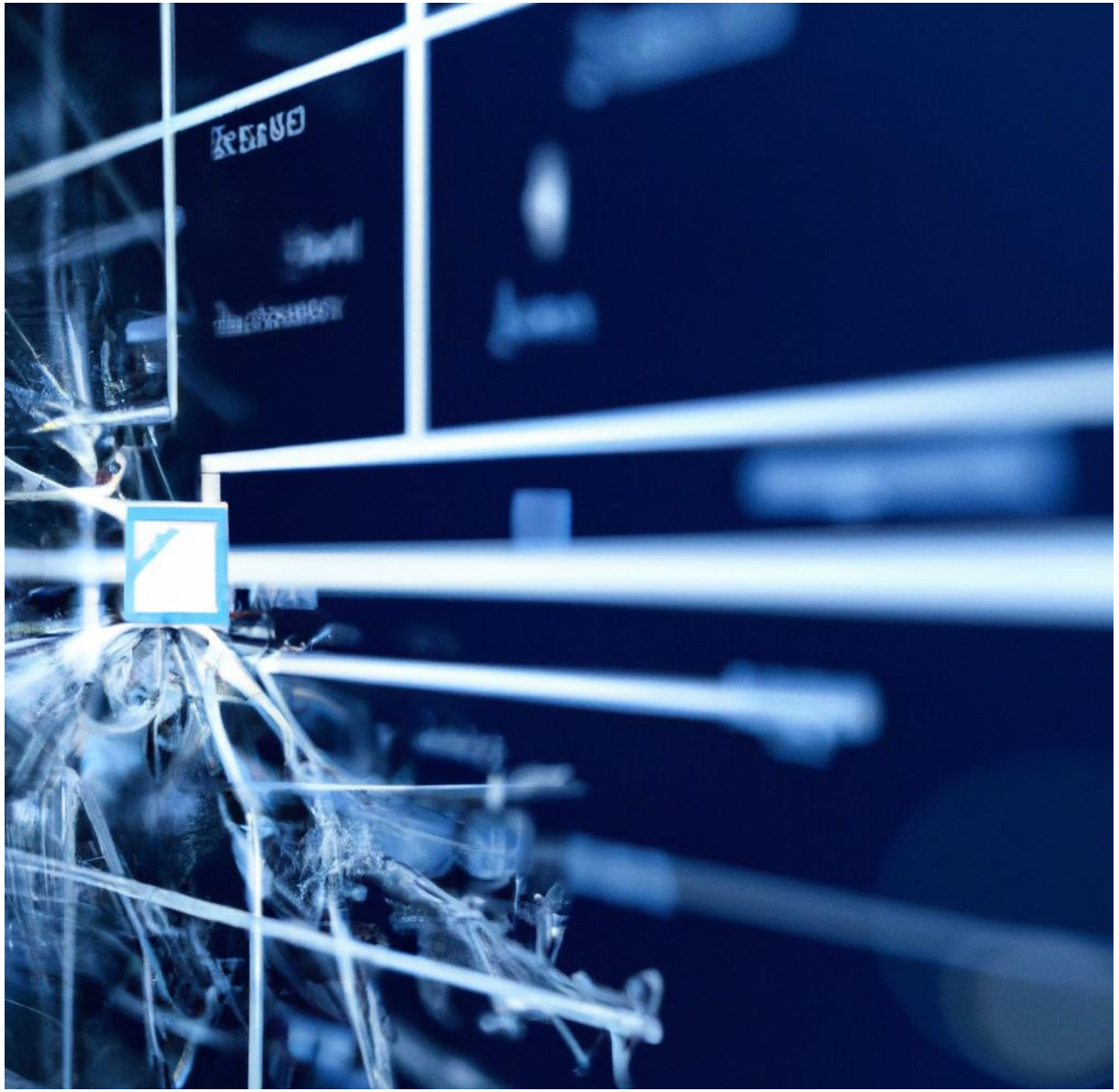


1 MOTIVATION

2 DATEN: SOEP

3 ANALYSEERGEBNISSE

4 FAZIT & AUSBLICK



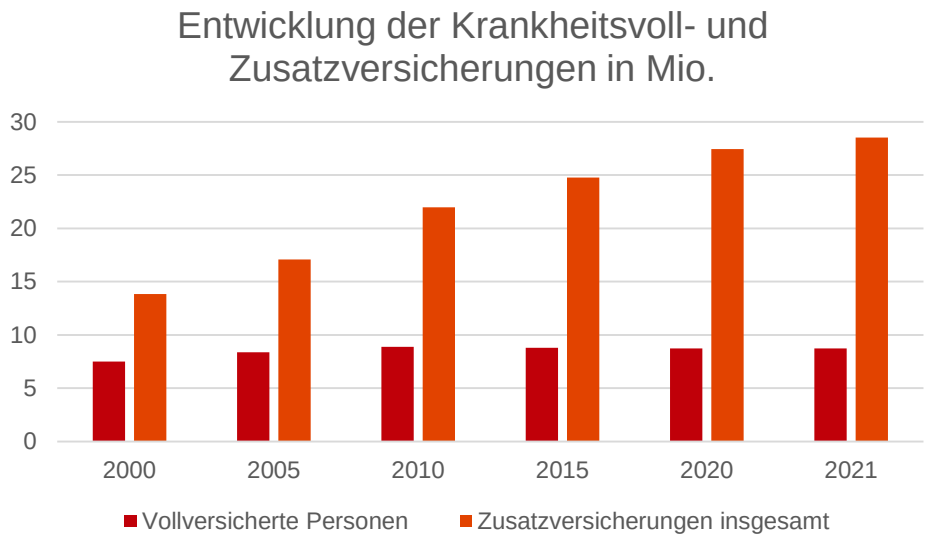
1 MOTIVATION

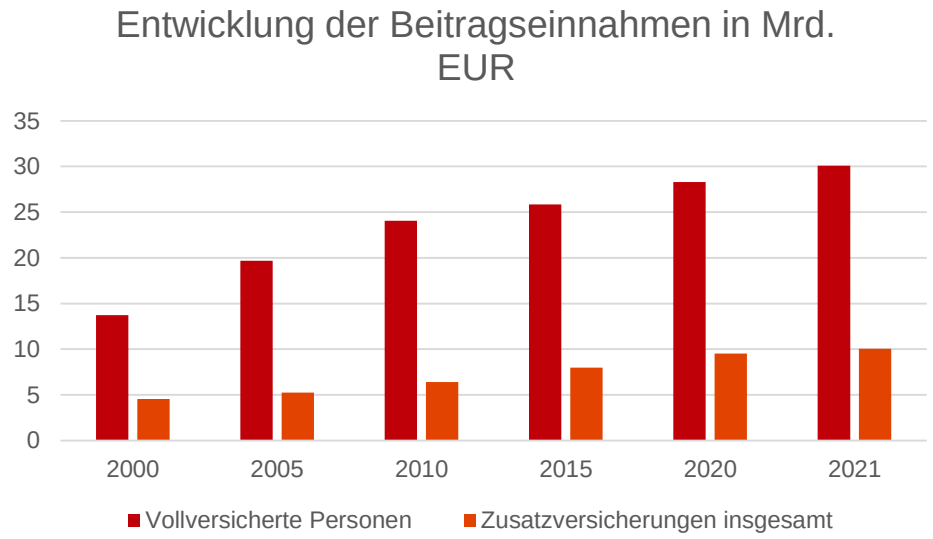
2 DATEN: SOEP

3 ANALYSEERGEBNISSE

4 FAZIT & AUSBLICK

Motivation: Kennzahlen und Relevanz der privaten Krankenzusatzversicherung

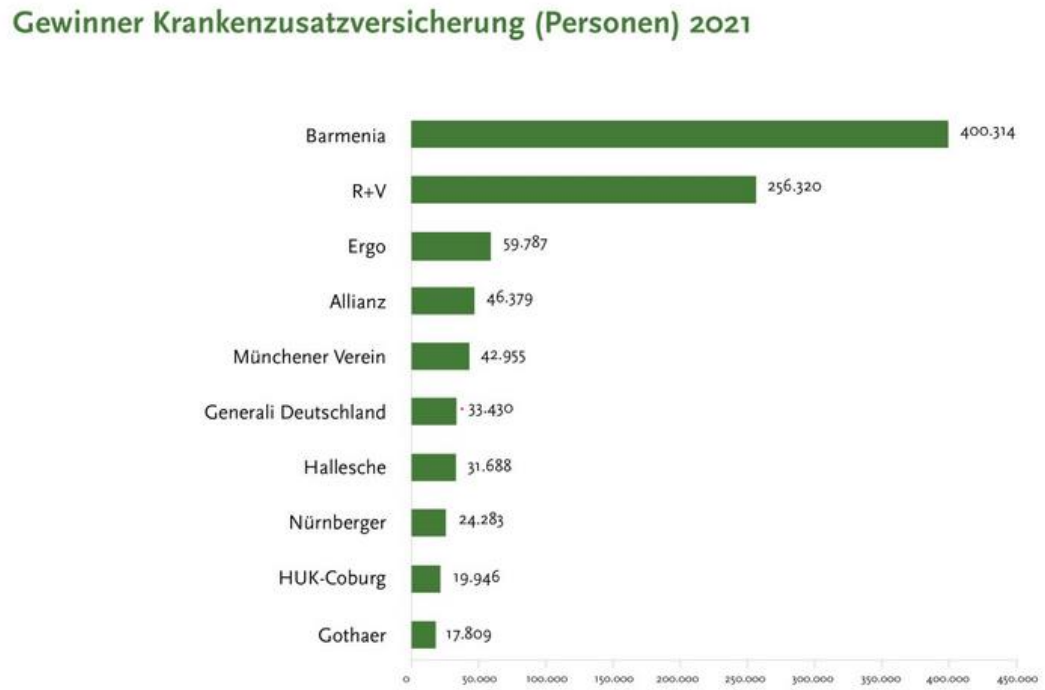




- Starkes Bestandswachstum in der privaten Krankenzusatzversicherung (KZV) in den letzten Jahren
- Beitragsvolumen der privaten KZV ebenfalls deutlich gestiegen

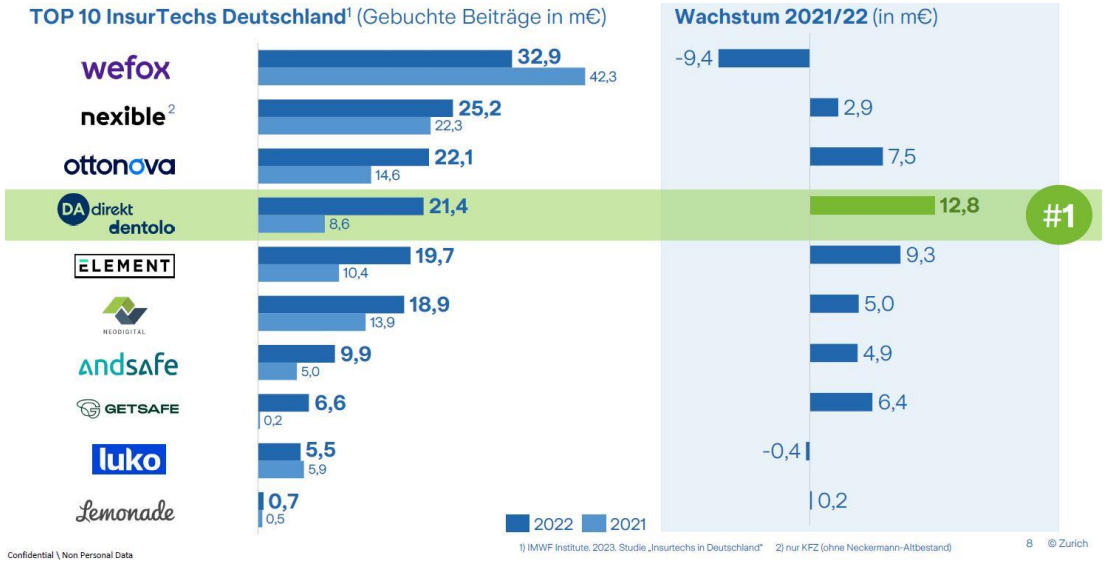
Quelle: Gesundheitsberichterstattung des Bundes (GBE)

Motivation: Kennzahlen und Relevanz der privaten Krankenzusatzversicherung



Quelle: Assekurata (November 2022)

Motivation: Kennzahlen und Relevanz der privaten Krankenzusatzversicherung



- Das auf Krankenzusatzversicherungen spezialisierte Insurtech *Dentolo* ist das wachsstumsstärkste Startup im deutschen Versicherungsmarkt (2021/22)

Quelle: Versicherungsmonitor (Juni 2023)

Motivation: Kennzahlen und Relevanz der privaten Krankenzusatzversicherung

Versicherer mit der Höchstnote hervorragend (FFF+) in mindestens 4 Kategorien

Gesellschaft	Zahn-ersatz	Zahn-behandlung	Stationär	Sehhilfen	Naturheil-verfahren	Vorsorge	Gesamt FFF+
SDK	FFF+	FFF+	FFF+	FFF+	FFF+	FFF+	6
Barmenia	FFF+	FFF+	FFF+		FFF+	FFF+	5
Münchener Verein	FFF+	FFF+	FFF+	FFF+		FFF+	5
Arag	FFF+	FFF+	FFF+		FFF+		4
DFV	FFF+	FFF+	FFF+			FFF+	4
DKV	FFF+			FFF+	FFF+	FFF+	4
Gothaer	FFF+	FFF+	FFF+			FFF+	4
Inter	FFF+	FFF+	FFF+	FFF+			4

- Franke und Bornberg (FB): Solides Niveau bei Zusatzversicherungen

Quelle: Ratingagentur Franke und Bornberg (FB) (März 2023)

Motivation: Kennzahlen und Relevanz der privaten Krankenzusatzversicherung

Pressemitteilung

AOK Bayern und Allianz koope-rieren bei Zusatzversicherungen

16.10.2023 • AOK Bayern • 3 Min. Lesedauer

Anhören

Unterzeichnung der neuen Kooperation (v. li.): Daniel Bahr, Vorstand der Allianz Private Krankenversicherung, Nina Klingspor, Vorstandsvorsitzende der Allianz Private Krankenversicherung, Dr. Irmgard Stippler, Vorstandsvorsitzende der AOK Bayern, Stephan Abele, stellvertretender Vorstandsvorsitzender der AOK Bayern.

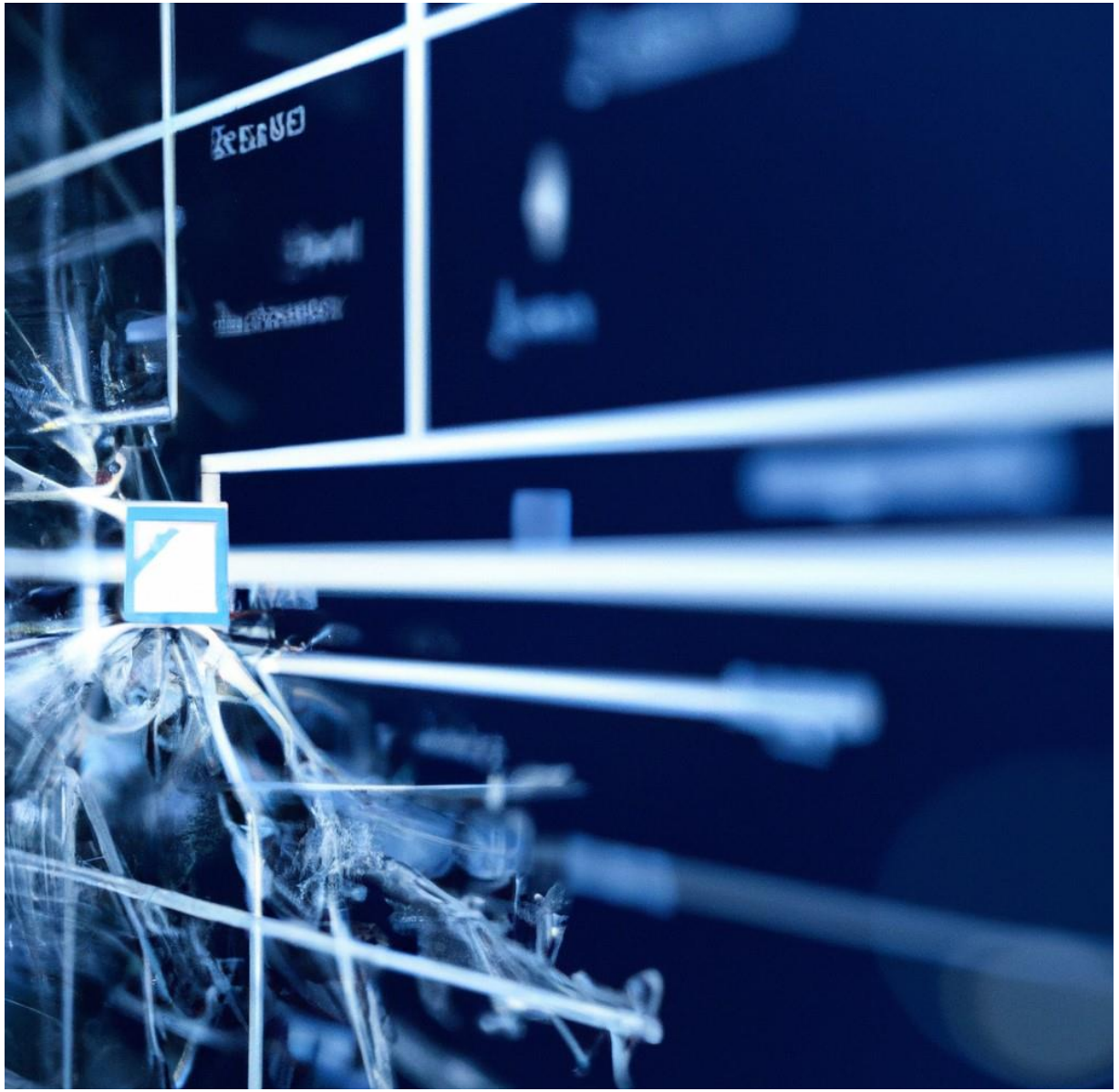
Quelle: Versicherungsmonitor / AOK Bayern (Oktober 2023)

- Die Bekanntgabe der Kooperation von AOK Bayern und der Allianz Private Krankenversicherung unterstreicht einmal mehr die Relevanz der Zusatz-versicherungen auf dem deutschen Markt

Motivation: Relevanz der privaten Krankenzusatzversicherung

Ähnliche Fragestellungen wurden bereits in der wissenschaftlichen Literatur untersucht:

- Lange et al. (2017): Analyse des Erwerbs privater Krankenzusatzversicherungen in Deutschland; Haupttreiber: Versicherungsneigung und Einkommen
- Bonsang und Costa-Font (2022): Anhand von Längsschnittdaten aus Deutschland wurde dargelegt, dass der Erwerb von der Kontrollüberzeugung des eigenen Einflussbereichs (*internal locus of control*) abhängig ist
- Eckert et al. (2021): Methodisch ähnliche Vorgehensweise; Thema: Identifikation von Faktoren des Abschlusses von Berufsunfähigkeitsver-sicherungen mit Methoden des Maschinellen Lernens



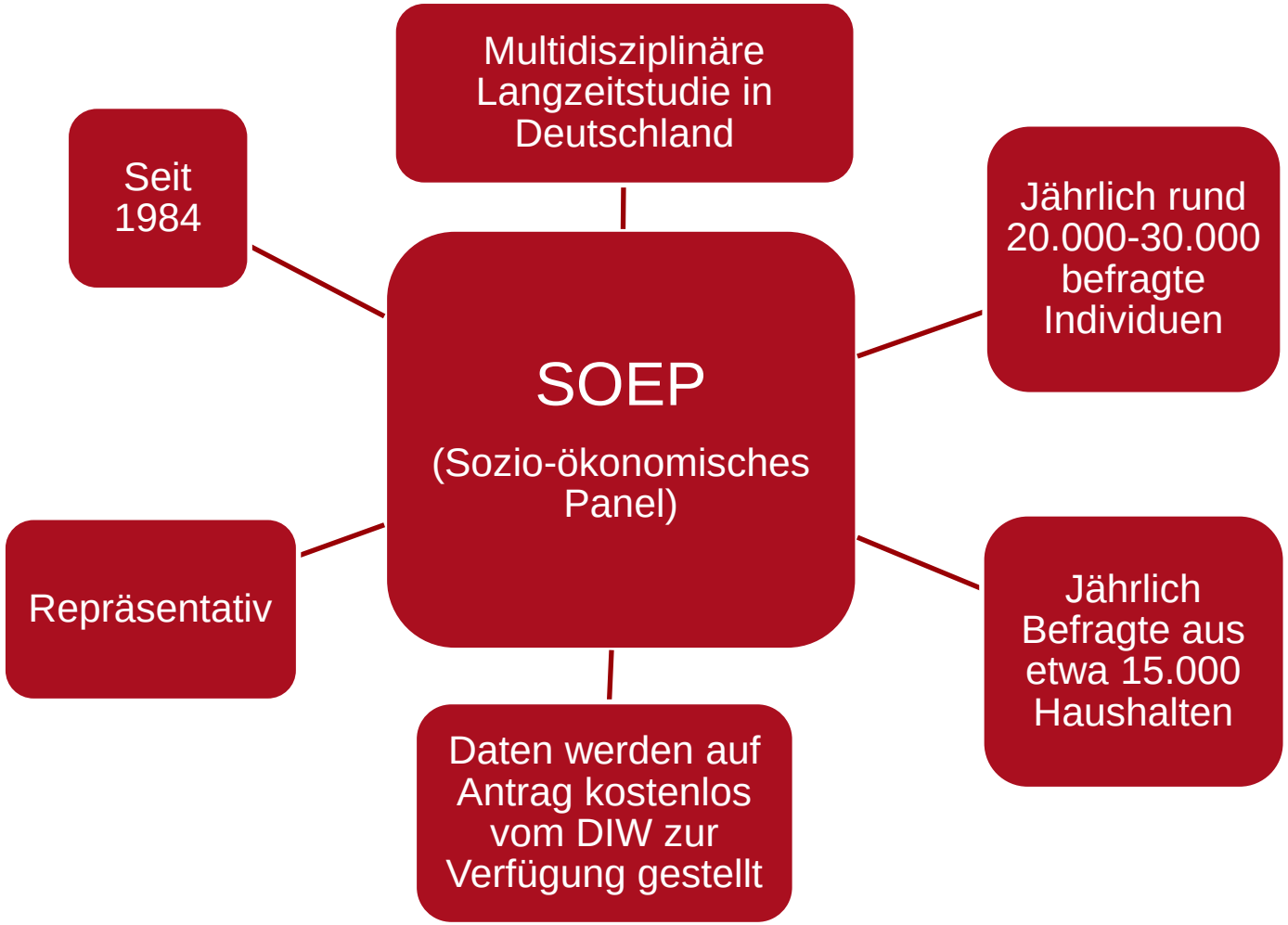
1 MOTIVATION

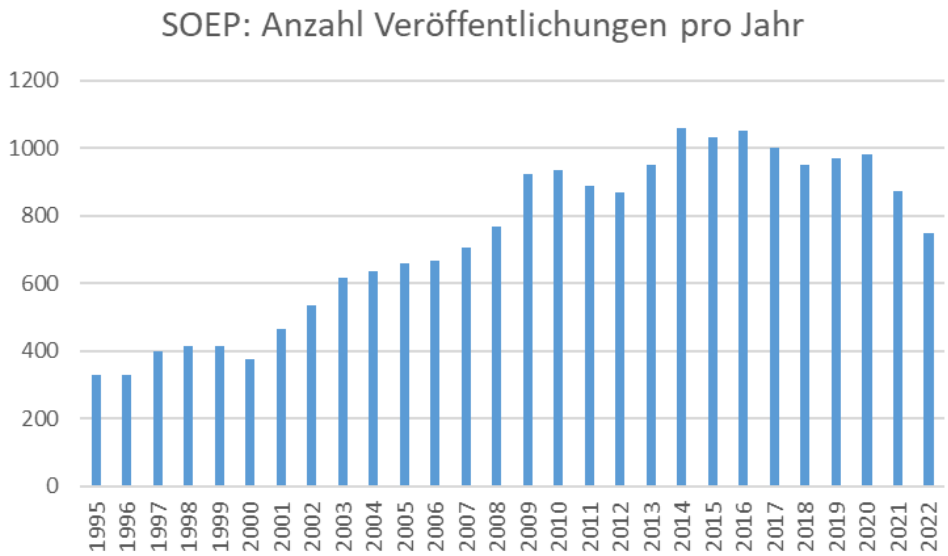
2 DATEN: SOEP

3 ANALYSEERGEBNISSE

4 FAZIT & AUSBLICK

Daten: SOEP-Studie





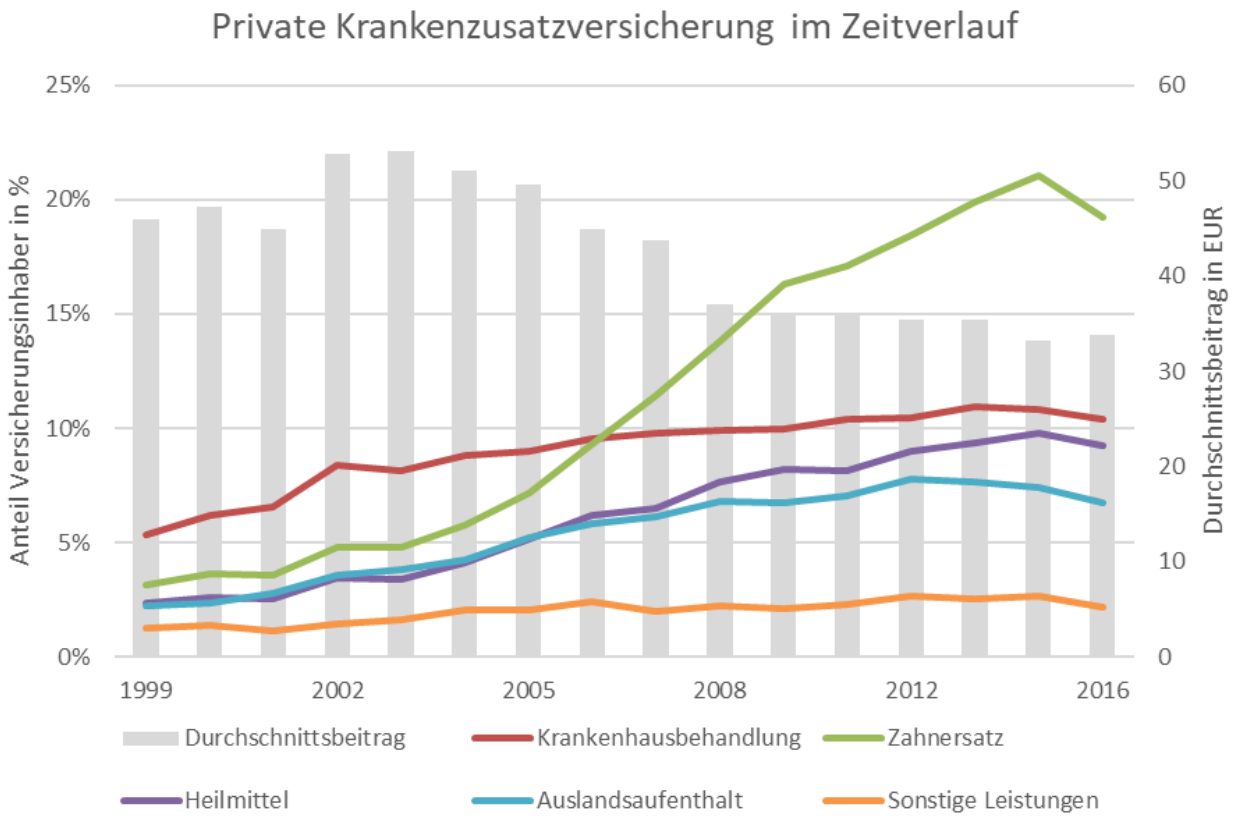
Quelle: DIW (diw.de)

Daten: SOEP-Studie

- Die Daten geben u. a. Auskunft über Einkommen, Erwerbstätigkeit, Bildung und Familienstand
- Für unsere Fragestellung v. a. Informationen über das Vorhandensein einer privaten KZV relevant
- Siehe z. B. Goebel et al. (2019) für eine detaillierte Beschreibung der Daten

Quelle: DIW (diw.de)

Daten: Trend der Zielvariablen



- Steigender Anteil des Vorhandenseins einer privaten KZV im Zeitverlauf; Haupttreiber: Zahnzusatzversicherungen
- Durchschnittsbeitrag rückläufig
- 45 % / 26 % / 18 % / 10 % / 2 % der VN haben eine / zwei / drei / vier / fünf Leistungsarten abgesichert
- Für die Analyse verwenden wir Daten für 2016 (Grund: Vergleichbarkeit zu Literatur, keine Kriseneffekte)

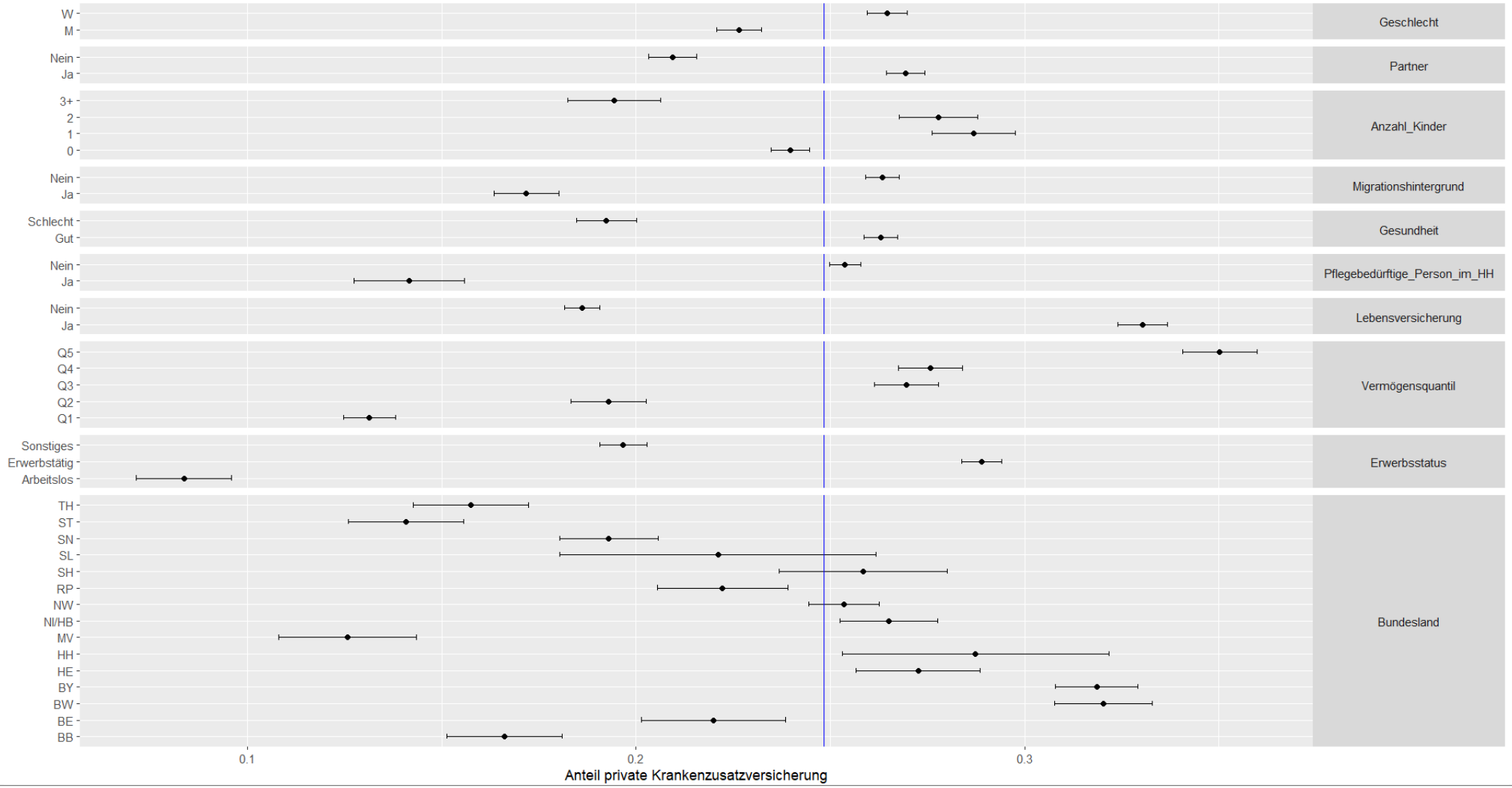
Quelle: SOEP v.34, Daten von 1999 bis 2016

Daten: Verwendete Merkmale

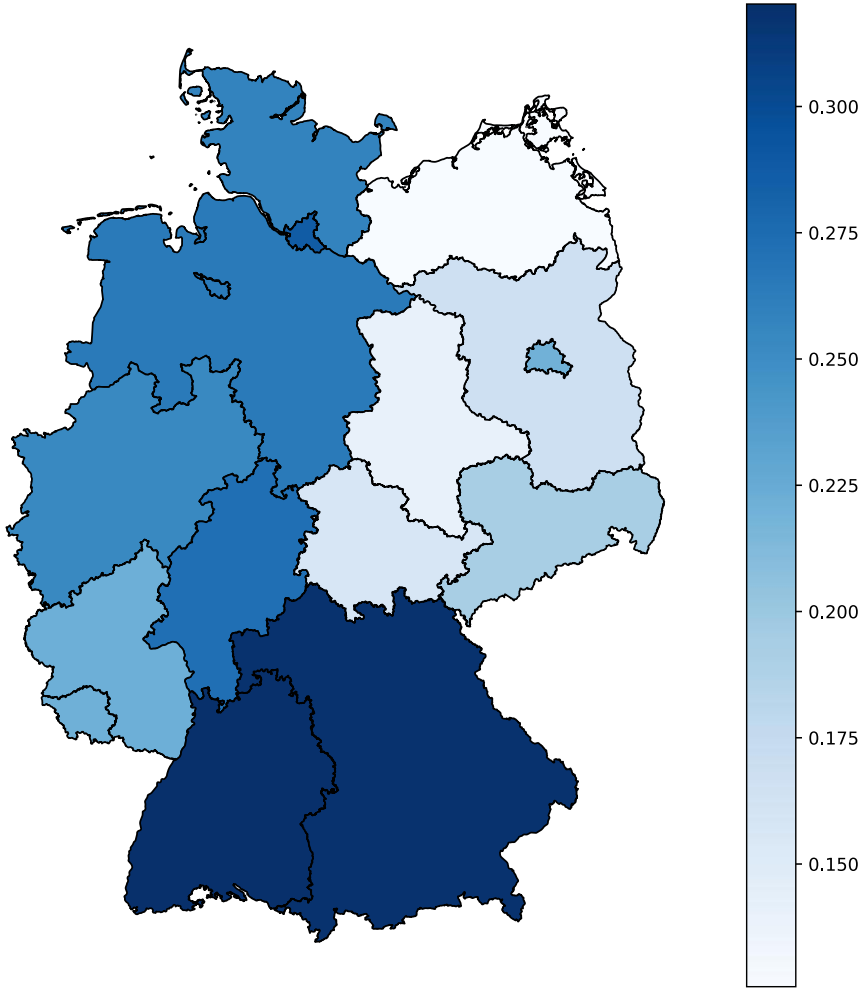
Kategorielle Variablen	Ausprägungen
Private Krankenzusatzversicherung	Ja / Nein
Geschlecht	M / W
Pflegebedürftige Person im Haushalt	Ja / Nein
Anzahl Kinder	Keine / 1 / 2 / 3+
Erwerbsstatus	Erwerbstätig / Arbeitslos / Sonstiges
Region	Dummy Variable je Bundesland
Haushaltsnettovermögen (Quintile)	Dummy Variablen je Quintil
Partner vorhanden	1 für verheiratet oder verpartnert, 2 sonst
Migrationshintergrund	Ja / Nein
Nationalität	Deutsch / Andere
Lebensversicherung vorhanden	Ja / Nein
Geburt eines Kindes	Ja / Nein
Tod des Partners	Ja / Nein
Scheidung / Trennung	Ja / Nein
Krankenhausübernachtung im letzten Jahr	Ja / Nein
Gesundheitszustand	„1“ falls „sehr gut“, „gut“, „zufriedenstellend“; „2“ sonst
Status GKV	„1“ Pflichtmitglied, „2“ Freiwilliges Mitglied, „3“ Familienmitglied, „4“ Versichert als Rentner, Arbeitsloser etc.
Berufliche Stellung	Sonstige, Arbeiter (ungelernt, angelernt, gelernt), Meister /Vorarbeiter, Selbstständige, Angestellter (einfache Tätigkeit), Angestellter (qualifizierte Tätigkeit), Angestellter (hochqualifiziert, Leitung)

Stetige / Diskrete Variablen	Ausprägungen
Haushaltseinkommen	Logarithmiertes, reales Haushaltsnettoäquivalenzeinkommen
Alter (in Jahren)	zwischen 25-90 Jahren
Dauer Ausbildung (in Jahren)	
Risikobereitschaft	diskret, höhere Werte bei höherer Risikobereitschaft
Locus of Control	stetig, höhere Werte für selbstbestimmtes Leben; Details s. Back-Up Items: Habe nicht das erreicht was ich verdiene, Mein Lebenslauf hängt von mir ab, Was man erreicht hängt vom Glück ab, Andere bestimmten über mein Leben, Zweifle bei Schwierigkeiten an meinen Fähigkeiten, Möglichkeiten werden von sozialen Umständen bestimmt, Wenig Kontrolle über mein Leben; Ansatz individueller, zeitinvarianter Durchschnittswert

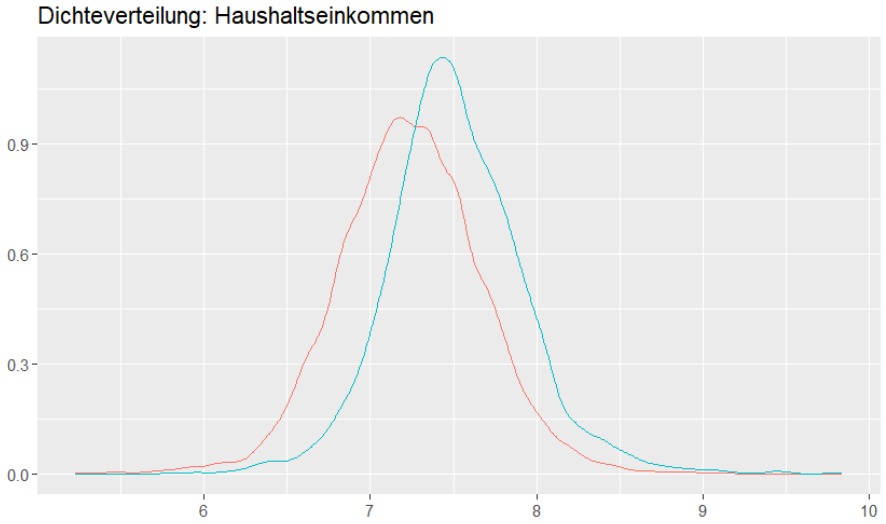
Daten: Deskriptive Statistiken ausgewählter Merkmale (1/4)

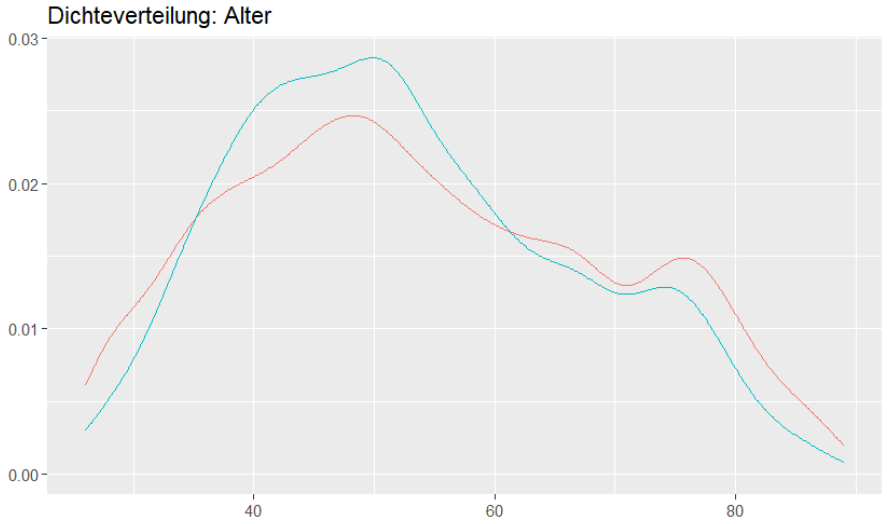


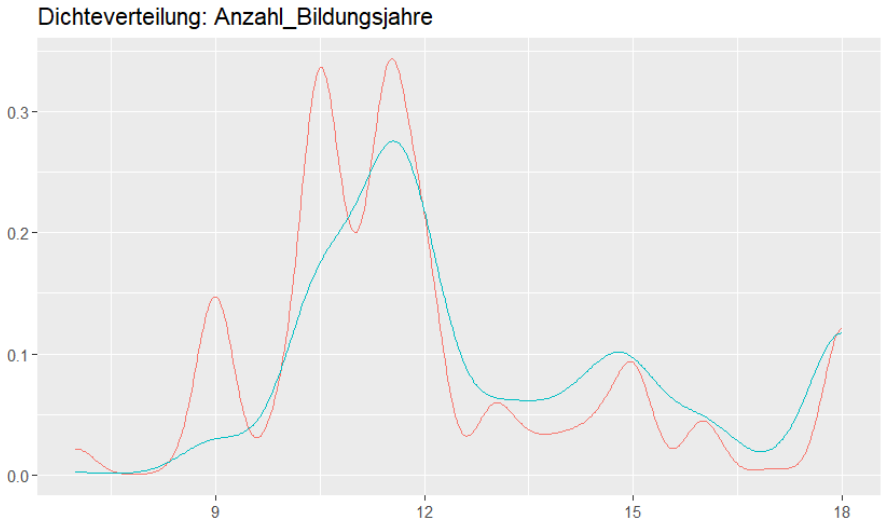
Daten: Deskriptive Statistiken ausgewählter Merkmale (2/4)

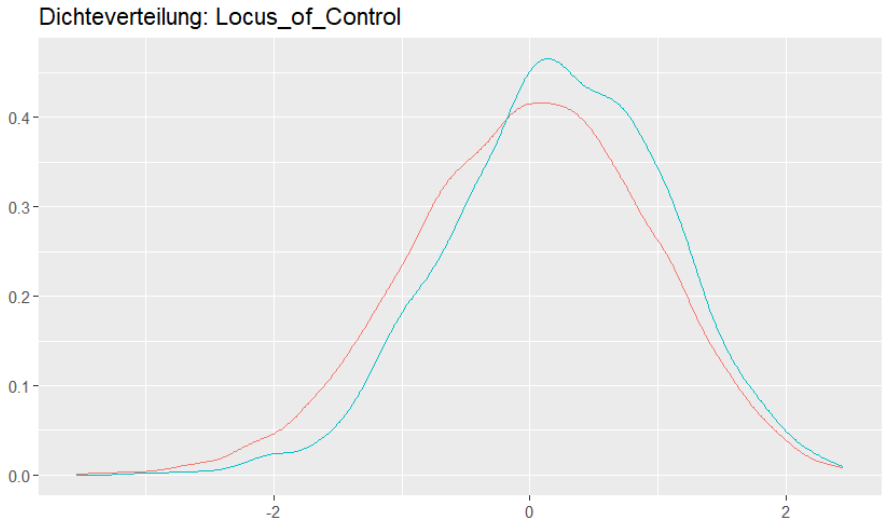


Daten: Deskriptive Statistiken ausgewählter Merkmale (3/4)

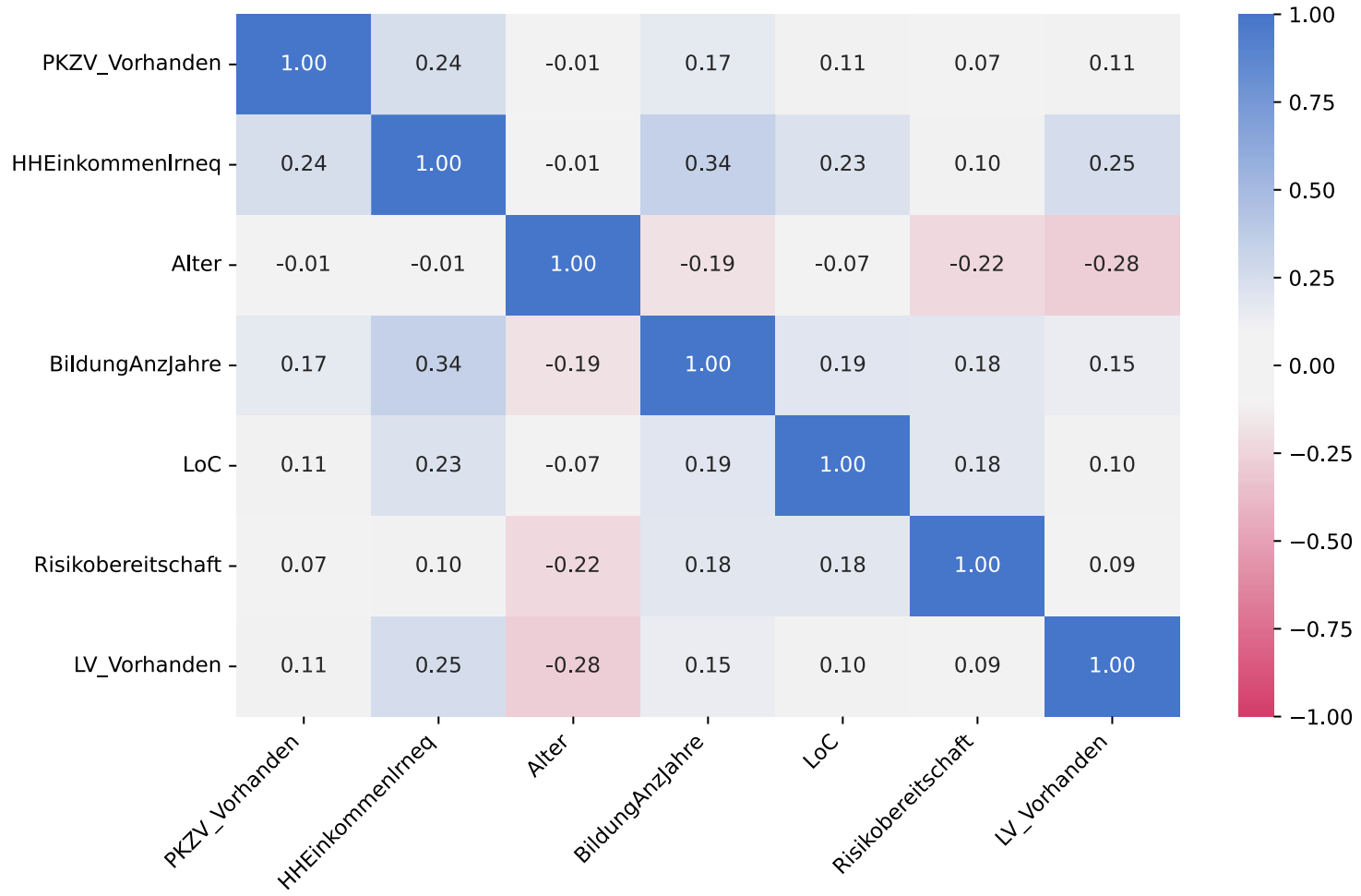


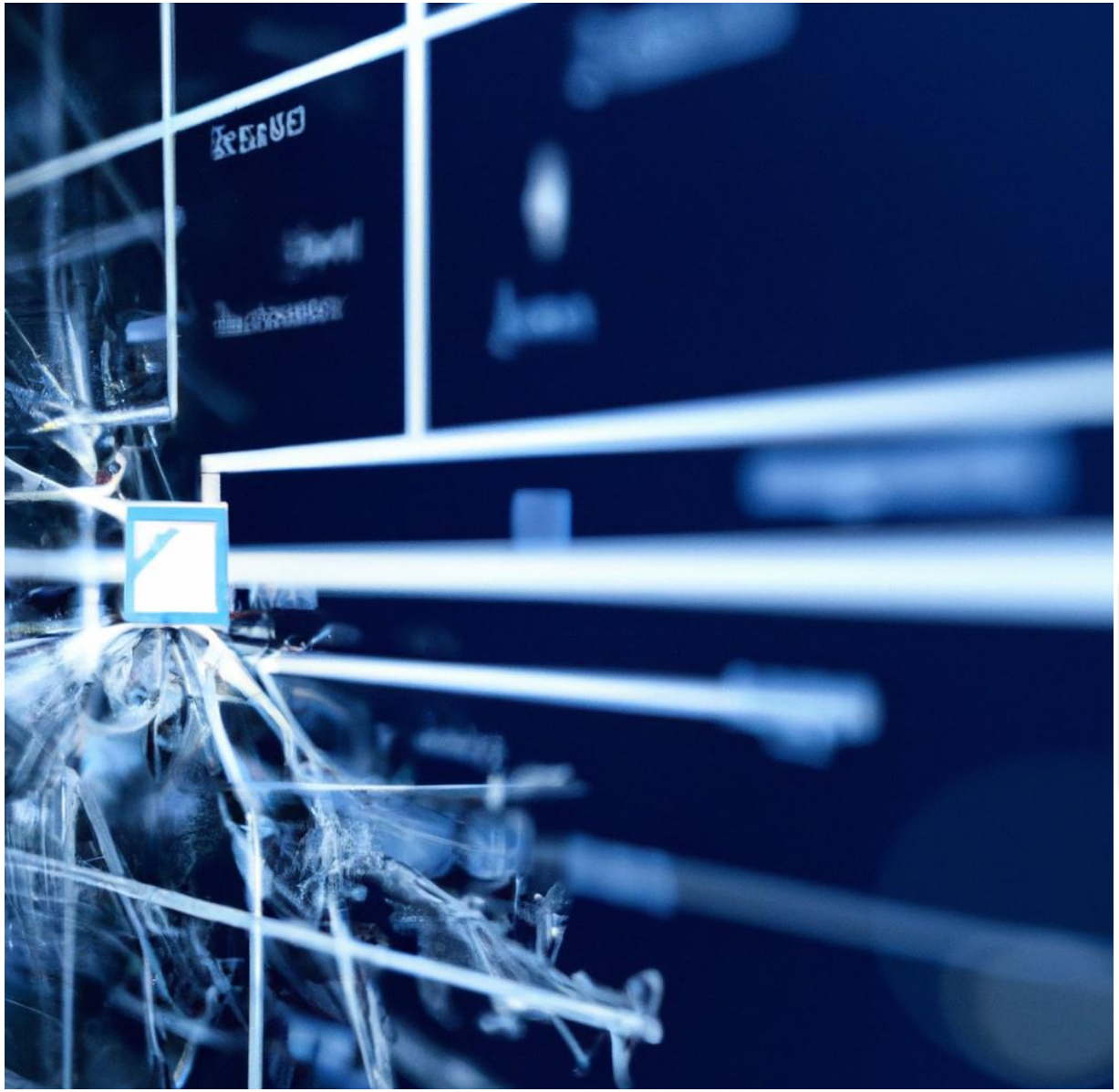






Daten: Deskriptive Statistiken ausgewählter Merkmale (4/4)





1 MOTIVATION

2 DATEN: SOEP

3 ANALYSEERGEBNISSE

4 FAZIT & AUSBLICK

Problemstellung

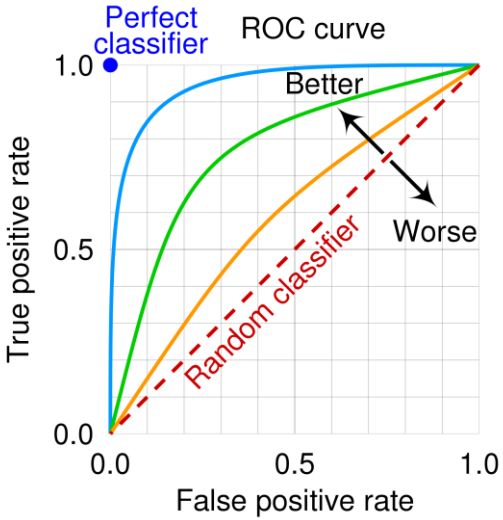
Binäres Klassifikationsproblem:

Prognostiziere das Vorhandensein einer privaten Krankenzusatzversicherung auf Basis von personenbezogenen Merkmalen wie z. B. Alter, Haushaltseinkommen, Locus of Control

BildungAnz Jahre	Risikobereit-schaft	Alter	HHEinkommen lrneq	LoC	Bundesland	Beschäftigung	LV_ Vorhanden	…	PKZV_ Vorhanden
10,5	6,667	41	7,485	1,205	BW	erwerbstätig	1	…	1
14,5	3,727	30	6,662	0,514	ST	erwerbstätig	0	…	0
12,0	2,200	46	7,748	0,513	NI/HB	erwerbstätig	1	…	1
10,5	7,000	53	7,534	-0,042	NI/HB	Sonstiges	0	…	0
12,0	0,636	72	7,268	-1,113	BY	Sonstiges	0	…	1
11,5	3,750	36	6,617	-2,023	BB	arbeitslos	1	…	0
12,0	2,000	28	7,483	1,209	NW	erwerbstätig	1	…	?
12,0	6,000	44	7,318	0,542	ST	erwerbstätig	0	…	?
9,0	5,800	37	7,416	1,337	RP	erwerbstätig	1	…	?

Gütemaß: Area under Curve (AUC)

- *Accuracy* (= Anteil korrekt klassifizierter Samples) als typisches Gütemaß wegen Klassenimbalance (ca. 1:4) nur bedingt geeignet
- Verwendung des Gütemaßes *AUC* (= „Area under (ROC) Curve“):
  - Berücksichtigung von Klassenimbalance
  - Unabhängigkeit von Festlegung des Klassifikationsschwellenwerts
  - Bemessung der Trennschärfe eines Modells
  - Gütemessung über eine einzige Kennzahl
- *ROC Curve* ergibt sich durch Abtragen von False-Positive-Rate zu True-Positive-Rate für unterschiedliche Klassifikationsschwellenwerte; AUC errechnet sich als Fläche unter der ROC Curve
- Interpretation: siehe Schema rechts



Quelle: Medium.com

AUC-Wert	Interpretation
0,5 – 0,6	unbrauchbar
0,6 – 0,7	kaum Trennschärfe
0,7 – 0,8	moderate Trennschärfe
0,8 – 0,9	gute Trennschärfe
0,9 – 1,0	exzellente Trennschärfe

Klassischer Ansatz aus der Statistik: Logistische Regression

Idee:

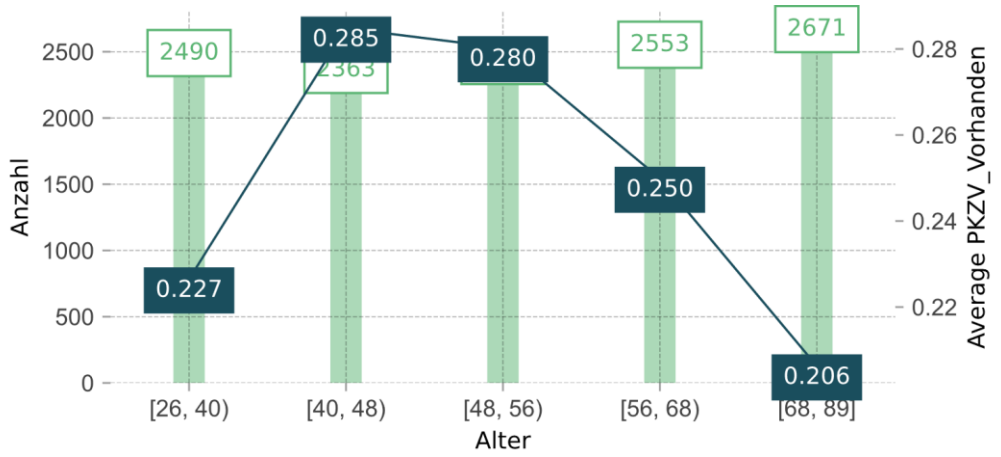
$$\log\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) = \alpha_0 + \sum_{j=1}^{n} \alpha_{i,j} \cdot X_{i,j}$$

Feature	Koeffizient
HHEinkommenlrneq	0,943
Vermoegen__Q5	0,524
LoC	0,066
…	…
Alter	0,002
…	…

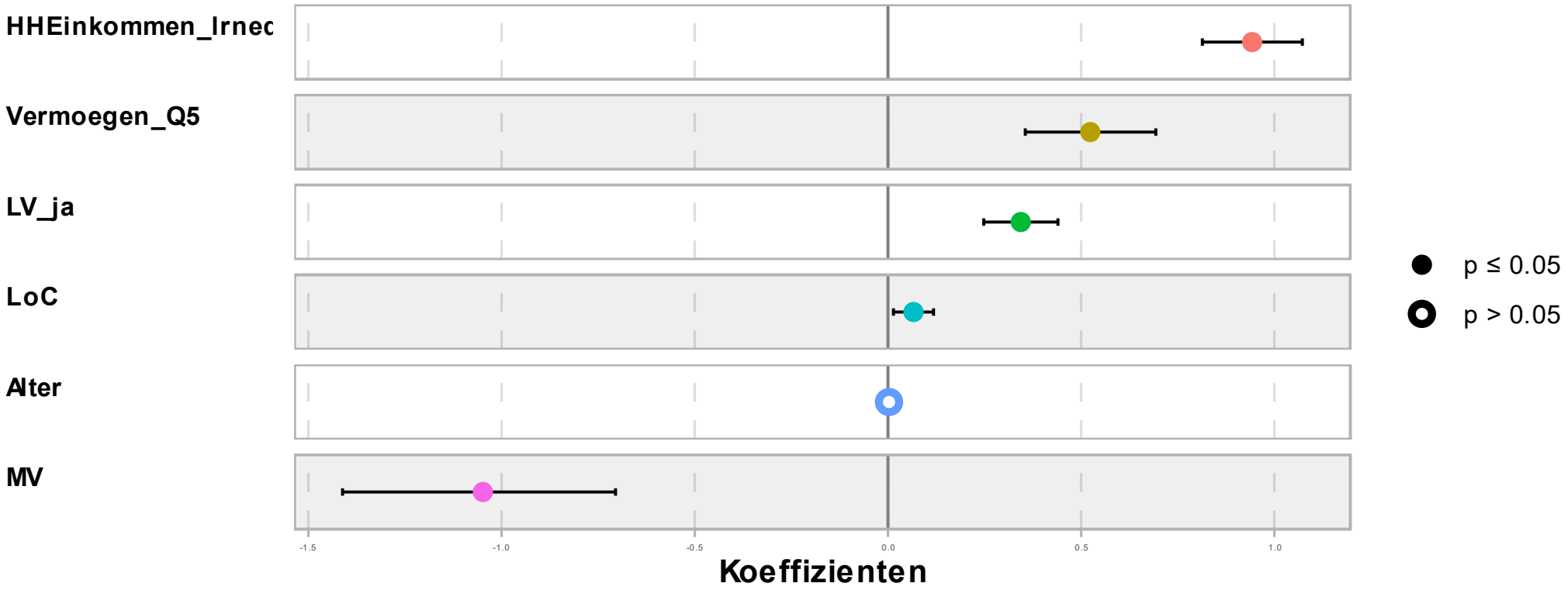
Ausgewählte Vor- und Nachteile:

+ etabliertes, wohlbekanntes Verfahren
+ (vergleichsweise) einfach zu erklären und zu interpretieren

– Nichtlinearitäten nicht adäquat abbildbar
– Prognosegüte oft lediglich moderat



Determinanten via Logistischer Regression



Machine-Learning-Workflow

1 **Datenaufbereitung**
(u. a. Feature Engineering, Skalierung, One-Hot-Encoding, Umgang mit Missing Values)

2 **Modellierung**
(u. a. Auswahl geeigneter Machine-Learning-Verfahren, Hyperparameter-Tuning, Modellfitting)

3 **Evaluation**
(u. a. Prognose auf Test-Daten, Berechnung unterschiedlicher Scores, Vergleich der Verfahren)

4 **Erkenntnisgewinnung**
(u. a. Anwendung von Explainable-AI-Methoden wie z. B. Feature Importances und SHAP)

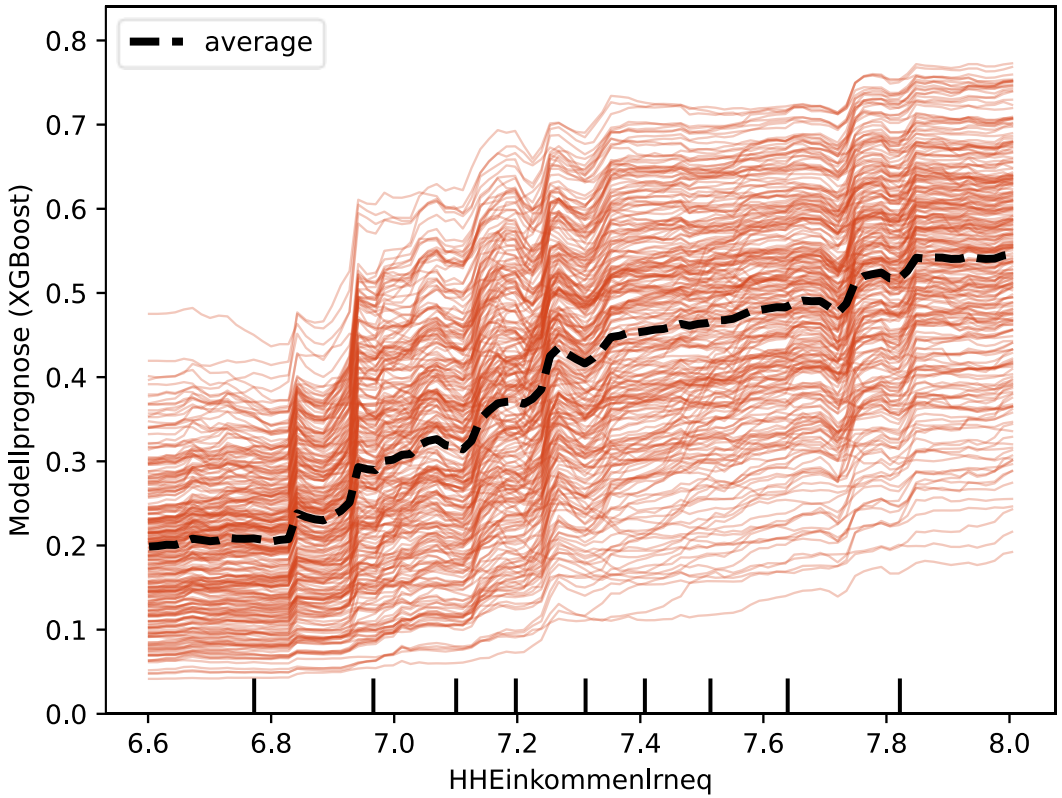
Vergleich der betrachteten Machine-Learning-Verfahren

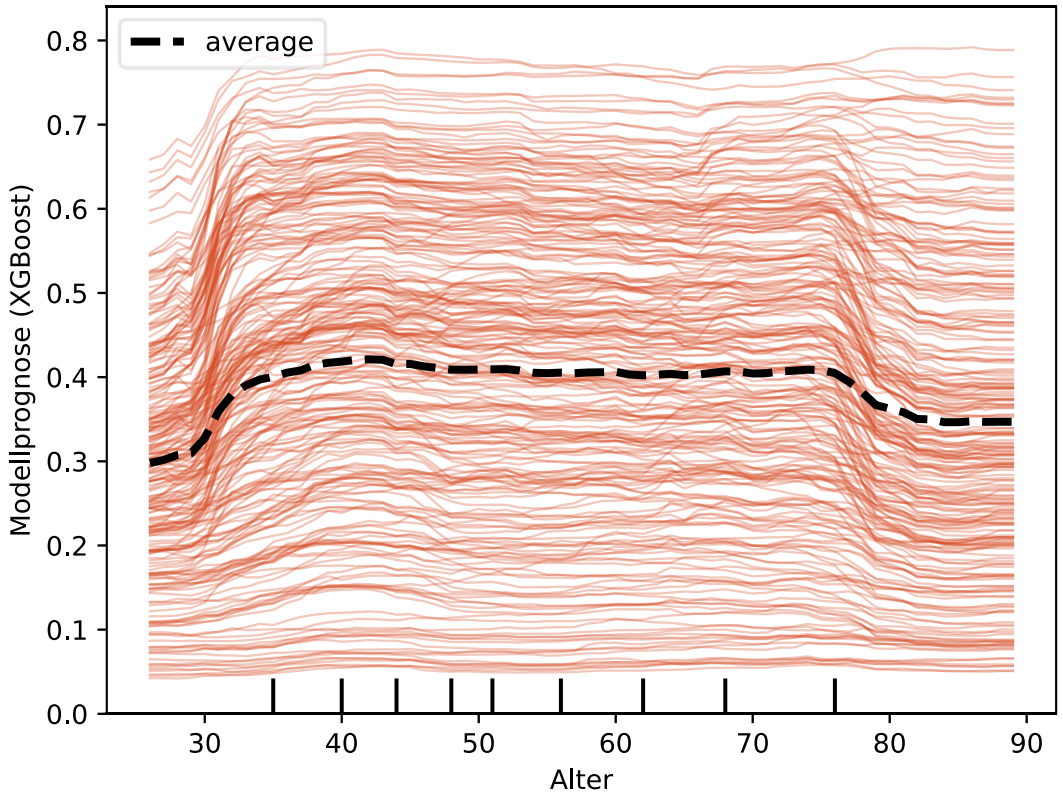
Machine-Learning-Verfahren	AUC		Accuracy	
	in-sample	out-of-sample	in-sample	out-of-sample
Logistische Regression (inkl. Reg.)	0,718	0,693	0,647	0,630
Random Forest	0,814	0,719	0,756	0,752
XGBoost	0,870	0,728	0,785	0,697
AdaBoost	0,784	0,712	0,666	0,634
Künstliches Neuronales Netz	0,744	0,706	0,755	0,750

Erkenntnisse:

- Performance der Logistischen Regression lediglich moderat
- fortgeschrittenere Machine-Learning-Verfahren schneiden besser ab, wenngleich nur geringfügig
- AUC-Ergebnisse werden über weitere Fehlermaße (z. B. Accuracy) bestätigt

Explainable AI: Partial Dependence Plots





Explainable AI: SHAP zur lokalen Interpretierbarkeit

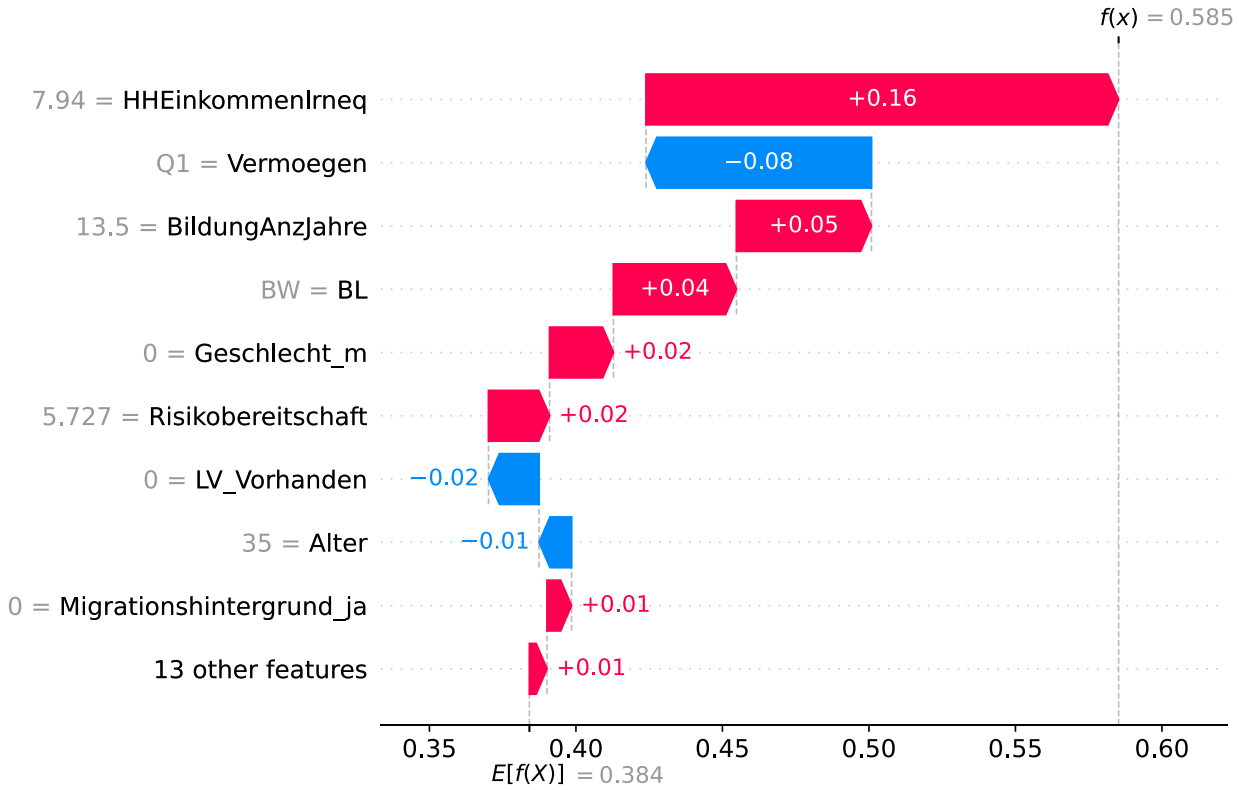
Grundgedanke zu SHAP:

Verwendung von Shapley-Werten zur additiven Zerlegung der Abweichung einer Vorhersage zum Mittel aller Vorhersagen; d. h.,

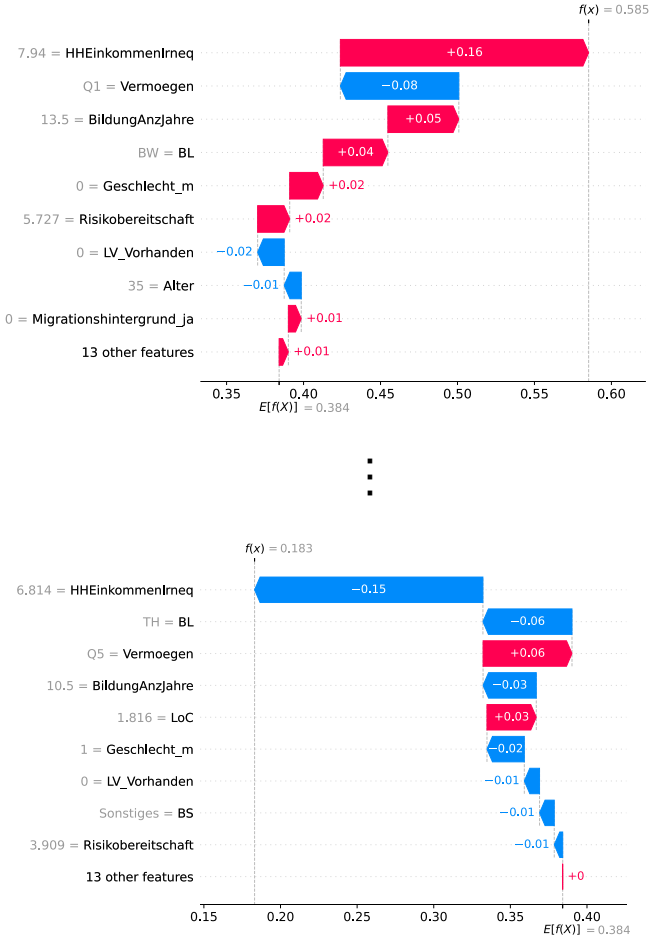
$$\hat{y}_i = \bar{y} + \sum_{j=1}^{m} \phi_{i,j}$$

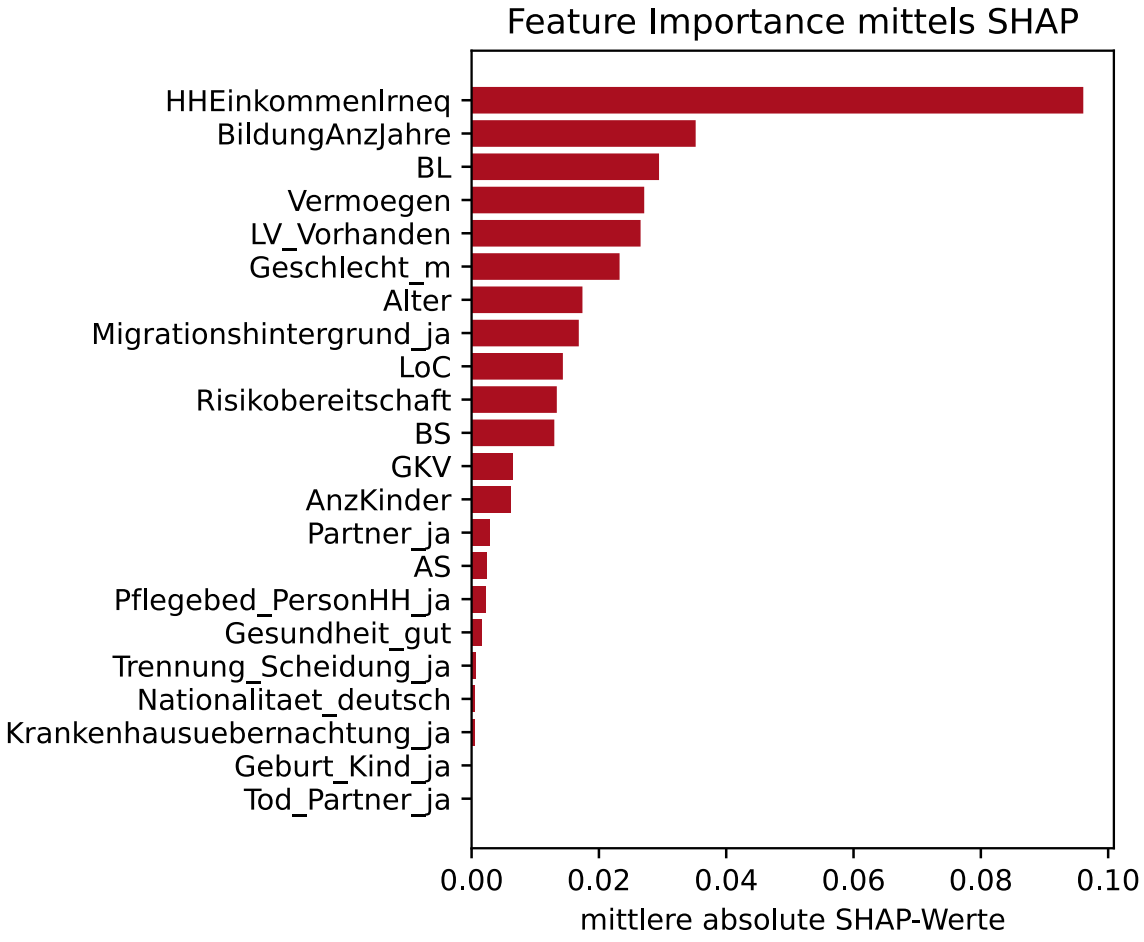
wobei

- $\hat{y}_i$ die Modellprognose zur i-ten Zeile ist,
- $\bar{y}$ der Durchschnitt über alle Modellprognosen ist und
- $\phi_{i,j}$ der Shapley-Wert des Merkmals j der i-ten Zeile ist.

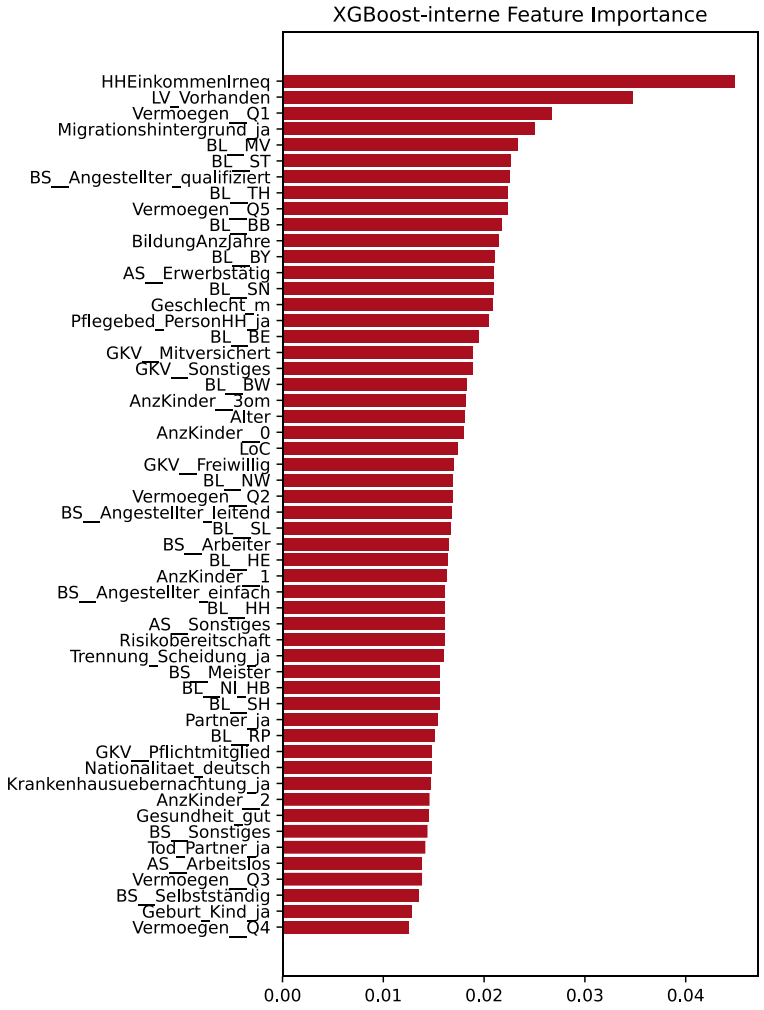


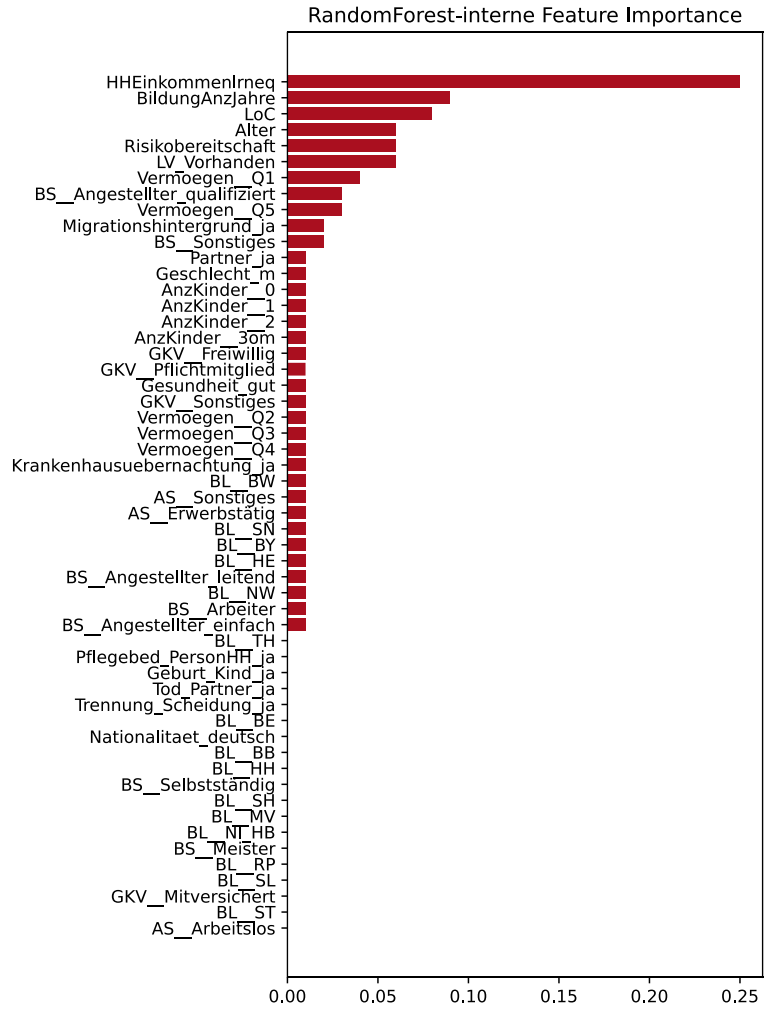
Explainable AI: SHAP zur globalen Interpretierbarkeit

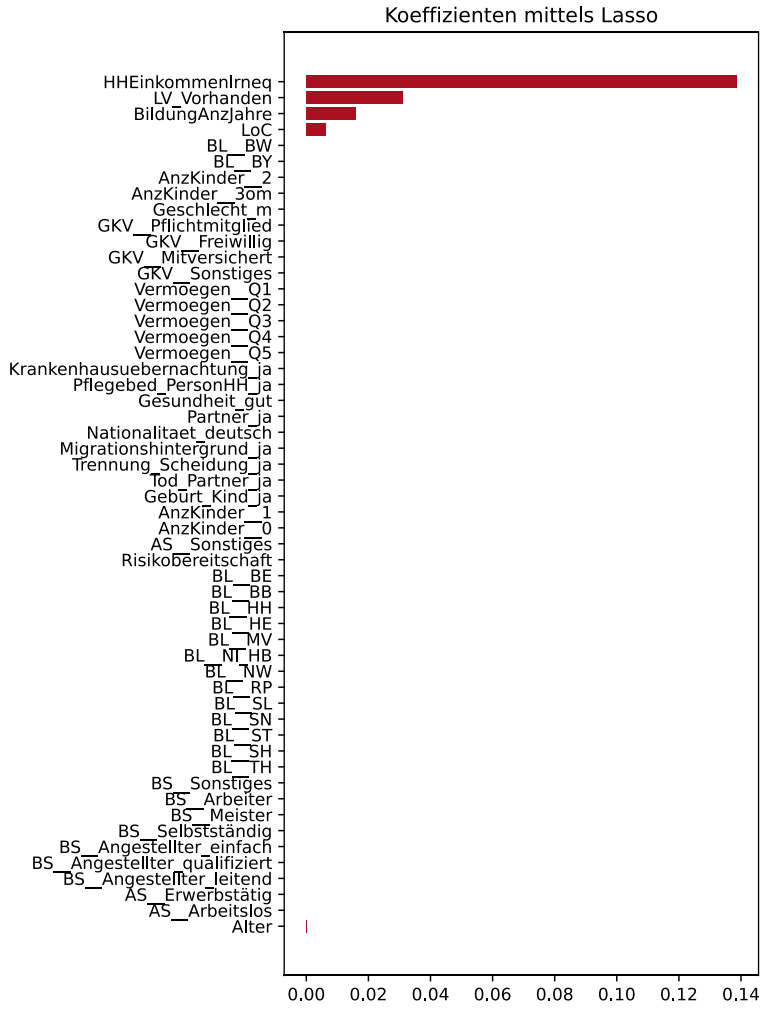


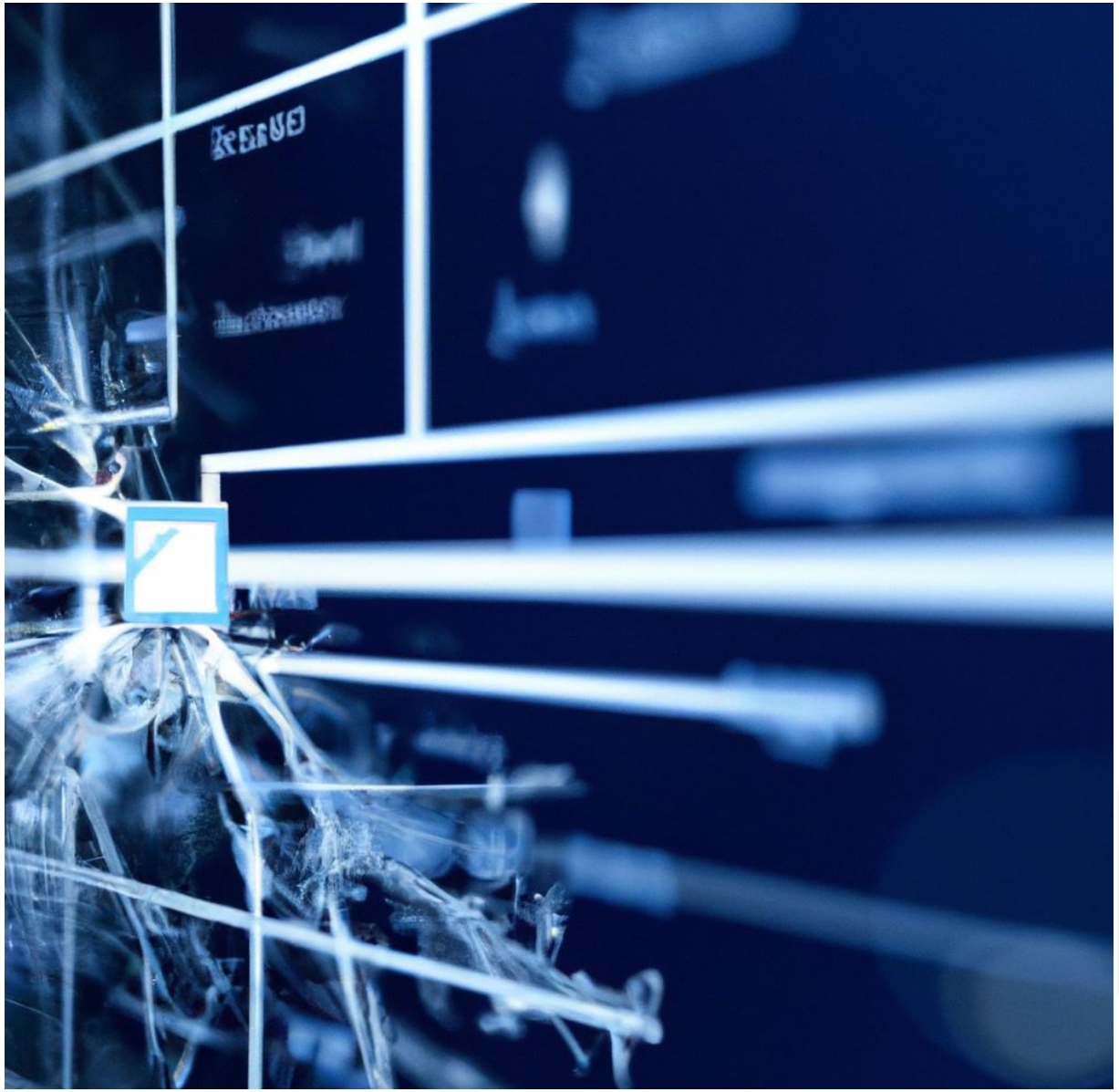


Explainable AI: Weitere Feature Importances









1 MOTIVATION

2 DATEN: SOEP

3 ANALYSEERGEBNISSE

4 FAZIT & AUSBLICK

Fazit & Ausblick

Fazit:

- Prognose in vorliegender Problemstellung herausfordernd
- Machine-Learning-Verfahren schneiden geringfügig besser ab als Logistische Regression (inkl. Reg.)
- identifizierte Determinanten (Haushaltseinkommen, Versicherungsneigung, LoC, Bundesland) decken sich i. W. mit bisherigen Ergebnissen aus der Literatur

Ausblick auf weitere Fragestellungen:

- Verwendung von zusätzlichem Feature Engineering und/oder zusätzlicher Machine-Learning-Modelle
- Fokussierung auf spezielle Krankenzusatzversicherungsarten (z. B. Zahnzusatz)
- Fokussierung auf andere Versicherungsarten (z. B. Lebensversicherung)
- Prognose von “Erwerb” statt von “Vorhandensein”
- Prognose Zahlungsbereitschaft in EUR

Backup-Folien

Locus of Control (LoC): Details

- LoC ist ein persönlichkeitspsychologisches Konstrukt, das sich auf die von einer Person wahrgenommene Verursachung von Situationen bezieht
- Personen mit einem sog. internalen LoC sind der Überzeugung, dass Ereignisse stark von ihren eigenen Fähigkeiten und ihren Anstrengungen gesteuert werden können
- Konstruktion: Hauptkomponentenanalyse auf Basis der nachfolgenden Items, welche jeweils auf einer Skala von 1 bis 7 gemessen werden (1: Stimme überhaupt nicht zu, 7: Stimme voll zu)
  - Habe nicht das erreicht was ich verdiene
  - Mein Lebenslauf hängt von mir ab
  - Was man erreicht hängt vom Glück ab
  - Andere bestimmten über mein Leben
  - Zweifle bei Schwierigkeiten an meinen Fähigkeiten
  - Möglichkeiten werden von sozialen Umständen bestimmt
  - Wenig Kontrolle über mein Leben
- Vorhersage des ersten Faktors liefert stetiges Maß für sog. internalen LoC (vgl. Bonsang und Costa-Font (2022) und Quellen darin)

SHAP (1/2)

Grundgedanke zu SHAP:

Verwendung von Shapley-Werten zur additiven Zerlegung der Abweichung einer Vorhersage zum Mittel aller Vorhersagen; d. h.,

$$\hat{y}_i = \bar{y} + \sum_{j=1}^{m} \phi_{i,j}$$

mit

$$\phi_{i,j} = \sum_{L \subseteq M \setminus \{j\}} \underbrace{\frac{|L|! \cdot (p - |L| - 1)!}{p!}}_{\text{Shapley-Gewicht}} \cdot \underbrace{\left[\hat{f}\left(X_{i,\, L \cup \{j\}}\right) - \hat{f}\left(X_{i,\, L}\right)\right]}_{\text{Beitrag von Merkmal } j},$$

wobei $\hat{y}_i$ die Prädiktion des i-ten Datenpunkts, $\bar{y}$ das Mittel aller Prädiktionen, $M = \{1, \ldots, p\}$ die Menge aller Merkmale und $\hat{f}\left(X_{i,\, L \cup \{j\}}\right)$ bzw. $\hat{f}\left(X_{i,\, L}\right)$ geeignete Prädiktionen unter Verwendung der Merkmale $X_{i,\, L \cup \{j\}}$ bzw. $X_{i,\, L}$ sind.

SHAP (2/2)

(A1) *Efficiency:* $v(\mathcal{M}) = \sum_{j=0}^{p} \phi_j$, where $\phi_0 = v(\emptyset)$ denotes the non-distributed payoff (often set to 0 in cooperative games).

(A2) *Symmetry:* If $v(\mathcal{L} \cup \{j\}) = v(\mathcal{L} \cup \{k\})$ for every $\mathcal{L} \subseteq \mathcal{M} \setminus \{j, k\}$, then $\phi_j = \phi_k$.

(A3) *Dummy player:* If $v(\mathcal{L} \cup \{j\}) = v(\mathcal{L})$ for all coalitions $\mathcal{L} \subseteq \mathcal{M} \setminus \{j\}$, then $\phi_j = 0$.

(A4) *Linearity:* Consider two cooperative games with gain functions v and w. Then, $\phi_j^{(v+w)} = \phi_j^{(v)} + \phi_j^{(w)}$ and $\phi_j^{(\alpha v)} = \alpha \phi_j^{(v)}$ for all $1 \leq j \leq p$ and $\alpha \in \mathbb{R}$.

Quelle: Wüthrich et al., SHAP for Actuaries: Explain any Model (2023)

Literatur

Inhaltlich

- Lange et al. (2017)
  - Ziel: Analyse von Selektionseffekten und Determinanten der privaten Krankenzusatzversicherung (Krankenhausbehandlung & Zahnersatz)
  - Datenbasis: SOEP, Jahre 2008 & 2010
  - Ergebnis: Einkommen und Versicherungsneigung (z. B. Vorhandensein weiterer Versicherungen) sind die wichtigsten Einflussgrößen für das Vorhandensein einer privaten KVZ
- Bonsang und Costa-Font (2022)
  - Ziel: Analyse der Determinanten der privaten Krankenzusatzversicherung
  - Datenbasis: SOEP, Jahre 1999-2016; HILDA (Australien), Jahre 2005-2014
  - Ergebnis: Positiver Effekt eines selbstbestimmt wahrgenommenen Lebens (LoC) auf das Vorhandensein einer privaten KVZ

Methodisch

- Eckert et al. (2021)
  - Ziel: Analyse der Determinanten der Berufsunfähigkeitsversicherung mit ML-Verfahren
  - Angewandte Methoden: Logistische Regression, Random Forest, Gradient Tree Boosting, Stochastic Gradient Descent (mit SVM)

Literaturverzeichnis

- Renate Lange, Jörg Schiller und Petra Steinorth, Demand and selection effects in supplemental health insurance in Germany, The Geneva Papers on Risk and Insurance - Issues and Practice (2017) 42: 5-30.
- Eric Bonsang und Joan Costa-Font, Buying control? ‘Locus of control’ and the uptake of supplementary health insurance, Journal of Economic Behavior and Organization (2022) 204: 476-489.
- Jan Goebel, Markus M. Grabka, Stefan Liebig, Martin Kroh, David Richter, Carsten Schröder und Jürgen Schupp, The German Socio-Economic Panel (SOEP), Jahrbücher für Nationalökonomie und Statistik (2019) 239: 345–360.
- Christian Eckert, Daniela Giesinger, Felix Müller und Antonia Schöning, Machine Learning in der Berufsunfähigkeitsversicherung? Eine Analyse von Risikofaktoren, Der Aktuar (2021), 2: 86-92.