

Erklärbare Künstliche Intelligenz: Eine Diskussion für Aktuarinnen und Aktuare

Anja Schmiedt (Technische Hochschule Rosenheim)
Simon Hatzesberger (Allianz Private Krankenversicherungs-AG)
Benjamin Müller (HDI AG)



DAV

DEUTSCHE
AKTUARVEREINIGUNG e.V.



DGVFM

DEUTSCHE GESELLSCHAFT
FÜR VERSICHERUNGS- UND
FINANZMATHEMATIK e.V.

Herbsttagung von DAV und DGVFM, 20./21.11.2023

Erklärbare Künstliche Intelligenz

Gleichnamige Arbeitsgruppe des Ausschusses
ADS „eXplainable Artificial Intelligence“



Kick-Off im Dezember 2022



Ziele – für heute?

Foto: Von links: Anja Schmiedt, Guido Grützner, Simon Steinbach,
Simon Hatzesberger, Florian Walla, Dariush Sadeghi-Yam, Zoran Nikolić
(Es fehlen: Martin Hüttemann, Benjamin Müller, Corinna Walk)
(Bild: [Anja Schmiedt](#))



Erklärbare Künstliche Intelligenz – Unsere Ergebnistypen

- **Ergebnisbericht: (Arbeitstitel)**

„eXplainable Artificial Intelligence“

– Ein aktueller Überblick für
Aktuarinnen und Aktuare

- Notebooks

1. Einleitung	
1.1. Motivation	
1.2. Zielsetzung	
1.3. Von AI zu ML	
2. Charakterisierung komplexer Modelle.....	
2.1. Definitionen von maschinellen Lernen.....	
2.2. Charakteristika von komplexen Modellen.....	
3. Erklärbarkeit komplexer Modelle	
3.1. Definitionen von Erklärbarkeit	
3.2. Charakteristika von Erklärbarkeit.....	
3.3. Erklärbarkeitsmethoden.....	
3.4. Aktuarielle Sorgfalt	
4. Wann gilt ein Modell als hinreichend erklärt?	
4.1. Kriterien an ein erklärbares Modell	
4.2. Determinanten eines erklärbaren Modells.....	
5. Zusammenfassung und Fazit.....	
Anhang	
A1: Aktuarielle Sorgfalt am Beispiel.....	
Literaturverzeichnis	

Erklärbare Künstliche Intelligenz – Warum das alles?

Branchenübergreifende Leitlinien

Ethik-Leitlinien der High-Level Expertengruppe der Europäischen Kommission (1)

Ethische Grundsätze

Achtung der
menschlichen Autonomie

Schadensverhütung

Fairness

Erklärbarkeit

**Fundamente einer
vertrauenswürdigen KI**

Versicherungsspezifische Leitlinien

AI Governance Principles der EIOPA (2)

Menschliche Aufsicht

Technische Robustheit und
Performance

Datenqualitätsmanagement
und -aufbewahrung

Transparenz und
Erklärbarkeit

Fairness und
Nichtdiskriminierung

Verhältnismäßigkeit

(1) Hochrangige Expertengruppe für künstliche Intelligenz (HEG-KI, 2019). *ETHIK-LEITLINIEN FÜR EINE VERTRAUENSWÜRDIGE KI*. Europäische Kommission, 08.04.2019, [Online](#)

(2) Consultative Expert Group on Digital Ethics in Insurance (2021). *Artificial Intelligence Governance Principles: Towards Ethical and Trustworthy Artificial Intelligence in the European Insurance Sector*. European Insurance and Occupational Pensions Authority, 17.06.2021, [Online](#)

- Künstliche Intelligenz – Definition(en)?



- Maschinelles Lernen (ML) – Definition(en)?



- Differenzierung zwischen ML-Modellen und klassischen statistischen Modellen
 - Definition(en)? Sinnvoll?



Abbildung: Eigene Darstellung

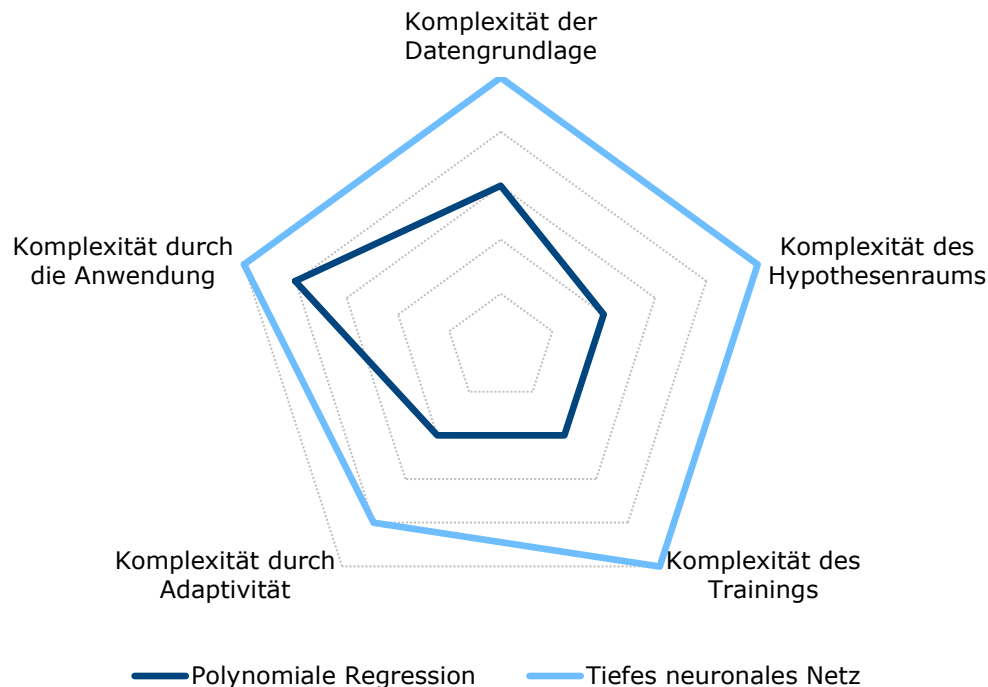
Komplexe Modelle – Charakteristika

- ML-Modelle und klassische statistische Modelle: **Kontinuum**

- Grundlage der Erklärbarkeitsdiskussion: **Modellkomplexität**

Abbildung: Schematische Darstellung zu Charakteristika von komplexen Modellen (eigene Darstellung, vgl. (1))

Charakteristika von komplexen Modellen



Erklärbare Künstliche Intelligenz – Worum geht es?

- Erklärbarkeit – Definition(en)?

“A model is explainable when its internal behavior can be directly understood by **humans (interpretability)** or when **explanations** (justifications) can be **provided** for the main factors that led to its output.”
(European Banking Authority, 2020, S. 6, [Online](#))



- Erklärbarkeit vs. Interpretierbarkeit
- Erklärbarkeit vs. Erklärung
- Erklärbarkeit vs. xAI



Abbildung: Eigene Darstellung

Charakteristika von Erklärbarkeit

- xAI als Forschungsfeld

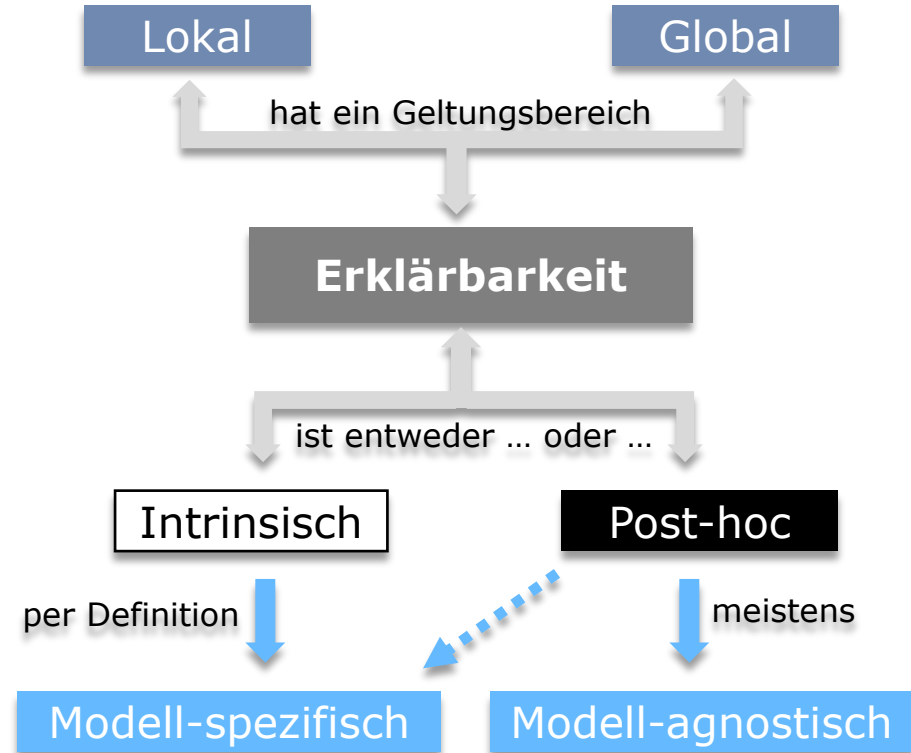
Adadi & Berrada (2018)
Altmann et al. (2020)
Molnar (2020)

...

Baeder et al. (2021)
Delcaillau et al. (2022)
Owens et al. (2022)

...

Abbildung: Schematische Darstellung zu Charakteristika von Erklärbarkeit (eigene Darstellung)



Erklärbare Künstliche Intelligenz – Unsere Ergebnistypen

- Ergebnisbericht

- Notebooks**

- Toy Example
- Reimplementation
- Simulation Study
- Use Case

Katalogisierung der Notebooks der DAV-AG "Erklärbare Künstliche Intelligenz"										
Stand: 20.10.2023										
#	Type	Name of Notebook (type_language_proble m_dataset_supplement .file_format)	Notebook Type	Programmi ng Language	Data set	Data source (Link)	Problem	ML Methods	XAI Method	Global/
1	Toy Example	toy_python_classification_ TravelInsurance.ipynb	Jupyter	Python	Travel Insurance	https://www.kaggle.com/datasets/bjashv14/travel-insurance-prediction-data	Classification	XGBoost, NN	Partial Dependence Plot (PDP) Global Surrogate Model Accumulated Local Effects (ALE) Permutation Feature Importance Shapley Additive Explanations (SHAP) Local Interpretable Model-Agnostic Explanations (LIME) Individual Conditional Expectation (ICE) Counterfactual Explanations Anchors	Global Global Global Global Local Local Local Local
2	Toy Example	toy_python_regression_M edicalCosts.ipynb	Jupyter	Python	Medical Costs	https://www.kaggle.com/datasets/mirichoi0218/insurance	Regression	XGBoost, NN	Partial Dependence Plot (PDP) Global Surrogate Model Accumulated Local Effects (ALE) Permutation Feature Importance Shapley Additive Explanations (SHAP) Local Interpretable Model-Agnostic Explanations (LIME) Individual Conditional Expectation (ICE)	Global Global Global Global Local Local Local
3	Reimplementation	reimpl_python_regression_ SwedishMotorInsurance_ ice-pdp-fi.ipynb	Jupyter	Python	SwedishMotor Insurance	https://www.kaggle.com/datasets/floer/swedish-motor-insurance	Regression	DecisionTree	Individual Conditional Expectation (ICE) Partial Dependence Plot (PDP) Feature Importance (scikit-learn)	Local Global Global
4	Reimplementation	reimpl_python_regression_ SwedishMotorInsurance_ shap.ipynb	Jupyter	Python	SwedishMotor Insurance	https://www.kaggle.com/datasets/floer/swedish-motor-insurance	Regression	DecisionTree	Shapley Additive Explanations (SHAP)	Local
5	Reimplementation	reimpl_python_classification_ on_hastie_ice-pdp.ipynb	Jupyter	Python	make_hastie_ 10_2	https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_hastie_10_2.html	Classification	Gradient Boosting	Individual Conditional Expectation (ICE)	Local
6	Simulation study	sim_python_regression_si mulated_data_ale_applica tion.py	Jupyter	Python	simulated		Regression	Gradient Boosting	Partial Dependence Plot (PDP) Accumulated Local Effects (ALE)	Global Global Global
7	Simulation study	sim_python_regression_si mulated_data_pdp_correl ated_features.py	Jupyter	Python	simulated		Regression	Gradient Boosting	Partial Dependence Plot (PDP) Accumulated Local Effects (ALE)	Global Global Global
8	SCR Use Case	report_scr_maineff.Rmd	Rmarkdown	R	scr_insurance	https://github.com/DAV-AG/DeutscheAktuarvereinigung/blob/master/scr_insurance/scr_maineff.Rmd	Regression	Linear main effects	Standardised regression coefficients permutation feature importance, Principal components Shapley allocation Functional ANOVA	Global Global Global Global Global
9	SCR Use Case	report_scr_quadratic.Rmd	Rmarkdown	R	scr_insurance	https://github.com/DAV-AG/DeutscheAktuarvereinigung/blob/master/scr_insurance/scr_quadratic.Rmd	Regression	Linear with interaction	Sobol indices / Permutation feature importance Shapley allocation	Global Global



Beispiel zur Illustration von xAI-Methoden (1/7)

Regressionsproblem:

Prognose der medizinischen Kosten einer Person basierend auf Alter, Geschlecht, BMI, Kinderanzahl, Raucherstatus und Region

Beispiel zur Illustration von xAI-Methoden (1/7)

Regressionsproblem:

Prognose der medizinischen Kosten einer Person basierend auf Alter, Geschlecht, BMI, Kinderanzahl, Raucherstatus und Region

AGE	SEX	BMI	CHILD-REN	SMO-KER	REGION	CHARGES
18	female	30.115	0	no	northeast	2203.47
24	male	29.830	0	yes	northeast	18648.42
54	female	27.645	3	no	northwest	11305.93
22	female	36.000	0	no	southwest	2166.73
35	female	34.210	1	yes	southeast	48885.23

Beispiel zur Illustration von xAI-Methoden (1/7)

Regressionsproblem:

Prognose der medizinischen Kosten einer Person basierend auf Alter, Geschlecht, BMI, Kinderanzahl, Raucherstatus und Region

AGE	SEX	BMI	CHILD-REN	SMO-KER	REGION	CHARGES
18	female	30.115	0	no	northeast	2203.47
24	male	29.830	0	yes	northeast	18648.42
54	female	27.645	3	no	northwest	11305.93
22	female	36.000	0	no	southwest	2166.73
35	female	34.210	1	yes	southeast	48885.23
44	female	38.060	0	no	southeast	
47	male	36.200	1	no	southwest	
50	female	27.830	2	yes	southeast	

Beispiel zur Illustration von xAI-Methoden (1/7)

Regressionsproblem:

Prognose der medizinischen Kosten einer Person basierend auf Alter, Geschlecht, BMI, Kinderanzahl, Raucherstatus und Region

AGE	SEX	BMI	CHILD-REN	SMO-KER	REGION	CHARGES
18	female	30.115	0	no	northeast	2203.47
24	male	29.830	0	yes	northeast	18648.42
54	female	27.645	3	no	northwest	11305.93
22	female	36.000	0	no	southwest	2166.73
35	female	34.210	1	yes	southeast	48885.23
44	female	38.060	0	no	southeast	?
47	male	36.200	1	no	southwest	?
50	female	27.830	2	yes	southeast	?

Beispiel zur Illustration von xAI-Methoden (1/7)

Regressionsproblem:

Prognose der medizinischen Kosten einer Person basierend auf Alter, Geschlecht, BMI, Kinderanzahl, Raucherstatus und Region

AGE	SEX	BMI	CHILD-REN	SMO-KER	REGION	CHARGES
18	female	30.115	0	no	northeast	2203.47
24	male	29.830	0	yes	northeast	18648.42
54	female	27.645	3	no	northwest	11305.93
22	female	36.000	0	no	southwest	2166.73
35	female	34.210	1	yes	southeast	48885.23
44	female	38.060	0	no	southeast	?
47	male	36.200	1	no	southwest	?
50	female	27.830	2	yes	southeast	?

Verwendetes Verfahren:

Künstliches Neuronales Netz

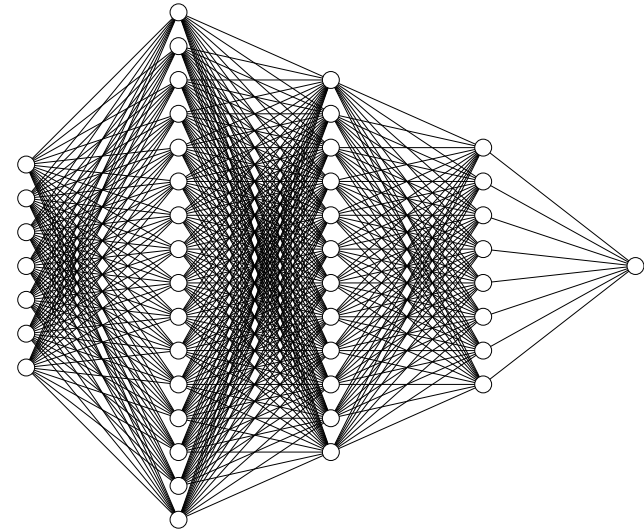
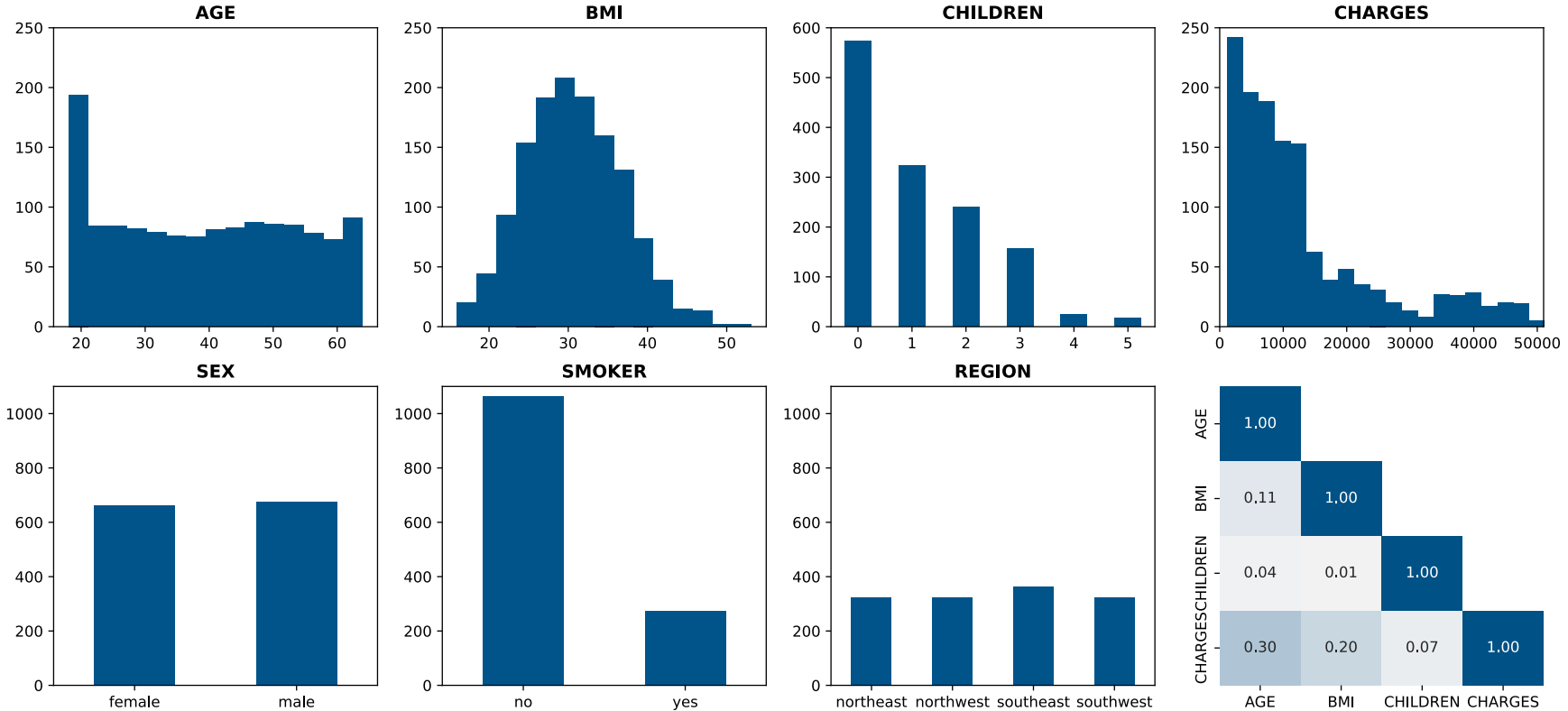


Abbildung: Schematischer Aufbau eines Künstlichen Neuronalen Netzes (*eigene Darstellung*)

Beispiel zur Illustration von xAI-Methoden (2/7)



Beispiel zur Illustration von xAI-Methoden (3/7)

Ausgewählte xAI-Methoden:

	global	lokal
modell-spezifisch		
modell-agnostisch		

Abbildung: Benennung und Klassifizierung ausgewählter xAI-Methoden zum vorliegenden Regressionsproblem ([eigene Darstellung](#))

Beispiel zur Illustration von xAI-Methoden (3/7)

Ausgewählte xAI-Methoden:

	global	lokal
modell-spezifisch	LRP (<i><u>L</u>ayer-wise <u>R</u>elevance <u>P</u>ropagation</i>)	
modell-agnostisch		

Abbildung: Benennung und Klassifizierung ausgewählter xAI-Methoden zum vorliegenden Regressionsproblem (*eigene Darstellung*)

Beispiel zur Illustration von xAI-Methoden (3/7)

Ausgewählte xAI-Methoden:

	global	lokal
modell-spezifisch	LRP (<i>Layer-wise Relevance Propagation</i>)	
modell-agnostisch	Permutation Feature Importance	

Abbildung: Benennung und Klassifizierung ausgewählter xAI-Methoden zum vorliegenden Regressionsproblem (*eigene Darstellung*)

Beispiel zur Illustration von xAI-Methoden (3/7)

Ausgewählte xAI-Methoden:

	global	lokal
modell-spezifisch	LRP (<i>Layer-wise Relevance Propagation</i>)	
modell-agnostisch	Permutation Feature Importance	SHAP (<i>Shapley Additive Explanations</i>)

Abbildung: Benennung und Klassifizierung ausgewählter xAI-Methoden zum vorliegenden Regressionsproblem (*eigene Darstellung*)



Beispiel zur Illustration von xAI-Methoden (4/7)

Grundgedanke zu SHAP:

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

Beispiel zur Illustration von xAI-Methoden (4/7)

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.



Beispiel zur Illustration von xAI-Methoden (5/7)

$$\hat{y}_i = 11488,11$$

$$\bar{y} = 14817,02$$

Beispiel zur Illustration von xAI-Methoden (5/7)

$$\hat{y}_i = 11488,11$$

$$\bar{y} = 14817,02$$

no = SMOKER

56 = AGE

28.595 = BMI

0 = CHILDREN

northeast = REGION

female = SEX

Beispiel zur Illustration von xAI-Methoden (5/7)

$$\hat{y}_i = 11488,11$$

$$\bar{y} = 14817,02$$

$$\phi_{i, \text{SMOKER}} = -5314,95 \quad \text{no} = \text{SMOKER}$$

$$\phi_{i, \text{AGE}} = 3667,35 \quad 56 = \text{AGE}$$

$$\phi_{i, \text{BMI}} = -1829,61 \quad 28.595 = \text{BMI}$$

$$\phi_{i, \text{CHILDREN}} = -454,58 \quad 0 = \text{CHILDREN}$$

$$\phi_{i, \text{REGION}} = 414,81 \quad \text{northeast} = \text{REGION}$$

$$\phi_{i, \text{SEX}} = 188,06 \quad \text{female} = \text{SEX}$$

Beispiel zur Illustration von xAI-Methoden (5/7)

$$\hat{y}_i = 11488,11$$

$$\bar{y} = 14817,02$$

$$\phi_{i, \text{SMOKER}} = -5314,95$$

no = SMOKER

$$\phi_{i, \text{AGE}} = 3667,35$$

56 = AGE

$$\phi_{i, \text{BMI}} = -1829,61$$

28.595 = BMI

$$\phi_{i, \text{CHILDREN}} = -454,58$$

0 = CHILDREN

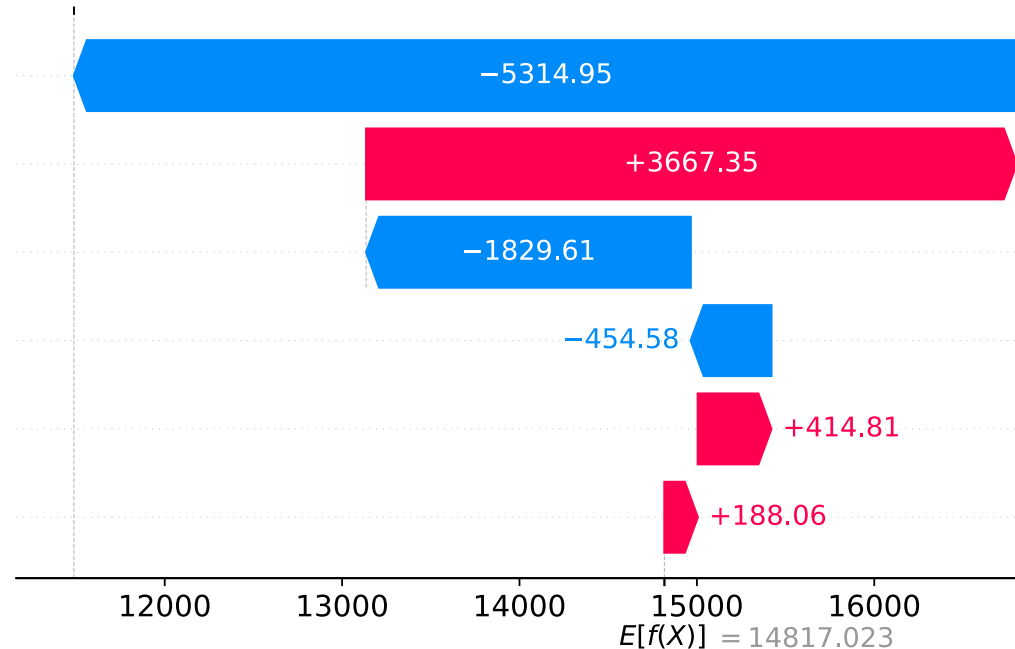
$$\phi_{i, \text{REGION}} = 414,81$$

northeast = REGION

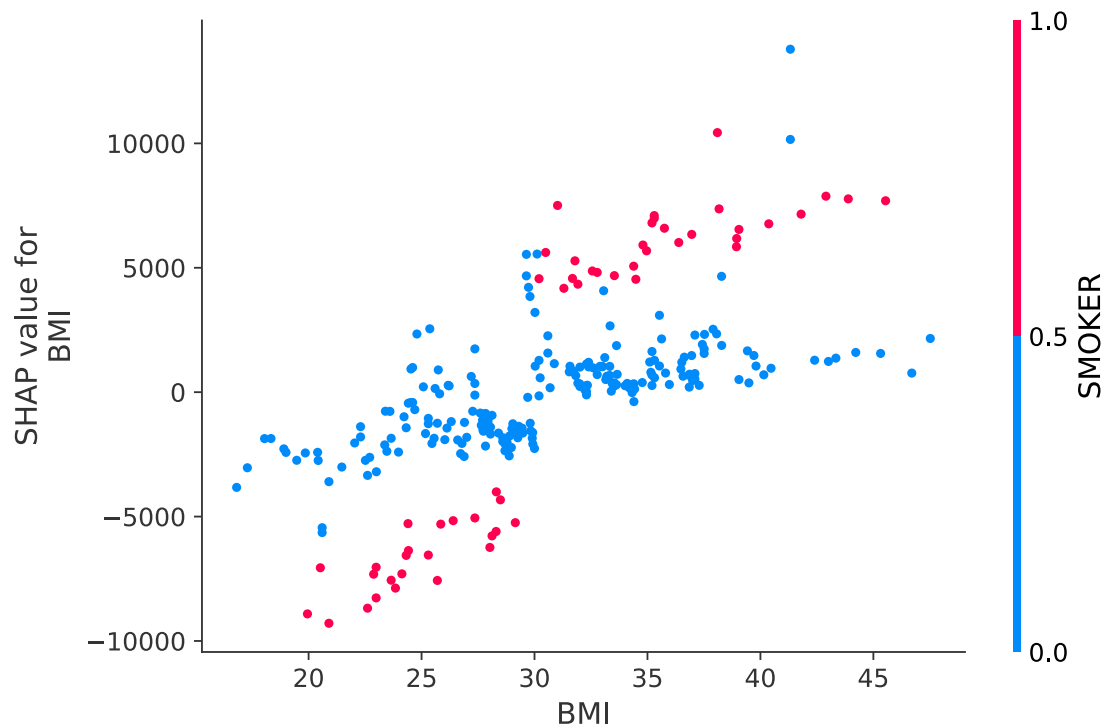
$$\phi_{i, \text{SEX}} = 188,06$$

female = SEX

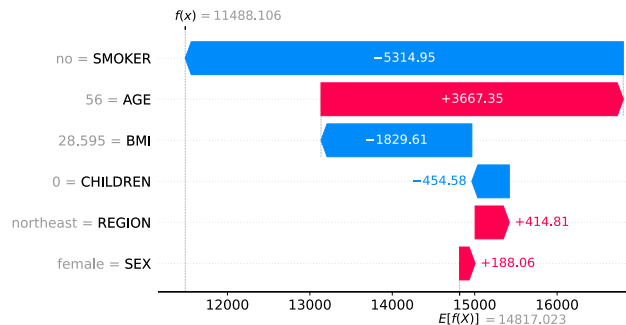
$$f(x) = 11488.106$$



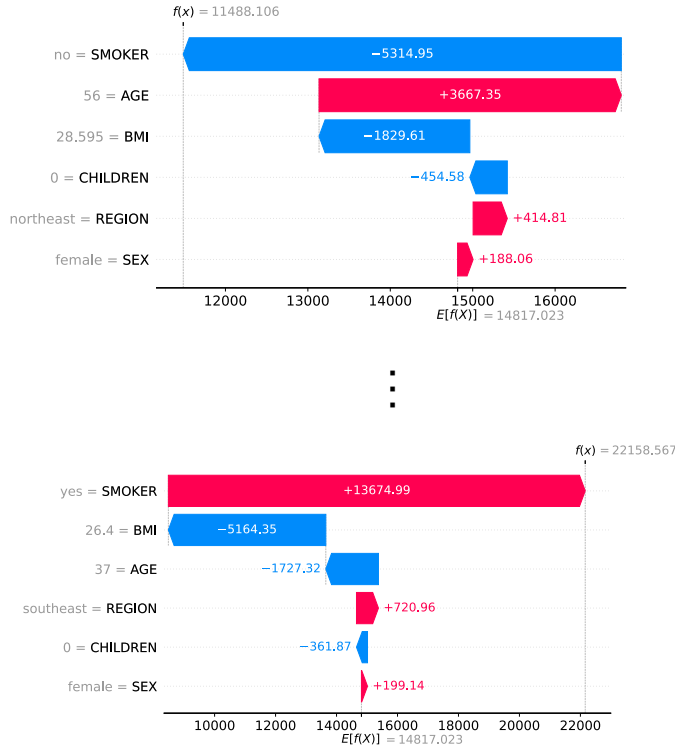
Beispiel zur Illustration von xAI-Methoden (6/7)



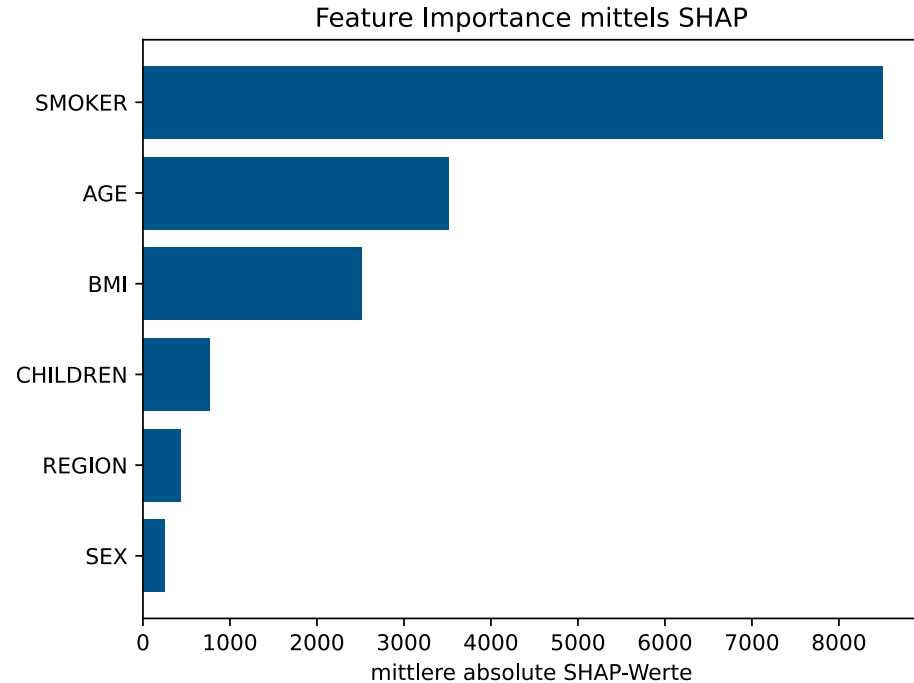
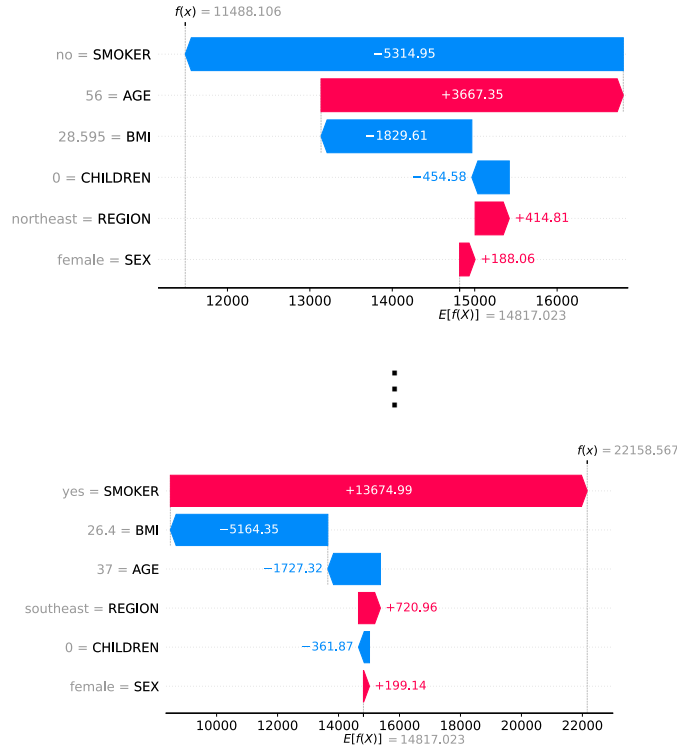
Beispiel zur Illustration von xAI-Methoden (7/7)



Beispiel zur Illustration von xAI-Methoden (7/7)



Beispiel zur Illustration von xAI-Methoden (7/7)



Erklärbare Künstliche Intelligenz – Unsere Ergebnistypen

- Ergebnisbericht

- Notebooks**

- Toy Example
- Reimplementation
- Simulation Study
- Use Case

#	Type	Name of Notebook (type_language_proble m_dataset_supplement .file_format)	Notebook Type	Programmi ng Language	Data set	Data source (Link)	Problem	ML Methods	XAI Method	Global/
1	Toy Example	toy_python_classification_ TravelInsurance.ipynb	Jupyter	Python	Travel Insurance	https://www.kaggle.com/datasets/bjashv14/travel-insurance-prediction-data	Classification	XGBoost, NN	Partial Dependence Plot (PDP) Global Surrogate Model Accumulated Local Effects (ALE) Permutation Feature Importance Shapley Additive Explanations (SHAP) Local Interpretable Model-Agnostic Explanations (LIME) Individual Conditional Expectation (ICE) Counterfactual Explanations Anchors	Global Global Global Global Local Local Local Local
2	Toy Example	toy_python_regression_M edicalCosts.ipynb	Jupyter	Python	Medical Costs	https://www.kaggle.com/datasets/mirichoi0218/insurance	Regression	XGBoost, NN	Partial Dependence Plot (PDP) Global Surrogate Model Accumulated Local Effects (ALE) Permutation Feature Importance Shapley Additive Explanations (SHAP) Local Interpretable Model-Agnostic Explanations (LIME) Individual Conditional Expectation (ICE)	Global Global Global Global Local Local Local
3	Reimplementation	reimpl_python_regression_ SwedishMotorInsurance_ ice-pdp-fi.ipynb	Jupyter	Python	SwedishMotorI nsurance	https://www.kaggle.com/datasets/floer/swedish-motor-insurance	Regression	DecisionTree	Individual Conditional Expectation (ICE) Partial Dependence Plot (PDP)	Local Global
4	Reimplementation	reimpl_python_regression_ SwedishMotorInsurance_ shap.ipynb	Jupyter	Python	SwedishMotorI nsurance	https://www.kaggle.com/datasets/floer/swedish-motor-insurance	Regression	DecisionTree	Shapley Additive Explanations (SHAP)	Local
5	Reimplementation	reimpl_python_classificati on_hastie_ice-pdp.ipynb	Jupyter	Python	make_hastie_ 10_2	https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_hastie_10_2.html	Classification	Gradient Boosting	Individual Conditional Expectation (ICE)	Local
6	Simulation study	sim_python_regression_si mulated_data_ale_applica tion.py	Jupyter	Python	simulated		Regression	Gradient Boosting	Partial Dependence Plot (PDP) Accumulated Local Effects (ALE)	Global Global Global
7	Simulation study	sim_python_regression_si mulated_data_pdp_correl ated_features.py	Jupyter	Python	simulated		Regression	Gradient Boosting	Partial Dependence Plot (PDP) Accumulated Local Effects (ALE)	Global Global
8	SCR Use Case	report_scr_maineff.Rmd	Rmarkdown	R	scr_insurance	https://github.com/DAV-AG/DeutscheAktuarvereinigung/blob/master/scr_insurance/scr_maineff.Rmd	Regression	Linear main effects	Standardised regression coefficients permutation feature importance, Principal components	Global Global Global
9	SCR Use Case	report_scr_quadratic.Rmd	Rmarkdown	R	scr_insurance	https://github.com/DAV-AG/DeutscheAktuarvereinigung/blob/master/scr_insurance/scr_quadratic.Rmd	Regression	Linear with interaction	Shapley allocation Functional ANOVA Sobol indices / Permutation feature importance Shapley allocation	Global Global Global Global

 :
:
:

Abbildung: Katalogisierung der AG-Notebooks,
Stand 20.10.23 (Arbeitsfassung)

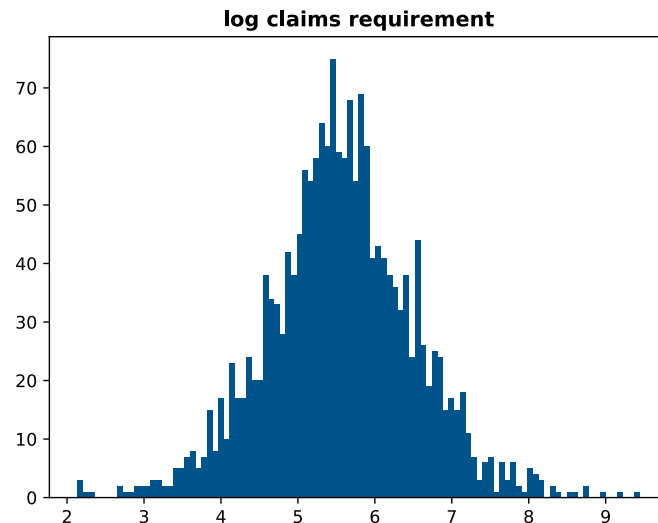
 :
:
:

Laborbeispiel

- Lerntyp: überwachtes Lernen
- Problemtyp: Regression
- Datensatz: abgeleitet von SwedishMotorInsurance¹, 1797 Zeilen, 5 Spalten
- Features: alle kategoriell

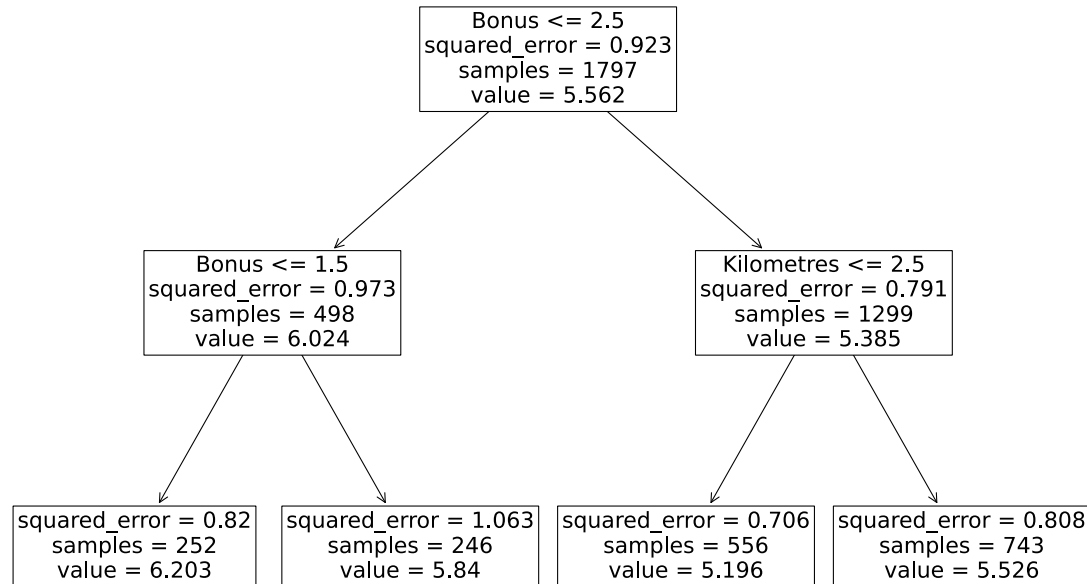
Zielvariable: Schadenbedarf
(logarithmiert)

Feature	Anzahl Ausprägungen
Kilometres	5
Zone	7
Bonus	7
Make	9



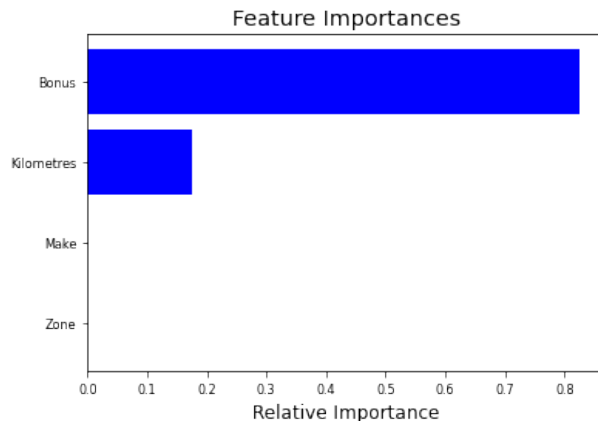
¹: <https://www.kaggle.com/code/ashwin8699/swedish-motor-insurance-simple-linear-regression/input>

Einfaches Modell (Entscheidungsbaum, scikit-learn)



Baumtiefe: 2

Anwendung von FI (scikit-learn, baumbasierte Verfahren)



```
pd.DataFrame(data=DT.feature_importances_,
             index=DT.feature_names_in_)
```

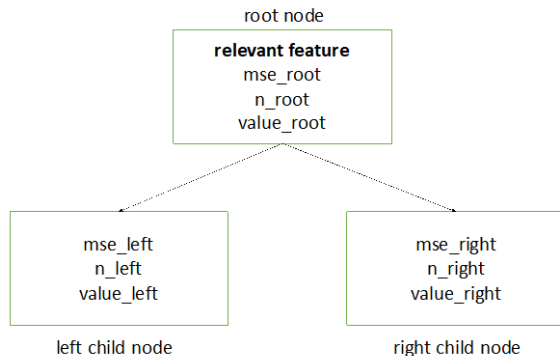
	0
Kilometres	0.174492
Zone	0.000000
Bonus	0.825508
Make	0.000000

```
def plot_feature_importances(feature_names, feature_importances):
    """
    help function:
    horizontal bar plot of known feature importances and corresponding feature names
    """
    indices = np.argsort(feature_importances)
    plt.title('Feature Importances')
    plt.barh(range(len(indices)), feature_importances[indices], color='b', align='center')
    plt.yticks(range(len(indices)), [feature_names[i] for i in indices], fontsize=8)
    plt.xticks(fontsize=8)
    plt.xlabel('Relative Importance')
    plt.show()

plot_feature_importances(DT.feature_names_in_, DT.feature_importances_)
```

Wie funktioniert Feature Importance in scikit-learn?

Binärer Split:



mit $n_{root} = n_{left} + n_{right}$

Gewinn für Knoten i:

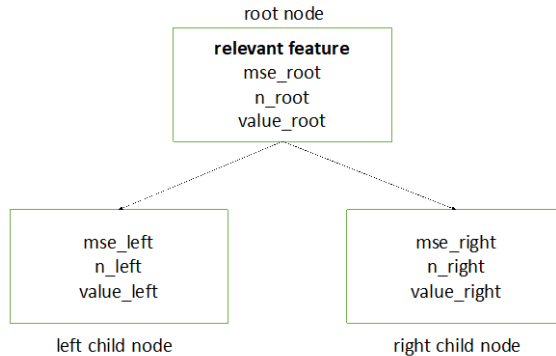
$$g_i = mse_{root} - \frac{n_{left}}{n_{root}} \cdot mse_{left} - \frac{n_{right}}{n_{root}} \cdot mse_{right}$$

Feature Importance (vor Normierung):

$$FI(feature) = \sum_{\{ \text{node } i: \text{feature is relevant feature} \}} \frac{n_i}{n_{total}} \cdot g_i$$

Wie funktioniert Feature Importance in scikit-learn?

Binärer Split:



mit $n_{root} = n_{left} + n_{right}$

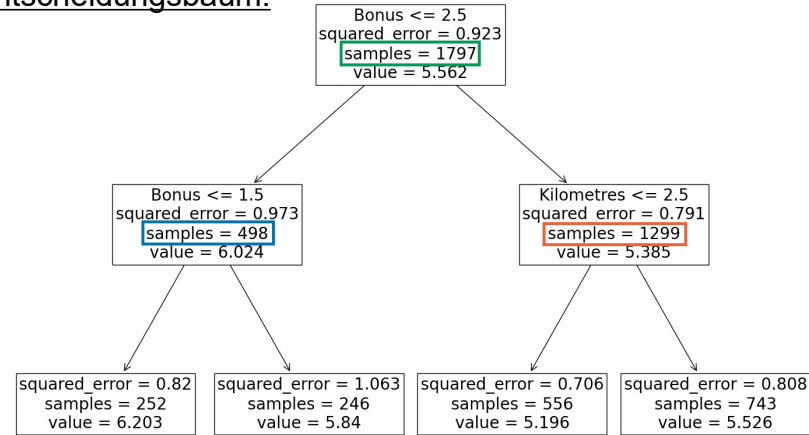
Gewinn für Knoten i:

$$g_i = mse_{root} - \frac{n_{left}}{n_{root}} \cdot mse_{left} - \frac{n_{right}}{n_{root}} \cdot mse_{right}$$

Feature Importance (vor Normierung):

$$FI(feature) = \sum_{\{node\ i: \text{feature is relevant feature}\}} \frac{n_i}{n_{total}} \cdot g_i$$

Entscheidungsbaum:

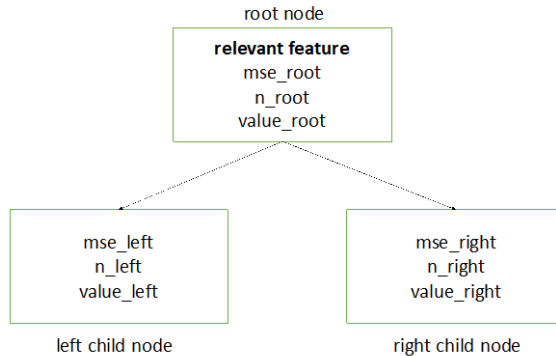


$$FI(Bonus) = \frac{1797}{1797} \cdot g_1 + \frac{498}{1797} \cdot g_2$$

$$FI(Kilometres) = \frac{1299}{1797} \cdot g_3$$

Wie funktioniert Feature Importance in scikit-learn?

Binärer Split:



mit $n_{root} = n_{left} + n_{right}$

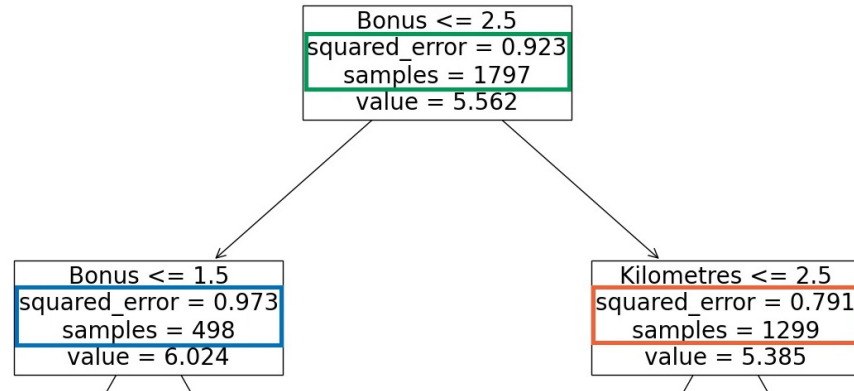
Gewinn für Knoten i:

$$g_i = mse_{root} - \frac{n_{left}}{n_{root}} \cdot mse_{left} - \frac{n_{right}}{n_{root}} \cdot mse_{right}$$

Feature Importance (vor Normierung):

$$FI(feature) = \sum_{\{node\ i: \text{feature is relevant feature}\}} \frac{n_i}{n_{total}} \cdot g_i$$

Knoten 1:



$$g_1 = 0.923 - \frac{498}{1797} \cdot 0.973 - \frac{1299}{1797} \cdot 0.791$$

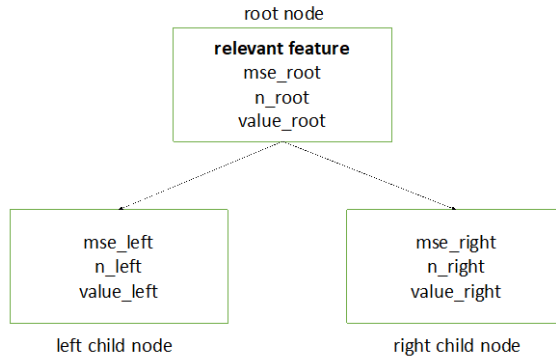
$$\approx 0.0816$$

$$FI(Bonus) \approx \frac{1797}{1797} \cdot 0.0816 + \frac{498}{1797} \cdot g_2$$

$$FI(Kilometres) = \frac{1299}{1797} \cdot g_3$$

Wie funktioniert Feature Importance in scikit-learn?

Binärer Split:



mit $n_{root} = n_{left} + n_{right}$

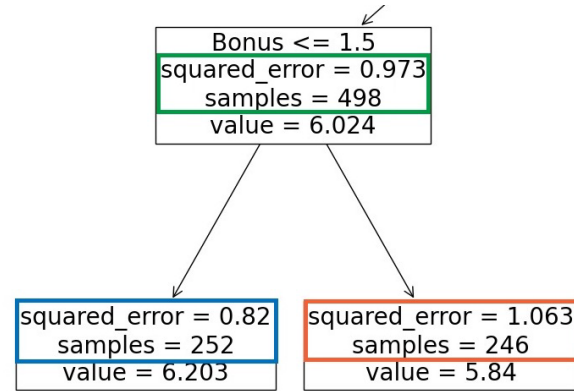
Gewinn für Knoten i:

$$g_i = mse_{root} - \frac{n_{left}}{n_{root}} \cdot mse_{left} - \frac{n_{right}}{n_{root}} \cdot mse_{right}$$

Feature Importance (vor Normierung):

$$FI(feature) = \sum_{\{node\ i: \text{feature is relevant feature}\}} \frac{n_i}{n_{total}} \cdot g_i$$

Knoten 2:



$$g_2 = 0.973 - \frac{252}{498} \cdot 0.82 - \frac{246}{498} \cdot 1.063$$

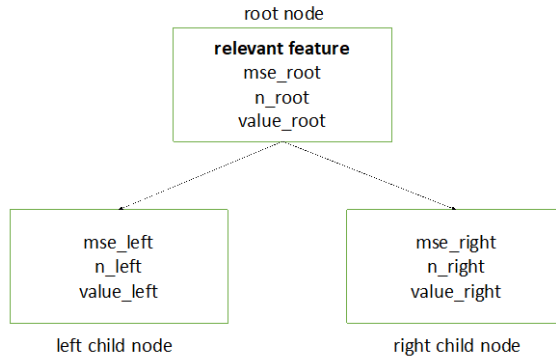
$$\approx 0.03296$$

$$FI(Bonus) \approx \frac{1797}{1797} \cdot 0.0816 + \frac{498}{1797} \cdot 0.03296$$

$$FI(Kilometres) = \frac{1299}{1797} \cdot g_3$$

Wie funktioniert Feature Importance in scikit-learn?

Binärer Split:



mit $n_{root} = n_{left} + n_{right}$

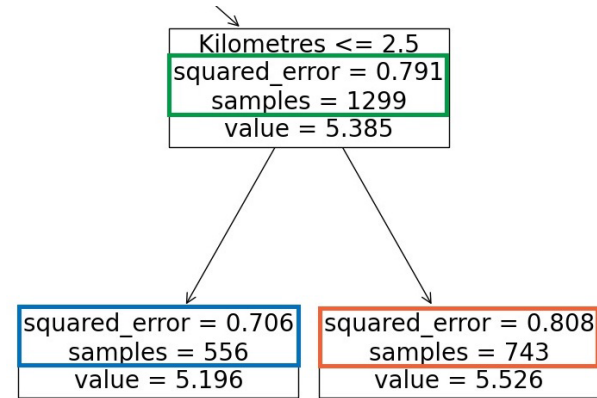
Gewinn für Knoten i:

$$g_i = mse_{root} - \frac{n_{left}}{n_{root}} \cdot mse_{left} - \frac{n_{right}}{n_{root}} \cdot mse_{right}$$

Feature Importance (vor Normierung):

$$FI(feature) = \sum_{\{node\ i: \text{feature is relevant feature}\}} \frac{n_i}{n_{total}} \cdot g_i$$

Knoten 3:



$$g_3 = 0.791 - \frac{556}{1299} \cdot 0.706 - \frac{743}{1299} \cdot 0.808 \approx 0.02666$$

$$FI(Bonus) \approx \frac{1797}{1797} \cdot 0.0816 + \frac{498}{1797} \cdot 0.03296 \approx 0.09073$$

$$FI(Kilometres) \approx \frac{1299}{1797} \cdot 0.02666 \approx 0.01927$$

Feature Importance – Normierung

Die Summe der berechneten Feature Importances ist:

$$\begin{aligned} \text{sum}_{FI} &:= FI(\text{Bonus}) + FI(\text{Kilometres}) \\ &\approx 0.09073 + 0.01927 = 0.11. \end{aligned}$$

Feature Importance (normiert):

$$\begin{aligned} FI(\text{Bonus})^{\text{normed}} &= \frac{FI(\text{Bonus})}{\text{sum}_{FI}} \approx \frac{0.09073}{0.11} \approx 0.82482 \\ FI(\text{Kilometres})^{\text{normed}} &= \frac{FI(\text{Kilometres})}{\text{sum}_{FI}} \approx \frac{0.01927}{0.11} \approx 0.17518. \end{aligned}$$

Erklärbare Künstliche Intelligenz – Aktuarielle Sorgfalt

Über einen sorgfältigen Einsatz von xAI-Methoden

- Was ist das Ziel und warum diese Methode?
- Wo und inwieweit ist die Methode dokumentiert?
- Inwieweit sind Voraussetzungen der Methode erfüllt?
- Wie sind die Ergebnisse zu interpretieren?
- Inwieweit ist Schätzunsicherheit berücksichtigt?

1. Einleitung	
1.1. Motivation	
1.2. Zielsetzung	
1.3. Von AI zu ML	
2. Charakterisierung komplexer Modelle.....	
2.1. Definitionen von maschinellen Lernen.....	
2.2. Charakteristika von komplexen Modellen.....	
3. Erklärbarkeit komplexer Modelle	
3.1. Definitionen von Erklärbarkeit	
3.2. Charakteristika von Erklärbarkeit	
3.3. Erklärbarkeitsmethoden.....	
3.4. Aktuarielle Sorgfalt	
4. Wann gilt ein Modell als hinreichend erklärt?	
4.1. Kriterien an ein erklärbares Modell	
4.2. Determinanten eines erklärbaren Modells.....	
5. Zusammenfassung und Fazit.....	
Anhang	
A1: Aktuarielle Sorgfalt am Beispiel.....	
Literaturverzeichnis	

Erklärbare Künstliche Intelligenz – Zusammenfassung

1. Einleitung

1.1. Motivation

1.2. Zielsetzung

1.3. Von AI zu ML

2. Charakterisierung komplexer Modelle.....

2.1. Definitionen von maschinellen Lernen.....

2.2. Charakteristika von komplexen Modellen.....

3. Erklärbarkeit komplexer Modelle

3.1. Definitionen von Erklärbarkeit

3.2. Charakteristika von Erklärbarkeit

3.3. Erklärbarkeitsmethoden.....

3.4. Aktuarielle Sorgfalt

4. Wann gilt ein Modell als hinreichend erklärt?

4.1. Kriterien an ein erklärbares Modell

4.2. Determinanten eines erklärbaren Modells.....

5. Zusammenfassung und Fazit.....

Anhang

A1: Aktuarielle Sorgfalt am Beispiel.....

Literaturverzeichnis

Katalogisierung der Notebooks der DAV-AG "Erklärbare Künstliche Intelligenz"

Stand: 20.10.2023

#	Type	Name of Notebook (type_language_proble m_dataset_supplement .file_format)	Notebook Type	Programmi ng Language	Data set	Data source (Link)	Problem	ML Methods	XAI Method	Global/
1	Toy Example	toy_python_classification_TravelInsurance.ipynb	Jupyter	Python	Travel Insurance	https://www.kaggle.com/datasets/tjashv14/travel-insurance-prediction-data	Classification	XGBoost, NN	Partial Dependence Plot (PDP) Global Surrogate Model Accumulated Local Effects (ALE) Permutation Feature Importance Shapley Additive Explanations (SHAP) Local Interpretable Model-Agnostic Explanations (LIME) Individual Conditional Expectation (ICE) Counterfactual Explanations Anchors	Global Global Global Global Local Local Local Local
2	Toy Example	toy_python_regression_MedicalCosts.ipynb	Jupyter	Python	Medical Costs	https://www.kaggle.com/datasets/mlr-ichol0218/insurance	Regression	XGBoost, NN	Partial Dependence Plot (PDP) Global Surrogate Model Accumulated Local Effects (ALE) Permutation Feature Importance Shapley Additive Explanations (SHAP) Local Interpretable Model-Agnostic Explanations (LIME) Individual Conditional Expectation (ICE)	Global Global Global Global Local Local Local
3	Reimplementation	reimpl_python_regression_SwedishMotorInsurance_ice-pdp-flj.ipynb	Jupyter	Python	SwedishMotorInsurance	https://www.kaggle.com/datasets/fljser/swedish-motor-insurance	Regression	DecisionTree	Partial Dependence Plot (PDP) Feature Importance (scikit-learn)	Global Global Global Global
4	Reimplementation	reimpl_python_regression_SwedishMotorInsurance_shap.ipynb	Jupyter	Python	SwedishMotorInsurance	https://www.kaggle.com/datasets/fljser/swedish-motor-insurance	Regression	DecisionTree	Shapley Additive Explanations (SHAP)	Local
5	Reimplementation	reimpl_python_classification_hastie_ice-pdp.ipynb	Jupyter	Python	make_hastie_10_2	https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_hastie_10_2.html	Classification	Gradient Boosting	Individual Conditional Expectation (ICE)	Local
6	Simulation study	sim_python_regression_simulated_data_ale_application.py	Jupyter	Python	simulated		Regression	Gradient Boosting	Partial Dependence Plot (PDP) Accumulated Local Effects (ALE)	Global Global
7	Simulation study	sim_python_regression_simulated_data_pdp_correlated_features.py	Jupyter	Python	simulated		Regression	Gradient Boosting	Partial Dependence Plot (PDP) Accumulated Local Effects (ALE)	Global Global
8	SCR Use Case	report_scr_maineff.Rmd	Rmarkdown	R	scr_insurance	https://github.com/DeutscheAktuarvereinigung/insurance-notebooks	Regression	Linear main effects	Standardised regression coefficients Permutation feature importance, Principal components Shapley allocation	Global Global Global
9	SCR Use Case	report_scr_quadratic.Rmd	Rmarkdown	R	scr_insurance	https://github.com/DeutscheAktuarvereinigung/insurance-notebooks	Regression	Linear with interaction	Functional ANOVA Sobol Indices / Permutation feature importance Shapley allocation	Global Global Global

Abbildungen: Gliederung des AG-Ergebnisberichts (links),
 Katalogisierung der AG-Notebooks (rechts),
 Stand 10.11.23 ([Arbeitsfassung](#))

Literaturauswahl (alph.)

- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138-52160.
- Altmann, T., Bodensteiner, J., Dankers, C., et al. (2020). *Limitations of Interpretable Machine Learning Methods*. [Online](#).
- Baeder, L., Brinkmann, P., & Xu, E. (2021). *Interpretable Machine Learning for Insurance – An Introduction with Examples*. Society of Actuaries, [Online](#).
- Bundesanstalt für Finanzdienstleistungsaufsicht (2021). *Maschinelles Lernen in Risikomodellen - Charakteristika und aufsichtliche Schwerpunkte*. Konsultationspapier. 15.07.2021, [Online](#).
- Consultative Expert Group on Digital Ethics in Insurance (2021). *Artificial Intelligence Governance Principles: Towards Ethical and Trustworthy Artificial Intelligence in the European Insurance Sector*. European Insurance and Occupational Pensions Authority, 17.06.2021, [Online](#).
- Delcaillau, D., Ly, A., Papp, A., & Vermet, F. (2022). Model transparency and interpretability: survey and application to the insurance industry. *European Actuarial Journal*, 12(2), 443-484.
- European Banking Authority (2020). *EBA Report on Big Data and Advanced Analytics*. EBA/REP/2020/01, [Online](#).
- Hochrangige Expertengruppe für künstliche Intelligenz (HEG-KI, 2019). *ETHIK-LEITLINIEN FÜR EINE VERTRAUENSWÜRDIGE KI*. Europäische Kommission, 08.04.2019, [Online](#).
- Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
- Molnar, C. (2020). *Interpretable Machine Learning – A Guide for Making Black Box Models Explainable*. [Online](#).
- Owens, E., Sheehan, B., Mullins, M., Cunneen, M., Ressel, J., & Castignani, G. (2022). Explainable artificial intelligence (XAI) in insurance. *Risks*, 10(12), 230.
- Python-Pakete: scikit-learn (<https://scikit-learn.org/stable/index.html>), shap (<https://shap.readthedocs.io/en/latest/index.html>)

[illegible]DEUTSCHE
AKTUARVEREINIGUNG e.V.

DEUTSCHE GESELLSCHAFT
FÜR VERSICHERUNGS-UND
FINANZMATHEMATIK e.V.

Herbsttagung von DAV und DGVFM, 20./21.11.2023

Anhang

Anja Schmiedt
Simon Hatzesberger
Benjamin Müller



DAV

DEUTSCHE
AKTUARVEREINIGUNG e.V.



DGVFM

DEUTSCHE GESELLSCHAFT
FÜR VERSICHERUNGS- UND
FINANZMATHEMATIK e.V.

Herbsttagung von DAV und DGVFM, 20./21.11.2023

Backup-Folie zu SHAP

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

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

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