Explainable Artificial Intelligence

Modell-agnostische Erklärungsansätze im Vergleich

Büsra Karaoglan

Fachbereich Mathematik, Naturwissenschaften und Datenverarbeitung
Studiengang Business Mathematics

29. September 2020

Inhaltsverzeichnis

Inhaltsverzeichnis

Motivation

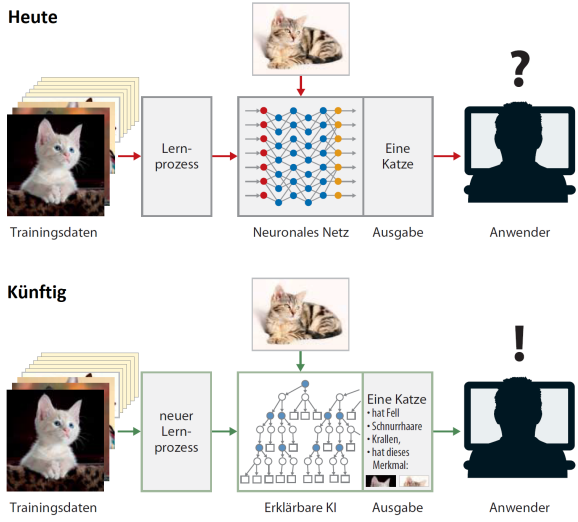

Abbildung: Konzept der erklärbaren Künstlichen Intelligenz (HA18)

Inhaltsverzeichnis

Grundlagen der XAI

- Clever-Hans-Effekt
- Scheinkausalität
- Algorithmischer Voreingenommenheit

Inhaltsverzeichnis

LIME-Erklärung

$$\xi(x) = \underset{g \in G}{\text{argmin}}\ L(f, g, \pi_x(z)) + \Omega(g)$$

- d-dimensionaler Merkmalsraum $X = \mathbb{R}^d$
- Ausgaberaum $Y = \mathbb{R}$
- Black-Box-Modell $f : \mathbb{R}^d \to \mathbb{R}$ mit $x \in \mathbb{R}^d$
- Interpretierbares Modell $g : \mathbb{R}^{d'} \to \mathbb{R}$
- Komplexität $\Omega(g)$
- Transformationsfunktion $IR : \mathbb{R}^d \to \mathbb{R}^{d'}$ mit $x' \in \mathbb{R}^{d'}$
- Ähnlichkeitsmaßes $\pi_x(z)$ mit permutierte Instanzen $z' \in \mathbb{R}^{d'}$
- Verlustfunktionen $L(f, g, \pi_x(z))$

LIME-Algorithmus

THM
TECHNISCHE HOCHSCHULE MITTELHESSEN

Algorithm 1 Spärliche lineare Erklärungen mit LIME

Require: Black-Box Modell f, Stichprobengröße N
Require: Instanz x, dazugehörige interpretierbare Darstellung x'
Require: Ähnlichkeitsmaß $\pi_x(z)$, Länge der Erklärung K

1: $\mathcal{Z} \leftarrow \{\}$
2: **for** $i \in \{1, 2, 3, ..., N\}$ **do**
3: $\quad z'_i \leftarrow sample\_around(x')$
4: $\quad \mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$
5: **end for**
6: $w \leftarrow$ K-LASSO$(\mathcal{Z}, K)$ ▷ mit z'_i als Merkmale, $f(z)$ als Zielvariable
7: **return** w

LIME-Algorithmus für Tabellendaten

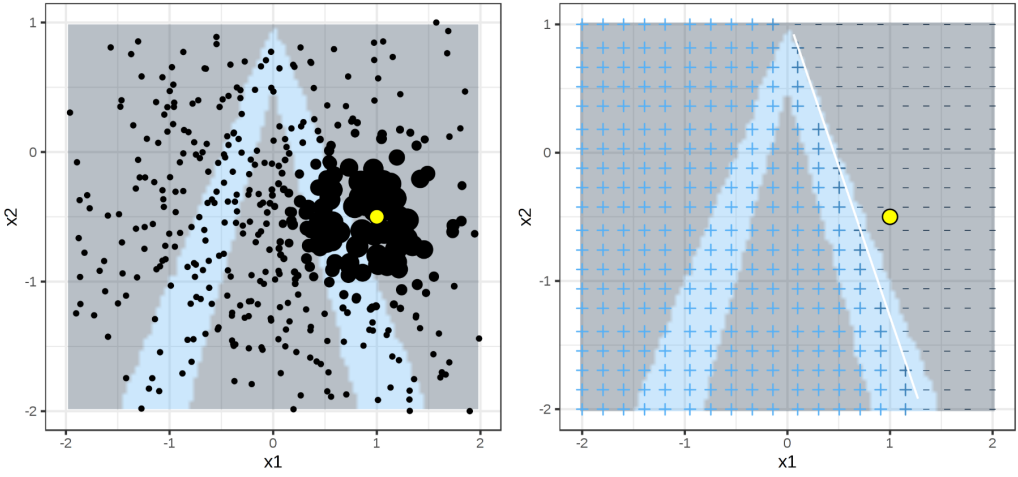

Abbildung: LIME-Algorithmus für tabellarische Daten (MC20)

Inhaltsverzeichnis

Shapley-Wert

Shapley-Wert für kooperative Spiele:

$$\phi_j(v) = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!\,(N - |S| - 1)!}{N!} \left(v(S \cup \{j\}) - v(S)\right)$$

Shapley-Wert für erklärbares maschinelles Lernen:

$$\phi_j = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!\,(M - |S| - 1)!}{M!} [f_x(S \cup \{j\}) - f_x(S)]$$

SHAP-Modell

Additive Merkmalszuordnung:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j$$

SHAP-Wert:

$$\phi_j(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!\,(M - |z'| - 1)!}{M!} \left[f_x(z') - f_x(z' \backslash j) \right]$$

Kernel SHAP

$$\Omega(g) = 0$$

$$\pi_{x'}(z') = \frac{(M-1)}{\binom{M}{|z'|}|z'|(M-|z'|)}$$

$$L(f, g, \pi_{x'}) = \sum_{z' \in Z} [f(h_x(z')) - g(z')]^2 \pi_{x'}(z')$$

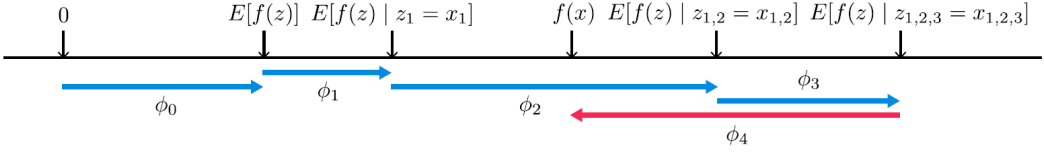

Abbildung: Schematische Darstellung der SHAP-Werte (LL17)

Inhaltsverzeichnis

Datenbeschreibung

THM
TECHNISCHE HOCHSCHULE MITTELHESSEN

Default Of Credit Card Clients Datensatz:

- ID
- LIMIT_BAL
- SEX
- EDUCATION
- MARRIAGE
- PAY_0 - PAY_6
- BILL_AMT1 - BILL_AMT6
- PAY_AMT1 - PAY_AMT6
- default payment next month

Vergleich von Black-Box Modellen

THM
TECHNISCHE HOCHSCHULE MITTELHESSEN

Modell	Precision	Recall-Wert	F1-Wert	Accuracy	ROC
SVM Classifier	0.52	0.52	0.52	0.79	0.69
RF Classifier	0.52	0.51	0.52	0.79	0.69
MLP Classifier	0.42	0.63	0.53	0.73	0.70

Tabelle: Übersicht der Klassifizierungsmetriken von angewendeten Modellen

LIME- und Kernel-SHAP-Erklärungen I

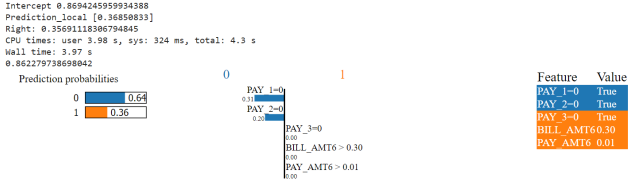

Abbildung: LIME-Erklärung zum SVM Classifier für Instanz 1

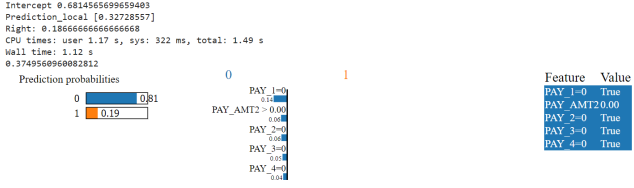

Abbildung: LIME-Erklärung zum RF Classifier für Instanz 1

LIME- und Kernel-SHAP-Erklärungen II

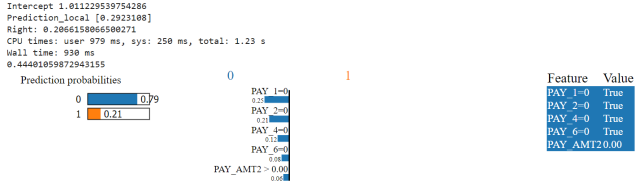

Abbildung: LIME-Erklärung zum MLP Classifier für Instanz 1

Abbildung: Kernel-SHAP-Erklärung zum MLP Classifier für Instanz 1

LIME- und Kernel-SHAP-Erklärungen III

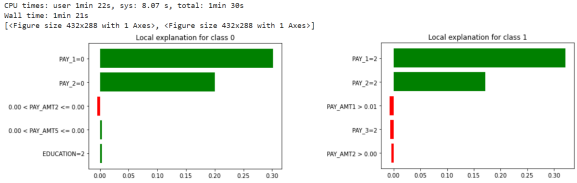

Abbildung: SP-LIME-Erklärung zum SVM Classifier

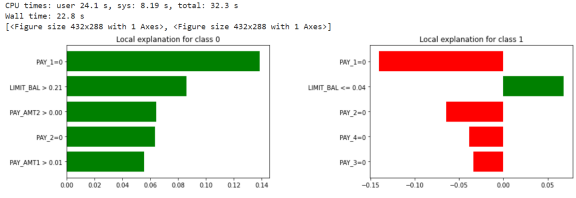

Abbildung: SP-LIME-Erklärung zum RF Classifier

LIME- und Kernel-SHAP-Erklärungen IV

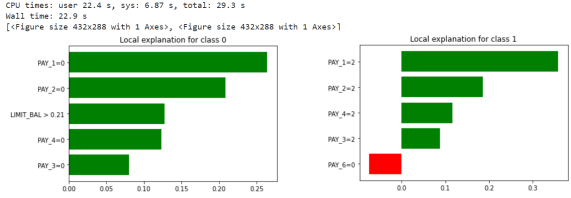

Abbildung: SP-LIME-Erklärung zum MLP Classifier

LIME- und Kernel-SHAP-Erklärungen V

CPU times: user 30min 48s, sys: 18min 35s, total: 49min 23s
Wall time: 26min 35s

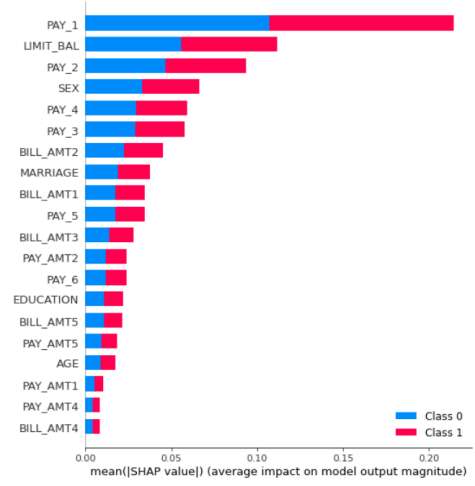

Abbildung: Globale Kernel-SHAP-Erklärung zum MLP Classifier

Literaturverzeichnis

- Holzinger, A.
  Interpretierbare KI - Neue Methoden zeigen Entscheidungswege künstlicher Intelligenz auf
  Heise Medien, Heft 22, 2018
- Lundberg, S. M. und Lee, S.-I.
  A unified approach to interpreting model predictions
  In: Advances in Neural Information Processing Systems, 2017
- Molnar, C.
  Interpretable Machine Learning - A Guide for Making Black Box Models Explainable
  Abgerufen am 20.08.2020 von
  https://christophm.github.io/interpretable-ml-book/
- Ribeiro, M. T., Singh, S. und Guestrin, C.
  „Why Should I Trust You?" - Explaining the Predictions of Any Classifier
  In: Knowledge Discovery and Data Mining, 2016

Danke für Ihre Aufmerksamkeit