Explainable AI

DHBW Mannheim

Jan Rüdt, 1737304, WII20DSA

Modul: Aktuelle Data Science Entwicklungen II - Reinforcement Learning

Dozentin: Janina Patzer

Gliederung

Was ist Explainable AI und warum ist sie wichtig?

Modelle und Techniken zur Erklärbarkeit von AI-Modellen

Häufige Anwendungsbereiche

Herausforderungen bei der Integration

Fallbeispiel: Explainable AI in der Praxis

Literaturverzeichnis

Was ist Explainable AI und warum ist sie wichtig?

„Erklärbare künstliche Intelligenz (XAI) ist eine Reihe von Prozessen und Methoden, die es menschlichen Anwendern ermöglichen, die von maschinellen Lernalgorithmen erzeugten Ergebnisse und Ausgaben zu verstehen und ihnen zu vertrauen.“ (IBM 2021)

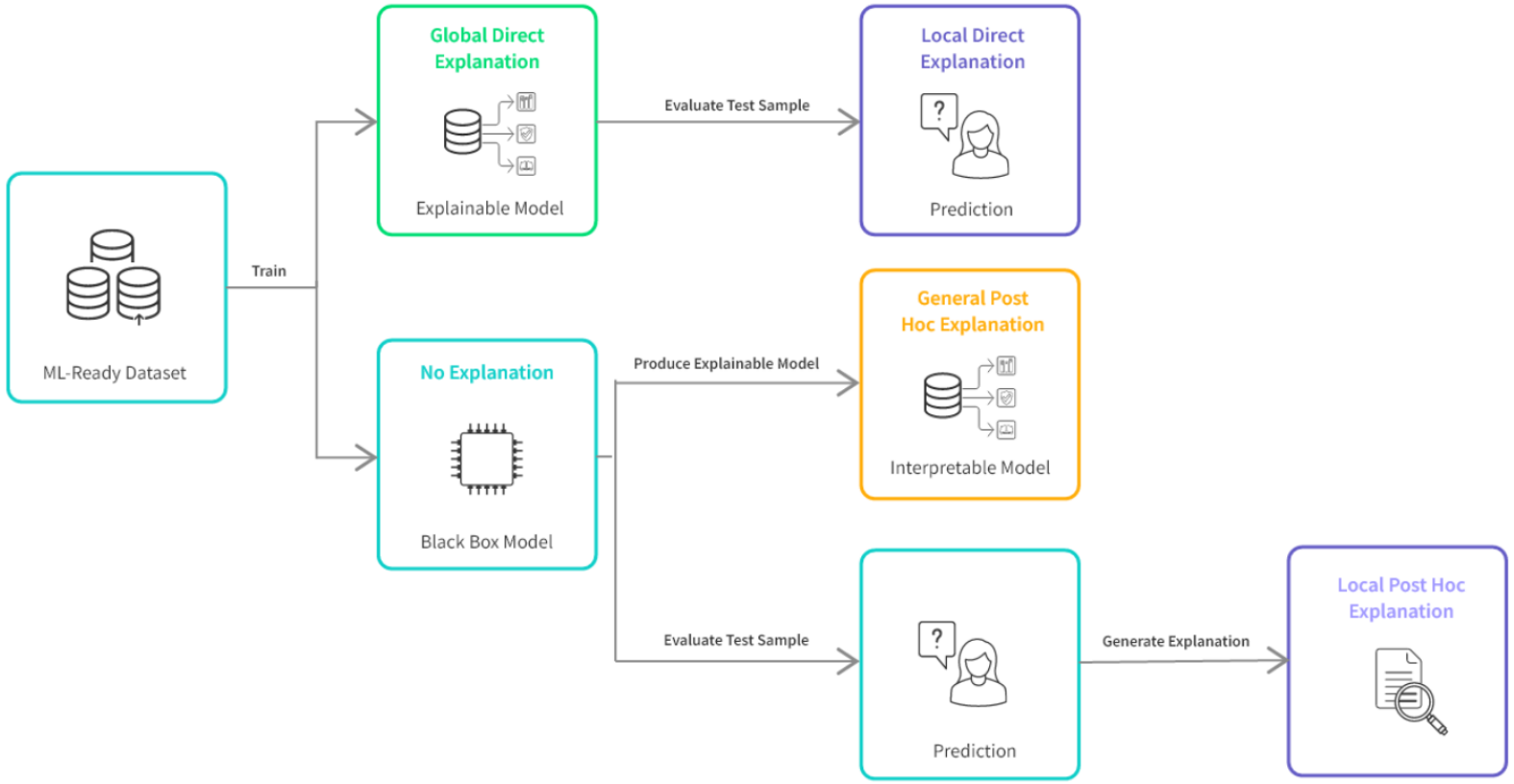
Bessere Entscheidungsfindung

Transparenz und Vertrauen

Rechtliche und ethische Gründe

Vermeidung von Bias und Diskriminierung

Modelle zur Erklärbarkeit von AI-Modellen



(QlikTech International AB)

Häufige Techniken:

- Local interpretable model-agnostic explanations (LIME)
- SHapley Additive exPlanations (SHAP)
- Layer-wise Relevance Propagation (LRP)

Häufige Anwendungsbereiche

- Finanzwesen: Kreditvergabe und Risikobewertung
- Gesundheitswesen: Diagnoseunterstützung und Behandlungsentscheidungen
- Rechtswesen: Vorhersage von Gerichtsurteilen und Strafmaß
- Autonome Fahrzeuge: Entscheidungsfindung im Straßenverkehr

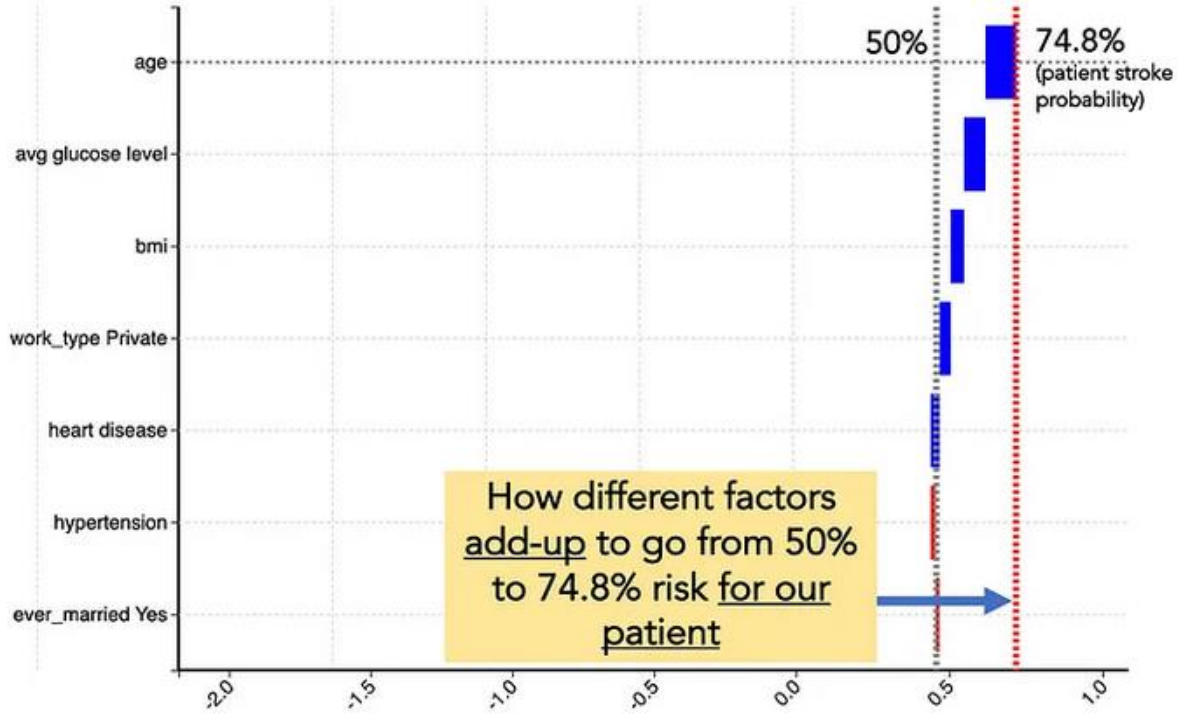
Herausforderungen bei der Integration

KOMPLEXITÄT DER MODELLE

TRADE-OFF ZWISCHEN ERKLÄRBARKEIT UND LEISTUNG

SCHUTZ SENSIBLER DATEN

Fallbeispiel: Explainable AI in der Praxis



Expected risk (for any patient in data)	50%
Age	+15.6%
Glucose level	+10%
BMI	+4.4%
Other risk reducing factors	-5.2%
Total Risk	74.8%

Expected risk (for any patient in data)	50%
Age	+15.6%
Glucose level	~~+10%~~
BMI	~~+4.4%~~
Other risk reducing factors	-5.2%
Total Risk	~~74.8~~ 60.4%

(Pranay 2021)

Literaturverzeichnis

- Explainable AI: Interpreting, Explaining and Visualizing Deep Learning (2019): Springer, Cham.
- Lecture Notes in Computer Science (2022). International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers: Springer, Cham.
- Proceedings of International Conference on Data Science and Applications (2023): Springer, Singapore.
- Bhuyan, Bikram Pratim; Srivastava, Sudhanshu (2023): Feature Importance in Explainable AI for Expounding Black Box Models. In: Proceedings of International Conference on Data Science and Applications: Springer, Singapore, S. 815–824. Online verfügbar unter https://link.springer.com/chapter/10.1007/978-981-19-6634-7_58.
- Dallanoce, Francesco (2022): Explainable AI: A Comprehensive Review of the Main Methods. In: *MLearning.ai*, 04.01.2022. Online verfügbar unter https://medium.com/mlearning-ai/explainable-ai-a-complete-summary-of-the-main-methods-a28f9ab132f7, zuletzt geprüft am 20.07.2023.
- Guestrin, Carlos; Ribeiro, Marco Tulio; Singh, Sameer (2016): Local Interpretable Model-Agnostic Explanations (LIME). Online verfügbar unter https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/, zuletzt geprüft am 20.07.2023.
- Holzinger, Andreas; Saranti, Anna; Molnar, Christoph; Biecek, Przemyslaw; Samek, Wojciech (2022): Explainable AI Methods - A Brief Overview. In: Lecture Notes in Computer Science. International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers: Springer, Cham, S. 13–38. Online verfügbar unter https://link.springer.com/chapter/10.1007/978-3-031-04083-2_2#Fig2.
- IBM (2021): Erklärbare KI. Online verfügbar unter https://www.ibm.com/de-de/watson/explainable-ai, zuletzt geprüft am 20.07.2023.
- Jiménez-Luna, José; Grisoni, Francesca; Schneider, Gisbert (2020): Drug discovery with explainable artificial intelligence. In: *Nat Mach Intell* 2 (10), S. 573–584. DOI: 10.1038/s42256-020-00236-4.
- Lindwurm, Eugen (2019): InDepth: Layer-Wise Relevance Propagation. In: *Towards Data Science*, 15.12.2019. Online verfügbar unter https://towardsdatascience.com/indepth-layer-wise-relevance-propagation-340f95deb1ea, zuletzt geprüft am 20.07.2023.
- Montavon, Grégoire; Binder, Alexander; Lapuschkin, Sebastian; Samek, Wojciech; Müller, Klaus-Robert (2019): Layer-Wise Relevance Propagation: An Overview. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning: Springer, Cham, S. 193–209. Online verfügbar unter https://link.springer.com/chapter/10.1007/978-3-030-28954-6_10.
- Pranay, Dave (2021): Top 5 techniques for Explainable AI. In: *Towards Data Science*, 27.11.2021. Online verfügbar unter https://towardsdatascience.com/top-5-techniques-for-explainable-ai-34349990cc83, zuletzt geprüft am 20.07.2023.
- QlikTech International AB: What is Explainable AI? Benefits & Best Practices. Online verfügbar unter https://www.qlik.com/us/augmented-analytics/explainable-ai, zuletzt geprüft am 20.07.2023.