Erklärbarkeit maschineller Lernverfahren - Feature Engineering mit Black-Box-Modellen

Christophe Krech

Hochschule Darmstadt - Fachbereich Mathematik & Naturwissenschaften und Fachbereich Informatik

Problemstellung

Durch die Einführung der Datenschutzgrundverordnung (DSGVO) ist die Notwendigkeit für erklärbare und nachvollzierbare, maschinelle Lernverfahren erneut gestiegen. Ziel muss es sein ein möglichst erklärbares Modell, mit einer möglichst hohen Vorhersagequalität zu entwickeln. Beide Anforderungen stehen sich - zumindest bei vielen Problemen - allerdings diametral gegenüber. Mittlerweile sind zahlreiche Methoden entstanden, die Entscheidungen eines Black-Box-Modells nachträglich zu erklären. Diese Methoden ermöglichen tiefe Einblicke in das Verhalten der Modelle, sind aber in Bezug auf ihre Interpretierbarkeit nicht mit intrinsisch erklärbaren Verfahren, wie einer logistischen Regression, vergleichbar.
In dieser Arbeit werden zwei Ansätze präsentiert, die bessere Trennschärfe der modernen, komplexen Verfahren, mit dem erklärbaren Charakter einer logistischen Regression zu kombinieren. Diese Ansätze führen ein performanceorientiertes Feature Engineering für erklärbare Modelle durch und greifen dabei auf die Erklärungen eines leistungsfähigeren Black-Box-Ansatzes zurück.

SHAP-Erklärungen

SHAP ist ein Verfahren zum Generieren von Erklärungen, das von Lundberg und Lee (2016) [3] vorgeschlagen wurde und auf Shapley-Values beruht. Dieses Verfahren erzeugt Erklärungen g eines Black-Box-Modells f für eine einzelne Beobachtung x. Konkret sucht es ein erklärendes, lineares Modell mit binären Variablen $z' \approx h_x(z')$:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i'$$

mit $z' \in \{0,1\}^M$, wobei M die Anzahl vereinfachter Features und $\phi \in \mathbb{R}$ ist. Diese Erklärbarkeitsverfahren ordnen jedem Merkmal i einen Effekt ϕ_i zu. Summieren der Effekte aller Feature ergibt ungefähr die Vorhersage $f(x)$ des Originalmodells. Diese Klasse der sog. **Additive Feature Attribution Methods** besitzt unter gewissen Annahmen eine eindeutige Lösung:

$$\phi_i(f,x) \sum_{S \in Z \setminus \{i\}} \frac{|S|!(M-|S|-1)!}{M!} [f_x(S \cup \{i\}) - f_x(S)] \quad (1)$$

Dabei bezeichne S die Indexmenge von z_i, wobei diese ungleich null sein müssen. Z sei die Menge aller M Inputfeatures d.h $x \in \mathbb{R}$

Feature Engineering mittels SHAP

Es werden zwei verschiedene Methodiken präsentiert, die unabhängig voneinander, aber auch gemeinsam angewendet werden können:
Die erste Methodik bedient sich der strukturellen Unterschiede zwischen White- und Black-Box-Verfahren in Bezug auf die funktionale Form der daraus resultierenden Modelle.

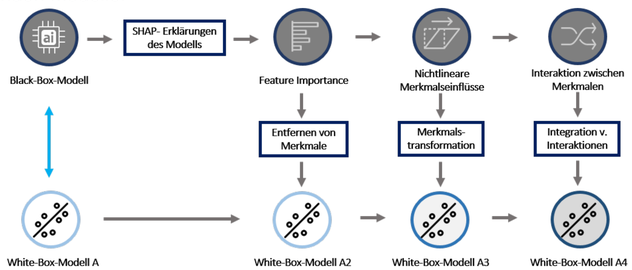

Fig. 1: Vorgehen zum Feature Engineering anhand von Erklärungen eines Black-Box-Modells

Daraus werden Transformationen des Merkmalsraums abgeleitet, die stets die Erklärbarkeit des Modells im Auge behalten. Das Vorgehen für diese Methode ist in Abbildung 1 exemplarisch dargestellt:

- Anhand verschiedener **Feature Importance**-Maße werden unwichtige Merkmale detektiert und aus dem White-Box-Modell entfernt, um zum einen deren Interpretierbarkeit und zum anderen deren Performance zu verbessern.
- Anschließend werden **Merkmalstransformationen** mittels SHAP-Erklärungen durchgeführt. Je nach Merkmalstyp und Gestalt des *SHAP Dependence Plot* erfolgt eine **Klassierung** oder **nichtlineare** Transformation anhand von **Polynomen** statt. Eine beispielhafte Transformation findet sich in Abbildung 2.

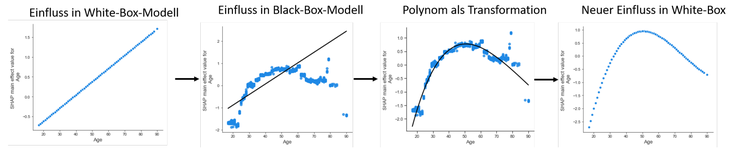

Fig. 2: Exemplarische Darsellung der Transformation mittels Polynomen: Der Merkmalseinfluss im White-Box-Modell (log. Regression) ist linear. Der Einfluss im Black-Box-Modell (Gradient Boosted Machine) hingegen nichtlinear. Mittels eines gefitteten Polynoms erfolgt eine Transformation des Merkmals, was auch den Einfluss in der log. Regression transformiert.

- Zu guter Letzt folgt die Integration von **Interaktionen** in das White-Box-Modell mittels den Methoden der Merkmalstransformation.

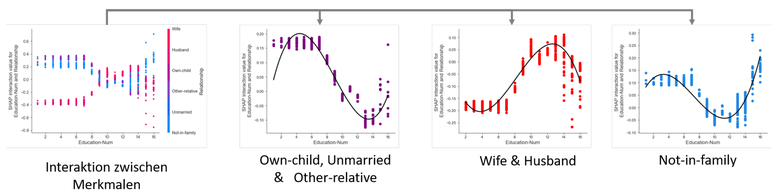

Fig. 3: Es sind unterschiedliche Funktionsverläufe zu erkennen. Statt einer Transformation für den kompletten Datenraum des betrachteten Merkmals, erfolgt eine abschnittsweise Defintion verschiedener Transformationsfunktionen im 2-D-Datenraum der interagierenden Merkmale. Diese verschiedenen Transformationen werden als eigenständige Merkmal zur Verfügung gestellt.

Analyse der Vorhersagedifferenz

Die Teilmenge der Daten, die bereits von einem einfachen, erklärbaren Modell korrekt klassifiziert wird, ist nicht besonders spannend; selbst wenn das komplexe Black-Box-Modell auf anderem Wege zu seiner Entscheidung kommt. Von deutlich größerem Interesse sind die Bereiche der Daten, in denen es zu einer **großen Differenz** zwischen den Vorhersagen des komplexen und des einfachen Modells kommt. Denn in diesem Bereich scheinen die vereinfachenden Annahmen des White-Box-Modells nicht adäquat, um die Komplexität des Problems abzubilden und das Black-Box-Modell liefert einen echten Mehrwert.

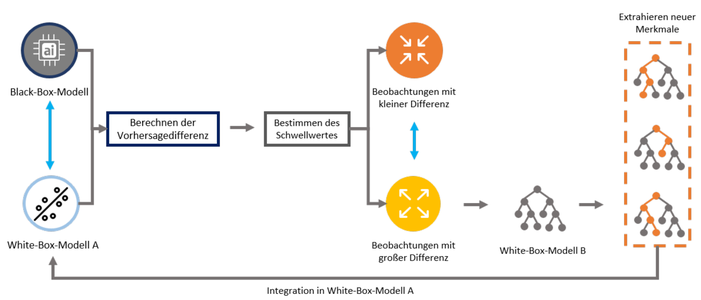

Fig. 4: Vorgehen zur Analyse der Vorhersagedifferenz

Durch Berechnen der **Vorhersagedifferenz zwischen White- und Black-Box-Modell** und Wahl eines geeigneten Schwellwerts, lässt sich der Datensatz in Beobachtungen mit kleiner und solche mit großer Vorhersagedifferenz aufteilen. Zum einen können nun die **Verteilungen** und **Einflüsse der Merkmale** in den beiden Datenteilen miteinander verglichen werden, wie Abbildung 5 illustriert.

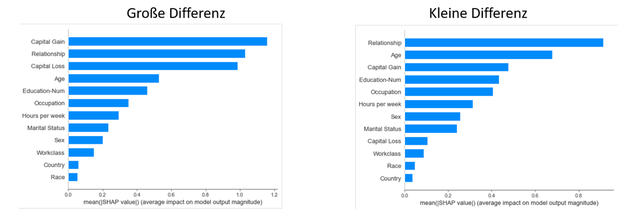

Fig. 5: Feature Importance bei kleiner bzw. großer Vorhersagedifferenz zwischen White- und Black-Box-Modell auf dem Adult-Datensatz. Es ist eine Veränderung der Reihenfolge und Effektstärke zu erkennen.

Zum anderen besteht die Möglichkeit auf den **Dateninstanzen mit großer Vorhersagedifferenz** ein **neues White-Box-Modell** zu trainieren und daraus Merkmale zu generieren (vgl. Abbildung 4)

Ergebnisse

Das Vorgehen beschreibt eine Kombination der Vorhersagepower moderner maschineller Black-Box-Verfahren, mit dem intrinsisch erklärbaren Charakter linearer Modelle. Damit liefert die Methodik zu gleich eine Orientierung, wie das Feature Engineering für White-Box-Modelle systematischer gestaltet werden kann. Momentan verbringen die Entwickler solcher Modelle einen Großteil ihrer Zeit, mit der Vorbereitung der Datenbasis für das Modelltraining [1]. Das in dieser Arbeit entwickelte Vorgehen kann diesen Prozess durch ein systematisches Feature Engineering unterstützen.
Beide Methoden wurde auf dem Adult-Datensatz [2] kombiniert angewendet und führten zu einer Verbesserung der Trennschärfe (in AUC evaluiert) um 1.5 Prozentpunkte von 0.906 auf 0.921. Die Differenz zwischen den AUC-Werten der logistischen Regression und der Gradient Boosted Machine (0.928) verringerte sich dabei um 68 Prozent.

Zukünftige Arbeit

In zukünftigen Arbeiten könnte eine systematische Untersuchung des hier entwickelten Vorgehens auf einer Vielzahl von Datensätzen erfolgen. Einzige Einschränkung ist die überlegene Performance des verwendeten Black-Box-Modells gegenüber dem erklärbaren Ansatz.
Eine Analyse der Methodik mit verschiedenen Black-Box-Modellen erscheint vielversprechend. Von besonderem Interesse ist der Vergleich der Transformationen verschiedener Black-Box-Modelle für ein und das selbe Merkmal. Daraus könnten sich Aknüpfungspunkte ergeben, dass Training von Black-Box-Verfahren so zu gestalten, dass der Einfluss eines Merkmals auf die Zielgröße wünschenswerte Eigenschaften vorzuweisen hat. Eine teilweise Automatisierung des Vorgehens wäre ebenfalls denkbar.

References

[1] CrowdFlower. “Data Science Report 2016”. In: (2016). URL: https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf.

[2] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017. URL: http://archive.ics.uci.edu/ml.

[3] Scott M Lundberg and Su-In Lee. “A Unified Approach to Interpreting Model Predictions”. In: *Advances in Neural Information Processing Systems 30*. 2017, pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.