ABSTRACT

Especially in regulated industries, linear models continue to enjoy great popularity. They have been tried and tested in practice and the regulatory authorities are familiar with their interpretation. Despite years of experience with these models, the developers are often not able to achieve the performance of modern machine learning methods like treebased ensembles or neural networks.
The introduction of the General Data Protection Regulation (GDPR) enhances the need for explainable and trustworthy machine learning methods. The aim is to develop a model with the highest possible prediction quality, that can provide the highest possible level of explainability. However, both requirements are diametrically opposed - at least for many problems. In the meantime, numerous methods have been developed to post-hoc explain the decisions of a black box model. These methods allow deep insights into the behaviour of the models, but cannot be compared with intrinsically explainable methods like logistic regression in terms of their interpretability.
In this thesis two approaches are presented to combine the better predictive power of the modern, highly complex models with the explainable character of a logistic regression. These approaches implement a performance-oriented feature engineering for white-box models using the explanations of a more powerful black box approach. For this purpose, two different methods are developed, which can be applied independently or together. Depending on the methodology and the problem, these either increase the performance of the explainable model or its explainability. In some cases even both is possible.
The first methodology makes use of the structural differences between white-box and black-box methods with regard to the functional form of the resulting models. This indicates that transformations of the feature space are derived which maintain the explanatability of the model. The second method only considers observations if the black box model comes to a different prediction than the white box model. By applying both methods, the performance of a logistic regression can be increased by 1.5 percentage points by SHAP explanations of a gradient boosted machine on an example data set. This reduces the difference between the ROC-AUC of the black box model and the logistic regression by 68 percent.

Keywords: *Explainable Machine Learning, Feature Engineering, Gradient Boosting Machines, Shapley Additive Explanations, Logistic Regression, Black Box Models*

ZUSAMMENFASSUNG

Insbesondere im regulierten Umfeld erfreuen sich lineare Modelle nach wie vor einer großen Beliebtheit. Sie sind im Einsatz erprobt und auch die Aufsichtsbehörden mit deren Interpretation vertraut. Trotz jahrelanger Erfahrung im Umgang mit diesen Modellen, sind die Entwickler häufig nicht imstande, die Performance moderner maschineller Lernverfahren aus den Bereichen der Ensemble Modelle oder Neuronalen Netze zu erzielen.
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, hochgradig komplexen Verfahren, mit dem erklärbaren Charakter einer logistischen Regression zu kombinieren. Diese Ansätze führen ein performance-orientiertes Feature Engineering für erklärbare Modelle durch und greifen dabei auf die Erklärungen eines leistungsfähigeren Black-Box-Ansatzes zurück. Dazu werden zwei verschiedene Methodiken entwickelt, die unabhängig voneinander, aber auch gemeinsam angewendet werden können. Je nach von Methodik und Problem, erhöht sich durch diese entweder die Performance des erklärbaren Modells, oder dessen Erklärbarkeit. In einigen Fällen ist sogar beides möglich.
Die erste Methodik bedient sich der strukturellen Unterschiede zwischen White- und Black-Box-Verfahren in Bezug auf die funktionale Form der daraus resultierenden Modelle. Daraus werden Transformationen des Merkmalsraums abgeleitet, die stets die Erklärbarkeit des Modells im Auge behalten. Die zweite Methodik betrachtet hingegen ausschließlich Beobachtungen, für die das Black-Box-Modell zu einer anderen Einschätzung wie das White-Box-Modell gelangt. Durch die Anwendung der beiden Verfahren kann die Performance einer logistischen Regression durch SHAP-Erklärungen einer Gradient Boosted Machine auf dem Beispieldatensatz um 1.5 Prozentpunkte erhöht werden. Die Differenz zwischen der ROC-AUC des Black-Box-Modells und der logistischen Regression sinkt dadurch um 68 Prozent.

Schlagwörter: *Erklärbarkeit, Explainable Machine Learning, Feature Engineering, Gradient Boosting Machines, Shapley Additive Explanations, Logistische Regression, Black-Box-Modelle*