Herausforderungen: Machine Learning

Dass beim Machine Learning Muster und Gesetzmäßigkeiten identifiziert werden können, bedeutet nicht, dass diese auch einer *theoretischen Fundierung* entsprechen müssen oder auf ein *reales Phänomen* schließen lassen. Vielmehr können Algorithmen scheinbare Muster und Gesetzmäßigkeiten herausstellen, die primär auf den Trainingsdatensatz und die darin enthaltenen Variablen und deren (zufällige oder nicht zufällige) Verteilung der Merkmalsausprägungen zurückzuführen sind. Wie in der Statistik gilt daher die zwingende *Unterscheidung zwischen Korrelation und Kausalität*: Diese können miteinander einhergehen, müssen es aber nicht! Die Auswahl der nachfolgend genannten Herausforderungen (und Lösungsmöglichkeiten) soll daher auf eine adäquate Anwendung von Machine Learning Algorithmen vorbereiten:

(1) Unzureichende Datenqualität

Statistische Analyseverfahren und Befunde sowie deren Derivate in Form von Machine Learning Algorithmen können nur unter Berücksichtigung der jeweiligen Datenqualität angemessen interpretiert werden. Hierzu ist nicht nur ein Fokus auf die *relevanten Variablen* erforderlich, sondern auch die Berücksichtigung von deren *Skalenniveaus*, *Lagemaßen* und *Streuungsmaßen*, sowie die Überprüfung auf *statistische Ausreißer* und *theoretisch nicht fundierte Verzerrungen* innerhalb der zugrundeliegenden Verteilungsfunktionen. Auch *Fehlwerte* von einzelnen Fällen (oder Gruppen von Fällen) in Bezug auf einzelne (oder mehrere) Merkmalsausprägungen innerhalb der Variablen können einen nicht zu vernachlässigenden Einfluss auf die *Performance* der Machine Learning Algorithmen nehmen.

(2) Übermäßige Optimierung der Performance

Dass ein Machine Learning Algorithmus in Bezug auf einen Trainings- und Validierungsdatensatz eine 100-prozentige Klassifikation oder korrekte Vorhersage ermöglicht, spricht zwar für eine sehr gute Performance, allerdings gilt diese *ausschließlich für den Trainings- und Validierungsdatensatz*. Zur Erinnerung: Wenn allgemeingültige Muster und Gesetzmäßigkeiten aufgeschlüsselt werden sollen, dann ist ein auf die besonderen Muster und Gesetzmäßigkeiten zugeschnittener Machine Learning Algorithmus unter Umständen weniger in der Lage, diese auch in Bezug auf reale Phänome und somit innerhalb vom Trainings- und Validierungsdatensatz abweichenden Datensätzen zu erkennen. Insbesondere mit Blick auf den Determinationskoeffizienten konnte beispielhaft herausgestellt werden, dass ein Anteil der Varianz in der Regel nicht vorhergesagt werden kann. Dies bedeutet zwar eine niedrigere Performance bei einem Machine Learning Algorithmus, bietet über den bisher unerklärten Anteil der Varianz aber die zusätzliche Möglichkeit, abweichende Muster und Gesetzmäßigkeiten in praxisnäheren Datensätzen zu erkennen. Schließlich ist ein *Dietrich als Werkzeug* zum „Knacken“ von Schlössern auch ein unspezifischeres Werkzeug als ein passgenauer Schlüssel eines Schlosses, dafür vermag dieser aber mehrere Schlösser zu öffnen.

(3) Limitation auf einen Machine Learning Algorithmus / Trainings- und Validierungsdatensatz

Auch aus dem Vergleich der Performance *unterschiedlicher Machine Learning Algorithmen* lassen sich wertvolle Informationen hinsichtlich der Passgenauigkeit zum Trainings- und Validierungsdatensatz ziehen. Wenn man zusätzlich *mehrere Trainings- und Validierungsdatensätze* zur Anwendung bringt (Stichwort: *Resampling*), dann ergibt sich an Stelle von einer möglichen Performance ein Intervall an

erwartbaren Performances, welches schließlich einen realistischeren Schluss auf ein reales Phänomen ermöglicht.

(4) Fehlende Mathematik- und Statistikkenntnisse

Dieser Punkt kommt scheinbar selbsterklärend daher, zielt aber auf eine fundamentale Herausforderung in der Anwendung von Machine Learning Algorithmen ab. Zunächst einmal sind viele *Analyseverfahren ursprünglich der Mathematik und Statistik zuzuordnen*. Als Beispiel seien die Faktorenanalyse und die Clusteranalyse genannt, welche im Rahmen der Mathematik und Statistik in der Regel unter Verweis auf Vektoren und die Transformation von Vektoren behandelt werden. Insbesondere in den letzten Jahren und mit zunehmender Verbreitung des Schlagwortes Machine Learning könnte allerdings der Eindruck entstehen, dass es sich ausschließlich um zwei unterschiedliche Machine Learning Algorithmen handelt, die über Packages bereitgestellt und mit wenigen Befehlen aufgerufen werden können. Der Rückgriff auf Packages und der Aufruf über reproduzierbare Befehle ist zwar richtig, die Interpretation der Befunde erfordert allerdings ein Verständnis der zugrundeliegenden Mathematik und Statistik. Ohne eine solche Kenntnis ist eine angemessene Reflektion der Befunde in vielen Fällen nicht möglich. Die gute Nachricht ist: Mit einer intensiven Auffrischung des *mathematischen Wissens aus der Oberstufe* ist man eigentlich schon gut vorbereitet. Als Empfehlung sei daher an dieser Stelle der zielführende *Mathematik-Crash-Kurs* auf Coursera der Kollegen Dye, Page, Cooper und Deisenroth des Imperial College London genannt.

Schlussendlich – diese Erfahrung bestätigt sich in vielen Beiträgen von Kolleginnen und Kollegen – führen in einigen Fällen die klassischen Analyseverfahren aus der Statistik, bspw. in Form einer linearen Regression, ebenso zu zielführenden Befunden, wie der Rückgriff auf Machine Learning Algorithmen. Das klingt dann vielleicht nicht so fancy, trägt aber in vielen Fällen maßgeblich zur *Transparenz* und somit zur *wissenschaftlichen Nachvollziehbarkeit* bei – und mit diesem Ziel sind wir unter dem Schlagwort *Wissenschaftliche Gütekriterien* ursprünglich in diesen Kurs gestartet.

In diesem Sinne sollte für alle Forscherinnen und Forschr gelten: *KISS – Keep it Short and Simple*!