Volker Reichenberger / Dirk Schieborn / Stephan Vorgrimler

Interpretierbarkeit maschineller Lernverfahren in der Kreditrisikomessung

Die leistungsfähigen Verfahren des maschinellen Lernens halten unaufhaltsam Einzug in die verschiedensten Anwendungsbereiche im Finanzsektor. Während sie von einer großen Gemeinschaft von Forschern und Anwendern laufend weiterentwickelt werden, nimmt sich auch die Bankenaufsicht dieses Themas aktiv an und bezieht in Richtlinien und Diskussionspapieren Stellung.

Eine zentrale Herausforderung besteht darin, die oft komplexen Verfahren und ihre Wirkungsweise interpretierbar zu machen und damit Entwicklern, Anwendern und Prüfern Möglichkeiten zur Plausibilisierung und Akzeptanz zu eröffnen.

Insbesondere Kreditrisikomodelle stellen einen naheliegenden Use Case für maschinelle Lernverfahren dar. Eine prominente Modellklasse sind auf internen Daten basierende Schätzverfahren für Risikoparameter wie Ausfallwahrscheinlichkeit (PD) und Verlustquote bei Ausfall (LGD), die sich für die Verwendung im IRBA (Internal Ratings Based Approach) zur Ermittlung des regulatorischen Eigenkapitals qualifizieren. Denn der bankenaufsichtliche Anspruch an sie ist hoch: hinsichtlich der Datenqualität, der Prozesse und nicht zuletzt der Methodik selbst. Für Letztere hat der Regulator den Detaillierungsgrad der Anforderungen in den vergangenen Jahren massiv erhöht (IRB Repair), sodass die regulatorischen Vorschriften nunmehr bis tief in die Modellstruktur reichen.

Türöffner Methodenfreiheit

Die seit der Einführung des IRBA bestehende weitgehende Methodenfreiheit bei der Wahl des mathematisch-statistischen Prognoseverfahrens – beispielweise zur Entscheidung, ob ein gegebener Kreditnehmer eine geringe oder hohe Ausfallwahrscheinlichkeit aufweist – bleibt dennoch größtenteils erhalten und stellt einen Türöffner für maschinelle Lernverfahren dar. Blickt man auf die methodischen Ansätze, die Banken in den letzten 15 Jahren im IRBA verwenden, so fällt auf, dass sich die Modelle zwar durchaus darin unterscheiden, aus welchen einzelnen Teilmodellen sie bestehen und wie diese Teilergebnisse miteinander kombiniert sind (Modellstruktur).

Auf Ebene der Teilmodelle selbst ist die methodische Bandbreite allerdings überschaubar: Neben eher selten verwendeten komplexen Ansätzen für Spezialfinanzierungen (zum Beispiel Cashflow-Modelle auf Basis von Monte-Carlo-Simulationen) kommen meist traditionelle multiple Regressionsmodelle zum Einsatz. Solche Regressionsmodelle liegen in den meisten Fällen auch den weit verbreiteten Scorecard-Modellen zu Grunde, bei denen die verschiedenen Ausprägungen der Risikofaktoren eines Kreditnehmers bepunktet und zu einem Score-Wert aggregiert werden, welcher in eine Ausfallwahrscheinlichkeit übersetzt wird. Ein großer Vorteil dieser Verfahren liegt in ihrer Nachvollziehbarkeit und – infolgedessen – Transparenz: Ursache und Wirkung lassen sich oft unmittelbar am Modell ablesen.

> „Kreditrisikomodelle stellen einen naheliegenden Use Case für maschinelle Lernverfahren dar."

Während die klassischen Regressionsverfahren zwar streng genommen ebenfalls unter den Begriff der maschinellen Lernverfahren fallen, so erscheinen die Namen der fortschrittlicheren Methoden, die gemeinhin mit dem Begriff maschinelle Lernverfahren verbunden werden, so geheimnisumwoben wie die Fähigkeiten, die ihnen zugeschrieben werden. Neuronale Netze, die der Struktur des menschlichen Gehirns nachempfunden sind und allein aufgrund dieser Konnotation gleichermaßen euphorische wie fatalistische Zukunftsvisionen heraufbeschwören – oder „Random Forests", die an Irrfahrten in Märchenwäldern denken lassen. Je komplexer diese Verfahren strukturiert sind, desto schwieriger wird es, ihre Wirkungsweise zu verstehen. Somit liegt die Analogie einer Blackbox nahe, deren innere Logik dem Anwender unbekannt bleibt. Stattdessen sind für ihn nur die Eingabe- und Ausgabewerte sichtbar, was zur Beurteilung der Prognosegüte ausreicht. Weist ein maschinelles Lernverfahren eine hohe Prognosegüte auf, so ist dies als positiv zu werten – unabhängig von der Mechanik, die in der Blackbox verbaut ist. Die kritische Frage allerdings ist: Wie kann eine Bank die Angemessenheit eines Verfahrens insbesondere für zukünftige, noch unbekannte Daten beurteilen? Interpretierbarkeit kann hierbei helfen.

Eine zentrale Voraussetzung für die Beurteilung von Prognosemodellen ist es somit, ihre Wirkungsweise nachzuvollziehen und interpretieren zu können. Dies gilt für Entwickler, Anwender und Prüfer gleichermaßen. In diesem Aspekt stehen die fortschrittlicheren maschinellen Lernverfahren in deutlichem Gegensatz zu den traditionellen, einfachen Regressionsverfahren. Letztere nämlich sind aufgrund ihrer mathematischen Struktur gut zu verstehen und zu überblicken.

Anpassungsfähigkeit und Overfitting

Maschinelle Lernverfahren heißen so, weil sie auf Basis von Fakten (Trainingsdaten) Wissen aufbauen (lernen), welches sie dann als Grundlage für Bewertungen oder Entscheidungen heranziehen. Der Wissensaufbau entspricht dabei der Justierung interner Stellschrauben (Parameter). Bei einigen der sogenannten überwachten Lernverfahren (zum Beispiel neuronalen Netzen) erfolgt diese Justierung in mehreren iterativen Schritten, bei denen die Ist-Ergebnisse der Prognosen mit den Soll-Ergebnissen verglichen und entsprechende Justierungen der Parameter angestoßen werden. Grob gesprochen gilt die Regel: Je höher die Anzahl der internen Stellschrauben, desto besser kann sich ein maschinelles Lernverfahren an die Trainingsdaten anpassen, was zunächst die Prognosegüte erhöht. Hierbei kann es jedoch auch über das Ziel hinausschießen, indem es sich an Kriterien in den Daten anpasst, die nur scheinbar zur Erklärung beitragen, sich aber beim Versuch, die gewonnenen Erkenntnisse auf die Bewertung neuer Daten zu übertragen, als irreführend beziehungsweise nicht transferierbar herausstellen. Dieses Phänomen wird als Overfitting bezeichnet und kann sich negativ auf die Prognosegüte des Verfahrens bei neuen, unbekannten Daten auswirken.

Bei der Suche nach dem geeigneten mathematisch-statistischen Verfahren beispielsweise für die Schätzung der Ausfallwahrscheinlichkeit von Kreditnehmern ist somit darauf zu achten, dass das Verfahren anpassungsfähig genug ist, um durch Justierung seiner Parameter die generalisierbaren Aspekte der Trainingsdaten (im Fall der Ausfallwahrscheinlichkeit: der historischen Kreditnehmer- und Ausfalldaten) zu extrahieren. Gleichzeitig sollte das Verfahren eine ausreichende strukturelle Starrheit aufweisen, um der Gefahr des Overfitting zu entgehen. Grundsätzlich gilt zudem: Je mehr strukturelle Starrheit (zum Beispiel Linearität) ein Modell aufweist, desto einfacher ist es zu verstehen – siehe die klassischen Regressionsmodelle. Am anderen Ende des Spektrums stehen beispielweise künstliche neuronale Netze. Diese weisen eine hohe Anpassungsfähigkeit auf, sind strukturell extrem flexibel, bergen dabei aber die Gefahr, nicht generalisierbaren Spezifika der Trainingsdaten durch Overfitting einen zu hohen Einfluss beizumessen. Solcherlei Modellrisiken können kontrolliert werden, indem das Modell und seine Wirkungsweise möglichst transparent gemacht und genau untersucht werden.

Aufsicht: Licht in die Blackbox!

Um die Bankenaufsicht in die Lage zu versetzen, ein Parameterschätzverfahren für die Verwendung im IRBA zuzulassen, muss eine Bank nicht nur nachweisen können, dass das verwendete Verfahren eine angemessene Prognosegüte aufweist. Zusätzlich muss sie plausibel darstellen, dass das Modelldesign für den Anwendungszweck angemessen ist, dass die Risikodifferenzierung angemessen erfolgt und dass die mit der Anwendung und Pflege betrauten Mitarbeiter das Modell inklusive seiner Leistungsfähigkeit und deren Einschränkungen verstehen. Diese Forderung nach Transparenz und Interpretierbarkeit steht in deutlichem Gegensatz zur Blackbox. Um fortschrittlichere maschinelle Lernverfahren wie neuronale Netze oder Random Forests im Rahmen von Kreditrisikomodellen unter aufsichtlicher Billigung zum Einsatz zu bringen, ergeben sich demnach gänzlich neue Herausforderungen an die Erklärbarkeit der Verfahren im Rahmen der Modellentwicklung und -validierung.

Die Bankenaufsicht (insbesondere EZB, Deutsche Bundesbank, BaFin und EBA) hat in verschiedenen Reports sowie Diskussions- und Prinzipienpapieren zur Verwendung maschineller Lernverfahren im Finanzsektor Stellung bezogen. Dies zeigt, dass vonseiten der Aufsicht grundsätzlich Offenheit gegenüber maschinellen Lernverfahren inklusive deren Blackbox-Eigenschaft besteht – sofern die

Prof. Dr.
Volker Reichenberger

Studiendekan MSc Operations Management,
ESB Business School, Reutlingen

Prof.
Dr. Dirk Schieborn

Studiendekan BSc Internationales
Wirtschaftsingenieurwesen – Operations,
ESB Business School, Reutlingen

Stephan Vorgrimler

Partner Risikomanagement,
msgGillardonBSM AG, Bretten

Kreditrisikomodelle stellen nach Ansicht der Autoren einen naheliegenden Anwendungsfall von maschinellen Lernverfahren dar. Ein Problem erkennen sie allerdings in der Komplexität dieser Verfahren, was die Analogie der „Blackbox" nahelege. Eine kritische Frage sehen sie daher darin, wie eine Bank die Angemessenheit eines Verfahrens insbesondere für zukünftige und somit noch unbekannte Daten beurteilen soll. Als hilfreich erachten die Autoren da die Interpretierbarkeit der Daten. Ziel dabei ist es, ein Verfahren durch quantifizierte Metriken beurteilen zu können. Es sei jedoch wichtig, dass man die Interpretationsmethoden versteht, sonst bestehe die Gefahr, Verfahren, die man nicht versteht, durch Methoden zu erklären, die man ebenfalls nicht versteht. (Red.)

Risiken sorgfältig unter Kontrolle gehalten werden (vgl. Erwägung Nr. 7 im Richtliniendiskussionspapier „The Use of Artificial Intelligence and Machine Learning in the Financial Sector", Deutsche Bundesbank, November 2020). Erklärbare Künstliche Intelligenz (Explainable Artificial Intelligence – XAI) wird dabei als möglicher Lösungsansatz zum Umgang mit der Blackbox-Eigenschaft gesehen (vgl. Erwägung Nr. 8 im o. g. Papier). Des Weiteren wird die Notwendigkeit strenger Validierungsverfahren betont, die zum jeweiligen Anwendungsfall passen (vgl. Erwägung Nr. 10 im o. g. Papier). In diesem Zusammenhang ist auch auf den aktuellen Entwurf für eine EU-Verordnung „zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz" zu verweisen, welcher Analysen der Kreditwürdigkeit als Hochrisikoanwendung einstuft, mit entsprechenden Folgen für die Governance dieser Verfahren (vgl. EU/COM/2021/206 final).

Die BaFin hat jüngst in einem Konsultationspapier (Maschinelles Lernen in Risikomodellen – Charakteristika und aufsichtliche Schwerpunkte, Juli 2021) mögliche aufsichtliche Schwerpunkte für die Prüfung von Risikomodellen auf Basis maschineller Lernverfahren zur Diskussion gestellt. Es überrascht nicht, dass auch hier die Erklärbarkeit der Verfahren im Fokus steht: Je komplexer das Modell, desto schwieriger ist es, den funktionalen Zusammenhang zwischen Input und Output verbal oder mittels mathematischer Formeln dem Prüfer oder Anwender verständlich zu machen. Erklärbarkeitsmethoden (XAI) erhöhen aber nicht nur das Verständnis der Wirkungsweise, sondern – und das ist zentral – bieten zudem vielversprechende Ansätze zur Validierung der Verfahren und zur Kontrolle der Modellrisiken, indem Modellschwächen wie Overfitting aufgedeckt werden können. Gleichzeitigt würdigt das Papier die oft höhere Prognosekraft komplexer maschineller Lernverfahren im Vergleich zu einfachen Modellen (zum Beispiel Regressionsmodellen).

Im Rahmen von Entwicklung und Validierung der aktuell bei IRBA-Modellen weit verbreiteten regressionsbasierten Schätzverfahren für PD oder LGD werden einerseits Analysemethoden eingesetzt, die insoweit spezifisch auf Regressionsverfahren zugeschnitten sind, als sie auf deren strukturelle Eigenschaften abstellen: dazu zählen zum Beispiel die Messung der Modellgüte mittels des Gütemaßes R^2, der F-Test auf Gesamtsignifikanz der Prädiktorvariablen sowie der t-Test zur Signifikanzmessung einzelner Regressoren – wobei kritisch anzumerken ist, dass die Voraussetzungen für den t-Test (Normalverteilung der Residuen) in vielen Fällen nicht erfüllt sind. Demgegenüber kommen andererseits Analysemethoden zum Einsatz, die in den Bereich der sogenannten modellagnostischen Verfahren fallen, da sie lediglich die Ein- und Ausgabewerte des Modells, nicht aber seine strukturellen Spezifika berücksichtigen. Weitverbreitete Beispiele sind die Receiver-Operator-Charakteristik (ROC) zur Messung der Trennkraft oder der Binomialtest zur Kalibrierungsmessung.

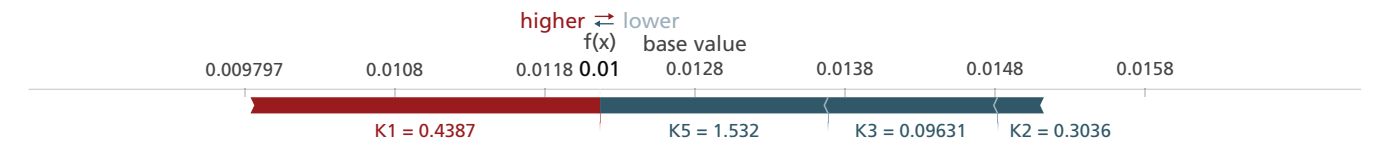
Unabhängig vom Einsatzgebiet ist die Erklär- und Interpretierbarkeit maschineller Lernverfahren für deren Verständnis und Akzeptanz von enormer Bedeutung. Infolgedessen stellt die Suche nach Erklärbarkeitsmethoden ein aktuell äußerst aktives Forschungsgebiet dar. Modellagnostischen Ansätzen kommt hierbei eine besondere Bedeutung zu, da sie verfahrensunabhängig sind und somit Vergleiche zwischen verschiedenen Modellen erlauben. Mittlerweile haben sich einige dieser Ansätze etabliert – nicht nur methodisch, sondern auch in Blick auf die technische Verfügbarkeit in statistischer Software. Beispiele dafür sind graphische Analysen wie Partial Dependence Plots oder Individual Conditional Expectation Plots, die Abhängigkeitsbeziehungen zwischen einzelnen Inputfaktoren und dem Output (das heißt der Prognose) des Modells visualisieren.

Ebenfalls verbreitet im Bereich der modellagnostischen Erklärbarkeitsmethoden sind Methoden der Feature Interaction zur Untersuchung der gemeinsamen Einflussnahme von Faktorenkombinationen auf die Prognose, sowie Ansätze, die das eigentliche Modell auf Basis einfacherer Ersatzmodelle (Surrogates) erklären. Hierbei unterscheidet man Global Surrogates, also Ersatzmodelle, die das Modell in Gänze ersetzen, von Local Surrogates, die das Modell lediglich lokal ersetzen, das heißt bezogen auf fest vorgegebene Werte der Eingabefaktoren. Prominenter Vertreter für letztere Klasse ist LIME (Local Interpretable Model-agnostic Explanations), eine Methode, die von der Bankenaufsicht explizit als Erklärbarkeitsmethode genannt wird – genauso wie die Methode der Shapley Values beziehungsweise die davon abgeleitete Methode SHAP (Shapley Additive Explanations), die im Folgenden näher beleuchtet werden sollen.

Interpretierbarkeitsanalysen als Teil künftiger Entwicklungen

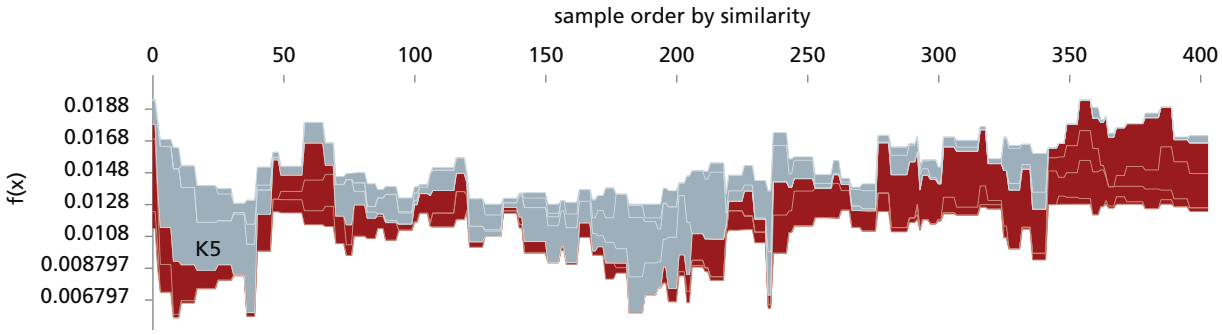
Aus Sicht der Autoren ist es lediglich eine Frage der Zeit, bis maschinelle Lernverfahren zumindest in Form von Teilmodellen in die Welt der (Kredit-)Risikomodelle Einzug halten werden. Eine entscheidende Hürde ist dabei die Akzeptanz durch die Bankenaufsicht, welche wiederum massiv davon abhängen wird, ob und wie

Abbildung 1: Force Plot



Quelle: V. Reichenberger, D. Schieborn, S. Vorgrimler

Abbildung 2: Force Plots auf Portfolioebene



Quelle: V. Reichenberger, D. Schieborn, S. Vorgrimler

es gelingt, im Rahmen der Modellentwicklung, der Initialvalidierung sowie des Validierungskonzepts für die Regelvalidierungen geeignete Analysen vorzusehen, die die Wirkungsweise des Modells hinreichend erklären und gleichzeitig Modellrisiken wie Overfitting oder unplausible Wirkungsrichtungen einzelner Faktoren angemessen aufzudecken beziehungsweise zu kontrollieren helfen. Am Beispiel von SHAP sollen solche möglichen Komponenten künftiger Validierungen im Folgenden beispielhaft illustriert werden.

Im Rahmen eines gemeinsamen Forschungsprojekts zwischen der ESB Business School der Hochschule Reutlingen und der msg GillardonBSM AG wurden in einem ersten Schritt Zeitreihen von Unternehmensdaten in Form von jeweils fünf Bilanzkennzahlzeitreihen (K1 bis K5) auf einem Zwanzig-Jahres-Horizont synthetisch erzeugt. Grundlage hierfür wiederum sind ein vektorautoregressives Simulationsmodell für makroökonomische Größen sowie ein einfaches zufallsbasiertes Verfahren, welches jedes Unternehmen einer (von zwei) Branchen zuweist. Die Unternehmen repräsentieren die Kreditnehmer eines Portfolios von Unternehmenskrediten einer fiktiven Bank. Auf Basis der fünf Bilanzkennzahlzeitreihen K1 bis K5 wurde anschließend simuliert, welches Unternehmen in welchem Jahr einen Kreditausfall erleidet (Kennzahl K4 wurde dabei so gewählt, dass sie keinen Einfluss auf den Kreditausfall hat, um dies in den späteren Analysen zu verifizieren). Die so simulierten Bilanzkennzahlen der einzelnen Unternehmen in Verbindung mit den simulierten Ausfallereignissen stellten die Grundlage für die weiteren Schritte dar.

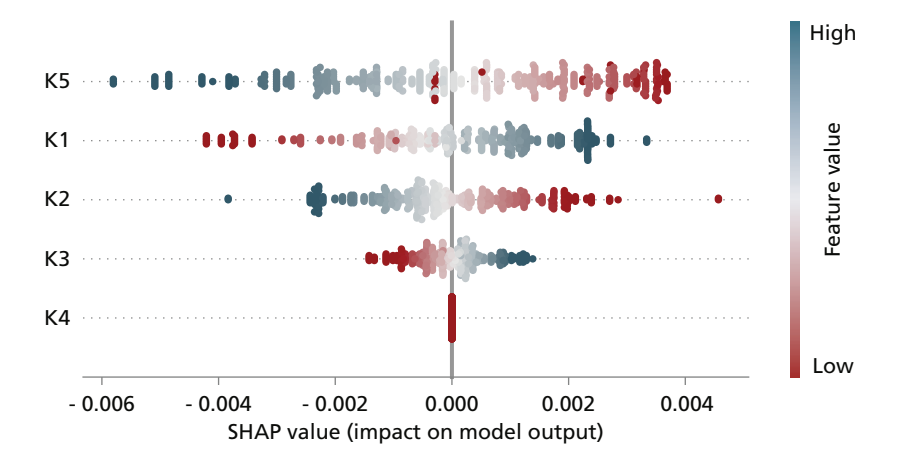
Neuronales Netz als exemplarisches komplexes Modell

Es ist anzunehmen, dass die fiktive Bank die simulierten Unternehmensdaten und die zugehörigen Ausfallereignisse als Trainingsdatensatz verwendet, auf welchem ein künstliches neuronales Netz zur Prognose von Ausfallwahrscheinlichkeiten trainiert wurde. Im Rahmen der Entwicklungs- und Validierungsdokumentation muss die Bank zunächst nachweisen, dass das Modell eine hinreichende Prognosegüte aufweist. Dies kann – ganz analog zu Regressionsverfahren – über klassische Trennschärfeanalysen wie der ROC-Kurve oder dem Gini-Koeffizienten gemessen werden und stellt somit keine Neuerung dar. Hier im Fokus steht indes die Frage, wie die Bank nachweisen kann, dass das Modell ein für das Anwendungsportfolio sowie für dessen Risikoprofil angemessenes und plausibles Verhalten zeigt. Hierzu wird die Bank Analysen durchführen müssen, die über die typischerweise in Entwicklungs- oder Validierungsdokumentationen enthaltenen Analysen deutlich hinausgehen. Beispiele dafür sind im Folgenden dargestellt.

SHAP ist eine modellagnostische Erklärungsmethode für maschinelle Lernverfahren, die auf den sogenannten Shapley Values (benannt nach Lloyd Shapley) aus der Spieltheorie beruht. Ihre Leistung besteht darin, für jede Wahl spezifischer Input-Werte den vom Modell gelieferten Output-Wert zu erklären, indem dieser in ursächliche Verbindung mit den einzelnen Input-Werten gebracht wird. Die Erklärung erfolgt dadurch, dass zunächst die Differenz zwischen dem aus den Input-Werten errechneten Output-Wert und dem bezüglich eines Referenzdatensatzes (zum Beispiel der Trainingsdaten) ermittelten durchschnittlichen Output-Wert (sogenannte Zentraltendenz) ermittelt wird. Diese Differenz wird dann in einzelne additive Beiträge (sogenannte SHAP-Werte) zerlegt – je einen pro Input-Faktor. Diese Beiträge können positiv oder negativ sein und erlauben eine Interpretation nach dem Schema: Input-Faktor Nr. X mit dem Wert Y verschiebt den Output-Wert des Modells um den SHAP-Wert Z relativ zur Zentraltendenz. Kennt man also für alle Input-Werte die zugehörigen SHAP-Werte, so lässt sich der Weg von der Zentraltendenz zum spezifischen Output-Wert Schritt für Schritt entlang der einzelnen Input-Faktorwerte nachvollziehen: er ergibt sich als Aggregation der einzelnen zugehörigen SHAP-Werte.

Die SHAP-Werte erklären ein Prognoseverfahren lokal, das heißt in Bezug auf einen spezifischen Satz von Inputwerten. Für andere Input-Werte können sich – je nach Komplexität des Verfahrens – völlig andere SHAP-Werte ergeben. Bezogen auf Prognoseverfahren für Ausfallwahrscheinlichkeiten beispielweise bedeutet das, dass die SHAP-Werte als Lupe für

Abbildung 3: SHAP Summary Plot



Quelle: V. Reichenberger, D. Schieborn, S. Vorgrimler

das Verhalten des Modells bezogen auf einen spezifischen Kreditnehmer dienen. Es lässt sich mit ihrer Hilfe nachvollziehen, welchen Einfluss jeder einzelne Input-Faktorwert für diesen Kreditnehmer auf dessen prognostizierte Ausfallwahrscheinlichkeit nimmt. Diese Einflüsse lassen sich grafisch als eine Folge aneinandergereihter Pfeile verschiedener Länge und Richtung veranschaulichen (auch Force Plot genannt). In der Abbildung 1 ist ein Force Plot zur Erklärung des neuronalen Netzes zur Schätzung der Ausfallwahrscheinlichkeit dargestellt, welches auf dem synthetisch erzeugten Unternehmensportfolio trainiert wurde.

Plausibilisierung von Richtung und Stärke der Wirkung

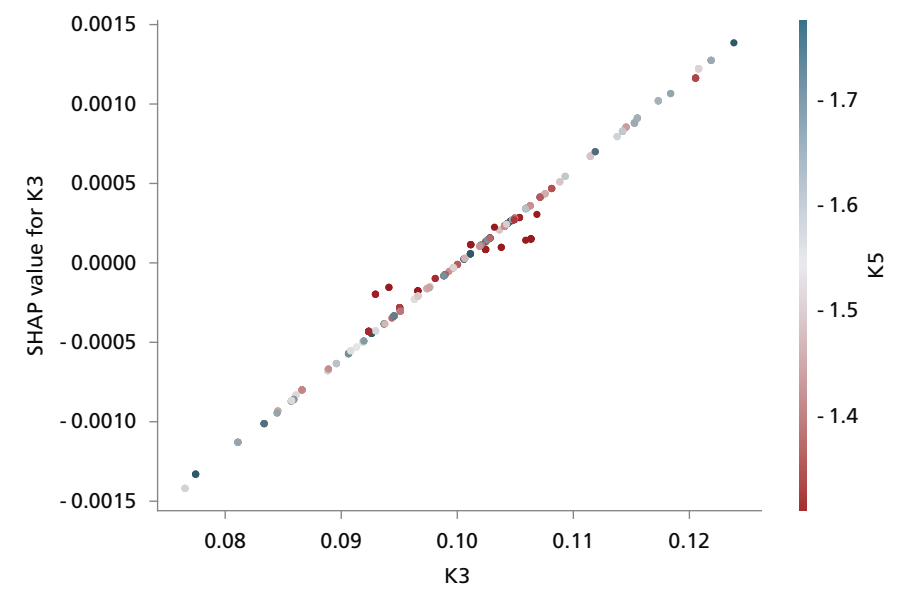
Er illustriert, wie sich die vom neuronalen Netz geschätzte Ausfallwahrscheinlichkeit (hier: 0,01 oder 1 Prozent) für ein zufällig ausgewähltes Unternehmen aus einer Verschiebung gegenüber der Zentraltendenz, das heißt der durchschnittlichen Ausfallwahrscheinlichkeit (base value 0,0128), ergibt. Diese Verschiebung resultiert aus einem Zusammenspiel mehrerer einzelner Kräfte. Drei davon sind nach links gerichtet und als blaue (negative) Pfeile dargestellt. Sie stellen die Beiträge (SHAP-Werte) der konkreten Werte der Bilanzkennzahlen K2, K3 und K5 für das zufällig gewählte Unternehmen dar, wobei sich die jeweiligen SHAP-Werte als die (negative) Länge der Pfeile ablesen lassen. Die zugehörigen Input-Werte von K2, K3 und K5 sind unter den Pfeilen notiert. Die vierte Kraft indes ist nach rechts gerichtet und als roter (positiver) Pfeil dargestellt. Seine Länge entspricht dem SHAP-Wert der Kennzahl K1. Die Kennzahl K4 liefert keinen eigenen Beitrag, da sie keinen Einfluss auf die simulierten Ausfälle nimmt und somit – korrekterweise – vom neuronalen Netz nicht bei der Ermittlung der Ausfallwahrscheinlichkeit berücksichtigt wird. Der Force Plot erlaubt also eine Plausibilisierung von Richtung und Stärke der Wirkung der spezifischen Werte der einzelnen Risikofaktoren auf die Modellprognose – allerdings nur lokal aus der Perspektive eines einzelnen ausgewählten Kreditnehmers.

Force Plots auf Portfolioebene

Berechnet man SHAP-Werte und Force Plots separat für jedes Unternehmen im gesamten Portfolio beziehungsweise in einem beliebigen Teilportfolio, so ergeben sich Erkenntnisse hinsichtlich der Wirkungsweise des Modells auf Gesamtbeziehungsweise Teilportfolioebene. Dies lässt sich in verschiedener Form visualisieren. In der Abbildung 2 sind die einzelnen Force Plots für ein Teilportfolio mit circa 400 der simulierten Unternehmen vertikal nebeneinander dargestellt.

Die Sortierung der Unternehmen kann hierbei nach verschiedenen Kriterien erfolgen. In diesem Beispiel erfolgt sie dergestalt, dass Unternehmen mit ähnlichen SHAP-Werteprofilen nebeneinander liegen – eine nützliche Darstellung zur Identifikation und weiteren Analyse potenzieller Cluster oder Subsegmente. Beispielsweise könnte für jedes Cluster ein Unternehmen als Repräsentant ausge-

Abbildung 4: Dependence Plot



Quelle: V. Reichenberger, D. Schieborn, S. Vorgrimler

wählt werden, anhand dessen die Wirkungsweise des neuronalen Netzes mittels eines individuellen Force Plots genauer auf Plausibilität untersucht wird.

Eine andere Form der Portfoliosicht erlaubt der sogenannte SHAP Summary Plot (Abbildung 3). Für jede der fünf Bilanzkennzahlen K1 bis K5, die zusammen die Input-Faktoren für das neuronale Netz bilden, ist jeweils jedes einzelne Unternehmen im Portfolio als Punkt dargestellt, wobei dessen Farbe den dazugehörigen Wert der Bilanzkennzahl codiert (blau entspricht einem niedrigen, rot einem hohen Wert). Die horizontale Position repräsentiert dagegen den SHAP-Wert, also den Verschiebungsbeitrag der betreffenden Kennzahl zur Ausfallwahrscheinlichkeitsprognose für das betreffende Unternehmen relativ zur Durchschnittsprognose. Damit die Verteilung der SHAP-Werte sichtbar wird, sind überlappende Punkte in vertikale Richtung auseinandergezogen. Auf Basis der Analyse der Farb- und Positionsverteilungen erlaubt der SHAP Summary Plot damit Rückschlüsse auf (lokale und globale) Wirkungsrichtungen und -stärken der einzelnen Inputfaktoren. Offenbar üben in unserem Beispiel die Kennzahlen K2 und K5 einen negativen Einfluss auf die Prognose aus: hohe Werte verringern, niedrige Werte erhöhen die Ausfallwahrscheinlichkeiten relativ zur Durchschnittsprognose. Die Kennzahlen K1 und K3 hingegen üben einen positiven Einfluss aus, während die Kennzahl K4 keine Auswirkung hat. Dies ist in unseren simulierten Beispieldaten die korrekte Interpretation – das neuronale Netz liefert somit plausible Prognosen.

Genauere Einblicke in die Wirkungsweise erhält man mit einem Dependence Plot (Abbildung 4). Hier sind für jedes Unternehmen im Portfolio der jeweilige Wert der Kennzahl K3 und der zugehörige SHAP-Wert als horizontale beziehungsweise vertikale Koordinaten eines Punktes dargestellt. Der oben erwähnte positive Einfluss von K3 ist gut zu erkennen. Die Farbcodierung der Punkte entspricht den Werten einer weiteren Kennzahl (hier: Kennzahl K5) und erlaubt Rückschlüsse auf die Abhängigkeit der beiden Kennzahlen K3 und K5 im Sinne einer Korrelation. Die dargestellten Analysen auf Basis von SHAP-Werten stellen ausgewählte Beispiele für viele mögliche Ansätze dar, die bei der Interpretation der Wirkung maschineller Lernverfahren helfen. Sie werden einen wesentlichen Teil der Analysen ausmachen, die im Rahmen der Entwicklung und Validierung von Kreditrisikomessverfahren durchzuführen sind, die auf maschinellen Lernverfahren basieren. Sie erlauben Entwicklern, Anwendern und Prüfern, die Wirkungsweise der Verfahren zu plausibilisieren und somit Licht in die Blackbox zu bringen.

Interpretierbarkeit: Was ist das?

Abschließend lohnt ein kurzer Blick auf den Begriff der „Interpretierbarkeit" maschineller Lernverfahren. Bislang hat sich keine einheitliche Definition für „Interpretierbarkeit" etabliert. Letztlich ist das Ziel, ein Verfahren durch quantifizierte Metriken zu beurteilen, um eine für Menschen nachvollziehbare Wirkungsweise des Verfahrens zu erschließen. Die Herausforderung der Interpretierbarkeit ist dabei mitnichten auf maschinelle Lernverfahren beschränkt. Auch ein „klassisches" multiples Regressionsmodell mit einer größeren Zahl von Inputfaktoren entzieht sich schnell einer direkten Interpretierbarkeit. Dasselbe gilt für einen tiefen Entscheidungsbaum, in dem beispielsweise das Merkmal „Beleihungsauslauf" an mehreren Stellen (mit unterschiedlichen Schwellenwerten) auftaucht: Auch hier „versteht" man nicht wirklich, warum das so an welcher Stelle geschieht.

Interpretationsmethoden werden an Relevanz, aber auch methodischer Bandbreite im Rahmen von Entwicklung und Validierung von Kreditrisikomodellen zunehmen – dies umso mehr, je mehr maschinelle Lernverfahren in die Welt dieser Modelle Einzug halten werden. Umso wichtiger ist es, die Interpretationsmethoden hinsichtlich ihrer Funktionsweise und Aussage genau zu verstehen. Andernfalls besteht die Gefahr, Verfahren, die man nicht versteht, zu „erklären" durch eine Methode, die man ebenfalls nicht versteht.