Tobias Mielich

Mit Integrated Gradients KI-Modelle verstehen

Durch den rapiden Fortschritt der Forschung, die Zunahme der Rechenleistung moderner Computer und den immer großflächigeren Einsatz in der Industrie gewinnt künstliche Intelligenz (KI) – insbesondere das Thema Machine Learning (ML) – mehr und mehr an Bedeutung. Auch in der Branche Banking wetteifern moderne ML-Methoden mit etablierten Verfahren und lösen diese aufgrund besserer Performance teils vollständig ab. Getrieben wird dieser Wandel durch den Kampf um Wirtschaftlichkeit und harschen Wettbewerb.

Doch durch den Fortschritt und die starke Medienpräsenz der ML-Methoden wird einer der Grundpfeiler des Bankwesens – das Vertrauen der Kunden in ihre Bank und indirekt auch in deren Geschäftsprozesse – auf die Probe gestellt. Wobei es die steigende Komplexität schwierig macht, diese Thematik im Detail zu verstehen.

Die Nachvollziehbarkeit von KI

Nationale und internationale Aufsichtsbehörden und Gesetze, wie die CRR/CRD und DSGVO, regeln bereits die Rahmenbedingungen zur klassischen Datenverarbeitung. Diese Rahmenbedingungen benötigen jedoch aufgrund der veränderten Funktionsweisen der modernen Methoden eine Überarbeitung beziehungsweise Erweiterung. Die BaFin, die Deutsche Bundesbank und auch die EBA (mit ihren Diskussions- und Thesenpapieren) nehmen als wichtige Aufsichtsorgane Stellung zum Einsatz und zur Entwicklung solcher Systeme zur Datenverarbeitung. Ein entscheidender Faktor ist die Erklärbarkeit beziehungsweise Nachvollziehbarkeit von künstlicher Intelligenz.

Zur Veranschaulichung stellen wir im Folgenden (stark vereinfacht) die Entscheidung einer Bank über eine Kreditvergabe auf Basis von Kundenscores dar, die von einem Machine-Learning-Modell gesteuert wird. Die Problemstellung lässt sich einfach zusammenfassen: Herr Mustermann möchte einen Kredit aufnehmen, der Bankangestellte gibt die Kundendaten in das Modell ein und erhält als Ergebnis einen hohen Score. Daraufhin gibt er grünes Licht für die Kreditvergabe. Doch wie genau entscheidet das Modell über diesen Score? Hält es sich an geltende Regeln oder Gesetze? Und hält diese Entscheidung einer genauen Prüfung der Revision stand? Um diese Fragen beantworten zu können, muss

die Bank in der Lage sein, nachzuvollziehen, aufgrund welcher Eigenschaften das Modell den ausgegebenen Score gebildet hat.

Am einfachsten lassen sich Modelle wie die logistische Regression oder Entscheidungsbäume erklären, denn sie besitzen eine inhärente Erklärbarkeit aufgrund ihrer simplen Funktionsweise. Bei der logistischen Regression besitzt jede Kundeneigenschaft ein Gewicht, und die Summe daraus zeigt das Ergebnis. Im oben genannten Beispiel besitzt Herr Mustermann Sicherheiten im vielfachen Wert der Kredithöhe. Das Modell gewichtet diesen Wert (nach Optimierung durch vergangene Kreditvergaben) mit 90 Prozent der Gesamtentscheidung und liefert somit einen hohen Score. Diese Gewichte können direkt abgelesen und interpretiert werden, wodurch eine hohe Transparenz und Nachvollziehbarkeit gegeben ist.

Ein entscheidender Nachteil bei solchen simplen Modellen ist jedoch ihre eingeschränkte Genauigkeit. Während viele Bereiche der Banking-Branche diese Methoden mit nennenswertem Erfolg im Tagesgeschäft verwenden, setzen mehr und mehr Institute auf komplexe Machine-Learning-Ansätze, um sich einen Wettbewerbsvorteil zu verschaffen oder mit anderen Instituten gleichzuziehen.

Moderne Modellarten, wie zum Beispiel neuronale Netzwerke, besitzen diese inhärente Erklärbarkeit nicht. Auf den ersten Blick erscheinen sie wie enorme Blackboxen, die keinen Einblick in die innere Funktionsweise zulassen. Dennoch existieren Techniken, die versuchen, die Rolle der Kundeneigenschaften anzugeben. Hierbei wird zwischen lokalen und globalen Methoden unterschieden: Lokale Methoden verändern geringfügig die Eingaben in das Modell und messen den Einfluss auf das Ergebnis (zum Beispiel zehn Prozent weniger Sicherheiten verringern den Score um fünf Basispunkte). Währenddessen versuchen globale Methoden, allumfassende Entscheidungskriterien des Modells herauszuarbeiten. In der Regel können nicht gleichzeitig präzise Aussagen über einen einzelnen Kunden und detaillierte Aussagen über das Zusammenspiel von Kundeneigenschaften für die gesamte Kundschaft getroffen werden. Es müssen Kompromisse gefunden werden, um anwendungsbezogen einen guten Mittelweg zu gehen.

Ein vielversprechender Ansatz – Integrated Gradients

Ein vielversprechender Ansatz, der lokale und globale Erklärbarkeit bestmöglich vereint, sind Integrated Gradients (IG). Dieses speziell auf bankingspezifische Anwendungen zugeschnittene Verfahren verwendet sowohl Informationen über das globale Entscheidungsverhalten des Modells als auch Einflüsse der Daten des einzelnen Kunden. So lässt sich eine bestmögliche Balance aus globaler und lokaler Entscheidungserklärung finden.

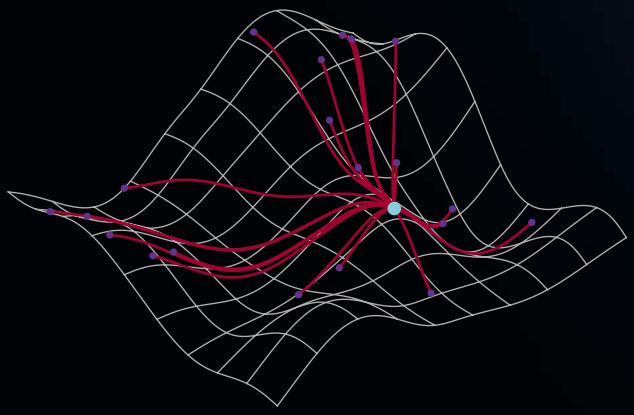

Abbildung 1: Zusammenspiel aus lokalen und globalen Einflussfaktoren bei Integrated Gradients

Grundlage für das IG-Verfahren sind repräsentative Baselines. Diese beschreiben Stichproben, die die verschiedenen statistischen Merkmale der vorhandenen Daten sowie Teilmengen davon so ausgeglichen wie möglich repräsentieren. Für das Beispiel der Bank, die einen Kredit an Herrn Mustermann vergeben möchte, würde eine ausgeglichene Baseline eine Menge an Kundendaten mit verschiedensten finanziellen und persönlichen Merkmalen enthalten. Das heißt, dass für die Untersuchung der Modellentscheidung und der damit einhergehenden Vergleiche mit der Gesamtkundschaft sowohl Kunden mit geringerem als auch hohem Einkommen sowie Kunden mit verschiedenen Zahlungsverläufen als auch Familienständen herangezogen werden müssen. Damit wird sichergestellt, dass alle möglichen Einflüsse auf die Modellentscheidung von der Wahl der Baseline abgebildet werden. Solche Datenauswertungen müssen dementsprechend auch aus der Sicht des Datenschutzes regelkonform sein. In diesem Beitrag sollen vorwiegend technische Aspekte beleuchtet werden, es sei jedoch gesagt, dass solche Datenschutzaspekte eine große Rolle bei der Prozess- und Modellbildung sowie der Validierung mit IG spielen.

Im darauffolgenden Schritt werden die einzelnen Kunden der Stichprobe gleichmäßig zu dem zu untersuchenden Zielkunden Herr Mustermann transformiert. Während dieser Transformation wird gemessen, wie das Modell seine Aussage zum Scoring verändert und ähnlich wie bei LIME (Local Interpretable Model-Agnostic →

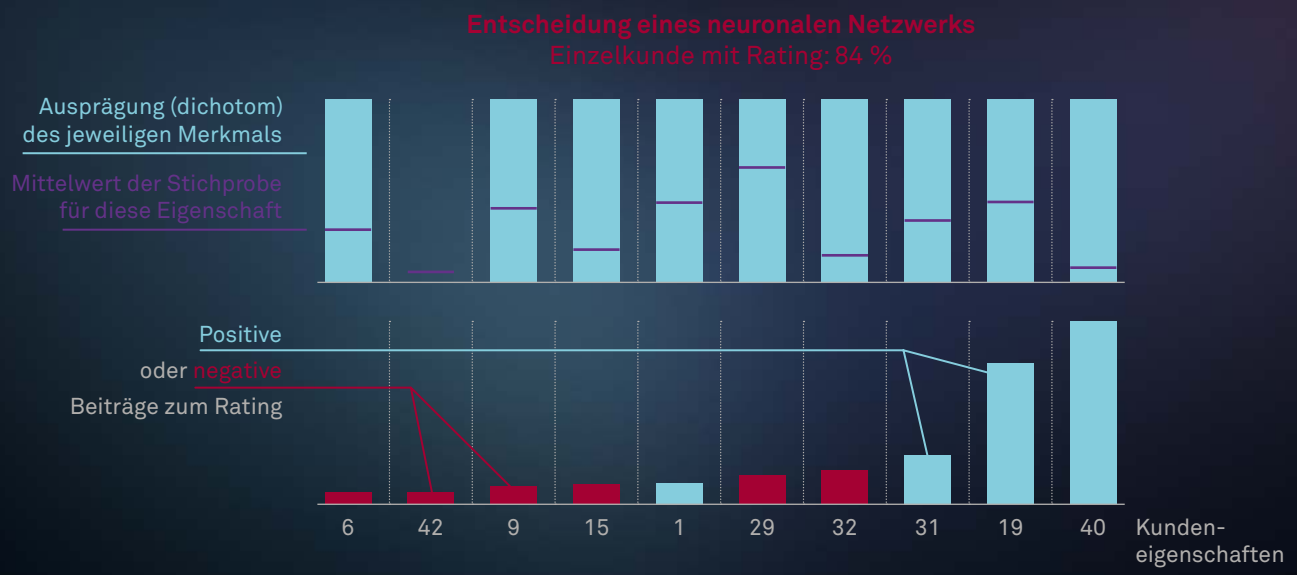

Abbildung 2: Beispielhaftes Entscheidungsprofil nach einer Analyse mit Integrated Gradients

Explanations) diese Änderung den einzelnen Eigenschaften zugewiesen wird. Dazu wird der mathematische Gradient der Entscheidungsfunktion bestimmt. Vergleichbar mit einer Fahrradtour durch die Berge wird für jede Eigenschaft ein Höhenprofil erstellt. Mit diesem Profil kann der Gesamteinfluss bestimmt werden.

Alle Gradienten der vollständigen Transformation werden aufsummiert und in einer Übersichtsgrafik dargestellt. Diese Grafik enthält die statistischen Informationen über die Baseline sowie die Ausprägung der Kundeneigenschaften des Zielkunden, hier Herr Mustermann. Damit lässt sich schnell einordnen, wie sehr er sich vom durchschnittlichen Kunden unterscheidet. Außerdem sind für alle Eigenschaften die Beiträge zum Kundenscore aufgetragen, diese können entweder einen positiven oder negativen Effekt auf das Ergebnis haben.

Mithilfe dieser Einflussfaktoren lassen sich schnell Rückschlüsse ziehen, ob die getroffene Entscheidung sinngemäß und möglichst nachvollziehbar ist. Integrated Gradients können somit durch geeignete Anwendung auf Einzelkunden sowie Kundengruppen ein umfangreiches Entscheidungsprofil eines Machine-Learning-Modells generieren. Für eine detaillierte Auswertung und Modellprüfung müssen zusätzlich mehrere Zielkunden mit derselben Baseline verprobt werden. Darüber hinaus müssen verschiedene Baselines mit derselben Grundmenge gebildet werden und die Untersuchungen wiederholt werden.

Fortgeschrittene Analysen beinhalten weiterhin noch Baselines, die nur auf Teilsegmenten der Kundenbasis bestehen, zum Beispiel nur aus Geschäftskunden oder Kunden mit sehr geringen Einlagen. Solche Teilsegmentanalysen geben Aufschluss, welche Dynamik in der Entscheidungsfindung des Modells existiert, wenn statistisch sehr unterschiedliche Domänen behandelt werden. Dies ist mitunter sehr wichtig für Stabilitätsanalysen oder Untersuchungen zur Verhinderung von Diskriminierung, welche von Aufsicht und Revision verlangt werden.

Fazit

Das Verständnis und die Nachvollziehbarkeit von komplexen Modellen stellt einen Grundpfeiler für zukünftige Datenverarbeitungen in der Banking-Branche dar. Zur Sicherung der Robustheit der Systeme und Einhaltung aller bisherigen und zukünftigen Regularien ist ein tiefes Verständnis der gelieferten Ergebnisse anhand lokaler und globaler Erklärungsmethoden – beispielsweise mit Integrated Gradients – unersetzbar. Umfangreiche Analysen und detaillierte Reports zum Modellverhalten stellen die Grundlage für das Vertrauen in diese modernen Systeme zur Datenverarbeitung dar – sowohl aus der Sicht des Kunden als auch aus der Aufsicht und der Revision.

Ansprechpartner

Tobias Mielich
Business Consultant
Tobias.Mielich@msg.group