Flossbach von Storch
RESEARCH INSTITUTE

GESELLSCHAFT & FINANZEN 06/04/2023

Daten – der Treibstoff für die Weiterentwicklung künstlicher Intelligenz

von SVEN EBERT

Zusammenfassung

Die Qualität von KI-Algorithmen, insbesondere auf dem Gebiet des maschinellen Lernens, steht und fällt mit der Qualität und Quantität der verfügbaren Daten. Das Thema Daten sollte somit auf der Agenda jedes Finanzmarktteilnehmers stehen.

Abstract

The quality of AI algorithms, especially in the field of machine learning, stands and falls with the quality and quantity of the available data. The topic of data should therefore be on the agenda of every financial market participant.

Daten – Rohstoff der Zukunft

In einem früheren Artikel haben wir bereits über bestehende Anwendungen künstlicher Intelligenz am Finanzmarkt geschrieben. Im Mittelpunkt stand die Methode des maschinellen Lernens und die zugehörigen Algorithmen.[1] Um ihre Wirkung entfalten zu können, benötigen diese jedoch Daten als Rohstoff. Daher beleuchten wir im Folgenden die Welt der Daten genauer.

Daten spielen eine zentrale Rolle bei den Fortschritten im Bereich maschinellen Lernens. Neben klassischen Datenquellen werden zunehmend auch alternative Daten für KI-Algorithmen genutzt. Insbesondere die Zunahme digitaler Sensorik sowohl in industrieller Produktion als auch in Alltagsgegenständen sowie die zunehmende Möglichkeit Nutzerverhalten aufzuzeichnen, werden die Basis für künftige Weiterentwicklungen sein. Wer keinen Zugang zu Daten hat, wird an dieser Entwicklung nicht teilhaben.

Gleichzeitig nutzen sich Daten nicht von allein. Als einer der technologischen Meilensteine von Chat-GPT wird die Fähigkeit der Programmierer genannt, den Algorithmus mit großen Datenmengen füttern zu können.[2] Die Fähigkeit Daten systematisiert zu erfassen und maschinenlesbar zu machen, wird daher auch in der Finanzindustrie eine Schlüsselrolle zukommen.

Charakteristika moderner Datenhaltung

Die Hoffnung, welche in künstliche Intelligenz (KI) im Allgemeinen und maschinelles Lernen (ML) im Speziellen gesetzt wird, besteht darin, dass Computer mehr Informationen als der Mensch in Verbindung miteinander setzen können und Muster finden, die dem menschlichen Auge und Verstand verborgen bleiben. Die Erhöhung der Rechenleistung von Computern ist hinlänglich bekannt und die methodischen Entwicklungen, beispielsweise im Bereich neuronaler Netze, haben wir bereits in einem vorigen Artikel dargelegt. Aber auch auf dem Gebiet der Daten, dem Rohstoff der Mustererkennung und dritten Baustein von KI, hat sich in der jüngeren Vergangenheit vieles getan.

Aktuell vermehren sich die weltweit verfügbaren Daten exponentiell. Im Jahr 2017 schätzte man, dass 90 Prozent dieser Daten innerhalb der letzten zwei Jahre entstanden seien.[3] Im Jahr 2020 betrug das Gesamtvolumen erzeugter

[1] Maschinelles Lernen an Finanzmärkten: gekommen, um zu bleiben - Flossbach von Storch (flossbachvonstorch-researchinstitute.com)
[2] The Technology Behind Chat GPT-3 (clearcogs.com)
[3] Marko Kolanovic und Rajesh T. Krishnamachari; Big Data and AI Strategies – Machine Learning and Alternative Data Approach to Investing, 2017.

und verwendeter Daten 64,2 Zettabytes. Schätzungen zu Folge sollen es im Jahr 2025 insgesamt 180 Zettabytes sein[4] – dies entspricht dem Gesamtspeicherplatz von 180 Milliarden IPhones der neuesten Generation. Allein das zur Verfügung stehende Volumen veranschaulicht, warum auch in der Finanzwirtschaft der Begriff „Big Data" in aller Munde ist. Der ehemalig Top-Manager von Google China Kai-Fu Lee formuliert dies in seinem Buch zu künstlicher Intelligenz so:

„Algorithms tuned by an average engineer can outperform those built by the world's leading experts if the average engineer has access to far more data."[5]

Neben der schieren Datenmenge sind zusätzlich zwei strukturelle Neuerungen bemerkenswert. Zum einen hat sich die Erhebungsgeschwindigkeit drastisch erhöht. So werden heute zum Beispiel Transaktionen an Börsen in elektronischen Orderbüchern in Echtzeit erfasst, gespeichert und publiziert. Kurse sind in Echtzeit für jedermann einsehbar.[6] Vor 30 Jahren konnte der normale Investor lediglich die Schlusskurse des Tages in den Abendnachrichten oder aus der Zeitung am Folgetag erfahren.

Zum anderen sind die verfügbaren Daten in ihrer Struktur vielfältiger geworden. So sind zum Beispiel höchstens 20 % der Daten, die von Internetnutzern produziert werden, überhaupt als strukturiert zu bezeichnen.[7] Feste Datenbankformate oder (Excel-) Tabellen, die standardisiert und maschinell lesbar sind, wie man sie beispielsweise vom statistischen Bundesamt kennt, bilden die Ausnahme. Unstrukturierte Daten wie z.B. Nachrichten und Videos repräsentieren den überwiegenden Teil der Daten. Insbesondere gehören hierzu auch von Nutzern über soziale Medien wie Twitter verbreitete Informationen. All diese Informationen sind Teil der sogenannten alternativen Daten, welche wir im Folgenden systematisiert untersuchen.

Alternative Daten – neue Datenquellen für neue Muster

Als alternative Daten gelten alle nicht-klassischen Daten. Für die Finanzwirtschaft können wir folgende spezielle Definition verwenden:

Alternative data is defined as non-traditional data that can provide an indication of future performance of a company outside of traditional sources, such as company filings, broker forecasts, and management guidance.[8]

---

[4] Daten - Volumen der weltweit generierten Daten 2025 (de.statista.com)
[5] Kai-Fu Lee; AI Superpowers, Houghton Mifflin Harcourt, 2018, Abschnitt: The Saudi Arabia of data.
[6] Offenes Orderbuch DAX | Börse Frankfurt (boerse-frankfurt.de)
[7] 30+ Big Data Statistics (2023) - Amount of Data Generated in The World (firstsiteguide.com)
[8] Alternative Data | Refinitiv

Für uns sind alternative Daten also alles, was nicht direkt aus Geschäftsberichten oder standardisierten Datenbanken wie dem Institutional Brokers' Estimate System[9] hervorgeht. Eine mögliche Untergliederung alternativer Daten in drei Kategorien ermöglicht uns eine positive Definition: Erstens individuelle von Personen erzeugte Daten, zweitens durch Geschäftsprozesse gewonnene Daten und drittens sensorische Daten.[10]

Abbildung 1: Klassifikation alternativer Daten

Individuelle Daten	Geschäftsprozessdaten	Sensoren
Soziale Medien	Transaktionsdaten	Satellitendaten
Nachrichten	Unternehmensdaten	Industrie 4.0
Suchmaschinenanfragen		Digitalisierung von Alltagsgegenständen

Quelle: Eigene Darstellung Flossbach von Storch Research Institute, J.P. Morgan QDS

Zu Daten von Personen zählen neben Posts in sozialen Medien auch durch Produktbewertungen erzeugte Informationen sowie die Untersuchung von Anfragen an Suchmaschinen. Das Ziel einer Auswertung kann das frühzeitige Aufspüren von Konsumentenstimmungen sein. Wird beispielsweise eine Restaurantkette kaum noch per Suchmaschine gesucht und erhält gleichzeitig permanent schlechte Bewertungen im Internet, liegt der Schluss nahe, dass der Aktienkurs der Firma sich aufgrund ausbleibender Gewinne zukünftig negativ entwickeln wird. Zentrale Frage hierbei ist, inwieweit die erhobenen Daten repräsentativ sind. Oder vereinfacht gesprochen: Abschätzige Kommentare über McDonalds in einem Internetforum über Sterne-Restaurants sagen vermutlich wenig über die Gewinnerwartung des Fast-Food Unternehmens aus.

Bedenkt man, dass nur zehn Prozent der weltweiten Daten einzigartig sind und der Rest aus Replikationen dieser Daten besteht,[11] bleibt zusätzlich die Frage, ob die Qualität der verwendeten Daten stimmt. Denn am Ende gilt beim maschinellen Lernen der Ausspruch „garbage in – garbage out"[12] in besonderem Maße – sind die Daten nicht repräsentativ für die Problemstellung, kann keine vernünftige Prognose entstehen. Da es sich bei den persönlichen Daten meist um Text handelt, werden diese meist mit Methoden des sogenannten Natural Language Processing bearbeitet. Die Software Chat-GPT ist das aktuell wohl bekannteste Beispiel für einen solchen Algorithmus.

[9] Institutional Brokers' Estimate System - Wikipedia
[10] Marko Kolanovic; Big Data and AI Strategies – 2019 Alternative Data Handbook, 2019, Abbildung 1.
[11] 30+ Big Data Statistics (2023) - Amount of Data Generated in The World (firstsiteguide.com)
[12] How Renaissance beat the markets with Machine Learning | by Neo Yi Peng | Towards Data Science

Die zweite Kategorie, d.h. durch Geschäftsprozesse gewonnene Daten, umfasst von Unternehmen und öffentlichen Einrichtungen produzierte oder gesammelte Daten. Hierunter fallen beispielsweise Scanner-Daten an Supermarktkassen oder Kreditkartentransaktionen. Auch hier ist die Anwendung für den Kapitalmarkt klar: Der Investor kann versuchen Unternehmensergebnisse vorherzusagen, bevor diese beispielsweise durch Bilanzpressekonferenzen zu öffentlich bekannten Informationen werden (was dann wiederum die Frage nach Insider-Information auslösen dürfte). Ein anderes Beispiel der politischen Sphäre: Wer glaubt, den Einfluss der Präsidentenwahlen auf den Kapitalmarkt zu kennen, der sollte während der Auszählung der Stimmen genau nach Ohio schauen. Kein Republikaner wurde je Präsident, ohne in Ohio siegreich zu sein.[13] Im Unterschied zu den von Personen produzierten Daten liegen aus Geschäftsprozessen gewonnenen Daten meist in strukturierter Form vor.

Aus Sensoren gewonnene Daten, die dritte Kategorie unserer Einteilung, ist vermutlich die am stärksten im Wachstum befindliche. Satellitenbilder werden ausgewertet, um Ernte-Erträge vorherzusagen oder den Container-Schiffsverkehr zu verfolgen.[14] Die Belegung der Parkplätze vor Einkaufszentren wird genutzt, um Vorhersagen über den zukünftigen Ertrag großer Supermarktketten zu treffen.

Weitere Dynamik entsteht aus dem alltäglichen Bereich. Unter dem Begriff „Internet der Dinge“ versammelt, enthalten immer mehr Alltagsgegenstände Mikroprozessoren und Netzwerktechnologie. Durch Smartphones aufgezeichnete Bewegungsprotokolle sind dabei nur die Spitze des Eisbergs. Mit dem Internet verbundene Lautsprecher, per Smartphone dimmbare Glühbirnen und „intelligente“ Thermostate - die Liste einstmals analoger Geräte, die heutzutage Daten sammeln, lässt sich beliebig erweitern. Und vernetzte Sensorik wird auch in der Industrie eingesetzt. Man denke nur an Produktionsstraßen, die selbstständig Nachschub an Teilen bestellen. Diese und andere Entwicklungen fallen unter den Begriff „Industrie 4.0“ oder „Industrielles Internet der Dinge“. Welche der neu gewonnenen Daten für einen Finanzinvestor am Ende relevant (und zugänglich) sein werden, lässt sich (noch) nicht abschließend bewerten. Aber schon jetzt lassen sich Eigenschaften benennen, die eine Anwendung von Machine-Learning-Methoden befördert.

[13] United States presidential elections in Ohio - Wikipedia
[14] Marcos Lopez de Prado und Alex Lipton; Three Quan Lessons from COVID-19, 2020.

Qualitätskriterien für Daten

Neben der Grundvoraussetzung in ausreichender Menge vorhanden zu sein, überzeugen für den Einsatz von ML in der Finanzwirtschaft geeignete Daten in dreifacher Hinsicht: Sie besitzen den richtigen Grad an Aufbereitung sowie die nötige Qualität und erfüllen idealerweise zusätzlich gewisse technische Kriterien.

Beim Grad der Aufbereitung reicht das Spektrum von Rohdaten bis zu vollständig vorprozessierten Daten. Vollkommen unbearbeitete Daten erfordern weitgehende Kenntnisse in der Datenverarbeitung. Rohe Satellitenbilder sind beispielsweise ohne Kenntnisse von (automatischer) Bildverarbeitung nicht ohne weiteres von einem ML-Modell zu verarbeiten. Komplett vorverarbeitete Daten haben andererseits eventuell schon entscheidende Informationen verloren. So kann eine saisonale Glättung den Einfluss der Jahreszeit entfernen, welcher aber entscheidend für die Prädiktion sein kann. Meist sind die Daten in der Praxis daher teilweise vorverarbeitet, das heißt in ein maschinell gut lesbares Format gebracht, nicht aber um vermeintliche Ausreißer, Fehler oder Trends bereinigt.[15]

Datenqualität hat mehrere Dimensionen: Grundsätzlich sind längere Datenreihen wünschenswert. Je nach Anwendung sollten eine Historie von mindestens drei Jahren und mindestens 50 Datenpunkten vorliegen.[16] Zusätzlich muss das Fehlen von Werten einer Zeitreihe ergründet werden. Ist es Zufall oder hat das Fehlen womöglich selbst Aussagekraft? Ein Beispiel sind Sensoren, die Messstände beim Unterschreiten eines bestimmten Grenzwertes nicht mehr melden (können).

Zusätzlich stellt sich die Frage welchen Verzerrungen, z.B. durch regionale Besonderheiten, die Daten ausgesetzt sind und inwieweit diese maschinell erkennbar sind. Ein gut trainiertes ML-Modell kann Verzerrungen eventuell selbstständig erkennen, wenn die Information in den Eingangsdaten enthalten ist. Weiß der Computer allerdings nicht, dass beispielsweise die Hälfte eines Datensatzes über Verkaufszahlen von Lebensmitteln aus einem kalifornischen Supermarkt und die andere aus Texas kommt, fehlt eine wichtige Information zur Prädiktion. Schließlich können auch schlicht unsauber arbeitende Sensoren Zufallsschwankungen produzieren, die eine wirksame Prädiktion verhindern.[17]

[15] Marko Kolanovic; Big Data and AI Strategies – 2019 Alternative Data Handbook, 2019.
[16] Marko Kolanovic; Big Data and AI Strategies – 2019 Alternative Data Handbook, 2019.
[17] John Paul Mueller und Luca Massaron; Machine Learning For Dummies, zweite Auflage, Wiley, 2021, Abbildung 8-2.

Abschließend gibt es noch technische Randbedingungen bei der praktischen Umsetzung, die die Qualität der Daten und damit auch des Algorithmus beeinflussen. Die Frequenz, mit der Daten zur Verfügung stehen, muss zur Anwendung passen. Während Geschäftsberichte nur einmal pro Jahr verfügbar sind, gibt es Aktienkurse von Unternehmen in Echtzeit. Möchte man den Zeitpunkt kurzfristiger Strukturbrüche voraussagen, scheint kein Weg an hochfrequenten Daten vorbeizuführen. Bei langfristigen Prognosen verstellen zu detaillierte Werte hingegen eventuell den Blick auf die großen Zusammenhänge.

Zusätzlich gilt es sich über die Latenzzeit der Daten im Klaren zu sein. Oder anders gesprochen: wieviel Zeit vergeht, bis die Daten nach ihrer Entstehung für einen Investor, anderen Marktteilnehmern und der breiten Öffentlichkeit bekannt sind. Bekommt der Investor gewisse Daten, die während eines Tages entstehen, erst am Tagesende als „Paket", dann können darauf keine Intraday-Trading Algorithmen aufgebaut werden. Erhält er Daten (legal) vor anderen Marktteilnehmern, hat er für eine gewisse Zeit einen Informationsvorsprung, der sich nutzen lässt – je größer der zeitliche Vorsprung, umso größer ist der potentielle Informationsvorsprung. Latenz ist damit eng mit der Anforderung funktionierender Schnittstellen zu externen Datenprovidern und schneller Ansteuerung eigener Datenbanken verknüpft.

Komplexere Modelle helfen nicht

Wenige oder schlechte Daten können nicht durch komplexere Modelle ausgeglichen werden, da beim Entwerfen von ML-Algorithmen zwei sich widerstrebende Trends aufeinandertreffen. Auf der einen Seite soll möglichst viel Information der vorhandenen Daten im Modell repräsentiert werden. Dies erreicht man im Allgemeinen durch eine Erhöhung der Modellkomplexität was zum Beispiel durch Erhöhung der Anzahl der Schichten eines neuronalen Netzes geschehen kann.

Je mehr ein Modell jedoch auf die Spezifika eines Datensatzes angepasst wird, umso größer ist das Risiko des sogenannten „Overfittings". Damit wird der Effekt bezeichnet, dass ein Modell sämtliche, das heißt auch die nicht strukturellen oder verallgemeinerbaren Besonderheiten eines Datensatzes berücksichtigt. Wendet man das Modell auf zukünftige Datensätze an, erzeugt es keine zufriedenstellenden Ergebnisse. Man spricht dann von Modellinstabilität. Die Kunst besteht darin, die Komplexität des Modells nur so lange zu erhöhen, bis die Modelinstabilität nicht überhandnimmt. Eine schematische Illustration des Sachverhaltes findet sich in der Literatur.[18]

[18] Marko Kolanovic und Rajesh T. Krishnamachari; Big Data and AI Strategies – Machine Learning and Alternative Data Approach to Investing, 2017, Abbildung 8.

Da der Gesamtfehler jedoch eine nur im konkreten Experiment zu bestimmende Größe ist, erkennen wir die zentrale Bedeutung großer Datenmengen aufs Neue. Je mehr Daten zur Verfügung stehen, desto besser kann man abwägen, ob die ideale Komplexität bereits erreicht ist.

Fazit

Der Zugang zu Daten und die Kompetenz mit diesen umzugehen sind zukünftig notwendige Voraussetzungen, um wettbewerbsfähig zu bleiben. Dabei sind nicht alle Daten gleich gut geeignet, zusätzliche entscheidungsrelevante Informationen zu liefern. Diese Feststellungen galten allerdings schon in Zeiten vor „Big Data“.

Neuartig ist, dass heutzutage die Masse an erfassten Daten exponentiell wächst und maschinelles Lernen den Menschen bei der Bewertung des Informationsgehalts der Daten unterstützen kann. Die Maschine kann ungleich mehr Daten durchforsten und findet eventuell Muster abseits klassischer Theorien, die der Mensch in den Daten nicht erwartet hätte.

Damit bietet extensive Datensammlung und experimentelle maschinelle Exploration dieser Daten ein weites Betätigungsfeld, um künftig Wettbewerbsvorteile zu erlangen. Das Thema Daten sollte somit auf der Agenda jedes Finanzmarktteilnehmers stehen.

RECHTLICHE HINWEISE

Die in diesem Dokument enthaltenen Informationen und zum Ausdruck gebrachten Meinungen geben die Einschätzungen des Verfassers zum Zeitpunkt der Veröffentlichung wieder und können sich jederzeit ohne vorherige Ankündigung ändern. Angaben zu in die Zukunft gerichteten Aussagen spiegeln die Ansicht und die Zukunftserwartung des Verfassers wider. Die Meinungen und Erwartungen können von Einschätzungen abweichen, die in anderen Dokumenten der Flossbach von Storch AG dargestellt werden. Die Beiträge werden nur zu Informationszwecken und ohne vertragliche oder sonstige Verpflichtung zur Verfügung gestellt. (Mit diesem Dokument wird kein Angebot zum Verkauf, Kauf oder zur Zeichnung von Wertpapieren oder sonstigen Titeln unterbreitet). Die enthaltenen Informationen und Einschätzungen stellen keine Anlageberatung oder sonstige Empfehlung dar. Eine Haftung für die Vollständigkeit, Aktualität und Richtigkeit der gemachten Angaben und Einschätzungen ist ausgeschlossen. **Die historische Entwicklung ist kein verlässlicher Indikator für die zukünftige Entwicklung.** Sämtliche Urheberrechte und sonstige Rechte, Titel und Ansprüche (einschließlich Copyrights, Marken, Patente und anderer Rechte an geistigem Eigentum sowie sonstiger Rechte) an, für und aus allen Informationen dieser Veröffentlichung unterliegen uneingeschränkt den jeweils gültigen Bestimmungen und den Besitzrechten der jeweiligen eingetragenen Eigentümer. Sie erlangen keine Rechte an dem Inhalt. Das Copyright für veröffentlichte, von der Flossbach von Storch AG selbst erstellte Inhalte bleibt allein bei der Flossbach von Storch AG. Eine Vervielfältigung oder Verwendung solcher Inhalte, ganz oder in Teilen, ist ohne schriftliche Zustimmung der Flossbach von Storch AG nicht gestattet.

Nachdrucke dieser Veröffentlichung sowie öffentliches Zugänglichmachen – insbesondere durch Aufnahme in fremde Internetauftritte – und Vervielfältigungen auf Datenträger aller Art bedürfen der vorherigen schriftlichen Zustimmung durch die Flossbach von Storch AG

IMPRESSUM

Herausgeber Flossbach von Storch AG, Research Institute, Ottoplatz 1, 50679 Köln, Telefon +49. 221. 33 88-291, research@fvsag.com; *Vorstand* Dr. Bert Flossbach, Kurt von Storch, Dirk von Velsen*; Umsatzsteuer-ID* DE 200 075 205; *Handelsregister* HRB 30 768 (Amtsgericht Köln); *Zuständige Aufsichtsbehörde* Bundesanstalt für Finanzdienstleistungsaufsicht, Marie-Curie-Straße 24 – 28, 60439 Frankfurt / Graurheindorfer Str. 108, 53117 Bonn, www.bafin.de*; Autor* Dr. Sven Ebert *Redaktionsschluss* 06. April 2023