prognostica
discover tomorrow

Vier Stellschrauben

Ihres ersten KI-Projekts

Inhaltsverzeichnis

Einleitung

Dieses E-Book ist insbesondere für all diejenigen interessant, die im Business-Umfeld erstmals aktiv mit **künstlicher Intelligenz (KI)** in Berührung kommen möchten. Die mit Hilfe von KI wiederkehrende Prozesse beschleunigen oder verschlanken möchten, diese kostengünstiger gestalten oder präzisere Ergebnisse erzielen möchten. Die die Automatisierung vorantreiben möchten, um den Beschäftigten monotone Arbeitsschritte abzunehmen, damit wieder mehr Zeit für die wirklich wichtigen Handgriffe, Entscheidungen oder kreativen Aufgaben ist. Die sich selbst und ihren Beschäftigten die Möglichkeit einer objektiven datenbasierten Zweitmeinung ermöglichen möchten, damit sich alle mit ihren Entscheidungen noch sicherer fühlen.

Der Begriff **“Künstliche Intelligenz” (KI)** (engl.: **“Artificial Intelligence”, AI**) ist allgegenwärtig. Immer mehr macht sich unter Unternehmen die Erwartung breit, KI heutzutage einfach nutzen zu müssen. Künstliche Intelligenz umfasst technisch ungeahnte Möglichkeiten. Der Begriff ist aber auch mit einer großen Angst konnotiert. Wir möchten hier mit Ihnen gemeinsam das Schreckgespenst KI auseinandernehmen, denn auch Sie können KI nutzen. Offensichtlich ist Ihr Interesse auch schon geweckt und Sie haben bereits den ersten Schritt gemacht, indem Sie unser E-Book heruntergeladen haben.

Ein KI-Projekt hat fast immer einen explorativen Charakter, bei dem der Ausgang nicht von vornherein feststeht und nicht alles von Anfang bis Ende durchdacht werden kann. Es erfordert, ein schnelles Scheitern zuzulassen und daraus zu lernen (“Fail fast”). Durch ein agiles Vorgehen kann man sinnvolle Abzweigungen auf dem KI-Weg rechtzeitig nehmen. Einfach mal ausprobieren, frühzeitig Zwischenstände prüfen, Vorgehen ggf. anpassen und kontinuierlich weitermachen. Daher finden Sie in diesem E-Book verschiedene Stellschrauben, die nicht notwendigerweise in einer festen Reihenfolge abgearbeitet werden müssen. Stattdessen ist es möglich und gewünscht, dass Sie nach ersten Erkenntnissen, falls nötig, Anpassungen an einer oder mehreren Stellschrauben vornehmen, was Sie näher an die Lösung und an den Erfolg Ihres KI-Projekts heranbringt.

Lassen Sie uns beginnen. Hier kommen die vier Stellschrauben Ihres ersten KI-Projekts.

Tipp: Fachbegriffe lassen sich beim Thema KI nicht ganz vermeiden. Am Ende dieses E-Books befindet sich ein Glossar mit wichtigen Begriffen – ein Auszug aus unserer umfassenden Knowledge Base rund um Data Science. Unser E-Book ist interaktiv gestaltet und Sie können sich bereits während des Lesens beim Klick auf die Fachbegriffe die Definitionen anzeigen lassen.

Vier Stellschrauben Ihres ersten KI-Projekts

Ein Schritt vorweg: Vertrautmachen mit dem Begriff “Künstliche Intelligenz”

Es ist unerlässlich, dass Sie wissen, was KI ist. Zugegebenermaßen ist künstliche Intelligenz ein breiter Begriff. Gut gefällt uns die simple Definition, unter künstlicher Intelligenz alle Eigenschaften eines IT-Systems zusammenzufassen, die menschenähnliche intelligente Verhaltensweisen zeigen[1]. Künstliche Intelligenz bedeutet aber nicht nur, dass Maschinen den Menschen nachahmen, sondern dass mit ihrer Hilfe auch Aufgaben bewältigt werden, die der Mensch alleine gar nicht, nicht so schnell, oder nicht so präzise schaffen würde. KI sollte immer einen operativen Charakter haben und somit die Menschen in ihrem Alltag unterstützen. Denken wir beispielsweise an Suchalgorithmen oder automatische Bilderkennung. Das Auge hätte bei der Bearbeitung vergleichbarer Aufgaben sehr viel zu tun und würde schnell müde werden. Künstliche Intelligenz dagegen ist unermüdlich, schnell und macht keine Flüchtigkeitsfehler.

Der im Business-Umfeld relevanteste Teilbereich der künstlichen Intelligenz ist das **Machine Learning (ML)**. Unter **Machine Learning** versteht man im Wesentlichen alle Verfahren, die es Maschinen ermöglichen, basierend auf Daten zu lernen und auf diese Weise Wissen zu generieren und ggf. Schlüsse zu ziehen. Spezieller lernen Algorithmen auf Grundlage von Daten, indem sie mit der Kenntnis jedes neuen Datenpunkts sich selbst hinterfragen, optimieren und anpassen. Dieser Lernaspekt, der der KI im Allgemeinen sehr wichtig ist, tritt im **Machine Learning** deutlich hervor. Da wir uns in diesem E-Book auf das Business-Umfeld konzentrieren, meinen wir – wenn wir von KI sprechen – i.d.R. Machine Learning. Eine häufige Anwendung von Machine Learning findet sich im Bereich **Predictive Analytics**, bei dem es darum geht, mittels Daten und analytischen Techniken Prognosen zu erstellen. Zu den wahrscheinlich bekanntesten Anwendungen von **Machine Learning** gehören Bilderkennung und Spracherkennung, bei denen **Künstliche Neuronale Netze** bzw. Deep-Learning-Algorithmen typische Aufgaben menschlicher Gehirne übernehmen. Durch die Medien ging, als eine künstliche Intelligenz des Unternehmens DeepMind 1997 zunächst den langjährigen Schachweltmeister[2], 2015/2016 schließlich Weltklasse-Go-Spieler[3] geschlagen hat. Auch wenn das wirklich revolutionär und mehr als nur bemerkenswert ist, müssen uns solche Fälle nicht beunruhigen: KI kann auch nur leisten, worauf man sie trainiert bzw. was man zulässt. Eine KI, die aufgrund von Training mit entsprechenden Fotoaufnahmen verschiedene Blumen unterscheiden kann, kann darin nicht einfach ein Gebäude oder Gesicht erkennen. Eine KI, die ausschließlich auf Basis der Produktabsatzzahlen Ihrer Konkurrenz Vorhersagen erzeugt, kennt und

1 Bitkom “Künstliche Intelligenz: Wirtschaftliche Bedeutung, gesellschaftliche Herausforderungen, menschliche Verantwortung”, unter *https://www.bitkom.org/sites/default/files/file/import/171012-KI-Gipfelpapier-online.pdf,* aufgerufen am 22.07.2020.

2 *https://www.faz.net/aktuell/wirtschaft/kuenstliche-intelligenz-in-4-stunden-zum-weltklasse-schach-15330791.html*, aufgerufen am 25.03.2021.

3 *https://deepmind.com/research/case-studies/alphago-the-story-so-far*, aufgerufen am 25.03.2021.

berücksichtigt Ihre Produktabsatzzahlen und Verkaufsstrategien nicht. Das deutet aber auch gleich auf einen äußerst wichtigen Aspekt hin: Daten und Datenqualität. Aber dazu später mehr.

Stellschraube 1: Der Use Case

Zunächst müssen Sie ungefähr festlegen, was Sie machen möchten. Das ist nicht immer so einfach, wie es klingt. Das "Was tun" ist immer auch im Wechselspiel mit "Sind die Voraussetzungen hierfür (insb. in Bezug auf die Daten) überhaupt gegeben?" und muss deshalb auch gemeinsam betrachtet werden. Wir listen im Anschluss verschiedene Use Cases auf, die mittels KI angegangen werden können und auch in Ihrem Unternehmen Nutzen bringen können. Was alle diese Use Cases gemeinsam haben ist, dass sie Daten als Grundlage haben. Diese Daten sollen zu einem bestimmten Zweck in der einen oder anderen Weise verarbeitet werden, z. B. gefiltert, bereinigt, aggregiert usw., wobei zusätzlich an irgendeiner Stelle der Intelligenzaspekt ins Spiel kommt: Aufgrund einer gewissen Nutzung der Daten sollen intelligente Schlüsse gezogen werden, etwa über vorhandene Muster, Eigenschaften und Zusammenhänge in den Daten oder das wahrscheinliche zukünftige Verhalten des untersuchten Phänomens. Hierbei handelt es sich i.d.R. um Erkenntnisse, die der Mensch aufgrund der Datenmenge oder Datenkomplexität entweder nicht ohne Weiteres erfassen kann, oder die zu mühselig und zeitraubend wären, um sie (in der Menge) von einem Menschen durchführen zu lassen. Hinzu kommt, dass wir als Mensch viele Dinge oft nur subjektiv betrachten können, basierend auf unseren individuellen Erfahrungen, dem aktuellen Befinden oder unserem Hungergefühl. Die Maschine kann deutlich objektiver und nüchterner an Sachverhalte herangehen und gerade diese Nüchternheit ist häufig ein großer Vorteil. Des Weiteren fällt es uns als Mensch oft schwer, wenn Zufall und Unsicherheit ins Spiel kommen. Verschiedene mögliche Ausgänge zu bewerten und geeignet zu gewichten erfordert i.d.R. die ausführliche Analyse von vergangenen Gegebenheiten und Zusammenhängen sowie beispielsweise aktuellen Trends.

Lassen Sie sich von unseren Beispiel-Use-Cases inspirieren und/oder überlegen Sie sich Antworten auf die folgenden Fragen, um Ihre Pain Points zu identifizieren:

- Welche (datenbasierten) Aufgaben führen Sie im Unternehmen immer und immer

wieder in ähnlicher Art und Weise durch? Welche Aufgaben erscheinen Ihnen repetitiv?
- Die Behandlung welcher Daten fällt regelmäßig unter den Tisch, weil niemand Zeit hat, sich darum zu kümmern?
- Bei welchen datenbasierten Aufgaben unterscheiden sich die Ergebnisse, wenn sie durch unterschiedliche Mitarbeiter*innen durchgeführt werden?
- Welche (datenbasierten) Aufgaben werden vernachlässigt, weil sie zu komplex und unübersichtlich erscheinen?
- Welche Aufgaben wären deutlich lohnenswerter, wenn sie statt durch manuelle Arbeit durch Maschinen durchgeführt werden könnten?
- Was macht Ihre Konkurrenz oder was machen andere (z. B. in der Größe oder Arbeitsweise) vergleichbare Unternehmen im Bereich KI?
- In welchem Anwendungsbereich vermuten Sie stark bisher unbekannte Zusammenhänge und Muster, konnten diese aber bisher nicht konkret identifizieren?
- Wo könnte durch Datenanalyse oder Automatisierung potenziell ein ganz neues Geschäftsfeld in Ihrem Unternehmen entstehen?

Priorisieren Sie die verschiedenen möglichen Use Cases nach Nutzen und kreisen Sie beispielsweise die drei Use Cases mit den für Sie größten Nutzen ein. Falls es möglich ist, gehen Sie noch einen Schritt weiter und erstellen Sie ein Aufwand-Nutzen-Diagramm. Kreisen Sie dann diejenigen Use Cases mit höchstem Nutzen und geringstem Aufwand ein. Wenn Ihre Auswahl an dieser Stelle noch nicht 100%-ig feststeht, ist das noch völlig unbedenklich aufgrund der oben genannten Wechselwirkung mit den Daten. Wenn Sie Anwendungen ausschließen oder runter priorisieren konnten, ist das schon sehr viel wert. Nehmen Sie Ihre Auswahl mit zu Stellschraube 2.

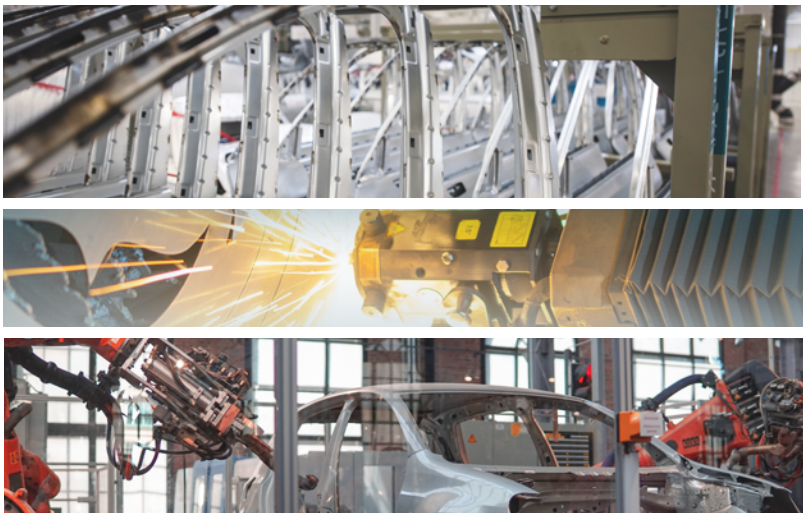
Action Points:
- Mindmap mit Pain Points bzw. möglichen Use Cases erstellen
- Use Cases priorisieren
- Top 3 Use Cases identifizieren

Use Cases

Use Case: Bilderkennung

Fragestellung

Im Karosserie- und Fahrzeugbau wird eine Vielzahl von Werkstücken mittels Lötstellen verbunden. Lötstellen sind mangelhaft, wenn sie beispielsweise Brüche oder Poren enthalten. Die Tauglichkeit von Lötstellen kann weitestgehend optisch beurteilt werden. Kann die Bewertung von Bildaufnahmen der Lötstellen ausreichend gut und ausreichend schnell mittels künstlicher Intelligenz vorgenommen werden? Zu beachten ist, dass in den Trainingsbildern defekte Lötstellen klar unterrepräsentiert sind im Vergleich zu qualitativ hochwertigen Lötstellen.



Lösung

Die Bilder der Lötstellen werden geeignet zugeschnitten und zentriert. Mittels geeigneter Data Augmentation wird der unausgewogene (imbalanced) Datensatz angereichert. Ein neuronales Netz wird auf dem Trainingsdatensatz trainiert, welches Lötstellen aufgrund der Bilder in verschiedene Kategorien klassifiziert (“gut”, “enthält Bruch”, “enthält Poren”, “enthält Spanreste”, ...). Die auf diese Weise erzeugte künstliche Intelligenz wird in Form eines eingebetteten Systems (embedded system) zusammen mit einer Kameralösung unmittelbar in den Produktionsprozess eingebunden. Das Gesamtsystem ist derart konstruiert, dass Expertenfeedback bzgl. Klassifikationen und Kategorien einbezogen werden kann, und auf diese Weise ein sich sukzessive verbesserndes, selbstlernendes System entsteht.

Benefits

- Zeitersparnis bei Qualitätschecks
- Einsparung von Produktionskosten
- Qualitätssteigerung in der Produktion durch Vermeidung menschlicher Fehler und Versäumnissen bei monotonen Aufgaben

Use Case: Supply Chain

Fragestellung

Ein produzierendes Unternehmen produziert eine große Zahl an Artikeln unterschiedlichen Typs. Wie viel wird von welcher Artikelgruppe in den nächsten Wochen abgesetzt werden?

Lösung

Ein automatisiertes Vorhersagesystem wird entwickelt, das unter Verwendung von KI-Verfahren Vorhersagen auf Artikelgruppenebene für die nächsten Wochen erzeugt und damit den unterschiedlichen Typen von Planungsobjekten gerecht wird. Dabei werden bereits bestehende offene Aufträge in die Prognose einbezogen.

Die Daten werden dem ERP-System des Unternehmens entnommen und Endergebnisse wieder dorthin zurückgespielt. Das Unternehmen nutzt eigene Reporting-Tools zur Anzeige und Visualisierung der Ergebnisse für Endnutzer. Ein Dashboard visualisiert Zwischenergebnisse und liefert Experteninformationen für Data Scientists.

Benefits

- Produktionsplaner gezielt darauf hinweisen, welche Artikelgruppen wie gut planbar sind und welche weiterhin manuellen Input benötigen
- Senken von Lager- und Abschreibungskosten
- Sicherstellen der Lieferfähigkeit
- Einsparung von Zeit für manuelle Planung

Use Case: Energy

Fragestellung

Bei der Vorhersage des Gasbedarfs muss darauf geachtet werden, dass das Netz immer optimal ausgelastet ist. Für objektive Vorhersagen können kurzfristige, feingranulare Daten auf Stunden- und Tagesbasis analysiert werden. Auch soll berücksichtigt werden, dass das aktuelle Wetter einen Einfluss auf den Energiebedarf hat. Auf wie viel Energiebedarf in den nächsten Wochen muss man sich in etwa einstellen?

Lösung

Mit einem automatisierten Online-Prognosesystem ist es beispielsweise möglich, dass einmal wöchentlich tagesgenaue Gasvorhersagen für die kommenden Wochen erzeugt werden. Die Vorhersagen erfolgen unter Einbezug von Wetterdaten, multiple Saisonalitäten werden mitmodelliert. Mithilfe von **Prognoseintervallen** werden Best- und Worst-Case-Abschätzungen vorgenommen. Statistische Unsicherheit wird mithilfe von Prognoseintervallen quantifiziert.

Benefits

- Kosteneinsparung und Investitionsvermeidung
- Basis für die Gestaltung von Kontrakten mit Energieanbietern
- Vermeidung eines kurzfristigen Energiezukaufs vom Spotmarkt

Use Case: Predictive Maintenance

Fragestellung

Eine große Anlage ist mit einer Vielzahl an Sensoren ausgestattet, die in kurzen Zeitabständen Messwerte erfassen. Von diesen Messwerten erhofft man sich, frühzeitig Informationen zu bekommen, wenn in der Anlage ein ungewöhnlicher Zustand herrscht. Solche Fehlerzustände sollen frühzeitig erkannt und angegangen werden. Dabei ist zu beachten, dass die Werte der Sensoren häufig stark von äußeren Bedingungen und voneinander abhängig sind. Diese Abhängigkeiten alle zu erkennen und stets den Überblick zu behalten, fällt selbst langjährigen Anlagenfahrern schwer. Eine KI soll Abhilfe leisten und im Arbeitsalltag die Beantwortung folgender Fragen unterstützen: Welche Wartungsarbeiten sind notwendig und wann sollten diese sinnvollerweise durchgeführt werden?

Lösung

Mittels geeigneter KI-Verfahren werden erwartete Zustände von Sensoren in Abhängigkeit weiterer Sensorwerte und Steuergrößen modelliert und ihre Abweichungen vom Ist-Zustand bestimmt. Multivariates Monitoring unterstützt die Erkennung von Fehlerzuständen. Im Falle von unerwartetem Verhalten werden automatisch Warnsignale ausgegeben. Ein Nutzer-Dashboard sowie eine mobile Anzeige ermöglichen das Anzeigen von Signalen sowie das Zurückverfolgen der Fehlerursache bzw. -position.

Benefits

- Frühzeitiger Austausch von defekten Teilen
- Bestimmung der Restlebensdauer von Teilen
- Vermeidung von ungeplanten Produktionsausfällen

Use Case: Finance

Fragestellung

Die Finanzkennzahlen eines Konzerns mit mehreren Tochterunternehmen sollen überwacht werden. Die vierteljährliche Bottom-up-Planung durch die Tochterunternehmen sorgt allerdings für Inkonsistenzen in den aggregierten Zahlen und macht die Gesamtfinanzplanung schwierig. Mit welchen Umsätzen der Tochterunternehmen und des gesamten Konzerns kann in den nächsten Monaten gerechnet werden?

Lösung

Ein automatisiertes Planungs- und Kontrollsystem erzeugt in kürzester Zeit objektive und präzise Finanzprognosen für die einzelnen Tochterunternehmen. Hochqualitative Forecasts ermöglichen eine sichere Planung, sodass positive wie negative Entwicklungen frühzeitig erkannt werden können und darauf reagiert werden kann. Dabei werden Wirtschafts- und Branchenindikatoren einbezogen und Best-Case-Worst-Case-Abschätzungen im Rahmen präziser Prognosen ermittelt. Zudem können im Falle außergewöhnlicher Situationen von den Finanzexperten Änderungen an den Prognosen vorgenommen werden. Statt vierteljährlich, werden die Vorhersagen jetzt auf Monatsbasis erstellt.

Benefits

- Objektive Zweitmeinung
- Konsistenz der Prognosen über verschiedene Unternehmensteile hinweg
- Die eigenen Unternehmenszahlen mit der wirtschaftlichen Lage systematisch in Beziehung setzen können
- Auf Trendwenden vorbereitet sein und reagieren können

Stellschrauben

Stellschraube 2: Die Daten

KI benötigt immer Daten, um loslegen zu können. Das müssen nicht immer unbedingt sehr viele sein – das kommt ganz auf die Anwendung an. Generell gilt jedoch: Werfen Sie erstmal keine Daten weg, wenn Sie mit einem KI-Projekt beginnen. Alles ist potenziell relevant. Ausgesiebt wird später. Machen Sie sich eine Liste, welche Daten direkt oder indirekt mit Ihrem Use Case zu tun haben. Werfen Sie einen Blick in die Vergangenheit, denn KI lernt aus historischen Daten. Was hat sich angesammelt?

Schauen Sie zunächst **intern**:

- Möchten Sie Ihre zukünftigen Absatzzahlen für Ihre Produkte vorhersagen? Dann werden Sie vermutlich in Ihrem ERP-System die vergangenen Absätze Ihrer Produkte finden.
- Möchten Sie mittels KI Ihre Maschine überwachen, um frühzeitig Ausfällen vorzubeugen? Prüfen Sie, ob Ihre Maschine mit Sensoren ausgestattet ist und wo die dazugehörigen Daten abgelegt werden.
- Möchten Sie mittels KI gewisse Objekte oder Objekteigenschaften auf Bildern erkennen? Dann sind Bilddateien dieser relevant, zu denen optimalerweise Vermerke (sog. Labels) existieren, welche Objekte bzw. Objekteigenschaften auf den Bildern benennen.

Weitere typische Datenquellen sind Excel-Tabellen, CRM-Systeme, Urlaubskalender, Buchhaltungssoftware, ...
Selbst wenn es aktuell umständlich und kompliziert erscheint die Daten abzuziehen: Verzichten Sie zunächst darauf, sich über Schnittstellen zu Systemen Gedanken zu machen. Dieses Problem ist fast immer lösbar und kann noch gut zu einem späteren Zeitpunkt angegangen werden. Evtl. beschränken Sie sich zu Beginn auf eine Auswahl oder Stichprobe der Daten.

Keine Daten vorhanden? Gerade Produktionsmaschinen erfassen zwar häufig die aktuellen Produktionsdaten, speichern diese aber nicht immer ab. Fangen Sie heute mit dem Speichern und Sammeln an!

Schauen Sie jetzt **extern**:

Unternehmensinterne Daten stehen sehr häufig nicht für sich alleine, sondern haben Bezug zu externen Einflussgrößen. Beispielsweise ist der Umsatz eines Eiscafés wesentlich abhängig von Jahreszeit und Wetter. Besorgen Sie sich in so einem Fall einen für Ihre Region gültigen Feiertagskalender sowie historische und aktuelle Wetterdaten mit numerischen Informationen wie z. B. Tageshöchsttemperaturen, Niederschlag, etc. Eine mögliche Quelle für Deutschland ist beispielsweise der Deutsche Wetterdienst: *https://www.dwd.de/DE/klimaumwelt/klimaueberwachung/deutschland/datenundprodukte/datenundprodukte_node.html*.

In der industriellen Produktion nehmen z. B. Rohstoffpreise Einfluss auf den Unternehmensumsatz. Listen Sie auf, welche Rohstoffe in Ihrem Unternehmen eine Rolle spielen und tragen Sie die Marktpreise der Rohstoffe zusammen. Zu möglichen Quellen zählen beispielsweise *finanzen.net, boerse.de* oder *indexmundi.com.*

Sie produzieren Konsumgüter, sind im Gastgewerbe unterwegs, spüren stark die Bewe-

gungen in Ihrer Branche oder die allgemeine wirtschaftliche Lage? Dann könnten Branchen- und Wirtschaftsindikatoren in Ihrem KI-Projekt eine Rolle spielen. Schauen Sie z. B. in der Datenbank des Statistischen Bundesamtes (*https://www-genesis.destatis.de/genesis/online*), welche dort vorhandenen Kategorien einen Zusammenhang zu Ihrer Fragestellung haben könnten.

Weitere oft benötigte Daten sind z. B. Marktdaten oder Social-Media-Daten. Es gibt keine allgemeingültige Aussage darüber, welche externen Daten Sie heranziehen sollten, denn es kommt ganz auf die Fragestellung an. Es sollte sich das gesamte Projektteam einbringen – unsere Stellschraube 3.

Action Points:
- Datenverfügbarkeit für ausgewählte Use Cases prüfen
- Relevante unternehmensinterne Daten(-auswahl) zusammenstellen
- Evtl. externe Daten beschaffen

Stellschraube 3: Das Projektteam

Es ist typisch für ein KI-Projektteam, dass es interdisziplinär zusammengestellt ist. Wahrscheinlich sind die benötigten Skills bereits im Unternehmen vorhanden. Je nach Fragestellung kann ein abteilungsübergreifendes Projektteam vorteilhaft sein. Im Folgenden werden die Teammitglieder beschrieben, die in typischen KI-Projekten benötigt werden:

Mind. ein/e Vertreter*in der Fachabteilung:

Die Fachabteilung kennt die sachlichen Gegebenheiten, aktuellen Prozesse und v.a. die aktuellen Probleme. Diese Personen sind unerlässliche Experten im KI-Projektteam. Was bereitet ihnen in der alltäglichen Arbeit Bauchschmerzen? Was muss die spätere Lösung können? Welche Zusammenhänge und Abhängigkeiten müssen berücksichtigt werden? Nicht zuletzt wird das KI-Projekt die Fachabteilung am Ende unterstützen und ihnen die Arbeit erleichtern. Daher ist es wichtig, dass die Fachabteilung von Anfang an die Möglichkeit hat, ihre Anforderungen mitzuteilen und das Projekt und die Lösung aktiv mitzugestalten. Das sorgt nicht nur dafür, dass die Lösung später sachlich korrekt funktioniert, sondern es erhöht auch die Akzeptanz der Lösung.

Datenbeauftragte*r bzw. IT-ler*in:

Diese Person ist die Hands-on-Person im Projekt: Der/die Datenbeauftragte kommt an die benötigten Daten ran. Er/sie hat Zugangsdaten zu den verschiedenen Systemen oder kennt die Personen und Wege, die helfen können. Er/sie ist in der Lage, relevante Daten in einem sinnvoll verarbeitbaren Format aus den Systemen/Datenbanken zu exportieren (z. B. csv-Format). Oftmals kann diese Person bereits erste wichtige Datentransformationsschritte durchführen, z. B. durch geeignete SQL-Abfragen: Daten aus verschiedenen Systemen verknüpfen, geeignet filtern oder aggregieren. Diese Person kennt i.d.R. auch im Unternehmen vorhandene Software, die zur Analyse oder Anzeige von Daten und Analyseergebnissen herangezogen werden kann.

Citizen Data Scientist:

Hier handelt es sich um jemanden, der Datenanalyse intensiv nutzt, nicht notwendiger-

weise aber eine entsprechende Ausbildung hat. Mitarbeiter*innen aus einigen Fachabteilungen beschäftigen sich oftmals sehr ausführlich mit ihren Daten und bauen dadurch auch ohne tieferes mathematisches oder statistisches Fachwissen ein Gespür für Daten und Kenntnisse zur Nutzung von Datenverarbeitungsprogrammen auf. Eine analytische Denkweise und Spaß daran, sich in neue Dinge einzuarbeiten, sind hierbei entscheidende Faktoren, um sich in die Lage zu versetzen, datenanalytische Methoden anzuwenden. Fragen Sie im Unternehmen nach, wer Python oder R installiert hat oder ein Excel-Add-In mit statistischen Funktionalitäten. Diese Person könnte das richtige Teammitglied in Ihrem ersten KI-Projekt sein.

Projektleiter*in:
Bei dem/der Projektleiter*in laufen die Fäden zusammen. Er/sie darf den Überblick und das Ziel nicht aus den Augen verlieren. Der/die Projektleiter*in sollte in der Lage sein, zwischen den unterschiedlichen Projektmitgliedern zu übersetzen und die Erkenntnisse in eine für andere Spezialisten verständliche Weise zu transformieren. Nicht zuletzt hat er/sie den Überblick über Projektbudget und Kapazitäten.

Die Funktionen können auch (gerade am Anfang) zusammenfallen, z. B. wenn der/die Citizen Data Scientist im Unternehmen genau aus der Fachabteilung stammt, die der zu behandelnde Use Case betrifft.

Wichtig ist, dass die Beteiligten explizit Freiräume erhalten, sich mit dem Thema KI zu beschäftigen. Die Anwendung von KI ist ein innovatives Thema, dem große Wichtigkeit beigemessen wird, aber nicht unbedingt immer Dringlichkeit. Es kann durchaus einen hohen Initialaufwand bedeuten, ein KI-Projekt durchzuführen. Aber der Aufwand wird in den allermeisten Fällen belohnt: durch schnellere und schlankere Prozesse, genauere Ergebnisse, innovative Produkte, neue Erkenntnisse, sparsameren Rohstoffeinsatz, weniger Abschreibungen …

Das Projekt sollte nicht bis ins letzte Detail durchgeplant werden, sondern es empfiehlt sich ein agiles Vorgehen. Bei KI-Projekten geht es darum, die Daten sprechen zu lassen. Dementsprechend sollen die Daten auch mitentscheiden dürfen, wo die Reise hingeht. Nicht immer geben Daten genau das her, was man sich am Anfang vorstellt oder von ihnen erhofft. Daher sollte eine gewisse Flexibilität vorgesehen werden, sowohl was den zeitlichen Horizont des KI-Projekts angeht als auch den genauen Inhalt. Wir empfehlen, dass sich das Team fürs erste Justieren der Stellschrauben nicht mehr als vier Wochen nimmt und ein realistisches Ziel für diese Zeitspanne setzt. Anschließend gilt es zu evaluieren, wo das Team steht und wie es weitergehen soll.

Action Points:
- Projektbeteiligte ansprechen
- Projektbudget auftreiben
- Zeitnahen Zeitpunkt für Projektstart festlegen
- Flexibel bleiben

Stellschraube 4: Die Methodik

Jetzt wird es spannend: Es geht darum, die Daten zu analysieren und auf die Anwendung der KI vorzubereiten. Der/die Citizen Data Scientist kommt zum Einsatz. Diese/r sollte sich zunächst die vorhandenen historischen Daten ansehen und zunächst eine Datenbereinigung durchführen, um z. B. offensichtliche Datenfehler zu korrigieren. Beachten Sie, dass die Datenqualität eine entscheidende Rolle

für die Qualität der Ergebnisse spielt. Den Beginn machen die unternehmensinternen Daten. Geeignete Graphiken vermitteln gerade zu Beginn oft schnell und anschaulich Dateninhalte. Einfache Graphiken sind hierbei ausgefallenen Graphiken fast immer vorzuziehen.

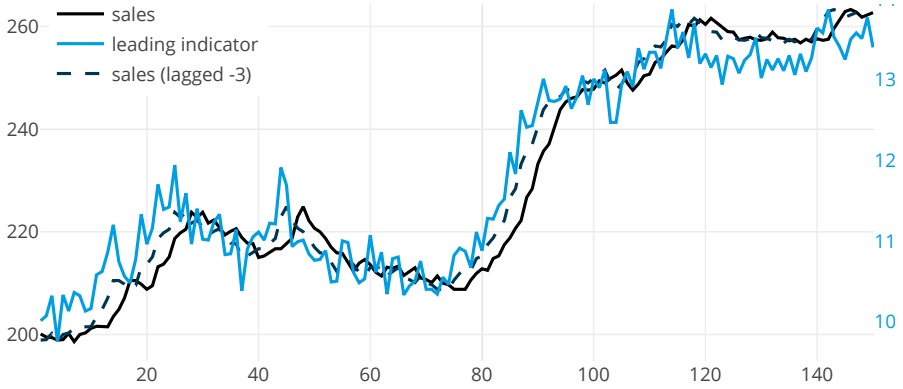
Zeitreihendaten (d.h. Daten, die einen Zeitbezug haben und über die Zeit erfasst wurden), sind z. B. am besten mit einem Linienplot zu visualisieren:

Nr. 1: Fig: Monatliche PKW-Neuzulassungen (in t) von Januar 2014 bis Februar 2021.
Quelle: *https://www.kba.de/DE/Statistik/Fahrzeuge/Neuzulassungen/MonatlicheNeuzulassungen/n_monatl_neuzulassungen_inhalt.html*, Neuzulassungen von Personenkraftwagen nach Marken und Modellreihen (FZ 10)

In unserem Beispiel *(Graphik Nr. 1)* sieht man die monatlichen Neuzulassungen von PKWs in Deutschland von 2014 bis Februar 2021. Auf den ersten Blick ist z. B. eine **Saisonalität** erkennbar – im Sommer gibt es i.d.R. mehr Neuzulassungen als im Winter –, ein Einbruch im September 2018 sowie ein für die Monate Januar bis März ungewöhnlicher Rückgang Anfang 2020 als Folge der Corona-Krise.

Abhängigkeiten kann man gut mit Hilfe eines Scatterplots visualisieren. Das ist sehr häufig sinnvoll, um einen ersten Eindruck von potenziellen (unternehmensinternen wie -externen) Einflussfaktoren zu bekommen. In unterer Graphik *(Nr. 2)* sieht man beispielsweise die Werte eines **Indikators** *(leading indicator)* aufge-

Nr. 2: Scatterplot der berühmten "leading indicator and sales"-Daten von Box & Jenkins (1976);
Quelle: *https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/BJsales.html*



Nr. 3: Die um drei Zeitpunkte verschobene Umsatz-Zeitreihe (sales lagged -3) und der Indikator (leading indicator) sind gleichlaufend, d.h. es existiert ein Lag von drei Monaten zwischen den beiden Zeitreihen.

tragen gegen Umsätze *(sales)* und kann einen klaren positiven linearen Zusammenhang zwischen beiden Variablen erkennen. Bei näherer Analyse stellt man fest, dass Zeitverschiebungen (sog. Lags) zwischen den beiden Variablen vorliegen, d.h. ein Aufschwung im Indikator mit etwas Verzögerung als Aufschwung in den Umsätzen erkennbar ist. Ein modifizierter Scatterplot (vorlaufende Variable zeitlich um den Lag verschoben) oder ein Linienplot mit zwei oder mehr (potenziell zeitlich verschobenen) Variablen kann den Sachverhalt graphisch darstellen. Numerisch kann er z. B. durch Bestimmung der Kreuzkorrelation analysiert werden. *(Siehe Graphik Nr. 3)*

Erste Datenvisualisierungen wie die obigen lassen bereits erste Erkenntnisse über die Daten zu und geben Hinweise, welche Art von KI potenziell Anwendung finden kann. In den obigen Fällen ist eine Prognose für die Zukunft das Ziel und zeitabhängige Daten spielen eine Rolle. Die KI sollte in solchen Fällen z. B. in der Lage sein, **Autokorrelationen** zu beachten. Der zweite Fall zeigt, dass das Heranziehen von externen Einflussfaktoren (in letzterem Fall ein geeigneter Wirtschaftsindikator) sinnvoll sein kann.

Auch eine KI ist immer nur so gut wie die Daten, mit der sie gefüttert wird. Nach diesem ersten Visualisierungsschritt kann es nützlich sein, noch eine Datenbereinigungsrunde anzuschließen, etwa zur Bereinigung von Ausreißern, zum Herausfiltern von Daten, die der Analyse nicht dienlich sind, oder um die Daten andersartig zu aggregieren.

Sind die Daten so groß, dass sie im ersten Schritt nicht auf einen Rechner passen? Zu diesem Zeitpunkt sollte man sich nicht die Mühe machen, einen großen Datensatz auf einmal zu bearbeiten, um nicht viel Zeit durch zu lange Berechnungsdauern zu verlieren. Zu Beginn bietet sich eine Stichprobe oder Teilmenge der Daten an. Bei zeitunabhängigen Daten lässt sich eine zufällige Stichprobe i.d.R. einfach entnehmen. Bei zeitabhängigen Daten können entweder zeitlich lange zurückliegende Daten ignoriert werden oder der Fokus auf die neueren Datenpunkte gelegt werden. Des Weiteren bietet es sich manchmal an, eine höhere Aggregationsstufe zu wählen, z. B. die Monatssummen statt der Werte auf täglicher oder stündlicher Basis.

Jetzt nähern wir uns immer mehr der KI. Jetzt soll der Rechner erste echte Berechnungen starten. Auch hier sollte es zunächst nicht zu kompliziert angegangen werden.

Ziel: Prognose

Wenn eine numerische und weitestgehend kontinuierliche Zielgröße vorliegt und das Ziel die Prognose ist, beginnen Sie mit einer **Regression** (Funktion *lm* im R-Paket *stats* bzw.

im Python-Paket *StatsModels*)! Sie ist eine Art "sichtbare KI". Eine **multiple Regression** oder eine Abwandlung davon ist die Grundlage vieler Machine-Learning- und damit KI-Verfahren, u.a. auch von dem wohl bekanntesten KI-Verfahren, dem **Neuronalen Netz** oder von **Multivariate Adaptive Regression Splines.**
Im ersten Schritt sollten nicht allzu viele potenzielle Einflussfaktoren (Regressoren) verwendet werden und darauf geachtet werden, dass sich diese möglichst stark voneinander unterscheiden, um zum jetzigen Zeitpunkt **Kollinearitäten** zu verhindern.

Die Regression kann bereits einige für eine KI typische Erkenntnisse erzielen:

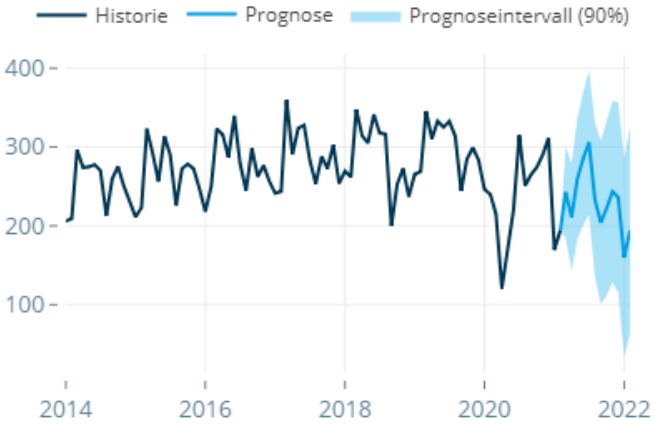
- Sie vermittelt einen Eindruck, welche der herangezogenen Regressoren tatsächlich Einflussfaktoren sind.
- Sie ermittelt von selbst Gewichte für die Einflussfaktoren und damit, ob die verschiedenen Einflussfaktoren eine eher wichtige oder unwichtige Rolle spielen.
- Sie ermöglicht die Erzeugung von Prognosen. Gute Prognosen sind die wesentliche Grundlage für datenbasierte Entscheidungen. (Siehe Definitionen von **Predictive Analytics/Prescriptive Analytics**)

Ein Beispiel für eine einfache lineare Regression können Sie Abbildung Nr. 4 entnehmen.

Nicht nur für zeitunabhängige Daten, sondern auch für Zeitreihendaten kann eine Regression gemacht werden. Hierbei fungiert die eigene Datenhistorie als Regressor der vorherzusagenden Zeitreihe. Indem einer Beobachtung (Wert) geeignete zurückliegende Beobachtungen als Regressoren zugeordnet werden, können Autokorrelationen, Saisonalitäten mit geeigneter Saisonlänge (z. B. 12 für monatliche Daten und Jahressaison) etc. modelliert werden. Üblichere Verfahren im Kontext von Zeitreihen sind Verfahren, welche Autokorrelationen bereits von vornherein in Betracht ziehen. Dazu gehören beispielsweise die Verfahren **ARIMA** und **Exponentielle Glättung**, die sich gut als Einstiegsmethoden für Zeitreihenprognosen eignen. Die wichtigsten Zeitreiheneigenschaften wie Trend und Saisonalitäten können durch diese Methoden erlernt und abgebildet werden. Erweiterungen erlauben den zusätzlichen Einbezug von Einflussfaktoren (ARIMAX bzw. Exponentielle Glättung mit Kovariaten), worauf bei einer ersten Berührung mit den Methoden aber verzichtet werden kann.

Nr. 4: Fig: Visualisierte lineare Regression mit einem Einflussfaktor, nämlich des um drei Monate verschobenen Indikators (leading indicator) auf die Umsätze (sales) der "leading indicator and sales"-Daten von Box & Jenkins (1976).

In nebenstehender Graphik (Nr. 5) ist beispielhaft eine Prognose zu sehen, die durch exponentielle Glättung entstanden ist. Die dunkelblaue Linie – die Punktprognose – ist eingefasst von einer Folge aus Prognoseintervallen (hellblau), welche die Prognoseunsicherheit quantifizieren, und damit der Tatsache gerecht werden, dass eine Prognose in den seltensten Fällen den wahren Wert exakt treffen wird.



Nr. 5: Fig: Prognose der monatlichen PKW-Neuzulassungen (in t) für März 2021 bis Februar 2022 durch exponentielle Glättung. Prognoseintervall zum Konfidenzniveau 90 %.

Ziel: Klassifikation/Mustererkennung

Was im Falle von Zeitreihen eine Kunst für sich ist, ist bei dem Thema "Klassifikation" Standard: die Aufteilung des Datensatzes in **Trainings- und Testdatensatz**. Das Verhältnis 2 : 1 (Trainingsdatensatz : Testdatensatz) ist ein guter Richtwert. Das Training der KI inkl. Training von Modellparametern, Hyperparametern sowie der Auswahl von Einflussfaktoren sollte hierbei ausschließlich auf dem Trainingsdatensatz stattfinden. Der Testdatensatz dient einer unabhängigen Überprüfung der Ergebnis- und Prognosegüte der KI. Weichen die Ergebnisse auf Trainings- und Testdatensatz stark voneinander ab, könnte das ein Zeichen von Overfitting sein (siehe auch Abschnitt "Reflektion und Expertenmodus").

Haben Sie eine Zielvariable vorliegen, die nur einige wenige konkrete Werte als mögliche Ausprägungen annehmen kann (im Extremfall nur die Werte 1 und 0 bzw. "ja" und "nein" = binäres Klassifikationsproblem), und Ihr Ziel die Klassifikation ist, beginnen Sie mit **CART** (Funktion *rpart* im R-Paket *rpart* bzw. im Python-Paket *scikit-learn*)! Breimans Klassifikationsverfahren ist ein einfaches Entscheidungsbaumverfahren, bei dem mehrmals hintereinander die Daten durch jeweils aufeinanderfolgende binäre Entscheidungen klassifiziert werden.

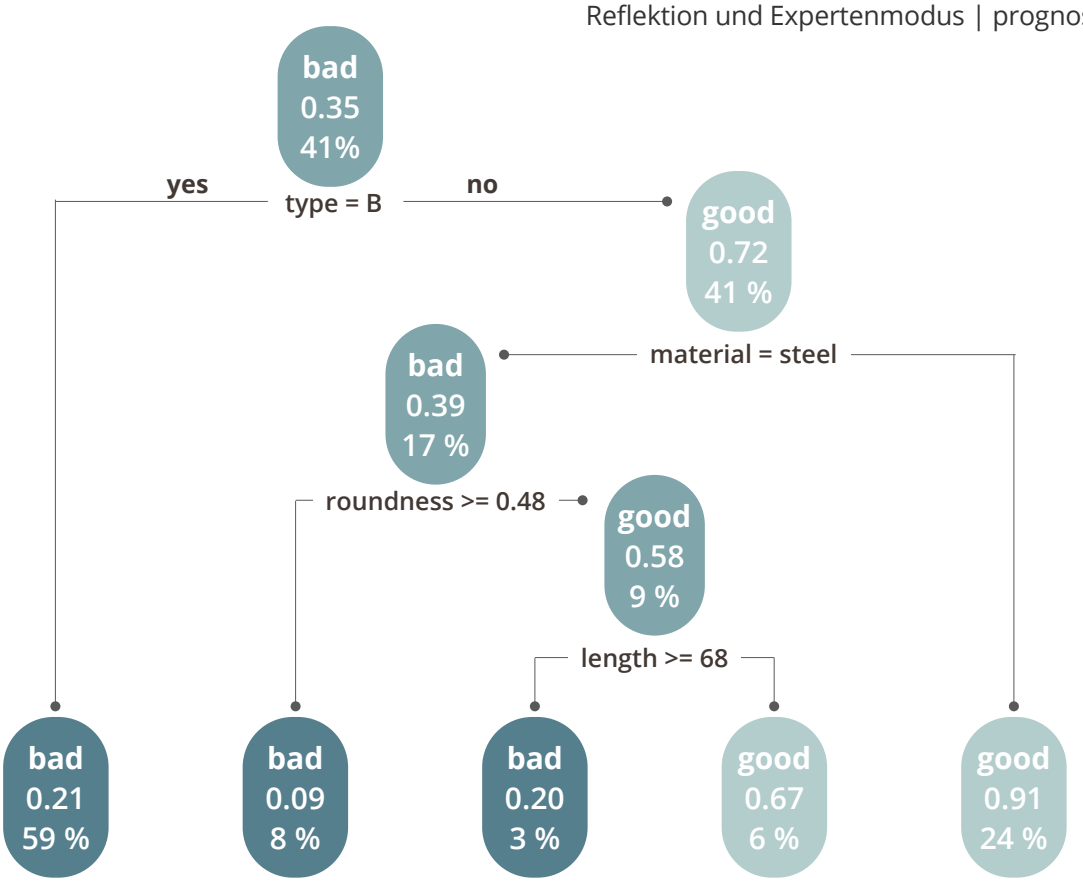
Sie erhalten:

- Erkenntnisse zu Klassifikationsergebnissen unter bestimmten Eigenschaften Ihrer Inputdaten
- Erkenntnisse zur Relevanz von Einflussfaktoren
- Eine Gruppierung Ihrer Daten in die jeweiligen Gruppen
- Erkenntnisse zur Güte der Klassifikation bzw. zur Falschklassifikationsrate

Die Ergebnisse sind in einem Entscheidungsbaum übersichtlich darstellbar und gut nachvollziehbar und geben auf diese Weise einen guten ersten Einblick in das Prinzip von Klassifikationsverfahren. Ein Beispiel finden Sie in Abbildung Nr. 6.

Action Points:

- Geeigneten Trainingsdatensatz auswählen
- Daten visualisieren
- Einfaches KI-Verfahren anwenden



Nr. 6: Fig: Entscheidungsbaum bei einem binären Entscheidungsproblem (mögliche Ausgänge: “good”, “bad”) bei einer industriellen Anwendung

Reflektion und Expertenmodus

In diesem Stadium ist es wichtig, dass Sie reflektieren, ob und wenn ja, welche Benefits Sie aus den ersten KI-Schritten ziehen konnten. Entsprechend gestaltet sich das weitere Vorgehen.

Überprüfen Sie die Güte der Ergebnisse: Welche Ergebnisse hätten Sie erzielt, wenn Sie auf den Einsatz von KI verzichtet hätten und wenn stattdessen z. B. Beschäftigte die gleiche Arbeit verrichtet hätten? Wer war schneller, wer langsamer? Welche Aspekte können durch den Einsatz der KI überhaupt erst in Betracht gezogen werden, auf welche menschlichen Aspekte müssen Sie dagegen vielleicht verzichten? Versuchen Sie eine Bewertung des Nutzens und Bezifferung des Einsparpotenzials. Halten Sie sich dabei vor Augen, dass nach diesen ersten Schritten in Richtung KI vermutlich immer noch viel Luft nach oben ist, selbst wenn Sie jetzt schon viel erreicht haben.

Wenn Sie bei den ersten Schritten statt des vollständigen Datensatzes nur einen Teildatensatz bearbeitet haben, dann nehmen Sie sich jetzt den vollständigen Datensatz vor. Evtl. ist es hierfür notwendig, die Rechenkapazität auszubauen, und mehr Geduld bei der Ergebniserzeugung mitzubringen. Hierfür ist es heutzutage nicht mehr nötig, die Rechenkapazität im eigenen Unternehmen vorzuhalten. Cloud-Services bieten bedarfsgerechte Rechen- und Speicherkapazitäten und bringen sogar gelegentlich datenanalytische bzw. Machine-Learning-Funktionalitäten mit.

Es ist nicht unwahrscheinlich, dass sich durch die von uns in den vorherigen Kapiteln vorgeschlagenen Schritte ergeben hat, dass die

zu analysierende Datenmenge gar nicht so groß ist, wie ursprünglich gedacht. Vielleicht hat eine geeignete Filterung und eine sinnvolle Aggregation die Daten in einen sehr gut handhabbaren kleineren Datensatz transformiert, der nicht weniger Informationen als der ursprüngliche enthält, um die es geht. Möglicherweise erscheint der Datensatz durch die Visualisierung und schrittweise Behandlung schlicht nicht mehr unübersichtlich und groß.

Der/die Citizen Data Scientist im Team sollte sich jetzt mit weiteren Methoden und Algorithmen beschäftigen, beispielsweise **Ridge Regression**, **Exponentielle Glättung** (mit und ohne Kovariaten), **ARIMA**, **Multivariate Adaptive Regression Splines (MARS)**, **Support Vector Machines, Random Forests, Artificial Neural Networks** oder **Deep Learning**, evtl. unter Zuhilfenahme vortrainierter neuronaler Netze. Folgender Data-Science-Grundsatz sollte hierbei Beachtung finden: Je komplizierter ein Modell oder Verfahren, desto anfälliger ist es i.d.R. für **Overfitting**, was im Sinne der Ergebnisgüte drastische Folgen für eine Prognose und die Schlussfolgerungen haben kann, die aus den Daten gezogen werden. Wir möchten dazu ermutigen, es so simpel zu halten wie möglich: Wenn zwei Verfahren annähernd das gleiche Ergebnis erzeugen, dann empfiehlt sich das einfachere, auch im Hinblick auf die zukünftige Wartung der Algorithmen. Ein transparenterer Algorithmus heißt nicht selten auch bessere Erklärbarkeit des Verfahrens und sorgt damit für eine höhere Akzeptanz im Unternehmen als ein Black-Box-Verfahren.

Das Thema **Overfitting** ist im Zusammenhang mit KI ein wichtiges Thema. Auf eine gewisse Art und Weise angewendet, kann man z. B. durch hochkomplizierte Modelle fast alle gewünschten Ergebnisse erzeugen lassen, die man möchte – oft ein Trugschluss. Es kann leicht passieren, dass man durch immer kompliziertere Modelle/Algorithmen den Wald vor lauter Bäumen nicht mehr sieht und scheinbare Muster erkannt werden, die so gar nicht existieren. Daher ist es wichtig, diese immer auch auf den Prüfstand zu stellen. Hier ist die **Kreuzvalidierung** ein wichtiges Mittel.

Bei der Bearbeitung der vorangegangenen Stellschrauben ist bestimmt aufgefallen, welcher Punkt besondere Aufmerksamkeit fordert und wo es potenziell noch Engpässe gibt. Hinterfragen Sie, welche Kompetenz im Team noch gestärkt werden sollte. Möglicherweise sind gerade die Aufgaben der Person, die sich um die Analysen kümmert (Citizen Data Scientist) anspruchsvoll und verlangen weitere Expertise: Der Einsatz eines Expert Data Scientist könnte sinnvoll sein.

Die Stärke einer KI ist das sukzessive Lernen. Beim Hinzukommen neuer Daten lernt der Algorithmus, indem er sich selbst hinterfragt und den neuen Gegebenheiten anpasst. Auf Dauer ist das manuelle Hinzufügen von Daten mühselig. Es lohnt sich, dass Sie sich um automatisierte Schnittstellen Gedanken machen.

Vergessen Sie zum Schluss nicht, von den Vorteilen und Erleichterungen Gebrauch zu machen, die der Einsatz der KI mit sich bringt. Was die Gesamtlösung angeht, sollten Sie sich stets folgenden Leitspruch zu Herzen nehmen: Gut muss nicht immer kompliziert sein. Wichtig ist es, überhaupt mal zu beginnen.

Action Points:

- Erste Ergebnisse bewerten
- Analysen auf größere Datenmenge ausweiten
- Expert Data Scientists für eine "Experten-KI" engagieren
- Schnittstellen schaffen / KI in die Alltagsprozesse integrieren
- Die Früchte der KI ernten

Fazit

Herzlichen Glückwunsch!

Sie haben jetzt erfolgreich die ersten Stellschrauben Ihres KI-Projekts justiert. Wir hoffen, wir konnten Ihnen vermitteln, dass KI keine Zauberei ist, sondern beim gezielten Angehen durchaus schnelle Erfolge erzielt werden können. Uns war wichtig, Ihnen den Schrecken zu nehmen und Ihnen zu zeigen, was hinter KI steckt und wie Sie sich dem annähern können. Wir glauben, dass Sie jetzt in der Lage sind, notwendige nächste Schritte besser einordnen und angehen zu können.

Viele KI-Anwendungen machen uns Angst, weil wir nicht genau wissen was dahinter steckt. Transparenz ist aber nicht immer vollständig umsetzbar, etwa bei vielen **Neuronalen Netzen.** bzw. **Deep-Learning-Mechanismen**, die oft eine riesige Anzahl an Parametern erzeugen, indem sie für verschiedene Inputfaktoren und -daten Verknüpfungen erzeugen, die oft nicht mehr interpretierbar sind. Aber: Mehr als aus den eingegebenen Daten etwas machen, kann die KI nicht. Und auch wenn man sich mit offenen Augen an andere KI-Anwendungen wagt, kann man sich, obwohl der Mensch vielleicht nicht mehr alles überblicken kann, sicher sein, dass gilt: Das, was man dem System liefert, wird für die Analyse verwendet. Wie so oft gilt auch hier, dass es in Unternehmen heutzutage nicht darauf ankommt, menschliche Aufgaben auf Biegen und Brechen Maschinen zu überlassen. Stattdessen sollten die größten Vorteile beider Seiten zum Einsatz kommen, sodass sich im intelligenten Zusammenspiel eine Gesamtlösung ergibt, die für alle den gewünschten Nutzen bringt.

Anhang

About us

Über prognostica

prognostica ist ein junges Beratungs- und Softwareunternehmen aus Würzburg, das auf den Bereich “Predictive Analytics” und “Data Science” spezialisiert ist.

Seit der Gründung 2014 ist prognostica mittlerweile auf 17 Mitarbeiter gewachsen und verfolgt Projekte branchenübergreifend. Das interdisziplinäre Team entwickelt innovative Lösungen, die den Anwendern helfen, beispielsweise ihre Finanz-, Produktions-, Absatz- oder Bedarfsplanung zu verbessern.

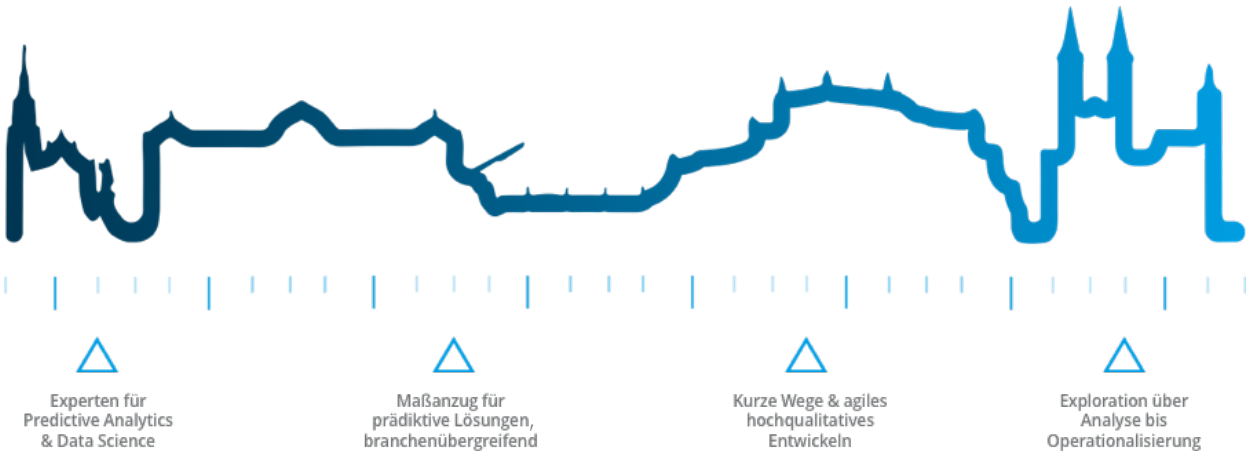
Die Stärke von prognostica ist die Erstellung von präzisen und objektiven Prognosen. Aus Zeitreihen werden mit Hilfe von statistischen Methoden und künstlicher Intelligenz (Machine-Learning-Verfahren) die relevanten Strukturen extrahiert, um die künftigen Entwicklungen sehr genau vorherzusagen.

Die Vorhersagelösungen von prognostica sind hoch automatisiert und individuell auf die Fragestellungen kleiner und großer Datenmengen zugeschnitten. Eine Besonderheit ist dabei die Erstellung von Prognosen unter Hinzunahme von geeigneten Wirtschaftsindizes und internen KPIs.

Ganzheitliche Lösungen stehen im Fokus: prognostica unterstützt ihre Kunden von ersten Analysen über Prototyp-Evaluierungen mit maßgeschneiderter, interaktiver Visualisierung bis hin zu einer integrierten, operativen Lösung.

Wer in einer digitalisierten Welt auf die Zukunft vorbereitet ist, trifft die besseren Entscheidungen.

Dabei hilft prognostica.



prognostica Knowledge Base – Ein Auszug

Predictive Analytics

Predictive Analytics erkennt Muster, Strukturen und Zusammenhänge in Daten und nutzt diese, um Aussagen über Ereignisse in der Zukunft zu treffen. Dabei kommen Methoden der klassischen Statistik wie auch des maschinellen Lernens (Machine Learning) zum Einsatz. Predictive Analytics lässt sich von Descriptive und Prescriptive Analytics folgendermaßen abgrenzen:

Descriptive Analytics analysiert und beschreibt historische Daten, erkennt Muster, zeigt Zusammenhänge auf oder entdeckt Anomalien in den Daten.

Predictive Analytics trifft darauf aufbauend Aussagen über die Zukunft, erstellt datenbasierte Prognosen und bemisst Eintrittswahrscheinlichkeiten von möglichen künftigen Ereignissen.

Prescriptive Analytics übersetzt die gewonnene Einschätzung der Zukunft in konkrete Handlungsempfehlungen und identifiziert optimale Entscheidungen.

Machine Learning (ML)

Machine Learning ist ein Teilbereich von künstlicher Intelligenz. Unter Machine Learning versteht man im Wesentlichen alle Verfahren, die es Maschinen ermöglichen, basierend auf Daten zu lernen und auf diese Weise Wissen zu generieren und daraus ggf. Schlüsse zu ziehen. Spezieller lernen Algorithmen auf Grundlage von Daten, indem sie mit der Kenntnis jedes neuen Datenpunkts sich selbst optimieren und anpassen. Machine-Learning-Verfahren können i.d.R. gut mit vielen erklärenden Variablen umgehen. Dafür benötigen sie im Vergleich zu statistischen Verfahren häufig mehr Trainingsdaten. Man unterscheidet generell zwischen Supervised und Unsupervised Learning.

Supervised Learning

Bei Supervised Learning wird mit gelabelten Daten gearbeitet. Das bedeutet, dass die Lösungen, die das System erzeugen soll, im Trainings- und Testdatenset bekannt sind. Es kann also entschieden werden, ob der Algorithmus im Einzelfall mit seiner Entscheidung gut/richtig oder weniger gut/falsch liegt. So kann gemessen werden, wie gut das trainierte Modell funktioniert. Selbstverständlich wird hier ein möglichst kleiner Fehlerwert angestrebt. Zu den Supervised-Learning-Verfahren gehören beispielsweise Entscheidungsbaumverfahren (CART und Random Forests), regressive Verfahren (SVM und MARS) sowie auch Künstliche Neuronale Netze (KNNs). Die genannten

Verfahren können sowohl zur Regression wie auch zur Klassifikation von Daten verwendet werden.

Classification And Regression Tree (CART)

CART ist ein vom amerikanischen Statistiker Leo Breiman geprägter Begriff für Entscheidungsbaum-Algorithmen, bei welchen über Binärbäume die Klasse, zu der Daten gehören, bestimmt werden. Im Fall der Zeitreihenvorhersage folgt aus dieser Klasse die Prognose. Diese Verfahren werden häufig auch im Machine-Learning-Bereich eingesetzt und dienen als Grundlage für Random Forests.

Random Forest

Ein Random Forest ist ein Supervised-Learning-Verfahren zur Klassifikation und Regression von Daten, in dem verschiedene, möglichst unterschiedliche Entscheidungsbäume generiert werden. Die Werte bzw. Klassen, die aus den verschiedenen Entscheidungsbäumen resultieren (siehe hierzu auch CART), werden hierbei zu einem Ergebnis kombiniert und können dadurch genauere Ergebnisse liefern als ein einzelner Entscheidungsbaum.

Multivariate Adaptive Regression Splines (MARS)

Multivariate Adaptive Regression Splines sind ein Verfahren, das basierend auf Daten selbstständig ein Modell entwickelt, in dem Nicht-Linearitäten und Interaktionen zwischen verschiedenen erklärenden Größen berücksichtigt werden.

Sukzessive wird ein Modell aus sogenannten hinge functions (und ihren Produkten) aufgebaut. Hinge functions sind bis zu einer gewissen Schwelle Null und gehen dann in eine Gerade mit positiver oder negativer Steigung über. Sie sehen aus wie ein Hockeystick. Durch eine geschickte Verkettung solcher hinge functions lassen sich komplexe Zusammenhänge besser approximieren als nur mit linearen Termen.
MARS-Verfahren besitzen eine interne Strategie, um sich für die beste Kombination aus hinge functions und zur Verfügung gestellter erklärenden Variablen zu entscheiden.

Support Vector Machine (SVM)

Support Vector Machine ist ein Supervised-Learning-Verfahren zur Klassifikation von Daten.
Um die Datenpunkte verschiedener Klassen möglichst gut voneinander zu trennen, sucht SVM nach solchen Klassengrenzen, die einen möglichst großen Abstand zu den Datenpunkten der verschiedenen Klassen besitzen und so einen möglichst breiten Bereich um die Klassengrenze frei von Datenpunkten lassen.
Bei den Klassengrenzen arbeitet SVM zunächst mit Geraden oder Ebenen. Mit dem sogenannten Kernel-Trick, d.h. einer geschickten Datentransformation, lassen sich aber auch sehr gut komplexe, nicht-lineare Trennlinien oder -flächen finden.
SVM kann auch zur Regression verwendet werden. Man spricht hier auch von Support Vector Regression.

Künstliches Neuronales Netz (KNN)

Ein künstliches neuronales Netz fußt auf einem dem menschlichen Gehirn nachempfundenen Netz aus Knoten (Neuronen) und ihren Verbindungen (Synapsen). Die Knoten eines neuronalen Netzes sind in hintereinander geschalteten Schichten angeordnet. Die Inputdaten werden dabei durch die einzelnen Schichten gereicht. Jeder Knoten gewichtet die Ausgaben von Knoten vorangestellter Schichten und gibt seinen Aktivierungsgrad nach bestimmten Aktivierungsregeln an Knoten in nachfolgenden Schichten weiter. So können komplexe Zusammenhänge abgebildet und Informationen aus verschiedenen Inputgrößen verknüpft werden. Auf bekannten Daten erlernt ein neuronales Netz zunächst in einer Trainingsphase passende Gewichte und Aktivierungsregeln, bevor es auf neue, unbekannte Daten angewandt werden kann.

Liegt eine komplexe Netzarchitektur mit vielen, tief ineinander geschalteten Schichten vor, so spricht man von **Deep Learning** bzw. **Deep Neural Networks**.

Typische Anwendungsfelder von neuronalen Netzen sind z. B. die Bild- oder Spracherkennung. An dieser Stelle kommen oft sog. **Convolutional Neural Networks (CNNs)** zu tragen. Neuronale Netze können aber auch im Zeitreihenkontext gut zur Mustererkennung oder Prognose eingesetzt werden. Besonders zu erwähnen sind hier die **Long Short-Term Memory Networks (LSTMs)**.

Training und Validierung

Machine-Learning- oder Prognosemodelle werden in der Regel auf bekannten Daten trainiert bzw. geschätzt. Hierbei gilt es ein solches Modell zu finden, das möglichst gut generalisiert, d.h. eine gute Leistung auf neuen, vorher nicht gesehenen Daten erbringt und dort robuste Ergebnisse liefert.

Overfitting

Overfitting liegt vor, wenn sich das Modell zu stark an die Trainingsdaten anpasst und Muster zu erkennen glaubt, die nicht existieren. Dies ist z. B. bei einem Zeitreihenmodell der Fall, wenn das Modell neben wichtigen zu erfassenden Strukturkomponenten wie Trend oder die Saisonalität auch die zufällige Streuung der Trainingsdaten mit modelliert. Solche Modelle sind in der Regel nicht generalisierbar, d.h. sie liefern keine stabilen Ergebnisse auf neuen, unbekannten Daten, die nicht zum Modelltraining verwendet wurden. Die Ursache von Overfitting ist, dass das Modell zu komplex ist, d.h. zu viele Parameter oder erklärende Variablen enthält, die zu stark auf die Trainingsdaten eingestellt wurden. Um Overfitting zu vermeiden oder zu reduzieren, gibt es verschiedene Strategien, z. B.:

- Out-of-Sample-Validierung auf einem unabhängigen Testdatensatz
- Kreuzvalidierung
- Bestrafung der Modellkomplexität durch Regularisierung (wie z. B. bei Elastic Net)

Out-of-Sample-Validierung

Um ein möglichst generalisierbares Modell zu finden, ist es üblich, einen vorliegenden Datensatz in ein **Trainings- und ein Testset** zu unterteilen. Ersteres wird
verwendet, um die passende Modellparameter zu erlernen, letzteres, um die Güte des trainierten Modells auf einem unabhängigen Datensatz zu überprüfen und die Vorhersagegenauigkeit zu bemessen.

Kreuzvalidierung

Kreuzvalidierung (engl. cross validation) ist eine verbreitete Validierungsstrategie im Machine-Learning-Bereich, um Overfitting zu vermeiden und robuste Ergebnisse auf unbekannten Daten zu liefern.
Eine einfache Out-of-Sample-Validierung, bei der der vorhandene Datensatz nur einmal in Trainings- und Validierungsdaten unterteilt wird, hat jedoch den Nachteil, dass relevante Muster aus den Validierungsdaten, die nicht in den Trainingsdaten zu erkennen waren, im Modell nicht berücksichtigt werden können. Wenn umgekehrt das Validierungsset die Realität nicht gut abbildet, kann man keine verlässlichen Erkenntnisse über die Generalisierbarkeit des Modells gewinnen.
Diesem Problem begegnet die Kreuzvalidierung, indem sie die Unterteilung des Datensatzes in Trainings- und Testset mehrfach wiederholt und zwar so lange, bis jedes Element genau einmal für die Testmenge verwendet wurde. Auf diese Weise wird das Modell am Ende auf (fast) allen Daten trainiert und auf (fast) allen Daten einmal validiert.

Regression

Regression ist ein statistisches Verfahren, mit dem der Zusammenhang von einer oder mehreren Größen (erklärende Variablen oder Regressor) auf eine Zielgröße (erklärte Variable oder Regressand) quantitativ modelliert wird. Liegen dabei nicht nur eine, sondern mehrere erklärende Variablen vor, so spricht man von multipler Regression.

Lineare Regression

Die lineare Regression ist die einfachste Form von Regression. Sie modelliert lineare Zusammenhänge à la "Für jedes Grad Celsius, um das die Tageshöchsttemperatur (erklärende Variable) steigt, steigt die Anzahl der Eisverkäufe pro Tag (Zielgröße) um eine feste Stückzahl." Eine lineare Regression schätzt für jede erklärende Variable einen passenden Koeffizienten (Faktor), so dass sie in Summe die Zielgröße möglichst gut beschreiben. Graphisch lässt sich die Funktionsweise einer linearen Einfachregression (d.h. nur eine erklärende Variable) folgendermaßen veranschaulichen: Zeichnet man die Datenpunkte in ein Koordinatensystem (x-Achse: erklärende Variable; y-Achse: Zielgröße), so wird nach einer solchen Gerade gesucht, die die Datenpunkte möglichst gut approximiert.

Kollinearität

Von (stochastischer) Kollinearität spricht man, wenn eine erklärende Variable stark mit einer anderen erklärenden Variablen korreliert. Kollinearität ist ein typisches Problem in Regressionsmodellen. Sind zwei Variablen stark miteinander korreliert, lässt sich aus Daten- oder Modellsicht schwer entscheiden, welche der beiden tatsächlich Einfluss auf die zu erklärende Größe ausübt. Möglicherweise ist eine von beiden redundant. Vielleicht sind aus sachlogischer Sicht aber auch beide in einer geeigneten Gewichtung relevant.
Kollinearität führt bei der Regression zu einer instabilen Schätzung der Modellkoeffizienten und erschwert generell die Interpretation des Modells. Korreliert eine erklärende Variable nicht nur mit einer sondern mit mehreren anderen erklärenden Variablen, so spricht man auch von Multikollinearität.

Regularisierte Regression

Regularisierte Regressionen sind spezielle Formen von Regressionen, bei denen Modellkomplexität bestraft wird mit dem Ziel ein möglichst robustes und generalisierbares Modell zu generieren und Overfitting zu vermeiden.
Um die Modellkomplexität bei der Schätzung des Modells zu berücksichtigen, werden neben den Abweichungen des Modells von den tatsächlichen Daten zusätzlich auch die Größenordnungen der Modellkoeffizienten betrachtet und geschickt kontrolliert. Beispiele für regularisierte Regressionen sind Ridge Regression, Lasso Regression und Elastic Nets.

- Bei einer **Ridge Regression** gehen neben den quadrierten Modellfehlern ebenfalls die quadrierten Koeffizienten in die Kostenfunktion zur Schätzung des Modells mit ein.
- Bei einer **Lasso Regression** werden an dieser Stelle die Absolutwerte der Koeffizienten betrachtet.
- **Elastic Net** kombiniert beide Arten der Penalisierung. Sowohl Ridge wie auch Lasso Regression sind Randfälle eines Elastic Nets.

Zeitreihenanalyse

Zeitreihe

Eine Zeitreihe beschreibt die zeitliche Entwicklung einer veränderlichen Größe wie zum Beispiel eines Umsatzes, eines Aktienkurses, eines Lagerbestands oder auch einer Temperatur. Die Beobachtungszeiträume einer Zeitreihe sind regelmäßig: Die Werte werden jährlich, monatlich, täglich, etc. erfasst. Zeitreihen dienen als Grundlage zur Analyse der Vergangenheitswerte aber auch für die Prognose der künftigen Entwicklung.

Saisonalität

Mit Saisonalität bezeichnet man eine für Zeitreihen typische Strukturkomponente. Saisonalität liegt vor, wenn sich in der Zeitreihe zyklische, wiederholende Figuren finden. Die Länge des Zeitraums, nachdem diese saisonalen Figuren wiederkehren, bezeichnet man als Saisonlänge. Monatsdaten weisen zum Beispiel oft eine Saisonalität mit einer Saisonlänge von 12 Monaten auf.

Indikator

Um eine Zeitreihe zu modellieren, sind neben intrinsischen Strukturen wie Trend und Saisonalität oft auch externe Kontextinformation und Einflussgrößen relevant. Enthält eine Einflussgröße relevante Information mit zeitlichem Vorlauf, so spricht man von einem Indikator. Ein Indikator antizipiert also künftige Entwicklungen in der zu prognostizierenden Zeitreihe. Den zugehörigen zeitlichen Versatz zwischen Indikator und der zu prognostizierenden Zeitreihe bezeichnet man als den Lag des Indikators.

Prognoseintervall

Eine (Punkt-)Prognose wird den künftigen, tatsächlich eintreffenden Wert selten ganz exakt treffen: Die Prognose ist stets mit einer gewissen Unsicherheit behaftet. Diese Unschärfe lässt sich mittels eines Prognoseintervalls quantifizieren. Das Prognoseintervall beschreibt einen Wertebereich um die statistische Punktprognose, der den tatsächlich eintreffenden Wert mit einer vorgegebenen Wahrscheinlichkeit, dem Prognosekonfidenzniveau, überdecken wird. Je größer das Prognosekonfidenzniveau, desto wahrscheinlicher, dass der künftige Wert vom Intervall überdeckt wird. Ein Prognosekonfidenzniveau von 95 Prozent bedeutet beispielsweise, dass von 100 auf eine bestimmte Weise berechneten Prognoseintervallen im Mittel 95 die wahren (zukünftigen) Zeitreihenwerte enthalten. In etwa 5 Prozent der Fälle dagegen liegen die wahren Zeitreihenwerte außerhalb.

Autoregressive Integrated Moving Average (ARIMA)

Ein ARIMA-Modell ist ein Modell zur Analyse und Prognose von Zeitreihen, in das vergangene Werte der Zeitreihe selbst sowie vergangene Fehlerterme eingehen. Die Analyse kann hierbei statt auf den Rohdaten auch auf (mehrfach) differenzierten Daten stattfinden. Saisonalitäten sowie exogene Einflussgrößen können in ARIMA-Modellen ebenfalls mitmodelliert werden.

Exponentielle Glättung mit Kovariaten (ESCov)

Das Verfahren der exponentiellen Glättung ist ein bewährtes Verfahren zur Analyse und Prognose von Zeitreihen, welches Niveau-, Trend- und multiple Saisonkomponenten in Betracht ziehen kann. Hierbei werden weiter zurückliegende Zeitreihenwerte üblicherweise weniger stark gewichtet als die jüngere Historie. Die Erweiterung „Exponentielle Glättung mit Kovariaten“ kann zusätzlich mit exogenen Einflussgrößen umgehen.

Statistische Grundbegriffe

Korrelation

Die Korrelation zwischen zwei Größen beschreibt den Grad ihres statistischen, typischerweise linearen, Zusammenhangs. Diese Beziehung ist ungerichtet, es wird also keine Aussage darüber getroffen, ob und wenn ja welche Größe die andere bedingt.
Wenn der Anstieg einer Größe tendenziell mit dem Anstieg einer anderen Größe einhergeht, sind diese Größen positiv korreliert, wenn umgekehrt eine Größe tendenziell sinkt, während die Andere steigt, spricht man von einer negativen Korrelation.
Ein typisches Beispiel: Die Tageshöchsttemperatur und die Eisverkäufe pro Tag sind in der Regel positiv korreliert.

Autokorrelation

Autokorrelation bedeutet, dass eine zeitlich veränderliche Größe mit sich selbst, verschoben um eine feste Zeiteinheit, korreliert. So sind z. B. die Höchsttemperaturen eines Tages mit den Höchsttemperaturen des vorherigen Tages positiv autokorreliert. Auf einen sehr heißen Tag folgt nämlich häufig ein Tag, der ebenfalls eine hohe Tagestemperatur aufweist.

Die Erläuterungen der Begriffe in unserer Knowledge Base waren gut verständlich und haben für Sie zum Verständnis komplizierter Sachverhalte im Umfeld von Data Science beigetragen? Unsere Kunden erhalten natürlich Zugang zu unserer ausführlichen, u.a. mit vielen erklärenden Graphiken angereicherten Knowledge Base. Sind Sie auch interessiert? Kontaktieren Sie uns gerne!

prognostica GmbH
Berliner Platz 6
97080 Würzburg

info@prognostica.de
+49 931 497 386 30

www.prognostica.de

prognostica
discover tomorrow


prognostica
discover tomorrow