Prof. Dr. Dr. Yoan Hermstrüwer HS 2023

Legal Tech

Lösungsvorschlag

1. Einige Gerichte greifen bei ihren strafprozessualen Haftentscheidungen auf KI-Systeme zur Vorhersage der Flucht- oder Wiederholungsgefahr von Beschuldigten zurück.

Welche Vor- und Nachteile haben solche Systeme im Allgemeinen? Bitte fassen Sie sich kurz!

Vorteile

Effizienz: KI kann grosse Datenmengen schnell analysieren und Muster erkennen, was zu schnelleren Entscheidungsprozessen führen kann.

Kostenersparnis: Hinzu kommt, dass das Training von KI-Systemen mittel- und langfristig kostengünstiger sein dürfte als der Einsatz von MI.

Genauigkeit: KI-basierte Vorhersagen und Klassifikationen sind bei hochqualitativen Datensätzen genauer als menschliche.

Objektivität: Grundsätzlich kann KI dabei helfen, subjektive Verzerrungen (biases) von menschlichen Entscheidungsträgern zu reduzieren.

Konsistenz: KI-Systeme können eine gleichmässigere Anwendung von Kriterien gewährleisten, was zu konsistenteren Vorhersagen, Klassifikationen und Entscheidungen führen kann.

Nachteile

Diskriminierung: KI-Systeme können bestehende Vorurteile in den Trainingsdaten widerspiegeln, was zu diskriminierenden Entscheidungen führen kann.

Transparenzmangel: KI-Entscheidungen haben bisweilen einen Black-Box-Charakter: der Entscheidungsprozess ist für Menschen oft schwer nachvollziehbar.

Beeinträchtigung der Waffengleichheit: Der Einsatz von KI in strafrechtlichen Entscheidungen wirft Fragen bezüglich der Fairness, der Rechenschaftspflicht und des Rechts auf eine angemessene Verteidigung auf.

Unvollständigkeit: Die KI berücksichtigt nur die Informationen aus den Trainingsdaten.

Verstärkung menschlicher Fehler: KI kann die Neigung zur unhinterfragten Übernahme der von ihr produzierten Ergebnisse führen (automation bias). Sie kann unter bestimmten Umständen auch dazu führen, dass ihre Ergebnisse vollständig ignoriert werden (algorithm aversion).

2. Das Bayes-Theorem beruht auf folgendem Satz:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Was beschreibt das Bayes-Theorem? Bitte erläutern Sie den Satz!

Das Bayes-Theorem ist ein fundamentales Prinzip in der Wahrscheinlichkeitstheorie, benannt nach dem englischen Statistiker Thomas Bayes. Es beschreibt die Art und Weise, wie vorhandenes Wissen über die Wahrscheinlichkeit eines Ereignisses aktualisiert wird, wenn neue Informationen verfügbar werden.

Formal ausgedrückt, ermöglicht das Bayes-Theorem die Berechnung der bedingten Wahrscheinlichkeit eines Ereignisses A, gegeben ein Ereignis B, auf der Grundlage der ursprünglichen Wahrscheinlichkeit von A, der Wahrscheinlichkeit von B gegeben A und der Wahrscheinlichkeit von B.

$P(A)$ ist die ursprüngliche Wahrscheinlichkeit von A (Prior bzw. Vorwahrscheinlichkeit), also die anfängliche Einschätzung der Wahrscheinlichkeit eines Ereignisses, bevor neue Beweise oder Daten berücksichtigt werden. Sie repräsentiert unser bestehendes Wissen oder unsere Annahmen über die Wahrscheinlichkeit des Ereignisses, basierend auf früheren Erfahrungen oder allgemeinem Wissen. In der Medizin kann der Prior etwa die Wahrscheinlichkeit sein, mit der eine Person eine bestimmte Krankheit hat, basierend auf der Häufigkeit der Krankheit in der Gesamtbevölkerung.

P(B) ist die Gesamtwahrscheinlichkeit von B (Marginalwahrscheinlichkeit). Sie berücksichtigt alle möglichen Wege, auf denen das Ereignis B eintreten könnte.

$P(A|B)$ ist die Wahrscheinlichkeit von A, gegeben B (Posterior bzw. Nachwahrscheinlichkeit), also die Wahrscheinlichkeit des Ereignisses A, nachdem die neuen Daten B berücksichtigt wurden.

$P(B|A)$ ist die Wahrscheinlichkeit von B, gegeben A (Likelihood), also die Wahrscheinlichkeit, die aktuellen Daten oder Beobachtungen zu erhalten, vorausgesetzt, das Ereignis (oder die Hypothese) A ist wahr. Sie quantifiziert, wie gut die neuen Daten mit einer bestimmten Hypothese übereinstimmen. In der Medizin kann die Likelihood die Wahrscheinlichkeit beschreiben, ein positives Testergebnis zu erhalten, wenn die Person tatsächlich erkrankt ist. (N.B.: So ausführliche Erörterungen wurden hier nicht erwartet. Die meisten Prüflinge haben hier ein gute Punktzahl erreicht.)

3. Am Paradeplatz wurde ein gelbes Velo gestohlen. Eine Kameraaufnahme hat unzweifelhaft ergeben, dass T am Tag des Diebstahls am Paradeplatz war. 10 von 100 gestohlenen Velos sind laut amtlicher Velostatistik gelb. T hält sich jede Woche mit einer Wahrscheinlichkeit von 2 von 100 am Paradeplatz auf. Ein Legal Tech-Tool hat über eine Vorhersage berechnet, dass T sich mit einer Wahrscheinlichkeit von 8 von 100 am Paradeplatz aufhält, wenn gelbe Velos gestohlen werden.

Wie wahrscheinlich ist es nach dem Bayes-Theorem, dass ein gelbes Velo gestohlen wurde, als sich T am Paradeplatz aufhielt?

Um die Wahrscheinlichkeit zu berechnen, dass ein gelbes Velo gestohlen wurde, als sich T am Paradeplatz aufhielt, können wir das Bayes-Theorem anwenden. Die gegebenen Informationen sind:

P(Gelbes Velo gestohlen) = 10 von 100 = 0,10 (Wahrscheinlichkeit, dass ein gestohlenes Velo gelb ist).

P(T am Paradeplatz | Gelbes Velo gestohlen) = 8 von 100 = 0,08 (Wahrscheinlichkeit, dass T sich am Paradeplatz aufhält, wenn ein gelbes Velo gestohlen wird).

P(T am Paradeplatz) = 2 von 100 = 0,02 (Wahrscheinlichkeit, dass sich T jede Woche am Paradeplatz aufhält).

Wir suchen P(Gelbes Velo gestohlen | T am Paradeplatz), also die Wahrscheinlichkeit, dass ein gelbes Velo gestohlen wurde, gegeben, dass sich T am Paradeplatz aufhält. Nach dem Bayes-Theorem ist diese Wahrscheinlichkeit gegeben durch:

P(Gelbes Velo gestohlen | T am Paradeplatz) = P(Gelbes Velo gestohlen) x *P(T am Paradeplatz | Gelbes Velo gestohlen) / P(T am Paradeplatz)*

Wenn wir diese Werte in die Bayes-Formel einsetzen, ergibt sich:

P(Gelbes Velo gestohlen | T am Paradeplatz) = 0,1×0,0.08/0.02 = 0.4 = 40 %

Die Wahrscheinlichkeit, dass ein gelbes Velo gestohlen wurde, als sich T am Paradeplatz aufhielt, beträgt also 40 %.

Wie verlässlich ist der Rückschluss, dass es sich bei T um den Täter dieses Diebstahls handelt?

Die Schlussfolgerung, dass T der Täter des Diebstahls ist, basierend allein auf der berechneten Wahrscheinlichkeit, ist nicht unbedingt verlässlich. Das Bayes-Theorem liefert nur eine Wahrscheinlichkeit dafür, dass ein gelbes Velo gestohlen wurde, während T am Paradeplatz war. Es sagt aber nichts Spezifisches über Ts Beteiligung am Diebstahl aus.

Erstens bedeutet die Tatsache, dass T am Tatort war, als ein gelbes Velo gestohlen wurde, nicht notwendigerweise, dass T der Dieb ist. Korrelation impliziert nicht Kausalität.

Zweitens hängt die Zuverlässigkeit des Rückschlusses stark von anderen verfügbaren Beweisen abhängen, wie Augenzeugenberichten, weiteren Kameraaufnahmen, physischen Beweisen oder Ts Alibi.

Drittens bezieht sich berechnete Wahrscheinlichkeit auf die Häufigkeit von Ereignissen in einer grösseren Population oder in einer Reihe von Situationen und ist nicht unbedingt auf den spezifischen Einzelfall anwendbar. Selbst wenn die Wahrscheinlichkeit, dass ein gelbes Velo

gestohlen wird, während T am Paradeplatz ist, relativ hoch ist, schliesst dies nicht aus, dass andere Personen ebenfalls in der Lage wären, den Diebstahl zu begehen.

Viertens sind strafprozessual unterschiedliche Wahrscheinlichkeitsmassstäbe einschlägig. Für die Einleitung eines polizeilichen Ermittlungsverfahrens ist ein Anfangsverdacht ausreichend (Art. 306 I StPO). Hierunter sind zureichende tatsächliche Anhaltspunkte zu verstehen, also die durch Tatsachen substantiierbare Möglichkeit einer Täterschaft. Hiernach dürfte die Wahrscheinlichkeit von 40 % ausreichen, um ein polizeiliches Ermittlungsverfahren einzuleiten. Für ein staatsanwaltliches Untersuchungsverfahren (Art. 309 I a. StPO) und die Erhebung einer Anklage (Art. 324 I StPO) ist hingegen ein hinreichender Tatverdacht erforderlich. Dabei handelt es sich um die überwiegende Wahrscheinlichkeit (>50 %), dass dem Täter nach dem bisherigen Stand der Ermittlungen die Tat nachgewiesen werden kann und er auch verurteilt werden kann. Hierfür würden die bisherigen Tatsachen und Anhaltspunkte nicht ausreichen. (N.B.: In der zweiten Teilaufgabe lag der Schwerpunkt. Das haben nur die allerwenigsten Prüflinge erkannt. Die Nennung der Rechtsnormen wurde hier nicht erwartet. Es ging um die Erörterung unterschiedlicher Wahrscheinlichkeitsmassstäbe und (kreative) Argumente. Daran mangelte es in den allermeisten Arbeiten.)

4. Der Betreiber des Letzigrund-Stadions möchte digitale Tickets für ein Konzert von Saylor Twift verkaufen. Im Rahmen der Sonderaktion «Soccer for Saylor» möchte der Betreiber jedem Käufer automatisch die Hälfte des Ticketpreises erstatten, wenn der FC Zürich sein Heimspiel am Vorabend des Konzerts gewinnt.

Mit welchem Vertragskonstrukt könnte der Stadionbetreiber dieses Ziel erreichen?

Die Kunden sollen hier einen Standardkaufvertrag für digitale Tickets für das Konzert von Saylor Twift abschliessen. Der volle Ticketpreis soll im Voraus bezahlt werden. In den Vertrag muss eine Klausel aufgenommen werden, die den Käufer dazu berechtigt, die Hälfte des Ticketpreises erstattungsweise zu verlangen, falls der FC Zürich sein Heimspiel am Vorabend des Konzerts gewinnt (etwa durch eine auflösende Bedingung). Diese Klausel definiert klar die Bedingungen, unter denen die Rückerstattung erfolgt, und die Methode der Rückerstattung. Die Bedingungen der Aktion müssen dabei klar und deutlich festgehalten werden, um Missverständnisse zu vermeiden. Dies sollte sowohl im Kaufprozess als auch in der Werbung für die Aktion erfolgen. Dazu sollte ein effizientes und zuverlässiges System zur Verarbeitung der Rückerstattungen eingerichtet werden.

Um den Prozess zu automatisieren und die Zuverlässigkeit zu erhöhen, könnte der Betreiber erwägen, einen Smart Contract (SC) auf einer Blockchain-Plattform zu verwenden. Dieser SC würde automatisch die Hälfte des Ticketpreises zurückerstatten, wenn die Bedingung (Sieg des FC Zürich) erfüllt ist. Dies eliminiert die Notwendigkeit für Vermittler und kann die Vertragsdurchführung beschleunigen und vereinfachen. Da SC die Notwendigkeit von Vermittlern oder traditionellen Durchsetzungsmechanismen reduzieren, könnte auf ihrer Grundlage die Rückerstattung kostengünstiger und schneller durchgeführt werden.

Was wäre dabei besonders zu beachten?

Der SC würde typischerweise in einer für Blockchain-Plattformen geeigneten Programmiersprache wie Solidity für die Ethereum-Blockchain geschrieben. Der SC hätte eine

Funktion zum Verkauf der Tickets. Bei jedem Kauf würde der Vertrag die Details des Käufers (z.B. die Wallet-Adresse) und die bezahlte Summe speichern. Der SC müsste eine Funktion enthalten, die die Bedingung für die Rückerstattung überprüft – in diesem Fall, ob der FC Zürich sein Heimspiel gewonnen hat. Dies könnte durch eine externe Datenquelle (ein sog. Orakel) sichergestellt werden, der das Ergebnis des Spiels verifiziert und der Blockchain-Plattform meldet. Wenn die Bedingung erfüllt ist (FC Zürich gewinnt), würde der SC automatisch die Hälfte des Ticketpreises an die Wallet-Adressen der Käufer zurückerstatten. Der SC sollte auch Sicherheitsmechanismen vorsehen, um Missbrauch zu verhindern und eine korrekte Ausführung zu gewährleisten.

Wie kommt ein gültiger Vertrag zustande, wenn der Betreiber die Bedingungen der Sonderaktion gegenüber einem unbestimmten Personenkreis anbietet?

(N.B.: Hier ging es darum, zu erklären, worin genau im konkreten Fall das Angebot und die Annahme liegen könnten. Das haben viele Prüflinge nicht erkannt.) Ein Vertrag kommt durch zwei übereinstimmende Willenserklärungen zustande, Angebot und Annahme. Dabei müssen die essentialia negotii hinreichend bestimmt in den Willenserklärungen zum Ausdruck kommen. Ein gültiger Vertrag in einer Situation, in der der Betreiber des Stadions die Bedingungen der Sonderaktion gegenüber einem unbestimmten Personenkreis anbietet, kann auf unterschiedlichen Wegen zustande kommen.

Invitatio ad offerendum: Der Betreiber gibt die Bedingungen der Sonderaktion bekannt, was rechtlich als Einladung an die Öffentlichkeit verstanden wird, ein Angebot zum Kauf von Tickets unter diesen Bedingungen zu machen. Diese Bekanntmachung stellt noch kein verbindliches Angebot dar, sondern lädt potenzielle Käufer ein, Angebote abzugeben.

Ein Kunde, der ein Ticket kaufen möchte, gibt ein Angebot ab, indem er das Ticket gemäss den vom Betreiber festgelegten Bedingungen bestellt. Dies kann beispielsweise durch Auswahl eines Tickets auf der Website des Stadions und die Durchführung des Bezahlvorgangs geschehen.

Der Betreiber nimmt das Angebot an, wenn er den Kauf bestätigt und das Ticket ausstellt. Die Annahme kann automatisiert erfolgen, zum Beispiel durch eine Bestätigungs-E-Mail oder die Ausstellung eines digitalen Tickets. Die Bedingungen der Sonderaktion (z.B. die Rückerstattung der Hälfte des Ticketpreises bei einem Sieg des FC Zürich) sind Teil des Vertrags und müssen von beiden Parteien erfüllt werden.

Offerta ad incertas personas: Denkbar wäre auch, die Bewerbung der Tickets als offerta ad incertas personas zu qualifizieren. Dabei handelt es sich um ein rechtsverbindliches Angebot an die Kunden, das diese durch die Bestellung der Tickets annehmen könnten. Das Angebot müsste dann unter die Bedingung gestellt werden, dass noch Tickets vorrätig sind. Dabei handelt es sich rechtlich um eine Einschränkung des Angebots bzw. einen vertragsrechtlich zulässigen Vorbehalt (Privatautonomie!).

Um rechtliche Probleme zu vermeiden, sollten die Bedingungen der Sonderaktion klar und eindeutig formuliert sein, damit keine Missverständnisse über die Verpflichtungen des Betreibers und der Kunden entstehen.

5. Eine Richterin möchte wissen, ob Autofahrer langsamer fahren, wenn höhere Geldbussen verhängt werden. In Zusammenarbeit mit der Legal Tech-Spezialistin des Gerichts möchte sie folgendes Modell schätzen:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Was beschreibt dieses Modell allgemein?

Es handelt sich hierbei um die grundlegende Darstellung eines linearen Regressionsmodells. In diesem Modell wird angenommen, dass eine abhängige Variable (y_i) linear von einer unabhängigen Variablen (x_i) abhängt, zuzüglich eines Fehlerterms (ϵ_i). Dieses Modell wird häufig in der Statistik und Ökonometrie verwendet, um Beziehungen zwischen Variablen zu analysieren und Vorhersagen zu treffen.

y_i: Dies ist die abhängige oder zu erklärende Variable. Sie repräsentiert den Wert, den wir zu modellieren oder vorherzusagen versuchen, und kann für jede Beobachtung i unterschiedlich sein.

β_0: Dies ist der Achsenabschnitt der Regressionsgeraden. Es repräsentiert den Wert von yi, wenn $x_i = 0 x_i = 0$ ist. Mit anderen Worten, es ist der erwartete Wert von y_i, wenn die unabhängige Variable nicht vorhanden ist oder den Wert Null annimmt.

x_i: Dies ist die unabhängige oder erklärende Variable. Sie ist die Variable, von der angenommen wird, dass sie einen Einfluss y_i hat.

β_1: Dies ist der Steigungskoeffizient. Er gibt an, wie sich die abhängige Variable y_i ändert, wenn sich die unabhängige Variable x_i um eine Einheit ändert. Er beschreibt die Stärke und Richtung des Zusammenhangs zwischen $x_i$und y_i.

ϵ_i: Dies ist der Fehlerterm für jede Beobachtung i. Er repräsentiert alle anderen Faktoren, die yi beeinflussen könnten, aber nicht im Modell enthalten sind. Der Fehlerterm wird als zufällig und normalverteilt mit einem Mittelwert von Null angenommen.

Insgesamt beschreibt dieses lineare Regressionsmodell, wie die abhängige Variable durch eine unabhängige Variable linear beeinflusst wird, unter Berücksichtigung zufälliger Fehler oder Störungen, die nicht durch das Modell erklärt werden können. (N.B.: So ausführliche Erörterungen wurden hier nicht erwartet.)

Kann es eingesetzt werden, um den Zusammenhang zwischen Geldbussen und Geschwindigkeit zu schätzen?

Ja, das lineare Regressionsmodell kann verwendet werden, um den Zusammenhang zwischen Geldbussen und Geschwindigkeit zu schätzen. Die Richterin ist hier am Effekt von Geldbussen auf die Geschwindigkeit interessiert. In einem linearen Regressionsmodell ist die Überschreitung der zulässigen Höchstgeschwindigkeit als abhängige und die Höhe der Geldbusse als unabhängige Variable zu behandeln:

Geschwindigkeit(-süberschreitung) = β_0 + β_1 × Geldbusse + ϵ_i

Die abhängige Variable y_i misst hier, wie schnell jemand fährt bzw. um wie viel die Geschwindigkeitsbegrenzung überschritten wurde. Die unabhängige Variable x_i beschreibt die Höhe der Geldbusse. β_0 ist der Achsenabschnitt, der die Basisgeschwindigkeitsüberschreitung angibt, wenn keine Geldbusse verhängt wird. β_1 ist der Steigungskoeffizient, der angibt, wie sich die Geschwindigkeit ändert, wenn sich die Höhe der Geldbusse ändert. ϵ_i ist der Fehlerterm und erfasst andere Faktoren, die sowohl die Geschwindigkeitsüberschreitung als auch die Höhe der Geldbusse beeinflussen könnten, aber nicht im Modell berücksichtigt werden. In diesem Modell würden man untersuchen, ob höhere Geldbussen tatsächlich mit geringeren Geschwindigkeiten korrelieren.

(N.B.: Viele Studierende haben die Höhe der Geldbusse zur abhängigen und die Geschwindigkeitsüberschreitung zur unabhängigen Variable gemacht:

Geldbusse = $\beta_0 + \beta_1 \times$ Geschwindigkeit(-süberschreitung) + ϵ_i

Die abhängige Variable beschreibt hier die Höhe der Geldbusse für Verkehrssünder, während die unabhängige Variable die gemessene Überschreitung der Geschwindigkeitsbegrenzung beschreibt. Ein positiver Wert von β_1 würde bedeuten, dass höhere Geschwindigkeitsüberschreitungen zu höheren Bussen führen.

Dieser Ansatz ist hier aber aus zwei Gründen unpassend. Erstens war gefragt, «ob Autofahrer langsamer fahren, wenn höhere Geldbussen verhängt» werden. Die Richterin interessiert sich dafür, ob Geldbussen wie erwartet wirken, nicht aber dafür, wann Geldbussen verhängt werden. Zweitens ist die letztgenannte Frage empirisch auch nicht besonders interessant. Denn es ergibt sich bereits aus gesetzlichen und administrativen Vorgaben, welche Geldbussen bei bestimmten Überschreitungen der zulässigen Höchstgeschwindigkeit verhängt werden. Wenn die Verhängung von Bussgeldern hier, wie praktisch meist der Fall, automatisiert erfolgt, würde man zwingend einen mechanischen, positiven Effekt von Geschwindigkeitsüberschreitungen auf Bussgelder erwarten. Daraus würde man nur lernen, dass die rechtlichen und administrativen (Buss-geld-)Vorgaben von den zuständigen Behörden befolgt werden. Interessanter ist aber – auch für die Richterin hier –, ob sich auch die Autofahrer an diese Vorgaben halten, sich also von Geldbussen abschrecken lassen.)

Muss sich die Richterin Sorgen machen, wenn der Koeffizient β_1 negativ ist?

Nein. Ein negativer Koeffizient β_1 würde bedeuten, dass höhere Geldbussen mit geringeren Überschreitungen der Geschwindigkeitsbegrenzung verbunden sind. Dies würde auf eine abschreckende Wirkung der Geldbusse hindeuten. Genau dies entspricht dem Zweck von Geldbussen bzw. dem Telos der Ermächtigungsgrundlagen zum Erlass von Geldbussen.

Angenommen der p-Wert von β_1 ist geringer als 0.05: Was bedeutet das?

Der p-Wert (oder Wahrscheinlichkeitswert) ist ein Begriff, der in der Statistik verwendet wird, um die Stärke der Evidenz gegen eine Nullhypothese zu bewerten. Er misst, wie wahrscheinlich es ist, die beobachteten Daten (oder etwas noch Extremeres) zu beobachten, vorausgesetzt, dass die Nullhypothese wahr ist.

Nullhypothese (H_0): Dies ist die Ausgangshypothese, die normalerweise besagt, dass kein Effekt oder kein Unterschied besteht (z.B., dass es keinen Unterschied in der Wirkung zweier Behandlungen gibt).

Der p-Wert wird durch statistische Tests berechnet, die auf den gesammelten Daten basieren. Er ergibt sich aus der Wahrscheinlichkeitsverteilung unter der Annahme, dass die Nullhypothese wahr ist.

Ein kleiner p-Wert (typischerweise kleiner als ein vorher festgelegtes Signifikanzniveau (alpha-level, wie 0,05) deutet darauf hin, dass die beobachteten Daten sehr unwahrscheinlich sind, unter der Annahme, dass die Nullhypothese wahr ist. In diesem Fall lehnen Forscher oft die Nullhypothese ab und schliessen, dass der Effekt statistisch signifikant ist.

Ein grosser p-Wert deutet darauf hin, dass die beobachteten Daten mit der Nullhypothese vereinbar sind. In diesem Fall gibt es nicht genügend Beweise, um die Nullhypothese abzulehnen.

Ein häufiges Missverständnis (N.B.: auch in der Prüfung) ist, dass der p-Wert die Wahrscheinlichkeit misst, dass die Nullhypothese wahr ist. Tatsächlich gibt er nur an, wie gut die Daten mit der Nullhypothese übereinstimmen. Er sagt nichts darüber aus, ob die alternative Hypothese wahr ist oder wie gross der beobachtete Effekt ist.

Was kann die Richterin hieraus über den Kausalzusammenhang zwischen Geldbussen und Geschwindigkeit lernen?

Aus dem p-Wert selbst lässt sich kein Rückschluss auf den Kausalzusammenhang zwischen Geldbusse und Geschwindigkeitsüberschreitungen ableiten. Der p-Wert gibt lediglich an, wie wahrscheinlich es ist, die beobachteten Daten zu erhalten, wenn die Nullhypothese (es besteht kein Effekt) wahr ist. Er misst die Übereinstimmung zwischen den Daten und der Nullhypothese, nicht aber die kausale Beziehung zwischen Variablen.

Ein statistisch signifikantes Ergebnis (ein kleiner p-Wert) kann auf eine Korrelation oder Assoziation zwischen Variablen hinweisen. Es sagt aber nichts darüber aus, ob eine Variable die andere verursacht. Zum Beispiel könnte es sein, dass Autofahrer nicht langsamer fahren, weil Geldbussen verhängt werden, sondern weil an besonders gefährlichen Stellen häufiger auf Geschwindigkeitsbegrenzungen hingewiesen wird. Selbst wenn eine starke statistische Assoziation vorliegt, können also andere, nicht berücksichtigte Variablen (omitted variable bias) für den beobachteten Effekt verantwortlich sein.

Um Kausalität festzustellen, sind oft experimentelle oder quasi-experimentelle Studiendesigns erforderlich, in denen Forscher aktiv eine Variable manipulieren und andere Faktoren kontrollieren. Selbst in solchen Designs muss man sorgfältig auf das Design, die Durchführung und die Analyse achten, um kausale Schlüsse zu rechtfertigen.

Kann die Richterin aus dem Modell eine Vorhersage ableiten?

Ja, aus dem genannten linearen Regressionsmodell können Vorhersagen abgeleitet werden. Das Modell ermöglicht es, für einen gegebenen Wert von Geldbussen einen geschätzten Wert der gefahrenen Geschwindigkeit zu berechnen.

Zunächst müssen die Koeffizienten β_0 und β_1 basierend auf den vorhandenen Daten geschätzt werden. Dies erfolgt üblicherweise durch eine Methode wie Ordinary Least Squares (OLS). Danach setzen wir den Wert von x, für den wir eine Vorhersage treffen möchten, in das Modell ein. Wir berechnen den Wert von y mit der Formel $\hat{y}=\beta_0+\beta_1 x$. Dieser Wert $\hat{y}$ ist die Vorhersage der abhängigen Variablen basierend auf dem Wert der unabhängigen Variablen x.

Es ist wichtig, sich bewusst zu sein, dass es eine gewisse Unsicherheit gibt, insbesondere durch den Fehlerterm ϵ_i im Modell. Die Genauigkeit der Vorhersagen hängt auch davon ab, wie gut das Modell die tatsächlichen Beziehungen zwischen den Variablen widerspiegelt und ob die Annahmen der linearen Regression (z.B. Linearität, Unabhängigkeit der Fehler) erfüllt sind.

6. Sie möchten ein Machine Learning-basiertes Legal Tech-Tool in der Schweiz und in der EU in Verkehr bringen. Ihr Tool soll auch von der öffentlichen Verwaltung rechtssicher genutzt werden können. Dafür müssen Sie sicherstellen, dass die von Ihrem Tool empfohlenen Entscheidungen transparent und begründbar sind.

Aus welchen europäischen und nationalen Vorschriften ergeben sich (in Zukunft) Transparenz- und Begründungspflichten?

Art. 21 I DSG
Art. 21 II DSG
Art. 22 III DSGVO
Art. 13 EU KI-Gesetz
Art. 52 EU KI-Gesetz

Worin unterscheiden sich diese Vorschriften?

Diese Vorschriften unterscheiden sich in ihrem territorialen und sachlichen Geltungsbereich. Das DSG gilt grundsätzlich nur für Sachverhalte, die sich in der Schweiz auswirken, auch wenn sie im Ausland veranlasst werden (Art. 3 I DSG).

Die DSGVO gilt hingegen die Verarbeitung personenbezogener Daten, soweit diese im Rahmen der Tätigkeiten einer Niederlassung in der EU erfolgt; Personen in der EU Waren oder Dienstleistungen angeboten werden; das Verhalten von Personen in der EU beobachtet wird; oder der Verantwortliche kraft Völkerrechts an das Recht eines EU-Mitgliedstaates gebunden ist (Art. 3 DSGVO).

(N.B. Hier hatten die Studierenden viele Freiräume. Ebenso zulässig war es, einige konkrete Unterschiede des Geltungsbereichs bzw. der Tatbestandsvoraussetzungen zu erörtern. Diese Erörterungen mussten präzise sein und sich an den konkreten Rechtsnormen orientieren. Punktabzüge mussten für allgemeine Aussagen hingenommen werden. Bsp.: «Das EU KI-Gesetz ist eher Regulierung, das DSG gewährleistet Datenschutz.»)

Was macht gute Erklärungen oder Begründungen von Machine Learning-basierten Entscheidungen im Allgemeinen aus?

Verständlichkeit: Gute Erklärungen sollten in einer klaren und verständlichen Sprache formuliert sein, angepasst an das Wissen und die Erfahrung des jeweiligen Publikums. Fachjargon sollte vermieden oder erklärt werden.

Kontraststärke: Es sollte erkennbar sein, welche Variablen einen Einfluss relativ zu anderen Variablen haben und wie sich unterschiedliche Variablen im Zusammenspiel auswirken.

Relevanz: Gute Erklärungen fokussieren auf die Aspekte des Modells, die für die spezifische Entscheidung oder Vorhersage am relevantesten sind.

Genauigkeit: Gute Erklärungen sollte die Funktionsweise des ML-Modells genau widerspiegeln, ohne wichtige Details zu übersehen oder zu vereinfachen, die das Verständnis der Entscheidungsfindung beeinflussen könnten.

Vollständigkeit: Gute Erklärungen berücksichtigen alle relevanten Faktoren, die zur Entscheidung beigetragen haben, einschliesslich der Daten, die in das Modell eingegeben wurden, und wie das Modell diese Daten verarbeitet hat. Sie müssen aber der Gefahr von Informationsüberlastung (information overload) entgegenwirken.

Aktionsorientierung: Insbesondere im Kontext privatrechtlicher Anwendungen sollten Erklärungen so gestaltet sein, dass sie den Nutzern helfen zu verstehen, wie sie ihr Verhalten anpassen können, um andere Ergebnisse zu erzielen.

Situationsbezogenheit: Die Erklärung sollte im Kontext der spezifischen Anwendung und des Zwecks des ML-Modells erfolgen.

Transparenz bezüglich Unsicherheiten und Grenzen: Gute Erklärungen sollten auch die Grenzen des Modells und eventuelle Unsicherheiten in den Vorhersagen offenlegen.

Was sind kontrafaktische Begründungen?

Kontrafaktische Begründungen im ML-Kontext beziehen sich auf die Erstellung und Analyse von "Was-wäre-wenn"-Szenarien, um zu verstehen, wie unterschiedliche Eingaben die Vorhersagen oder Entscheidungen eines ML-Modells beeinflussen. In der ML-Praxis helfen kontrafaktische Erklärungen dabei, die Black-Box-Natur komplexer Modelle wie neuronaler Netze transparenter zu machen, indem sie Einblicke in die Entscheidungsfindung des Modells bieten.

Kontrafaktische Begründungen setzen die Generierung von Eingabedaten voraus, die minimal von den ursprünglichen Daten abweichen, aber zu einer signifikant anderen Vorhersage oder Entscheidung durch das ML-Modell führen. Solche Beispiele helfen zu verstehen, welche Merkmale oder Variablen für das Modell am wichtigsten sind.

Kontrafaktische Erklärungen können verwendet werden, um die Funktionsweise und Entscheidungsfindung von ML-Modellen zu erläutern. Sie bieten eine menschenverständliche Erklärung dafür, wie eine Veränderung der Eingabedaten zu einer anderen Ausgabe oder Entscheidung des Modells führen könnte.

Worin bestehen ihre Vor- und Nachteile?

Vorteile

Intuitive Nachvollziehbarkeit treibender Variablen: Sie helfen dabei, komplexe Modelle verständlicher zu machen, indem sie aufzeigen, wie Veränderungen in den Eingabedaten die Vorhersagen des Modells beeinflussen.

Verbesserung der Modellrobustheit: Durch das Untersuchen kontrafaktischer Szenarien können Entwickler von ML-Modellen Schwachstellen und Bias in ihren Modellen identifizieren und korrigieren, was zu faireren und robusteren Systemen führt.

Unterstützung bei Entscheidungsfindung: In Anwendungen wie Kreditwürdigkeitsprüfungen oder bei medizinischen Diagnosewerkzeugen können kontrafaktische Erklärungen Endnutzern helfen zu verstehen, welche Faktoren geändert werden müssten, um ein anderes Ergebnis zu erzielen (z.B. welche Veränderungen nötig wären, um eine Kreditgenehmigung zu erhalten).

Aufdecken von Biases: Sie ermöglichen es, Verzerrungen (bspw. Diskriminierung) in ML-Modellen zu identifizieren, indem sie aufzeigen, wie bestimmte Eingaben systematisch zu unterschiedlichen Ergebnissen führen.

Entwicklung robusterer Modelle: Durch die Analyse, wie geringfügige Änderungen der Eingabedaten die Vorhersagen beeinflussen, können Entwickler die Robustheit ihrer Modelle verbessern.

Förderung von Vertrauen und Akzeptanz: Kontrafaktische Erklärungen können das Vertrauen in und die Akzeptanz von ML-Modellen bei Anwendern erhöhen, indem sie Transparenz schaffen.

Nachteile

Rechenintensität: Das Generieren kontrafaktischer Erklärungen kann rechenintensiv sein, besonders bei komplexen Modellen wie tiefen neuronalen Netzwerken.

Manchmal eingeschränkte Plausibilität: Die Nützlichkeit kontrafaktischer Erklärungen hängt stark von ihrer Relevanz und Plausibilität ab. Unrealistische oder irrelevante kontrafaktische Beispiele können mehr Verwirrung als Klarheit schaffen.

Nur ein Ausschnitt der treibenden Variablen: Kontrafaktische Erklärungen betrachten nur einzelne Szenarien und geben möglicherweise kein vollständiges Bild der Entscheidungsfindung des Modells.

Risiko von Fehlinterpretationen: Es besteht das Risiko, dass Anwender kontrafaktische Erklärungen falsch interpretieren, insbesondere wenn sie nicht über ausreichendes Hintergrundwissen verfügen.

Datenschutzrecht: Die Erstellung kontrafaktischer Beispiele kann datenschutzrechtliche Bedenken aufwerfen, besonders wenn sie sensible oder persönliche Daten betreffen.

(N.B.: So ausführliche Erörterungen wurden hier nicht erwartet. Allerdings waren hier durchaus (kreative) Argumente gefordert. Insgesamt führten die allerwenigsten Prüflinge hier mehr als zwei oder drei Stichworte ins Feld.)