Johner Institut

# Interpretierbarkeit von Machine Learning von Modellen

Mit Christoph Molnar, Prof. Dr. Christian Johner

## Transkript

00:00:05 Sprecher 1

Medical Device Insights, ein Podcast des Jona Instituts für Medizinproduktehersteller, Behörden und benannte Stellen.

00:00:17 Sprecher 1

Immer mehr Medizinprodukte verwenden Technologien des maschinellen Lernens, also einer Untergruppe der künstlichen Intelligenz, und gleichzeitig sind auch die Regulatoren fleißig unterwegs.

00:00:29 Sprecher 1

beispielsweise hat die chinesische N.

00:00:31 Sprecher 1

N.

00:00:31 Sprecher 1

P.

00:00:31 Sprecher 1

A.

00:00:32 Sprecher 1

gerade einen Entwurf veröffentlicht, wo sie auch weitere Anforderungen an das Thema Machine Learning stellt.

00:00:39 Sprecher 1

Ich selber bin unterwegs in einem W.

00:00:41 Sprecher 1

H.

00:00:41 Sprecher 1

O.

00:00:42 Sprecher 1

Arbeitskomitee, das sich auch zum Ziel gesetzt hat, so eine Guideline mitzuentwickeln.

00:00:47 Sprecher 1

Und bei diesen Anforderungen, die wir in diesen Guidelines finden, taucht auch das Thema Interpretierbarkeit immer mehr auf.

00:00:56 Sprecher 1

und so hab ich den Christoph Molnar hier eingeladen gehabt, der sich in diesem Bereich seit vielen, vielen Jahren engagiert.

00:01:04 Sprecher 1

Hallo Christoph, möchtest du dich ganz kurz mal vorstellen, was du machst und wie du in diesem Bereich der Interpretierbarkeit und des maschinellen Lernens du dich in der Vergangenheit schon engagiert hast?

00:01:14 Sprecher 2

Ja, hallo Christian, danke für die Einladung.

00:01:16 Sprecher 2

Ja, ich beschäftige mich jetzt seit dreieinhalb Jahren mit dem Thema.

00:01:21 Sprecher 2

Angefangen habe ich bisschen ungewöhnlich, dass ich mit einem Buch angefangen habe, darüber zu schreiben.

00:01:26 Sprecher 2

Also, mein Hintergrund ist auch in Statistik.

00:01:27 Sprecher 2

Also, hatte ich schon sozusagen viel, auf was ich aufbauen konnte, auch und momentan mache ich ein P.H.D.

00:01:35 Sprecher 2

gerade zu dem Thema auch.

00:01:37 Sprecher 2

Also, ich forsche auch selber und ja, bin jetzt auch gerade dabei, mehr oder weniger fertig zu werden.

00:01:42 Sprecher 2

Man weiß ja dann hinten raus nie genau, wie lange es dauert.

00:01:45 Sprecher 2

Genau, aber beschäftige mich auch da sehr intensiv mit dem Thema Interpretierbarkeit im maschinellen Lernen.

00:01:50 Sprecher 2

und hab auch zum Beispiel schon Softwareprodukte, also ein ein Softwarepaket in in R geschrieben da-

für, mit dem man dann Modelle des maschinellen Lernens interpretieren kann.

00:02:00 Sprecher 1

Über dieses Buch, das du gerade erwähnt hast, bin ich auch auf dich gestoßen.

00:02:04 Sprecher 1

Also für alle Hörer, wir haben das unten in den Begleitmaterialien mit verlinkt.

00:02:09 Sprecher 1

Bevor wir da einsteigen, könntest du uns ganz kurz mal erklären, was Interpretierbarkeit jetzt hier genau bedeutet.

00:02:15 Sprecher 2

Ja, da kriegt man immer unterschiedliche

00:02:18 Sprecher 2

unterschiedliche Antworten, je nachdem wen man fragt.

00:02:21 Sprecher 2

So eine ganz allgemeine Antwort wäre vielleicht so, das ist so der Grad zu dem ein Mensch eine Maschine verstehen kann, also das nachvollziehen kann, wie eine Entscheidung getroffen wurde.

00:02:33 Sprecher 2

Aber das ist natürlich sehr, eine sehr weiche Erklärung, also eine sehr weiche Definition und das ist auch eine, ich würde mal sagen, der größeren Kriegpunkte noch an dem Feld oder auch der Schwierigkeiten,

00:02:45 Sprecher 2

dass das nicht immer einfach ist zu zeigen, dass jetzt eine Erklärung oder eine Interpretation auch die korrekte ist von einem Modell.

00:02:52 Sprecher 1

Mhm, damit bist du jetzt schon bei den Fallstricken angekommen.

00:02:55 Sprecher 1

Du hattest mir neulich auch mal ein wunderbares Paper mitgegeben gehabt, die haben die Interpretierbarkeit noch mal aus 2 Richtungen sich angesehen gehabt oder dessen 2 Teile zerlegt, nämlich einmal in den Bereich der Explainability, also dass man als

00:03:11 Sprecher 1

Anwender irgendwie erklärt bekommt, was da geschieht, also quasi was die Blackbox da macht und das andere die Transparency, wo man dann wirklich im Modell innen drin noch mal ein genaueres Verständnis sich vermittelt.

00:03:24 Sprecher 1

Aber diese Definition sind jetzt auch nicht überall angenommen geworden und deswegen haben wir wahrscheinlich in der Tat noch ‚ne Weichheit in diesen Definitionen.

00:03:33 Sprecher 1

Ja, jetzt hast du schon drüber so bisschen gesprochen gehabt, dass es da auch noch ein paar Fallstricke gibt, auch jenseits der Definition.

00:03:42 Sprecher 1

Ja, wo steht man denn da im maschinellen Lernen und seiner Interpretierbarkeit und welche Fallstricke siehst du da vor allem?

00:03:48 Sprecher 2

Genau, ja, da haben wir auch vor kurzem eben ein eigenes Paper dazu geschrieben, ganz kurzes, wo wir bisschen mal aufzählen, was gibt es denn so für Möglichkeiten, das ein Modell falsch zu interpretieren.

00:04:00 Sprecher 2

In dem Fall haben wir uns ein bisschen überlegt, so wenn jetzt ein ein Wissenschaftler eben dieses Modell, ein ein Maschinenmodell des maschinellen Lernen benutzt, um eigene Daten zu analysieren,

00:04:10 Sprecher 2

Aber es trifft natürlich auch darauf zu, dann wenn man ein Modell benutzt in einem Produkt und da gibt es eben viele Dinge, die man falsch machen kann bei der Interpretation, dann unter anderem auch.

00:04:21 Sprecher 2

Also es fängt an bei ganz einfachen Sachen, wie dass ein Modell, wenn man es schlecht trainiert, also also Overfitting vorliegt, dass dass man sozusagen sich zu stark anpasst auf die Trainingsdaten, aber dann im Endeffekt für neue Daten gar nicht die, die dann das Modell ganz schlecht nur vorhersagen kann.

00:04:40 Sprecher 2

Wenn man jetzt ein solches Modell interpretiert, dann bekommt man natürlich, was man da dann sozusagen rausinterpretiert aus dem Modell, also was jetzt die wichtigsten Features waren und wie sie die Vorhersage beeinflussen.

00:04:50 Sprecher 2

Das kann natürlich dann in die Irre führen, wie es in Wirklichkeit eigentlich sozusagen der Sachverhalt ist.

00:04:56 Sprecher 1

Versteh ich dich richtig?

00:04:57 Sprecher 1

Dann hätten wir eigentlich ein doppeltes Problem im Bereich Overfitting.

00:04:59 Sprecher 1

Wir hätten einmal ein Modell, das nachher in der echten Welt, mit den echten Daten im Feld, nicht so genau

00:05:07 Sprecher 1

arbeitet oder die Vorhersagen nicht so genau erstellt und uns in der in dem Glaube irreführt, wie das innen drin funktioniert.

00:05:15 Sprecher 1

Also wäre das sozusagen ein Doppelproblem.

00:05:17 Sprecher 2

Nicht ganz, weil das also sozusagen die Interpretation darüber oder wenn wir jetzt zum Beispiel einen uns ein Ranking ausgeben lassen, welches die wichtigsten Feature waren,

00:05:26 Sprecher 2

Dann das, was ganz oben steht, ist dann schon das, was für das Modell das Wichtigste war.

00:05:30 Sprecher 2

Was jetzt aber der sozusagen der Falschweg wäre, ist dass wir dann, weil wir wir denken jetzt fälschlicherweise, dass das Modell korrekt ist und auch die korrekten Features benutzt hat.

00:05:39 Sprecher 2

Aber weil wir overgefüttert haben, wär dann sozusagen das Wichtigste, nicht unbedingt das Wichtigste, in sozusagen in dem unterliegenden Phänomen, das wir untersuchen.

00:05:47 Sprecher 2

Also in dem Sinne, da da kommt auch der, kommt es auf den Fokus an, ob man nur das Modell als sich betrachtet, dann sozusagen.

00:05:55 Sprecher 2

wieder dieses Feature Ranking nehmen, das wichtigste Feature tatsächlich auch das Wichtigste gewesen, vielleicht für das Modell, aber eben nicht das Wichtigste in in der Realität.

00:06:03 Sprecher 1

Welche von den Modellen, vielleicht wenn wir noch mal ‚n Schritt zurückgehen, würdest du als besonders leicht und welche als besonders schwierig betrachten bezüglich der Interpretierbarkeit.

00:06:13 Sprecher 2

Da gibt es natürlich auch Unterschiede, da also ganz generell wird oft so ‚ne, dass die einfachen linearen Regressionsmodelle als

00:06:22 Sprecher 2

einfach interpretierbar betrachtet oder auch Entscheidungsbäume.

00:06:25 Sprecher 2

Also lineare Regressionsmodelle ist ja eigentlich immer, dann multipliziert man eigentlich die die Inputdaten und summiert das auf und bekommt dann eine Vorhersage.

00:06:36 Sprecher 2

Oder auch in den Entscheidungsbäumen kann man ja sich den Baum aufzeichnen und sich so dann erklären, sozusagen, oder sogar visuell dann anschauen, wie die Entscheidung zustande kommt, indem

man sozusagen dem Pfad folgt in dem Baum.

00:06:47 Sprecher 2

bis zu der Entscheidung, zum Beispiel bis zur Klassifikation.

00:06:50 Sprecher 2

Und im Linearmodell kann man eben diese Gewichte, die das Modell schätzt, interpretieren und dann auch sehen, wie sie die Vorhersage beeinflussen.

00:07:01 Sprecher 2

Was dann und je komplexer natürlich diese interne Struktur wird von einem Modell, also wir denken zum Beispiel an neuronales Netzwerk, da haben wir eigentlich ja auch solche Summen, aber da haben wir halt ganz, ganz viele, also sozusagen mathematische Multiplikationen auch.

00:07:15 Sprecher 2

und Summen, die dann hintereinander geschalten sind in den sogenannten Layern.

00:07:19 Sprecher 2

Klar, die einzelnen Operationen sind einfach zu verstehen.

00:07:21 Sprecher 2

Also es ist auch sehr transparent, wenn man, wenn ich dir jetzt alle Gewichte geben würde und die Architektur von dem von dem neuronalen Netzwerk.

00:07:28 Sprecher 2

Allerdings ist es nicht mehr nachvollziehbar, wie die Entscheidung zustande kommt am Ende dann von dem Modell.

00:07:34 Sprecher 1

Mhm, und deswegen brauchen wir eben.

00:07:37 Sprecher 1

Verfahren, die uns helfen, es trotzdem zu verstehen, auch wenn wir jetzt nicht in der Lage sind, irgendwie 100000 oder gar Millionen von Gewichten zu verstehen oder deren Auswirkung nachher auf die Entscheidung des Modells.

00:07:49 Sprecher 1

Jetzt werden wir noch mal kurz zurückkehren zu der Frage, was kann jetzt da schief gehen bei der Interpretation und du hast gerade angesprochen gehabt, ihr habt auch ein Paper dazu veröffentlicht, das wir übrigens auch unten verlinkt haben.

00:08:02 Sprecher 1

Könntest du uns da noch ein paar plastische Beispiele geben, wie

00:08:06 Sprecher 1

man glaubt, etwas verstanden zu haben, aber es vielleicht dann doch nicht richtig verstanden hat.

00:08:11 Sprecher 2

Ein großes Thema ist sozusagen die kausale Interpretation.

00:08:15 Sprecher 2

Also, wenn man jetzt ein Modell hat, es macht bestimmte Vorhersagen, dann möchte man ja auch gern, wenn man sich anguckt, O.

00:08:24 Sprecher 2

K., wie, wie, wie beeinflusst jetzt ein bestimmter Input Feature meine Vorhersage, möchte man das ja vielleicht auch kausal interpretieren, ob das, das der auch in der echten Welt einen kausalen Zusammenhang hatte,

00:08:36 Sprecher 2

mit mit dem, was rauskommen soll.

00:08:38 Sprecher 2

Das darf man aber nicht immer machen, automatisch.

00:08:41 Sprecher 2

Also es müssen ganz gewisse Annahmen erfüllt sein.

00:08:43 Sprecher 2

Also ein ganz einfaches Beispiel wäre, wenn ich jetzt ein Modell mache oder mir baue, das mir vorhersagt, ob es morgen regnen wird.

00:08:50 Sprecher 2

Und ein gutes Feature könnte jetzt tatsächlich sein, dass ich mir angucke, ist heute der Boden nass?

00:08:57 Sprecher 2

Weil wenn der Boden heute nass ist, dann weiß ich, dann wird es morgen wahrscheinlich regnen.

00:09:01 Sprecher 2

Also das ist jetzt auch ein.

00:09:03 Sprecher 2

Beispiel, wo sozusagen ein Konflikt entstehen kann zwischen finden ein Feature, ein ein ein Feature ist sozusagen nützlich für eine Vorhersage, aber sobald wir es ins Modell aufnehmen, dürfen wir es nicht mehr kausal interpretieren.

00:09:15 Sprecher 2

Also das Modell würde lernen, Boden ist heute nass, morgen wird es regnen.

00:09:19 Sprecher 2

Das dürfen wir aber nicht kausal interpretieren, weil wenn wir das machen würden, dann würde es bedeuten, wir können Wasser auf den Boden kippen und in die Regenshows für morgen.

00:09:26 Sprecher 2

In dem Fall ist das Problem, dass das eigentliche kausale Feature ist ja, dass es heute regnet.

00:09:32 Sprecher 2

Und das ist schuld, sozusagen, dass der Boden nass ist und dass es morgen eher regnen könnte.

00:09:36 Sprecher 1

Das heißt also, was passieren könnte, ist, dass jemand glaubt, eine Kausalität entdeckt zu haben oder dass die, nachdem man das Modell interpretiert, zu diesem Schluss kommt, ja.

00:09:47 Sprecher 1

hab ich ein Feature, das entscheidend ist für die Vorhersage, aber an für sich hat man das falsche Feature erwischt gehabt.

00:09:54 Sprecher 1

Ja, man hat also, wenn ich dich richtig verstehe, schon interpretiert, dieses Feature ist entscheidend, aber der Rückschluss draus aus diesem Feature, der ist ja leider ein falscher.

00:10:02 Sprecher 1

Hättest du ein Beispiel aus der Medizin, weil glaube ich, viele unserer Hörer im Bereich Medizin, Medizinprodukte unterwegs sind.

00:10:09 Sprecher 2

Ja, also da hab ich auch ein Beispiel aus sozusagen meiner eigenen Forschung oder beziehungsweise noch, wo ich

00:10:16 Sprecher 2

eher als Statistiker gearbeitet hab, auch da haben wir uns aus Beobachtungsdaten angeschaut, ob ein bestimmtes Medikament, also sogenannte TNF-Alpha-Hemmer, wirksam sind und speziell haben wir uns da die Progression der Verknöcherung in der Wirbelsäule angeschaut, weil man hat da sehr gute Medikamente gegen die Entzündung an sich.

00:10:34 Sprecher 2

Diese, das sind eben diese TNF-Alpha-Hemmer, aber es ist noch unklar, oder oder was wir untersucht haben, ist eben, ob diese auch einen Einfluss haben und um diese Progression zu stoppen.

00:10:45 Sprecher 2

haben wir eben auch ein, dann eher so diese statistischen Modelle, also ein lineares Regressionsmodell benutzt dafür.

00:10:51 Sprecher 2

Da ist aber der Zusammenhang ebenso, dass diese Medikamente helfen halt sehr gut, diese Entzündung zu zurückzufahren.

00:10:59 Sprecher 2

Aber die Entzündung hat wiederum Einfluss natürlich auf die Progression, also von dieser Verknöcherung in der Wirbelsäule.

00:11:06 Sprecher 2

Jetzt, wenn man sich Entzündungs, diese Entzündungswerte mit ins Modell nimmt,

00:11:11 Sprecher 2

dann lernt das Modell sich, das ist eben ein sehr, sehr guter Vorhersagefeature für wie stark die Progression in einem bestimmten in der Wirbelsäule stattfinden wird.

00:11:21 Sprecher 2

Und der Effekt von dem Medikament, sozusagen geschluckt von dadurch, der geht komplett fast über über diese Hemmung der Entzündung.

00:11:30 Sprecher 2

Das heißt, in der Analyse, wo wir beides mit aufnehmen und also beides in unserem Modell haben, kommt dann am Ende heraus, dass dass das Medikament nicht

00:11:37 Sprecher 2

relevant war für das sozusagen nicht hilft bei der gegen diese Progression von von der Verknöcherung.

00:11:43 Sprecher 2

Aber aber das ist eben ,n Fehler, weil es hilft eben schon, indem es die Entzündung senkt.

00:11:49 Sprecher 2

Also das ist wieder da so ,n so ,n Fall, wo man eben, wenn man sozusagen die falschen Features mit aufnimmt, nicht die korrekt, also man muss eben auch darüber nachdenken, was möchte man denn für Aussagen haben am Ende, also nicht nicht haben, sondern was möchte man untersuchen, wie muss man sein Modell aufstellen, um die

00:12:06 Sprecher 2

Frage zu beantworten, die einen interessiert.

00:12:07 Sprecher 1

Wenn man jetzt das ganz kurz noch mal nach den Lösungen schaut, also in dem ersten Beispiel, das du genannt hast, mit der Wettervorhersage, hab ich dich so verstanden, hätte man den nassen Boden als Feature rausgenommen, weil das wichtigere Feature, nämlich des Regen, der Regen am aktuellen Tag gewesen wär, als Faktor, dass sich besonders stark auf die Prognose für das morgige Wetter auswirkt.

00:12:27 Sprecher 1

was hätte man jetzt in dem Beispiel deines Modells nehmen müssen, bei dem jetzt das Medikament und diese Entzündungsparameter als Feature mit eingeflossen sind und die ja vorhergesagt hab, wie sich dann die Krankheit weiterentwickelt.

00:12:41 Sprecher 1

Das heißt dann diese, diese Verknöcherung stattfindet, wie hätte man da reagieren müssen, um das zu korrigieren?

00:12:46 Sprecher 2

Also in dem Fall war die Antwort dann, dass wir den Entzündungshemmer rausnehmen aus dem Modell.

00:12:52 Sprecher 2

Und wir haben auch noch weitere Analysen gemacht, da das das sind sogenannte Mediationsanalysen, wo man sich anguckt, sozusagen, wie viel von dem Effekt geht über den Effekt der des der Entzündungssenkung und wie viel ist sozusagen der direkte Effekt von dem Medikament.

00:13:07 Sprecher 2

Aber wenn man das jetzt noch ein bisschen allgemeiner sieht, sozusagen, was macht man jetzt generell um oder wie kann man generell an solche Probleme rangehen, da hilft und das haben wir auch in dem Fall gemacht, einen sogenannten

00:13:19 Sprecher 2

so ‚n Graphen zu zeichnen.

00:13:20 Sprecher 2

Das heißt, man malt sich die Features eigentlich hin, also und malt sich sozusagen kausale Pfeile.

00:13:26 Sprecher 2

Das ist auch etwas, wo man dann Domain-Experten wirklich braucht dafür und da stecken Annahmen drin und das ist auch nicht, was man unbedingt immer aus den Daten automatisch lernen kann.

00:13:34 Sprecher 2

Also da, das ist eben dieses, man sieht die zwar die Korrelation, aber die Kausalität, dafür muss man oft Annahmen dann noch zusätzlich reinstecken.

00:13:41 Sprecher 1

Wenn ich dich richtig verstehe, ist also deine Empfehlung, nicht nur versuchen, die Modelle dahingehend

00:13:48 Sprecher 1

gehen, zu verstehen, dass man rausfindet, welche Feature, also Inputwerte, sind jetzt entscheidend für ‚ne Vorhersage, sondern dass man sich genau anschaut, wie hängen diese Feature untereinander voneinander ab, um dann gegebenenfalls irgendwelche Zwischenwerte rauszunehmen, damit man da nicht zu falschen Kausalitäten kommt.

00:14:07 Sprecher 2

Also jetzt, wenn man auch an Produkte denkt, ein Verband von nicht kausalen Features macht einem ein ein Machine Learning Modell natürlich auch anfälliger, zum Beispiel

00:14:17 Sprecher 2

gegenüber Attacken, zum Beispiel, wenn man jetzt sowas denkt wie ein ein Schufa-Score, wenn die jetzt zum Beispiel nehmen würden, wie viele Kreditkarten man hat, das ist vielleicht gar nicht kausal, wenn wenn jetzt es ist nur ein Beispiel, also es muss nicht stimmen, aber wenn jetzt zum Beispiel mehr Kreditkarten, die aber alle im Positiven sind, gut wären für einen guten Schufa-Score, dann könnte jetzt jemand hergehen und einfach ganz viele Kreditkartenkonten aufmachen.

00:14:38 Sprecher 2

Aber es ist natürlich nicht kausal, seine.

00:14:41 Sprecher 2

sozusagen seine die Wahrscheinlichkeit einen Kredit zurückzuzahlen oder so, hat sich ja nicht verbessert.

00:14:45 Sprecher 2

Deswegen, das wär jetzt nicht kausales Feature und das macht das Modell natürlich anfälliger auch gegen Missbrauch, wenn es nicht kausale Features verwendet.

00:14:53 Sprecher 1

Wenn du jetzt haben wir gerade sozusagen über die Interpretierbarkeit selber gesprochen gehabt, wie wird man jetzt rausfinden, wie gut ein Modell jetzt auch wirklich ist?

00:15:04 Sprecher 1

Also, wenn ich dich richtig verstehe, wär eine Möglichkeit einfach mal die

00:15:08 Sprecher 1

Kiste zu öffnen, schauen was da innen drin passiert, ‚n tieferes Verständnis mitzuentwickeln.

00:15:14 Sprecher 1

Kann man rein anhand von Qualitätsparametern, also ich sag jetzt mal vielleicht Sensitivität oder Accuracy, schon erkennen, wie leistungsfähig ‚n Modell ist oder gibt es da auch noch mal Fallstricke, auf die man achten sollte?

00:15:27 Sprecher 2

Also das Wichtige ist natürlich immer, also die Interpretivbarkeit ist natürlich die eine Sache.

00:15:32 Sprecher 2

Die andere Sache ist,

00:15:34 Sprecher 2

natürlich, dass man es richtig evaluiert.

00:15:35 Sprecher 2

Das heißt, dass man es auf Testdaten evaluiert und eben auch die sich Performancezahl dem anschaut, was natürlich auch gut ist, wenn man sich da zusätzlich auch die sozusagen die Unsicherheit davon anschaut.

00:15:46 Sprecher 2

Also, wie stark streut es denn?

00:15:48 Sprecher 2

Weil oft sieht man sich, also das ist auch ein sehr generelles Problem, oft guckt man sich immer nur eine Zahl an, aber diese Zahlen sind immer geschätzt mit Daten.

00:15:55 Sprecher 2

Also das kann jetzt sowohl sowas sein wie die Sensitivität, die wir, die wir messen, aber es kann natürlich auch sein, wenn wir jetzt die Feature Importance berechnen, da

00:16:03 Sprecher 2

das schätzen wir alles mit Daten und das ist mit Unsicherheit behaftet.

00:16:07 Sprecher 2

Deswegen ist eigentlich auch immer die Empfehlung, dass man sich da auch immer diese Spannbreite anschaut oder eben die Varianz von solchen Werten.

00:16:14 Sprecher 2

Weil vor allem, wenn die stark streuen, kann es natürlich sein, dass man mal einen höheren Wert erwischt oder ein sozusagen das dann falsch interpretiert.

00:16:21 Sprecher 1

Auch wenn ich es nochmal zusammenfasse, die Empfehlungen, die du umsetzt und damit eben auch Medizinprodukteherstellern gibst, habe ich jetzt, glaube ich, 3 gehört.

00:16:31 Sprecher 1

Das erste ist

00:16:32 Sprecher 1

zu versuchen, diese Modelle auch wirklich zu verstehen, als sich sozusagen sich des Themas Interpretierbarkeit anzunehmen.

00:16:39 Sprecher 1

Und da ist dein Buch, glaub ich, ‚ne großartige Hilfestellung.

00:16:43 Sprecher 1

Das Buch gibt es übrigens kostenfrei im Web anzusehen bei GitHub.

00:16:47 Sprecher 1

Der zweite Tipp, den ich gehört hab, ist, wenn man das Modell jetzt sich genauer angeschaut hat, nicht irgendwelchen Irrtümern zu erliegen, dass man glaubt, man hätte es verstanden, hat es aber dahingehend nicht, dass man

00:17:01 Sprecher 1

Korrelation und Kausalitäten verwechselt oder wirkliche Kausalitäten nicht entdeckt, beispielsweise Abhängigkeiten von Feature.

00:17:10 Sprecher 1

Und der dritte Tipp, den ich gehört hab, ist, wenn man Qualitätsparameter, du hast jetzt gerade Sensitivität noch mal genannt hat, bestimmt hat, dass man sich nicht auf dieses Maß alleine verlässt,

00:17:23 Sprecher 1

sondern einmal rausfindet, mit welchen Unsicherheiten sind denn diese Parameter tatsächlich verbunden.

00:17:29 Sprecher 1

Also, was ist das Konfidenzintervall und wahrscheinlich muss man das jetzt auch nicht nur als ein Wert mit einem Konfidenzintervall angeben, sondern das noch mal abhängig machen, beispielsweise von der Patientenpopulation.

00:17:43 Sprecher 1

Also, dass man sich möglicherweise diesen Wert mit seiner Unsicherheit über den kompletten Altersbereich oder für verschiedene Altersintervalle

00:17:52 Sprecher 1

anschaut.

00:17:53 Sprecher 1

Hab ich das richtig verstanden oder waren das die Dinge, die du auch tatsächlich gesagt?

00:17:56 Sprecher 2

Ja, genau, das es gibt natürlich noch viel mehr.

00:17:58 Sprecher 2

Sie haben jetzt natürlich nur sehr ein paar kleinere Dinge uns angeschaut, aber ja, ich denke, es sind zumindest wichtige Themen.

00:18:05 Sprecher 1

Auch für alle, die jetzt noch mehr darüber wissen wollen, wär mein Tipp, die beiden eben erwähnten Quellen zu studieren.

00:18:13 Sprecher 1

Einmal dieses Paper, das die häufigsten Fallstricke

00:18:18 Sprecher 1

nennt bei der Interpretierbarkeit von Machine-Learning-Verfahren.

00:18:22 Sprecher 1

Und das zweite wäre das Buch von Christoph Molnar: ‚Interpretierbarkeit von maschinellem Lernen'.

00:18:27 Sprecher 1

Das ist auf Englisch und kostenfrei bei GitHub veröffentlicht.

00:18:31 Sprecher 1

Es gibt noch weitere Artikel, die wir in ebenfalls verlinken, die wir hier im im Blog auf unseren Fachartikel mit haben.

00:18:37 Sprecher 1

Das sind beispielsweise Artikel auch zum Thema Validierung von Machine Learning Libraries, ein Artikel

zu den regulatorischen Anforderungen.

00:18:45 Sprecher 1

Wir haben auch

00:18:47 Sprecher 1

Videotrainings im Audit gerannt, genau zu diesen Themen und eine einen Leitfaden, den Christoph und ich gemeinsam erstellt haben und der jetzt momentan eben bei der W.H.O.

00:18:57 Sprecher 1

mit weiterentwickelt wird.

00:18:59 Sprecher 1

Und ich glaube, damit haben sie eine ganze Menge an Quellen und Ideen, wie sie ihre maschinellen Lernenverfahren, ihre Modelle in ihren Medizinprodukten prüfen können, damit wir das erreichen, was wir uns ja alle wünschen, nämlich Medizinprodukte, die sicher sind und die

00:19:17 Sprecher 1

eine möglichst hohe klinische Wirksamkeit haben.

00:19:19 Sprecher 1

Christoph, ganz herzlichen Dank, dass du mit dabei warst.

00:19:22 Sprecher 2

Gerne, danke für die Einladung.