

Also die erste Frage, die schon vorhin aufkam, war, wie die stattfinden, vor welchem Publikum. Also wir machen das privat. Das heißt, Manuel und ich sind da und die Gruppe und sonst niemand. Ihr könnt es euch auch ein bisschen vorstellen wie eine mündliche Prüfung. Ihr seht ja, es sind jeweils 25 Minuten eingeplant pro Gruppe. Also hier sieht man es zum Beispiel.

Und die setzen sich zusammen wie folgt. Ich schreibe natürlich dazu auch nochmal eine E-Mail. Die Idee ist, die 25 Minuten hier aufzuteilen, dass man sich in den ersten ungefähr zwei Minuten vorstellt, das Management von MyVC wäre im Raum und möchte gerne wissen, was habt ihr da eigentlich gemacht und warum werden wir jetzt noch reicher, als wir schon sind. Ja?

Das heißt, wirklich fürs Management adressiert und verständlich. Ich schätze mal so ein bis zwei Folien sollten reichen. Und stellt euch vor, das Management geht dann nach zwei Minuten wieder, weil die haben natürlich viel zu tun und ihr müsst euch prägnant drüber bringen. Also ungefähr zwei Minuten für MyVC Management. Und insgesamt wollen wir euch nur zehn Minuten reden lassen.

Und den Rest der 25 Minuten fragen wir dann. Dazu muss ich euch gleich noch erklären, wie eure Abgabe, eure Deliverables sozusagen strukturiert sein sollten. Also ungefähr acht Minuten bleiben dann noch für, ich sage mal, Details. Nein, ich nenne es nicht Details, das ist das falsche Wort. Also es soll nicht detailliert sein, es soll High Level sein, weil ihr habt ja auch nur acht Minuten Zeit.

sagen wir die technische Sicht. Also technische Sicht, ja, und fast so Dinge wie, anderen blauen Stift auch gewählt. Was für Features habt ihr gewählt? Was waren eure Gedanken dabei? Wie habt ihr die Auswahl eines Modells gemacht und warum? Und was waren eure Gedanken hinsichtlich Evaluationsstrategie? Was kam dabei raus? Welches Modell ist am besten und warum?

Also technische Sicht, nenne ich es mal. Ich mache es in Anführungszeichen. Also im Wesentlichen hilft es natürlich immer, das machen wir jetzt gleich nochmal ein bisschen detaillierter, in die Aufgabenstellung reinzuschauen. Das ist es nicht. Wo ist es denn? Genau. Also hier seht ihr die Lieferobjekte. Und das ist eigentlich das, dazu sage ich gleich nochmal. Und dann haben wir 15 Minuten Zeit.

Fragen von Manuel und mir und was es dazu braucht, sind einerseits diese Lieferobjekte. Hier steht wirklich eigentlich alles drin, insbesondere die letzten drei Punkte. Gebt ihr einfach als Dateien ab. Also, wenn ihr Tableau Prep benutzt habt, den Workflow und

Wenn ihr irgendwas mit Teilchen gemacht habt oder so, dann den Code oder was auch immer ihr da verwendet habt, sodass wir die Datenaufbereitung nachvollziehen könnten oder ausführen könnten. Wir werden das nicht für jede Gruppe machen, aber falls wir irgendwo Zweifel haben, können wir. Dann den Orange-Workflow.

Und dazu muss natürlich die Input-Datei, also das, was hier aus der Datenabbereitung rauskam und was ihr dann in Orange vorne reinladet, das bitte auch mit abgeben, damit wir das dann auch ausführen können in Orange Workflow. Das werden wir wahrscheinlich sicher bei einigen Gruppen auch ausprobieren und gucken, was da passiert.

Und natürlich auch unbedingt diese Sachen offen und fertig dabei haben, wenn die Abschlusspräsentation startet. Also nicht sagen, Moment, jetzt muss ich erstmal die richtige Datei finden und so, sondern das sollte offen sein, weil wir haben wenig Zeit und wir wollen möglichst in die Tiefe gehen. Also jetzt steht da oben Präsentation. Also das bedeutet, wenn da steht Präsentation, ihr müsst keinen Bericht abgeben und ihr sollt auch keinen Bericht abgeben.

sondern nur die Präsentation. Und die Präsentation sollte sich in zwei Teile teilen. Also ich nenne es mal Präsentationsdatei. Also das, was ihr abgebt. Also den offiziellen Teil sozusagen oder den Präsentationsteil für die ersten zehn Minuten. Und da, wie gesagt, technische Sicht in Klammern High Level. Also nicht

für jedes Feature, was ihr euch überlegt habt, die genaue Formel, die ihr in Tableau Prep eingegeben habt und alle sonstigen Details oder genaue Konfigurationen von irgendwelchen Modellen, die ihr in Orange trainiert habt, sondern wirklich High Level, was war die Idee dahinter und was kam drauf. Und dann macht doch noch ein, ich nenne ihn mal Backup-Teil, wo ich solche technischen Details dann auch noch reinpackt und

Die Idee ist, dass wir den Präsentations-Teil für die ersten zehn Minuten, oder dass ihr den nutzt und uns nochmal präsentiert und darbietet. Und wir werden uns natürlich alles, was ihr abgebt, vorher anschauen.

oder währenddessen anschauen und dann können wir damit in die Tiefe gehen. Also wir werden ja nicht die Zeit haben, jede dieser Folien im Detail anzuschauen, aber wir können dann wählen, welche Aspekte wir nochmal genauer anschauen wollen und da hilft es natürlich, wenn ihr da schon Details auf einer Folie habt. Und ja, ich meine, was da auf diesen Folien sein sollte, ist eigentlich hier beschrieben. Also

All die Sachen, die hier stehen, ihr müsst euch jetzt einfach nur noch überlegen, sozusagen, all die Punkte, die ja auch hier oben nochmal im Detail aufgelistet sind, was davon oder in welcher Form kommt das in diesen Präsentationsteil und was kommt eher in den Backup-Teil. Ist das für euch so ungefähr klar, was wir vorhaben mit euch?

Ja, es ist klar, für mich eine Frage zu dem Backup-Teil vor allem. Also es kann ja sein, dass wir Features gemacht haben, die wir schlussendlich nicht verwenden, weil es hat nichts gebracht etc. Und ist dann auch die Idee, dass man auch zeigt, was man gemacht hat und schlussendlich nicht verwendet hat oder...

Nur wirklich auf das fokussieren, was funktioniert und was wir schon darauf verwendet haben. Ja, also jetzt vielleicht gerade bei dem Beispiel, ich finde das schon interessant auch. Man kann ja gute Gedanken haben und es passiert oft im Leben, dass man einen Gedanken hat, der eigentlich nachvollziehbar und gut ist und dann funktioniert es trotzdem nicht. Ja, würde mich auch interessieren in dem Fall.

Okay, danke. Wenn es jetzt darum geht, sozusagen sämtliche Parameter, die ihr durchprobiert habt bei einem Modell und dann zu gucken, was war am Ende das Beste, da würde mir eigentlich reichen, wenn ihr sagt, wir haben verschiedene Werte probiert und das war das Beste. Bei den Features, da finde ich das interessant, einfach zu gucken, was habt ihr euch so überlegt und habt ihr verstanden, wie man Feature Engineering macht. Es geht ja immer darum,

Aus unserer Sicht zu beurteilen, was habt ihr gelernt. Wenn ihr das so im Hinterkopf behaltet, dann hilft es vielleicht auch zu entscheiden, was da rein soll und was nicht. Sonst noch Fragen? Okay. Jetzt muss ich mal gucken, habe ich noch was vergessen? Ich glaube nicht. Okay, dann habe ich versprochen, dass ich euch was zur Interpretierbarkeit erzähle. Und fangen wir doch an mit einem kleinen Gedankenexperiment.

Einfach nochmal, um das Ganze zu motivieren. Also stellt euch vor, ihr habt einen Tumor und der muss operiert werden. Lieber nicht, aber trotzdem ist es so. Jetzt habt ihr zwei Möglichkeiten.

Entweder ihr habt die Möglichkeit, euch von Menschen operieren zu lassen und andererseits gibt es da einen Roboter, der das alles automatisch macht und die Operation durchführen kann. Und ihr seht hier die Statistiken, also

Bei den Menschen war die Überlebenschance im Schnitt, oder war 85 Prozent, bei den Robotern 98. Also, was macht ihr? Roboter. Machen alle. Okay. Ja, okay. Das ist ja langweilig bisher. Jetzt, was machen wir hiermit? Wenn ihr jetzt die Gelegenheit hättet, vorher noch Fragen zu stellen, was würdet ihr gerne wissen?

Oder würdet ihr sagen, naja, das mit dem Roller, das klappt, das machen wir. Ja, Jeremy? Ob es immer die gleiche Person ist, bei der es 85% ist, oder ob es andere gibt, die haben einen höheren oder einen besseren Wert. Okay, also du würdest versuchen, die beste Person zu finden, die dich operieren soll. Du gehst daran aus, dass es schwankt, ja? Ja, ich bin noch notional gefestigt. Ja.

Noch was? Was für Fehler überhaupt passiert sind? Bei Menschen? Oder bei beiden? Warum willst du das wissen? Vielleicht hat der Mensch wirklich kleine Fehler gemacht, die korrigierbar sind und der Roboter macht große Fehler.

Naja, die waren schon groß, oder? Weil 15% der Patienten haben es nicht überlebt. Also ich meine, es muss ja auch nicht alles mit Fehlern zusammenhängen, oder? Marc, das würdest du fragen. Nein, das ist nicht so.

Also es könnte sein, dass die Testmenge hier nicht gleich groß war. Noch was? Jetzt gibt es noch den unteren Punkt. Wenn ihr jetzt mal annimmt, ihr habt eine Blutgerinnungsstörung, würden da noch zusätzliche Fragen bei euch aufkommen?

Und er kann schneller reagieren? Okay. Also ich meine, wenn du eine Blutgerinnungsstörung hast, dann weißt du es ja vorher. Also es ist jetzt nicht etwas, was spontan während der... Also was wir im Notfall direkt haben können, oder? Also nein. Du hast die Blutgerinnungsstörung und natürlich ist es eine Komplikation, die die Operation erschwert. Aber...

Nein, also der Tumor wurde diagnostiziert, dann plant man die OP und vorab füllst du den Fragebogen aus. Ja, ich habe das Gefühl, bei einer Blutgerinnungsstörung, da kann der Mensch vielleicht flexibel passieren, weil man da praktisch unerwartete Begriffe mit sich gibt. Ja, okay, also würdest du wissen wollen...

Also das ist deine Vermutung. Du würdest irgendwie wissen wollen, wie gut kann dieser Roboter darauf reagieren, oder? Hatte sich noch jemand gemeldet? Also mich würde dann auch interessieren, ob der Roboter überhaupt jemals Patienten operiert hat, die es so wie ich, diese Blutgerinnungsstörung haben, oder? Ich bin ja dann größerem Risiko ausgesetzt. Und dann würde ich gerne wissen, wurde der auch darauf trainiert sozusagen? Seht ihr das? Also wenn man jetzt sozusagen

sich überlegt, wie sowas trainiert wird, dann würde man ja bestimmte Risikofälle vielleicht erstmal ausschließen, aus ethischen Gründen auch. Also man kann ja, wenn sowas noch jung ist und noch nicht so perfekt, wie es hier erscheint, nicht einfach zum Beispiel Patienten mit Blutgerinnungsstörungen von so einer Maschine operieren lassen, die das noch nie gesehen hat, die gar nicht weiß, wie darauf zu reagieren ist. Man würde wahrscheinlich, wenn solche Risikofaktoren bestehen,

diese Patienten erstmal ausschließen. Seht ihr, dass da eine gewisse Unfairness entsteht? In dem Sinne, dass vielleicht, also das wäre auch so eine Frage, die ich hätte, wie vergleichbar

sind denn die Testmengen hier? Sind hier überhaupt solche schwierigen Fälle dabei gewesen? Oder hat man die, weil man eben sie ja nicht so als menschliche Versuchskaninchen behandeln wollte, sowieso von Menschen operieren lassen? Und das erklärt dann vielleicht auch, warum die Menschen mehr Probleme hatten, in Anführungszeichen. Also warum da schlechtere Zahlen dabei rauskommen. Weil hier sowieso mehr Risikofälle behandelt wurden. Versteht ihr, was ich meine? Das heißt...

Warum hat das was mit Interpretierbarkeit zu tun? Es geht ein bisschen darum, einerseits zu wissen, womit wurde das sozusagen konfrontiert, dieses System, und auch trainieren. Und andererseits hat das natürlich auch einen Einfluss darauf, was das kann und wie es mit bestimmten Ausnahmesituationen umgehen würde. Also mit bestimmten Fällen oder Instanzen, in dem Fall Patienten. Und sozusagen, wenn man Modelle des maschinellen Lernens einsetzt in so

Situationen, wo es wirklich, ja, hier geht es um Leben und Tod. Ja, manchmal geht es vielleicht auch zum Beispiel in der Justiz oder so, ob ich jemanden einsperre oder nicht. Kann ich auch und wird teilweise auch schon zumindest diskutiert, maschinelles Lernen einzusetzen. Also wirklich so wichtige Dinge, von denen viel abhängt. Dann sich auf solche Zahlen zu verlassen und gar nicht zu wissen, was tut dieses Ding eigentlich?

worauf stützt es seine Entscheidungen während der Operation, würde mich sehr nervös machen. Allein diese Zahl würde mir überhaupt nicht ausreichen. Ich würde noch sehr viel mehr wissen wollen, was das Innenleben von dieser Maschine angeht, bevor ich wirklich Vertrauen schöpfe. Und Interpretierbarkeit hat auch sehr, sehr viel mit Vertrauen zu tun. Wenn ich ein bisschen weiß, was so ein Modell tut, also das ist jetzt kein Modell, das ist ein mechanisches System, aber wenn wir es jetzt transferieren auf Maschine, also Modelle des maschinellen Lernens, dann

Geht es da auch um Vertrauen? Das schlägt mir Entscheidungen vor und ich muss mir überlegen, möchte ich den Vorschlägen folgen oder nicht? Okay, soweit zur Motivation. Ein bisschen dann noch unterstützende Argumente. Also Interpretierbarkeit ist uns auch schon nützlich gewesen beim, ich nenne es hier Debugging. Also zum Beispiel, erinnert ihr euch an das Days Open Problem?

Also wir hatten diesen Datensatz von dieser Softwarefirma mit ihrem Helpdesk und den Tickets, die in bestimmter Zeit beschlossen werden sollten, erinnert ihr euch? Und wir haben ein Modell gebaut, um zu versuchen, vorher zu sagen, wo das Service-Level-Agreement verletzt wird, bei welchen Tickets. Und dann hatten wir das Attribut Days Open, was eigentlich ein Fall von sogenannter Target Leakage war. Das heißt,

das hat die perfekte Information über das, was wir vorhersagen wollten, schon enthalten. Und man kommt immer wieder zu so Fällen, wo ein Modell sehr, sehr gut zu sein scheint. Und wenn man dann ein bisschen näher hinguckt, was es eigentlich tut, dann fällt einem auf, dass es vielleicht Informationen benutzt, die es eigentlich zum Vorhersagezeitpunkt nicht haben kann und die auch dann zu viel verraten über das, was ich vorhersagen will.

und die man nicht verwenden sollte, weil sie auch nicht verfügbar sein werden, wenn ich das Modell einbestehe. Also das ist sozusagen ein, also Data Leakage oder Target Leakage kann ich erkennen, wenn ich verstehe, was das Modell tut. Ausnahmefälle. Also ich habe hier mal so ein Beispiel

mitgebracht von Erklärung eines Bilderkennungsmodells. Also das Modell wurde darauf trainiert, Fahrräder zu erkennen.

Und Interpretierbarkeit bei solchen Modellen kann man darüber herstellen, dass man, also Interpretierbarkeit, man weiß immer noch nicht ganz genau, was das Modell macht, aber man weiß, worauf es fokussiert. Also je röter die Region im Bild ist,

Also es geht hier um neuronale Netze, aber es ist eigentlich auch egal. Wir haben irgendein Modell und der Input ist dieses Bild mit seinen Pixeln und die Farbschattierung sagt aus, also es ist ein Versuch, das interpretierbar zu machen, was das Modell tut, und die Farbschattierung sagt aus, wie sehr die Pixel des Bildes

zu der Klassifikation beitragen. Das heißt, hier sieht man, dass das Modell sich auf die Räder fokussiert oder dass es Fahrräder anhand der Räder erkennt. Also stellt euch vor, so ein Modell wird eingesetzt in einem autonom fahrenden Auto.

Dann will man bestimmte Hindernisse erkennen und Fahrräder als Fahrräder zu erkennen, ist auch nützlich, weil man dann weiß, aha, die können sich potenziell bewegen, im Gegensatz zu einem Mülleimer oder so. Also Mülleimer können sich auch bewegen, aber tun es nicht so oft. Genau, und hier kann man sich jetzt überlegen, was könnten Edge-Quenkel sein? Also hier steht, was passiert, wenn jetzt hier zum Beispiel Fahrradtaschen dranhängen?

die das Rad teilweise oder ganz verdecken. Oder wenn irgendwas anderes davor steht sozusagen. Das heißt, dieses Modell, das wird vermutlich in sehr vielen Fällen sehr gut funktionieren, weil es eben ein charakteristisches Merkmal von Fahrrädern, nämlich die zwei Räder,

gelernt hat, aber wenn durch irgendwelche Gründe die Räder verdeckt sind, dann wird es Schwierigkeiten bekommen. Und dann einfach das zu erkennen, ist schon wertvoll. Was bezieht sich das Modell auf? Also das Bild eines Fahrers? Also sagen wir mal so, das Modell bekommt als Input ein Bild von der Straße, der Verkehrssituation. Und es ist darauf trainiert, in dem Bild Fahrräder zu erkennen oder zu sagen, da ist ein Fahrrad.

Und die Heatmap ist nur der Versuch, das, was das Modell gelernt hat. Also das Modell kriegt als Trainingsmenge 100.000 Bilder und jeweils dazu gesagt, ob da ein Fahrrad drauf ist. Und vielleicht auch wo, also in welchem Bereich des Bildes das Fahrrad ist. Und dann lernt es irgendwas, aber wir wissen ja nicht was. Das Problem ist beim neuronalen Netz, das hat irgendwas gelernt, das sind irgendwie 20 Millionen Gewichte gelernt worden, aber was tun die eigentlich? Und diese Heatmap ist der Versuch, das

sichtbar zu machen, was das Modell tut. Also was es genau tut, wissen wir immer noch nicht. Wir wissen nur, wo es hinguckt. Aber schon das erlaubt uns eben ein bisschen drüber nachzudenken, was könnten für Probleme entstehen. Also wenn man über Bilder spricht, da gibt es auch noch die anderen lustigen Beispiele. Also kennt ihr das? Es gibt zwei prominente Beispiele.

Ja, ja, sowas hier. Ich meinte jetzt noch Beispiele mit Bildern. Also kennt ihr das mit den Hunden und den Wölfen? Oder Huskys und Wölfen?

Also da haben Leute angefangen zu trainieren mit Bildern von Huskys und Wölfen und wollten, dass das Modell die von anderen unterscheiden lernt. Und dann haben sie so eine Hitmap gemacht und geguckt, worauf das Modell schaut. Und zufälligerweise waren alle Bilder, auf denen Huskys drauf waren, hatten auch Schnee. Und dann haben sie gesehen, es guckt immer auf die Weißen.

Teile des Bildes. Und wenn es Schnee ist, dann sagt es Husky. Also es macht auf eine Art Sinn, weil Huskys eben meistens im Schnee unterwegs sind, aber könnte ja auch mal bei einem Wolf der Fall sein. War halt bei der Trainingsmenge nicht so. Da hat man gesagt, okay, das Modell hat eigentlich nichts Nützliches gelernt. Es hat den einfachsten Weg gefunden, auf dieser Trainingsmenge Wölfe von Huskys hinzukleiden. Oder es gab auch Beispiele von Röntgenbildern,

Es ging um Lungenentzündungen. Ich weiß nicht mehr genau, was da erkannt werden sollte. Der Fall war so, dass besonders schwere Fälle von Lungenentzündungen mit anderen, also die Patienten, die besonders schwer erkrankt waren, die wurden mit anderen Geräten geröntgt als die, bei denen es weniger schlimm war. Und

Immer unten auf dem Bild stand die Signatur des Geräts, mit dem das Bild gemacht wurde. Also irgendwie die Nummer oder was. Und dann hat man auch so eine Heatmap gemacht. Man hat gesehen, man hat ein Modell gelernt, was super unterscheiden kann oder vorhersagen kann, wie lange jemand überleben wird, glaube ich, war die Vorhersage, die das Modell machen sollte. Es war sehr gut. Und dann hat man eben auch geguckt, worauf schaut das Modell und hat gesehen, es guckt auf den Text. Das heißt...

ja, es sieht, mit welchem Gerät es gemacht wurde und wenn es eben das Gerät für die schwerer Erkrankten war, dann hat es vorher gesagt, dass die weniger lange überleben und das war natürlich dann oft korrekt. All solche Dinge kann man mit Interpretierbarkeit von Modellen hoffentlich erkennen. Oder eben, jetzt sind wir wieder bei strukturierten Daten, also nicht Bilddaten, zum Beispiel erkennen, was für Verzerrungen in den Daten sind. Also das, was du vorhin gesagt hast,

Wenn eben die Menschen immer männliche Bewerber bevorzugt haben, dann lernt das Modell auch das zu tun. Aber vielleicht wollen wir das ja gar nicht so weiterführen. Genau. Jetzt habe ich gesagt, das ist einmal sozusagen einfach erkennen, was das Modell tut und daran auch teilweise unerwünschte oder riskante oder falsche oder verzerrte Muster aufdecken, die das Modell gelernt hat.

Oder welche, die nicht generalisieren, die einfach sozusagen in diesen, also quasi Overfitting ist das ja, was ich vorhin auch beschrieben habe mit dem Wolf und dem Husky. Also was erkennen ist ein Punkt, der für interpretierbare Modelle spricht, das Vertrauen. Also letztlich wird das, was wir hier tun, ja irgendwo in den Unternehmen eingesetzt werden und es wird eingesetzt, um Entscheidungen zu unterstützen und

Die Menschen sind eigentlich doch am Ende meistens noch verantwortlich für die Entscheidung. Es gibt Beispiele, beispielsweise Marketing oder so, da ist das egal, ob ich nun einen mehr oder weniger kontaktiert habe. Da guckt keiner so genau, da kann man das auch machen, ohne dass der Mensch im Detail versteht, was das Modell tut. Aber oft, wenn es eben um wichtige Entscheidungen geht und die Menschen dafür verantwortlich sind, dann wollen sie auch ungefähr wissen, warum, weil sie dann besser dahinterstehen können. Und

Auch die Muster, die man ja erkennt, wenn man das Modell interpretiert, sind oft interessant. Also für ein Unternehmen kann es interessant sein, zu erkennen, zum Beispiel, warum bestimmte Kunden positiv reagieren auf meine Marketingkampagne oder welche Kunden das sind. Oder auch manchmal wissenschaftliche Erkenntnis, die ich aus den Mustern ziehen kann oder aus den Modellen und die dann für sich genommen wieder weitere Forschung ermöglichen oder so. Ja, also es gibt

Was wir schon kennengelernt haben, Modelle, die von sich aus interpretierbar sind. Also wir hatten gesehen, in Orange gibt es manche, da kann ich noch ein Widget hintendran hängen, was mir dann erklärt. Zum Beispiel ein Tree Viewer oder ein Nomogramm, ihr erinnert euch, oder? Das sind sozusagen die intrinsisch interpretierbaren Modelle. Und das, was wir uns heute anschauen, sind eher die Post-Hoc Modelle.

erklärbaren Modell oder nein nicht post hoc erklärbaren Modelle sondern post hoc Erklärung für nicht interpretierbare Modelle also letztlich geht es darum zu beobachten wie sich diese Modelle verhalten und daraus und das zu beschreiben ja also zum Beispiel zu sagen welche Features wohl vermutlich eine besonders große Rolle bei den Entscheidungen des Modells spielen und auch welche Feature-Werte zu welcher Art von Entscheidung führen, ohne dass man, das muss man sich immer vor Augen halten, wirklich weiß, was das Modell tut. Das sind nur Annäherungen. Wir werden vor allem über modellagnostische Verfahren reden, also solche, die man auf jedes beliebige nicht interpretierbare Modell, nicht intrinsisch interpretierbare Modell anwenden kann.

Das mit der Heatmap zum Beispiel war was Modellspezifisches. Also das funktioniert für Convolutional Neural Networks, also CNNs, habt ihr vielleicht schon mal von gehört, die man eben für Bilderkennung spezifisch einsetzt. Und sowas funktioniert natürlich nicht für ein Decision Tree oder so. Lokal und global unterscheiden wir noch. Irgendeine Idee, was das bedeutet? Was heißt lokale Erklärbarkeit? Was könnte das bedeuten?

Also lokal bedeutet das, ich eine bestimmte Entscheidung, also für eine Instanz kann ich erklären, wie die Entscheidung für diese Instanz getroffen wurde. Und global bedeutet das, ich das Modell als Ganzes oder die Muster, die das Modell enthält oder gelernt hat, als Ganzes beschreiben kann. Also wenn ich mir ein Tree Viewer aufmache, dann ist das eine globale Erklärung. Das erklärt mir nicht nur für eine Instanz, wie sie klassifiziert wird, sondern für alle.

Und jetzt eben nochmal der kurze Rückblick auf die intrinsisch interpretierbaren Modelle, die wir kennengelernt haben. Wir haben lineare Modelle kennengelernt, also zum Beispiel die logistische Regression. Letztlich gehört auch der Naive Bayes dazu, den haben wir aber auch gar nicht so im Detail angeschaut. Und lineare Regressionen natürlich auch. Das sind also letztlich

Letztlich ist die Vorhersage da immer eine gewichtete Summe von irgendwas. Also es können Logarithmen noch angewendet werden oder ich weiß nicht was, aber letztlich sieht es irgendwie so aus. Also ihr seht, hier gibt es noch Generalized Linear Models und Generalized Additive Models. Da kann man hier sozusagen auch diese Teile der Summe noch Funktionen anwenden. Ich will jetzt nicht ins Detail gehen, aber letztlich ist es so eine gewichtete Summe und man kann sich dann diese Gewichte anschauen. Also

Aus den Gewichten interpretiert man dann, was das Modell tut. Man kennt ja auch die Funktion. Das heißt, man weiß, die Vorhersage kommt über diese Formel zustande und die Gewichte sagen mir, was dabei zum Beispiel eine große Rolle spielt oder in welche Richtung etwas wirkt, also positiver oder negativer Koeffizient. Ihr erinnert euch.

Ja, das hat man schon sehr gut untersucht. Numerische Features werden damit besonders gut behandelt. Ich habe hier geschmeidig geschrieben. Also sie müssen nicht diskretisiert werden. Ich kann einfach numerische Feature hier reinspringen, mit einem Koeffizienten multiplizieren und muss nicht sagen, wenn die Schneehöhe größer als 83 Zentimeter ist, dann so und sonst so, sondern ich kann einfach...

Das ist jetzt vielleicht kein gutes Beispiel, aber kann die mit irgendwas multiplizieren. Was es eben nicht kann, das hatten wir ja auch schon kennengelernt, ist die Wechselwirkung zwischen Features zu erfassen oder multivariate Muster, also die Kombination von bestimmten Faktoren. Jedes für sich leistet einen Beitrag, aber die Kombination von verschiedenen Merkmalen leistet manchmal einen Extrabeitrag und der kann hier nicht erfasst werden. Die werden als unabhängig behandelt.

Und was man wissen muss, ich glaube, wir hatten das auch angeschaut bei der linearen Regression. Da ging es darum, vorher zu sagen, für wie viel ich mein Haus verkaufen konnte. erinnert ihr euch, da gab es diese nicht so intuitive Situation, dass das Modell gesagt hat, Häuser mit mehr Zimmern kannst du für weniger verkaufen als mit weniger Zimmern. Und dann haben wir festgestellt, man muss das erweitern.

Immer betrachten unter der Annahme, dass der Wert der anderen Features gleich bleibt. Also wenn ich zwei Häuser habe, die von der Quadratmeterzahl gleich groß sind, dann fängt das an, ein bisschen Sinn zu machen. Also zwei Häuser mit 200 Quadratmetern, das eine hat sechs Zimmer, das andere sieben, dann ist das mit sechs vielleicht attraktiver. Also eben, das muss man im Hinterkopf haben bei der Interpretation. Und ja, dann gibt es diese Spinner hier, die vielleicht besser funktionieren, weil sie auch nicht lineare Effekte erfassen können, wie es da oben steht.

Die werden dann aber auch gleich wieder weniger interpretierbar. Und dann hatten wir die Bäume und die Regellerner. Es gibt noch Entscheidungstabellen, Decision Tables. Im Prinzip geht es da immer darum, dass man so Wenn-Dann-Regeln hat. Also jeder Baum lässt sich ja auch als Menge von Wenn-Dann-Regeln darstellen. Also wenn ich den Weg von der Wurzel eines Baums runter zu einem Blattknoten gehe,

dann habe ich lauter Bedingungen, die ich antreffe auf dem Weg und die ergeben den Wenn-Teil und das, was in den Blattknoten steht, ist der Dann-Teil der Regel. Also das ist eigentlich im Prinzip alles das Gleiche. Letztlich sehr einfach interpretieren. Also das ist für uns intuitiv. Wir sagen, wenn das und das ist, dann gibt es SLA-Violation. Wenn andersrum, dann eben nicht.

kann auch offensichtlich Interaktionen zwischen Merkmalen, also Kombination oder multivariate Muster erfassen, eben durch diese Kombination von verschiedenen Bedingungen im Wenn-Teil. Und was aber schade ist, ist, dass es eben zwangsläufig, wenn ich so einen Baum zum Beispiel nehme, numerische Attribute immer diskretisiert. Und es ist auch instabil. Also tatsächlich habt ihr das vielleicht selber schon gemerkt.

Wenn man ein bisschen was ändert, zum Beispiel die Tiefe des Baumes, die man erlaubt oder irgendwelche anderen Parameter, kommen teilweise völlig andere Bäume raus. Ist euch das schon passiert, dass es sich stark ändert, auch bei kleinen Änderungen oder ihr Lastenattribut weg und plötzlich kommt ein ganz anderer Baum raus? Also das ist relativ häufig der Fall. Und dann, ja, Stichwort Vertrauen geht dann ein bisschen verloren.

Und natürlich können Bäume auch groß werden oder auch Regelmengen, sodass die Interpretierbarkeit dann doch wieder leidet. Ja, man kann vielleicht so ein bisschen sagen, wenn man viele numerische Features hat, dann würde man vielleicht eher dazu tendieren, so ein lineares Modell einzusetzen. Wenn man es viele kategorische hat und vielleicht auch erwartet, dass es multivariate Muster geben kann, dann eher ein logikbasiertes Modell. Okay.

Jetzt gehen wir weg von den intrinsisch interpretierbaren Modellen hin zu den Post-Hoc-Verfahren. Und ich will mit euch nur Modellagnostische anschauen. Also welche, die für jedes Modell

funktionieren. Das ist jetzt so eine Überblicksfolie. Also was haben wir da? Und wenn ich sage, was haben wir da, dann rede ich auch davon, dass wir es in Orange haben. Ihr könnt vielleicht das gleich schon in der Pause machen.

Ich zeige mal kurz was. Was bei mir? Keine Ahnung. Irgendwas, egal. Seht ihr das? Ich habe hier ein Explain-Tab. Habt ihr nicht, oder? Also da müsst ihr einfach hier zu den Options gehen und Add-ons und dann hier Explain auswählen. Könnt ihr in der Pause machen. Und ihr seht hier sind fünf Widgets, also überschaubare Sache. Und

Was ich euch zeige, ist zum größten Teil. Also die Partial Dependence Plots gibt es nicht in Orange, aber eigentlich sind die auch gut vertreten, werden gut vertreten durch Individual Conditional Expectations. Also ICE heißt das Widget, die Abkürzung hier von in Orange. Also, was haben wir? Na komm.

Es gibt die Permutation Feature Importance. Die ist global. Das heißt, die beschreibt global, welche Features sind wichtig für das Modell. Ja, was soll ich dazu sagen? Also es sieht, glaube ich, im Orange ein bisschen anders aus. Es ist einfach ein Bar Chart. Also je wichtiger das Feature, desto... Und es wird auch im Allgemeinen sortiert. Also das wichtigste Feature steht oben. Dann könnt ihr einfach mal sehen, welche Features...

sind wichtig in dem Modell. Und dann, sozusagen, will man vielleicht ins Detail gehen.

Typischerweise macht man die Analyse dann für die Wichtigsten. Also oft hat dieses Bar-Chart dann so eine Form. Also hier habe ich sozusagen die Features. Eins, zwei, drei. Und dann habe ich hier sozusagen die Importance. Also ich mache es jetzt mal als Kurve. Letztlich sind es natürlich Bars. Also oft sieht es dann so aus. Also da kann man sozusagen

Wenn man sich auf die ersten vier oder so konzentriert, hier kommt dann nicht mehr so wahnsinnig viel Importance zusammen. Und dann fokussiert man, also hier würde man auch sagen, okay, hier die ersten fünf, vielleicht das hier noch, würde man sich jetzt genauer anschauen. Weil was dieser Plot ja nicht sagt, ist, in welche Richtung wirkt das Feature. Also hier zum Beispiel Alter ist wichtig, ja.

Aber, also hier geht es um die Vorhersage von irgendeiner Art von Krebs, wenn ich es richtig in Erinnerung habe. Ja gut, es wird uns jetzt nicht groß überraschen, wenn ein höheres Alter eher das Krebsrisiko erhöht und ein jüngerer Alter es eher nicht erhöht. Und so wird es dann auch rauskommen. Also hier, das ist jetzt für das Alter mal gemacht. Ihr seht es hier, Age.

Das ist ein sogenannter Partial Dependence Plot, der jetzt wirklich nur für dieses eine ausgewählte Feature zeigt, also ich erkläre gleich, wie das funktioniert, aber es zeigt, wie die Vorhersage des Modells, also hier ist sozusagen die Cancer Probability, die Krebswahrscheinlichkeit, wie die abhängt vom Alter. Und ihr seht sozusagen hier so bei 42 ungefähr, steigt das Risiko stark an. So kann man das so interpretieren. Und das ist sozusagen...

werden wir gleich verstehen, wie das berechnet wird. Dieser Plot ist eine Mittelung über viele Patientinnen. Ich glaube, es sind Frauen alles. Und hier sieht man sozusagen jede Patientin einzeln. Also letztlich die gleiche Funktion. Es geht darum zu verstehen, wie hängt die Vorhersage vom Alter ab. Man sieht eben, das ist sozusagen im Mittel steigt das an, aber es gibt durchaus auch Patientinnen,

bei denen das Krebsrisiko nicht so signifikant ansteigt mit dem Alter. Und dann werden wir uns noch mit SHAP beschäftigen. Also SHAP basiert auf Shapley Values. SHAP steht für Shapley Additive

Explanations. Und die Shapley Values an sich sind eine lokale Methode und sagen mir sozusagen wirklich für eine einzelne Instanz,

Hier geht es, glaube ich, wieder um diesen Krebs. Das heißt, ich sehe das hier gar nicht für eine einzelne Instanz. Ich bilde dir nachher noch ein Beispielplot. Da gibt es Plots, die wirklich für eine einzelne Instanz mir erklären, warum wurde hier zum Beispiel ein hohes Krebsrisiko vorhergesagt. Also was sind die Risikofaktoren, die wichtigsten Features. Und der Summaryplot ist eigentlich global. Da sind sozusagen

alle Instanzen, also in dem Fall alle Patientinnen stecken da drin und sagen mir einerseits durch die Reihenfolge der Features hier, welche Features die wichtigsten sind. Also das sind nicht ganz die gleiche Reihenfolge wie hier, aber ungefähr. Das löst ja ein wenig Vertrauen ein. Und hier kann man dann auch, das zeige ich euch gleich wie, ablesen, welche Werte dieser

Abgemutet, zu welcher Vorhersage führen? Also es ist ein Betrieb, wenn ihr so wollt, eine Kombination aus allen dreien, die wir vorher hatten. Ich finde, man sieht es manchmal nicht so genau. Das ist nicht nur was, was ich finde. Ja, weil diese Kurven ja hier, ja, die sind ja nicht immer linear rauf oder runter. Und hier sehe ich sozusagen nur höherer Wert bedeutet weniger.

Höhere Krebswahrscheinlichkeit oder höherer Wert bedeutet niedrigere Krebswahrscheinlichkeit. Ich sehe aber nicht wirklich so eine Kurve, die mir sagt, ab welchem Wert zum Beispiel es stark ansteigt. Wollen wir Pause machen? Okay. Dann machen wir danach mit den Details weiter. So ein Viertel nach. Okay. Okay.

Also jetzt...

Ich will euch kurz die Ideen hinter den verschiedenen Ansätzen erklären, was auch dann hilft, hoffentlich die zu interpretieren.

Wenn man so ein bisschen weiß, wie es funktioniert. Also, bei Permutation Feature Importance funktioniert es so. Hier steht, die Wichtigkeit eines Attributs wird bemessen als der Anstieg des Fehlers. Also, Fehler ist variabel zu interpretieren. Ihr könnt angeben, welches Evaluationsmaß ihr verwenden wollt. Also, das kann Accuracy sein oder Area Under the Curve oder F1.

Ihr erinnert euch. Die ganzen Metriken, die wir gelernt haben, kann man hier nehmen und interpretiert es einfach als, wie viel schlechter wird das Modell, wenn ich ein Attribut zerstöre und die Zerstörung erfolgt durch Permutation. Also nehmen wir, was weiß ich, wieder das Alter von Patienten. Dann habe ich eine Testmenge von 100 Patienten,

auf der ich diese Permutation Feature Importance berechnen will. Und dann sage ich, okay, ich möchte wissen, wie wichtig ist das Attribut Alter. Und jetzt nehme ich einfach meine 100 Patienten und vertausche die Werte des Alters unter diesen Patienten. Irgendwie zufällig. Also jeder kriegt das Alter eines anderen Patienten zugewiesen zufällig. Also das ist einer der Nachteile, die hier aufgeführt werden, dass da unrealistische Instanzen erzeugt werden können. Zum Beispiel zwei Meter große Personen, die 30 Kilogramm wiegen.

Ist nicht unbedingt schlimm. Die Alterswerte werden vertauscht und dadurch wird dieses Attribut letztlich unbrauchbar. Da ich das Modell auf Daten trainiert habe, bei denen das Attribut Alter da war, muss ich jetzt auch irgendwas in die Spalte Alter reinschreiben und ich zerstöre sozusagen dieses Attribut, indem ich die Werte durch permutiere.

Und wenn Alter wichtig war für die Vorhersage des Krebsrisikos, sagen wir mal, dann sollte bei dieser Permutation das Modell deutlich schlechter werden. Wenn es sowieso nicht so ein wichtiges Attribut ist, dann wird sich an dem Score, was auch immer ich ausgewählt habe, zum Beispiel "Area under the curve", nicht viel ändern. Und das zerstört auch die Interaktion mit anderen Features. Also wenn ein Feature nur in Kombination mit anderen wichtig ist,

Dann wird es durch die Permutation auch zerstört und ich kann es sehen. Was ich eben brauche, was nicht bei allen anderen Modellen der Fall ist, ist der Zugang zum Ground Truth, also zu den Labels. Das heißt, ich brauche diese Testmenge und ich muss wissen, hatten die Patienten Krebs oder nicht am Ende zum Beispiel. Und dann kann ich das nutzen, um diesen Fehler oder diese Metrik zu berechnen.

die ja wichtig ist bei der Definition hier. Es ist ein randomisierter Prozess, das heißt, es kommt nicht immer genau das Gleiche raus. Meistens macht man es dann mehrmals und guckt, was passiert. Was passieren kann, das ist manchmal überraschend. Also überraschend drastisch, was da passieren kann. Wenn ich ein Feature hinzufüge,

was mit einem anderen Feature stark korreliert, dann kann die Wichtigkeit, also diese Permutation Feature Importance, des ursprünglichen Attributes extrem abnehmen. Warum ist das so? Also, okay, jetzt nehme ich das Beispiel aus dem Quiz, dann habe ich euch schon was verraten. Das habe ich eh schon gemacht. Da gibt es die Temperatur, von der irgendwas abhängt.

Und die gefühlte Temperatur. Natürlich korreliert das stark. Also die gefühlte Temperatur, da wird noch irgendwie die Luftfeuchtigkeit einberechnet und die Windstärke oder sowas. Aber korreliert natürlich stark miteinander. Jetzt könnt ihr euch vorstellen, wenn ich das Attribut gefühlte Temperatur nicht habe und dann die Werte bei Temperatur durch permutieren, permutiere und die Temperatur ist wichtig für die Vorhersage, dann wird dadurch...

die Performance des Modells stark leiden. Das heißt, die Permutation Feature Importance von Temperatur wird hoch sein. Wenn ich jetzt das andere Attribut hinzufüge, also die gefühlte Temperatur, was passiert, wenn ich die Temperaturwerte durchpermutiere? Das Modell wird sich einfach auf die gefühlte Temperatur stürzen. Das heißt, die Wichtigkeit von Temperatur kann dadurch drastisch geringer erscheinen, als sie vielleicht wirklich ist. Da muss ich ein bisschen aufpassen.

Gut, das war also das. Wie gesagt, wenn man sich nähern will der Frage, was tut mein Modell, dann ist es ein guter Startpunkt, erstmal zu wissen, was sind die wichtigen Features. Und dann fange ich an, auch mich dafür zu interessieren, wie hängt die Vorhersage des Modells von den konkreten Werten dieser wichtigsten Features ab. Da kann zum Beispiel der Partial Attendance Plot helfen. Wie wird der berechnet? Also,

Ich habe hier verschiedene Werte für das Alter zum Beispiel. Also das ist mein gewähltes Feature. Und das, was ich hier auf der Y-Achse abfrage, ist der durchschnittlich vorhersagewertes Modell, wenn alle Instanzen diesen Wert für das Alter hätten. Also ich nehme alle meine Patienten, also meine Instanzen, und ersetze deren wirkliches Alter durch das Alter 20 und dann 21 und 22 und so weiter und so fort.

Also ich habe gesagt, meinetwegen, wir haben 100 Patienten, mit denen wir das veranstalten. Dann sind die plötzlich alle 21 Jahre alt. Und ansonsten haben sie aber die gleichen Herz-Gut-Werte. Und dann kann ich sehen sozusagen, wie sich die hier vorhergesagte Krebswahrscheinlichkeit verändert.

Also dann sehe ich zum Beispiel den Anstieg bei etwas über 40 Jahren. Ist klar, wie es funktioniert? Das ist natürlich, steht hier drauf, Durchschnitt. Also es wird gemittelt.

Und das ist ein bisschen problematisch. Also erstens, erstmal ist es schön zu interpretieren. Man kann es auch so ein bisschen kausal interpretieren. Das ist okay sozusagen. Problem, ja, ich kann es für ein, ich könnte es dreidimensional für zwei Features machen, aber ich finde das nicht sehr interpretierbar. Ich würde es immer nur für eins machen. Und der letzte Punkt hier, die Bildung des Durchschnitts. Ja, wir werden es nachher sehen. Also bei den

ICE-Plots, da sehe ich wirklich dann für jede Instanz, für jeden Patienten zum Beispiel, die Kurve und dann sehe ich, dass es dann eine sehr große Varianz geben kann. Also es gibt vielleicht Patientinnen, bei denen das Krebsrisiko schon von Anfang an hoch ist und bei denen sich in diesem Alter um die 40 gar nichts verändert. Und das hier ist eben nur das Mittel und da kann es eine große Varianz geben, die durch dieses Mittel verdeckt wird oder Ausreißer oder wer auch immer, komische Verteilung.

Man sollte auch noch in Betracht ziehen, das ist hier unten abgebildet in diesem Plot, wie viele, also in dieser Testmenge von, sagen wir mal, 100 Instanzen, die ich verwendet habe, um den Plot zu erstellen, sehe ich, wie die Altersverteilung war. Das heißt, das, was hier hinten passiert, basiert nur auf sehr wenigen Instanzen. Das sollte mir klar sein. Das heißt, hier hinten wird der Plot etwas ungenauer. Hier vorne ist er vermutlich verlässlicher. Sollte man auch wissen. Ja, und es schaut eben immer nur ein,

Feature nacheinander an und Korrelationen beziehungsweise Interaktionen werden da ignoriert. Vielleicht können wir auch schnell noch, eigentlich sollte der jetzt kommen, finde ich. Das ist eben, letztlich ist es ähnlich. Ihr werdet sehen, in Orange gibt es noch eine gelbe Linie, die dann im Prinzip dem Partial Dependence Plot entspricht, also wieder das Mittel dieser Kurvenzeit. Und hier sehe ich ja sozusagen pro Instanz aus der Testmenge eine Linie.

Also nehmen wir mal, das hier ist eine Patientin und da sieht man, wie sich die Vorhersage, also die Krebswahrscheinlichkeit, die vorhergesagt wird, verändert, wenn man das Alter dieser Person verändert. Und im Mittel sieht man dann natürlich auch, dass es eben bei vielen Patientinnen so ist, dass um die 40,

diese Wahrscheinlichkeit, diese vorhergesagte Wahrscheinlichkeit ansteigt und hier sieht man eben auch, es gibt welche, bei denen ist es schon per se hoch. Da gibt es andere Risikofaktoren, da spielt das Alter keine Rolle oder keine große. Genau, also man sieht jetzt nicht nur den Mittelwert, sondern die ganze Verteilung. Es bleibt dadurch trotzdem intuitiv. Man sieht immer noch sehr schön diesen Knick hier bei 40. Natürlich bleibt es dabei, dass wir nur ein Feature darstellen können,

Und auch, dass es keine Interaktion einbezieht. Okay, jetzt mal zu den Shapley-Werten. Solltet ihr ein kleines Video gucken? Aber ich erkläre es erst mal und dann können wir gucken, ob das Video es besser erklärt. Das ist auf Englisch. Aber damit kommen wir, glaube ich, klar. Das ist gar nicht animiert, doch ist es wohl. Also die Shapley-Werte sind sozusagen die Basis für dieses Shap. Also Shap steht für

Shapley Additive Explanations. Der Mensch in dem Video sagt das auch gleich nochmal. Jedenfalls basieren die Shapley-Werte, das sind wirklich lokale Erklärungen, und die basieren auf spieltheoretischer Überlegung. Also da geht es um eine Koalition, die etwas zustande bringt. Und man muss das transferieren im Kopf sozusagen. Die Koalition, das sind die Features. Also Koalition, einfach ein Team sozusagen. Ein Team von verschiedenen

Features in dem Fall. In der Spieltheorie sind es vielleicht auch Menschen, die irgendwas zustande bringen. Und was hier zustande gebracht wird, ist der vorhergesagte Wert. Also zum Beispiel die Krebswahrscheinlichkeit. Und die Frage ist ja sozusagen, ich sehe das Ergebnis, also den Score, der rausgekommen ist. Und ich frage mich sozusagen, wie viel hat jedes Feature jetzt für diese eine Instanz, für diesen Patienten, zu diesem Score beigetragen? Und

Was man da macht, ist sozusagen, dass man verschiedene Koalitionen bildet. Also das ist auch ziemlich rechenintensiv, weil man sehr viele oder eigentlich alle Koalitionen durchprobiert, also alle möglichen Kombinationen von Features. Ich habe hier geschrieben, dass die Attributwerte einer Instanz einen Raum betreten, damit man es sich wirklich als Koalition oder als Team vorstellen kann. Und dann guckt man sozusagen, was passiert, wenn die alle da sind und die anderen aber nicht.

Und die Frage ist dann, wie sehr ändert sich die Vorhersage, wenn ein Attribut dabei ist oder eben nicht. Also wenn die Koalitionen, die ein Attribut enthalten, immer sehr viel erfolgreicher sind in der Vorhersage oder beziehungsweise, nein, nicht erfolgreich, um den Erfolg geht es nicht. Es geht darum sozusagen, ob sich der vorhergesagte Wert ändert. Also wenn eine Koalition ohne ein bestimmtes Feature immer sehr erfolgreich

sagen wir mal, geringe Krebswahrscheinlichkeit vorhersagt und dann, wenn das Attribut dazukommt, sich diese Wahrscheinlichkeit jeweils drastisch erhöht, dann ist es offensichtlich eins von diesen Features, die einen hohen Einfluss haben und ich kann auch die Richtung sagen. Also wenn eben die Wahrscheinlichkeit sich erhöht, dann ist es hier, deutet es nach rechts sozusagen, wo hier auf der X-Achse die Krebswahrscheinlichkeit abgebildet ist.

Ja, also wenn ihr dieses ganze Video schaut, dann kommt da auch noch eine Menge Mathematik und dann könnt ihr euch davon überzeugen, dass das eine solide Theorie ist. Da wurden auch Sachen bewiesen, wünschenswerte Eigenschaften. Es ist, wie gesagt, rechenintensiv. Man merkt das auch in Orange schon. Also es rechnet manchmal ein paar Sekunden und theoretisch müsste man alle Merkmale berücksichtigen.

macht man dann doch nie. Also man schaut sich dann die an, die am meisten beigetragen haben. Hier seht ihr eben sozusagen, unten gibt es dann so eine Zusammenfassung. 100 weitere Features haben also das beigetragen. Wollen wir das kurz anschauen? Ich glaube, jetzt muss ich kurz vielleicht so machen. Hier war das. Mal an den Anfang. Wenn ich jetzt teile...

Hi everyone, I'm Rob Giada, an engineer on the trusty AI team and I'm going to talk a little bit about my work with Shapley Additive Explanations or SHAP. Now, before we go into the specifics of SHAP and how it works, I first have to talk about the mathematical foundation it's built on, and that's Shapley values from Game Theory.

Schapley-Value wurde von Lloyd Schapley entwickelt, um eine richtige Lösung für die folgende Frage zu geben: Wenn wir eine Koalition C haben, die mit einer finalen Value V zusammenarbeitet, wie viel hat jeder einzelne Mitglied für diese finalen Value geholfen?

Was bedeutet das? Wir haben eine Koalition C , eine Gruppe von kooperierenden Mitgliedern, die zusammenarbeiten, um eine gewisse "Final Value V " zu erzeugen, die "Koalition Value" heißt. Das könnte etwas sein wie eine Kooperation von Arbeitnehmern, die zusammen einen bestimmten Profit erzeugen, oder ein Abendessengruppe, das ein Restaurantbill ausübt. Wir wollen genau wissen, wie viel jeder Mitglied zu dieser "Final Coalition Value" geübt hat. Welche Teil des Profits muss jeder Arbeitnehmer verdienen? Wie viel muss jeder Teil der Abendessenpartei für die Billung erzielen?

Übrigens ist das mit der Rechnung, finde ich, ein schlechtes Beispiel. Also es sei denn, man teilt sich die Sachen alle, aber sonst hat ja jeder was bestellt. Naja, vielleicht ist es doch interessant, wenn man sich alles teilt und man gucken muss, wer wie viel gegessen hat. Aber eben, wenn ich zusammenarbeite, um einen Profit zu erwirtschaften, dann ist es nicht so klar. Es geht um die Sachen, wo es auch darum geht,

was entsteht durch die Kooperation von zwei Koalitionsmitgliedern.

Er sagt dann nachher, dass diese Mitglieder, diese Koalitionsmitglieder, den Features entsprechen. Ich finde es gut, wenn man sich das von Anfang an vorstellt. Das hilft vielleicht. Wir schauen dann die jeweiligen Werte dieser beiden Koalitionen und vergleichen die Unterschiede zwischen den beiden.

Die Unterschiede sind die marginalen Beiträge von Teil 1 zu der Koalition, die von Teilen 2, 3 und 4 besteht. Wie viel Teil 1 zu diesem bestimmten Gruppen beiträgt. Wir enumerieren dann alle solcher Paare der Koalition, also alle Paare der Koalition, die nur abhängig davon differenzieren, ob Teil 1 eingeführt wird. Schauen wir uns die marginalen Beiträge für jeden an. Die mean marginalen Beiträge sind die Schapli-Werte für diesen Teil.

Wir können jetzt das gleiche Prozess für jedes Mitglied der Koalition machen. Wir haben eine richtige Lösung für unsere original Frage. Wir haben alle Schapli-Value. Mathematisch sieht der ganze Prozess so aus. Aber all das, was wir wissen müssen, ist, dass die Schapli-Value die gewünschte Anzahl der Beiträge ist, die ein einzelner Mitglied zu der Koalition macht.

Nun, das Konzept zu Modell-Erklärbarkeit zu übersetzen ist relativ einfach. Und das ist genau das, was Scott Lundberg und Sue In Lee 2017 mit ihrem Papier "Eine unifizierte Anwendung der Modell-Versprechungen" gemacht haben, wo sie "SHAP" entdeckt haben. "SHAP" verbringt das Problem der Schapli-Wertwerte, von dem, wie wir schauen, wie Mitglieder einer Koalition zu einer Koalition Werten beitragen, bis zum Punkt, wo wir schauen, wie individuelle Fähigkeiten zu Modell-Erzeugungen beitragen. Sie machen das so,

Also das sind jetzt Features. Ist klar, oder? Und da sieht man sozusagen, wenn das hier die Gesamtkrebswahrscheinlichkeit einer Patientin ist, wie viel jetzt die Farben dargestellt werden, die Features dieser Wahrscheinlichkeit. Ist es klar geworden soweit? Ich glaube, ungefähr da wollte ich aufhören. Das ist eine sehr spezifische Weise, eine, die wir mit dem Namen ihrer Algorithmen kennenlernen können. Schapli. Das ist eine sehr spezifische, additive Erklärung. Wir wissen, dass...

F-A-A-D-E. Okay. Ist es klarer geworden damit? Gut. Genau. Also das hier hatten wir jetzt besprochen. Das ist sozusagen die lokale Variante davon. Und dann kann man, wenn man einfach jetzt wieder zum Beispiel 100 Instanzen nimmt, so einen Plot hier erzeugen.

Hier haben wir jetzt wieder verschiedene Attribute. Hier sehe ich Punkte und jeder Punkt ist, wie es hier steht, ein Schapli-Wert für eine Instanz und ein Attribut. Also Schapli-Wert ist hier unten abgetragen. Wie viel trägt es bei zur Krebswahrscheinlichkeit? Je weiter rechts, desto eher ist die Vorhersage

Und also das ist sozusagen dieser Punkt hier, steht für den Werten des Attributs für eine Person in dem Fall, eine Instanz. Und hier seht ihr noch, ob dieser Punkt für dieses Attribut einen hohen oder einen niedrigen Wert hat. Also man macht das normalerweise für numerische Attribute, da ist klar, wie es funktioniert. Also je höher der Wert, desto roter, je kleiner der Wert, desto blauer. Und für

Kategorische Attribute macht man normalerweise dann, wir werden es sehen, es passiert von selbst in Orange, wird es OneHot encoded, also was werden wir nachher haben? Hier sind alle numerisch. Ja, was weiß ich, nehmen wir beispielsweise Monat.

hatten wir ja gesagt, wäre vielleicht oft gut, es als kategorisches Attribut zu behandeln. Dann gibt es da eins für Monat Januar, eins für Monat Februar, eins für Monat März und so weiter. Und dann gibt es eigentlich nur Rot und Blau. Also entweder es ist Februar oder es ist nicht Februar. Also Rot würde dann heißen, es ist Februar und Blau würde bedeuten, es ist nicht Februar. Und dann seht ihr auch, dann klumpen sich die Blauen und die Roten sehr zusammen. Und dann kann man ablesen, ob im Februar

Hier passt jetzt nicht. Im Februar ist nicht mehr Krebs als sonst. Das war jetzt mehr so die Anzahl Besucher im Skigebiet. Das hatten wir schon. Jetzt sind wir eigentlich durch. Jetzt habe ich eine Aufgabe für euch. Viel Text. Ich habe es diesmal nicht geschafft, so ein Dokument zu erstellen. Ja, habt ihr es installiert? Das wäre der erste Punkt hier. Wollt ihr das mal machen, falls ihr es noch nicht gemacht habt? Wer hat es schon gemacht?

Erfolgreich, ja. Also rechnen wir nicht mit Problemen. Genau. Der Josef fragt sich, wie man die verbindet, die Widgets. Also bedenkt immer, hier zum Beispiel sehen wir Kurven für verschiedene Instanzen. Das heißt, ein Input, den diese Widgets brauchen, sind immer Daten. Also einerseits

Die meisten dieser Widgets, seid ihr da? Die meisten dieser Widgets haben als Input natürlich ein Modell. Das ist das Modell, was wir erklären wollen. Und andererseits Daten. Und auf diesen Daten werden dann zum Beispiel Vorhersagen gemacht. Mit all diesen Verfahren auf verschiedene Arten. Das heißt, diese Daten werden benutzt, um die Erklärung zu erstellen. Das heißt, ihr müsst die auch mit Daten füttern.

Im einfachsten Fall nehmt ihr einfach die gleichen Daten, mit denen das Modell trainiert wurde. Das ist nicht so ganz die feine Art. Normalerweise müsste man erst ein Train-Test-Split machen. Könnt ihr auch machen mit einem Data-Sampler oder ihr nehmt einfach die Trainingsdaten. Ist für mich auch okay. Das ist mal die eine. Hier habe ich geschrieben, beachtet deren Input-Parameter. Also in der Kurzdokumentation steht es ja auch eigentlich immer drin, dass man noch Daten einfüttern muss. Manchmal braucht man noch was.

Zusätzliches, also zum Beispiel hierfür, das ist ja lokal, das heißt, hier wird eine spezifische Vorhersage erklärt, da muss ich natürlich noch eine einzelne Instanz reinfüttern. Seht ihr dann, aber das ist eigentlich erst auf der nächsten Folie. Jetzt machen wir erstmal globale Interpretation, das wäre dann lokal. Kommt dann auf der nächsten Folie hier. Fangen wir mit Nummer 1 an.

Also, ihr habt auf Moodle noch die Daten von der Übung mit Fixit. Wir nehmen einfach wieder diese Daten. Damals haben wir einen schönen Tree gebaut und vielleicht noch, weiß ich nicht, hattet ihr noch einen Rule-Learner ausprobiert? Und jetzt schreibe ich euch vor, ein Gradient-Boosting-Modell zu nehmen. Also, wenn ihr wollt, könnt ihr auch ein Random Forest oder ein neuronales Netz nehmen. Das ist mir letztlich wurscht. Hauptsache irgendwas, was blackboxy ist.

Und was schon gut wäre, dass ihr es so hinkriegt, dass es eine gute Performance hat. Also vielleicht erinnert euch so ungefähr oder ihr hängt nochmal irgendein Tree dran oder so, um den Vergleich zu haben an einem Test- und Score-Widget, dass ihr seht, dass es ein gutes Modell ist. Was ich hier geschrieben habe, verwendet dabei zunächst alle Attribute. Wir hatten damals letztlich nicht alle Attribute verwendet. Und dann gibt es sozusagen die Punkte B und C, wo ich euch frage sozusagen,

oder hier unten habe ich geschrieben, erstellt einen Steckbrief, in dem ihr beantwortet, einmal, welche Features sind da am wichtigsten in dem Modell und wie lässt sich das Verhalten des Modells beschreiben, in dem Sinne, wie hängen die Vorhersagen von den Werten der wichtigsten Features ab. Also jetzt müsst ihr euch erinnern, welche Plots euch dabei helfen und die halt entsprechend in der richtigen Weise dranhängen. Und dann hatten wir vorhin schon wieder über die Target-Leakage gesprochen. Ich weiß nicht, ob ihr euch erinnert.

mit dem Days Open. Naja, ihr werdet es schon euch wieder erinnern. Die Frage ist, könnt ihr das erkennen mit den Mitteln, die ihr da verwendet? Und ja, okay, Days Open steht hier sogar auch drin. Wenn ihr das erkannt habt oder auch nicht, dann bitte das Attribut rausnehmen und die Schritte A bis C, eigentlich nur B und C, müsst ihr nochmal wiederholen. Und dann, ja, bitte kurz euch aufschreiben, was ihr dabei beobachtet.

Und dann machen wir Pause und den anderen Teil nach der Pause. Wahrscheinlich weiß Orange das noch, oder? Nein. Normalerweise merkt sich das alles nicht. Okay. Also, macht doch Gruppen, nutzt den Raum.

Arbeitet zusammen, drei bis vier Leute. Wenn es mehr als drei sind, setzt euch um den Tisch rum, damit ihr miteinander sprechen könnt. Das Übliche. Ist klar, was zu tun ist? Schon, oder? Also jetzt gehen wir doch vielleicht

Ein bisschen an der Folie entlang. Also das Grading-Boosting-Modell, wenn ihr alle Attribute verwendet, hat eine sehr gute Performance, oder? Lass uns das mal kurz verifizieren. Ich tue spezifisch auch Days Open rein. Alles, alles, alles. Und jetzt, das ist vielleicht nochmal wichtig, das würde ich gerne nochmal sagen für alle.

weil ich es jetzt doch öfter wieder gesehen habe. Also die richtige Art, Test & Score zu verwenden, ist wirklich, dass man alle Daten da reinschickt und jetzt meinetwegen ein Gradient Boosting Modell da dran hängt. Dann machen wir es so. Und dann kann man gucken, wie gut das ist. Jetzt weiß ich nicht, warum der so lange rechnet. Am Ende stützt der sich sowieso nur auf Days Open. Aber mein Rechner ist auch gerade nicht so fit. Ich muss den wahrscheinlich neu starten. Ja.

Also, das hattet ihr ja wahrscheinlich auch, oder? Überall 1,0. Perfekt. Viel zu perfekt. Und wir wissen ja von damals noch, was das Problem war. Die Target Leakage, ja? Und das war jetzt so die erste Frage, ob man die jetzt irgendwie rausfindet. Und wie gesagt, die richtige Art, Test & Score zu verwenden, hier, damals hatten wir das auch mit dem Tree so gemacht, dass wir den hier abzweigen, ja? Und das würde ich jetzt hier auch euch bitten, so zu machen. Ihr lernt jetzt das Modell. Und jetzt, was habt ihr als erstes genommen?

Die erste Frage hier war ja, welche Features sind generell am wichtigsten? Welches Widget habt ihr genommen, um die Frage zu beantworten? Feature and Points. Feature and Points. Ja, genau. Und wie schließt man das an? Also einerseits braucht es das Modell, ja, und andererseits braucht es eben Daten, auf denen es das Modell dann wieder laufen lässt. Mit verschiedenen Permutationen von Attributwerten.

die es aus diesen Daten erstellt und dann guckt, ob es besser oder schlechter wird. Und wenn ihr das aufmacht, erstmal sehen wir noch nichts, hier habt ihr dieses Dropdown. Und ich hatte ja erklärt, irgendwo, dass die Wichtigkeit des Attributs abhängt davon, wie viel schlechter das Modell wird, wenn ich das Feature mit Permutationen zerstöre.

Und ob es schlechter oder besser wird, es wird immer schlechter, da kann ich hier verschiedene Maße auswählen. Das sind die üblichen Evaluationsmaße, die wir kennen. Area under the curve ist jetzt mal nicht schlecht wahrscheinlich. Und da sehe ich diese Targeted Leakage, oder? Also ich sehe, dass das Modell nur dieses eine Attribut verwendet. Nachher, wenn wir Days Open rausnehmen, dann werden wir sehen, dass ein echtes Bar Chart entsteht. Habt ihr noch was anderes gemacht? Oder könnten wir noch was anderes machen, um noch mehr zu verstehen, was das Modell tut?

Oder wart ihr damit fertig, Marc? Nicht Confusion Matrix? Distributions, okay. Was kriegst du dann? Ja, also dafür brauchst du nicht vorher einen Test und Score eigentlich. Das kannst du ja...

auch so geben lassen. Also gut, es tut nichts Schlimmes. Ich wollte aber eher hier unten was machen. Ich würde gerne noch verstehen, für welche Werte von Days Open sagt es was voraus? Wie würdet ihr das machen? Hat das jemand gemacht? Ich glaube schon. Ich habe es, glaube ich, gesehen. Okay, ich könnte noch Explain Model anhängen, ja. Doch, das machen wir. Das ist gut. Und dann finden wir noch was Besseres. Also machen wir Explain Model hier. Ach, komm.

Ich hasse diesen Computer. Also gleiche Weise es anzuschließen, das Modell und diese Daten. Okay, und dann sehe ich sowas hier. Kleiner Wert, großer Wert. Wisst ihr noch, wie es war damals im Tree? Was hat der Tree gelernt? Der hat gelernt, dass es eine Violation ist, wenn Days Open größer als 10 ist. erinnert ihr euch? Das war ja auch die Definition von dem Service Level Agreement. Hat jemand das geschafft, das sichtbar zu machen? Hey, kommt jetzt, ich habe es doch gesehen bei euch.

Also was man noch machen kann, ist doch dieses ICE. Also mit ICE können wir verstehen wirklich, wie die Vorhersage eines Modells von individuellen Attributen abhängt. Leider geht es nur für numerische, aber wir haben Glück, weil Days Open ist numerisch. Also machen wir das mal. Und jetzt seht ihr hier für die meisten dieser Attribute passiert gar nichts. Und bei Days Open sehe ich jetzt, lass uns hier oben mal auf Violation gehen.

Genau das, was ich auch damals im Tree gesehen habe. erinnert ihr euch? So grob. Also ich sehe hier bei 10 nichts. Also unter 10 Tage sagt es vorher in Range und dann Violation. Okay. Aber das ist ja alles total sinnlos. Das haben wir aber, also wenn man diese Widgets ausschöpft, kann man eigentlich erkennen, dass dieses Gradient Boosting Modell das gleiche macht wie damals der Tree. Seht ihr das?

Und jetzt lasst uns den Anweisungen folgen. Das habt ihr auch gemacht. Also Punkt D können wir mit Ja beantworten. Und jetzt lassen wir Days Open weg und gucken mal, was passiert. Bevor ich das mache, will jemand von euch zusammenfassen, wie die Antworten auf B und C lauten. Welches sind generell die wichtigsten Features? Was habt ihr rausgefunden? Problem.

Ja, also ja, ich habe es ja schon öfter gemacht, die Übung, ich erinnere mich auch, Problem Category war auch beim Tree zum Beispiel wichtig und da gab es aber auch gewisse Kategorien, die eher schwierig waren. Also können wir mal aufschreiben. Hat jemand noch weitere Attribute als sehr wichtig herausgefunden? Ja.

Genau. Mit welchem Widget nochmal? Habt ihr es rausgefunden? Genau, der rechnet noch, aber gleich können wir gucken. Also Feature Importance hatten wir vorhin dann gesehen, dass Days Open das einzige war.

Okay, hier, ich weiß nicht, vielleicht hast du Agent ID weggelassen und dann kommt die Experience. Bei mir ist die ID, die erst kommt und dann die Ticketzeit. Und Explain Model, sagst du, kann ich auch dafür verwenden? Tatsächlich ist es so. Und da sind auch die Attribute nach der Wichtigkeit geordnet. Was ich hier aber sehe, ist,

dass jetzt die kategorischen Attribute OneHot encoded wurden, ja, und dann sehe ich die verschiedenen Werte von ProblemCategory hier als Attribute auftauchen. Habt ihr auch sowas? Nein? Bei ExplainModel? Habt ihr ExplainModel verwendet? Und bei euch sieht es anders aus? Joseph, bei dir sieht es anders aus, oder? Massiv anders? Also beim Jeremy sieht es auch ganz verrückt aus. Ich habe immer noch nicht rausgefunden, wie er da hingekommen ist, aber...

Er hat noch viel mehr Farben, ist viel bunter. Okay. Also, was man auch feststellen kann, bei Feature Importance Widget und bei Explain Model ist die Reihenfolge nicht immer komplett gleich. Aber sie sollte auch nicht komplett abweichen. Also es ist normalerweise schon einigermaßen synchron sozusagen. Was...

Was sieht man jetzt zum Beispiel? Welche Problem-Categories führen eher dazu, dass es eine Violation gibt? Kann man das hier sehen in dem Plot? Ja, sag mal Anna. Also, Führen die alle eher zu einer Violation? Ja, bitte.

Also ich habe hier oben als Target Violation ausgewählt. Das heißt, je weiter rechts ein Punkt ist, desto eher spricht es für Violation. Und jetzt siehst du hier bei Hardware sind die Punkte, die weit rechts sind, rot. Das heißt, Hardware ja. Und die blauen sind Hardware nein. Aber erinnert ihr euch? Access Login war unproblematisch. Und das sehe ich hier auch. Also ich sehe hier sozusagen die

Die roten, bei denen Access Login ja ist, bei denen spricht es eher gegen eine Violation. Also die sind eher weiter links. Und die hier sollten ja wahrscheinlich auf der Null sein. Also die, bei denen es sich nicht um Access Login handelt, wobei die sind ja nicht auf der Null, die sind dann ja entweder Hardware oder System oder Software, die sind eher gefährdet für Violation. Also sehe ich hier auch

Und wir sehen auch zum Beispiel Ticket-Type, Request versus Issue. Da sind auch die Farben vertauscht. Seht ihr das? Und wir haben damals ja auch schon gemerkt, die Requests sind schwieriger, sozusagen, führen eher zur Beihilfe. Genau, also hier ist jetzt sozusagen diese Limitation auch nochmal ein bisschen augenfälliger. Ich sehe hier, jetzt will ich ja nicht Days Open, sondern ich will diese Attribute öffnen.

analysieren und die Agent Experience ist die einzige, die numerisch ist und für die anderen, für die Problem Category und Ticket Type, finde ich hier einfach keine Erklärung. Das heißt, da muss ich dann hier reingucken und ich persönlich finde das, ja, doch, es ist eigentlich schon einigermaßen klar. Man kann einigermaßen sehen, bei binären Attributen, wie es funktioniert und die sind ja jetzt hier binär gemacht.

Habt ihr noch was rausgefunden? Oder irgendeine Frage oder Sachen, die euch auffallen, die euch unklar sind? Bei welchem auf dem Buch?

Ja, müsste ich gucken, was du für ein Modell gelernt hast. Hängt natürlich auch davon ab. Das kann schon mal sein. Ja. Den hier? Den ICE? Ja.

Ja. Also P steht hier für die Wahrscheinlichkeit. Also wie verändert sich die Wahrscheinlichkeit, dass es in Range ist? Ich würde jetzt vielleicht hier auch wieder auf Violation umstellen, dann dreht sich alles um. Ja.

Also es ist nicht so konklusiv, sage ich mal. Also je erfahrener die Agents werden, also der Mittelwert geht runter sozusagen. Ja, das macht Sinn. Also erfahrenere Agents haben weniger Violations. Aber man sieht eben, Kate und Jane und Tom machen trotzdem Probleme. Ich weiß nicht mehr, wie sie, aber ihr erinnert euch.

Wobei es natürlich für, sagen wir mal, kategoriale Attribute mit mehr als 10 Werten oder so dann nicht mehr so cool ist wahrscheinlich. Hast du viele Farben. Ich gehe davon aus, dass die Hellsaun Access Login sind und die rechts, weiß nicht, welche Farbe die haben, sind die Hardware. Ja. Ja. Ja. Du kannst eine komplementäre Sichtweise einbringen. Ja. Wollt ihr mir noch was erzählen? Ja. Also wie du willst. Das war einfach ein Video, welches

dann kann man es wirklich perfekt machen. Es muss wirklich nur die Nullwerten davon sein. Und dann muss man versuchen, die Nullwerte zu verändern. Was mich noch bewundert hat, ist, der Fall ist noch aufgetreten, weil wir jetzt ein Negativbild gemacht haben. Wenn wir das Negativbild nicht gehabt hätten, hätten wir zwischen den beiden, also No-Base, noch ein Zwischenblatt gemacht, das wäre dann mit No

Und alle, die keinen Parkinson hatten, hatten dann natürlich Morph. Aber es gab auch welche, bei denen der Parkinson diagnostiziert worden war und du das Date einfach probiert hast. Und bei denen, glaube ich, hättest du 94% Accuracy oder so. Aber es war trotzdem eine Frage. Aber kann man damit wandeln? Oh nein.

und dass auch die selben Pläne ausgebucht werden. Eine weitere Frage an den Gast, du hast ja vorhin gesagt, dass wir so eine Art und Weise berechnen, die wir verwendet haben, die wir verwendet haben, die wir verwendet haben, die sollen wir ab

Also, ähm, duschen mit einem Alkoholpust, den wir zum Zeitpunkt der Verstehung schon geben, oder? Also, nimm doch einfach den Alkoholpust, der du aus der Aufwachung zählst. Ah, was? Aha, aus der Rückgabe-Stelle? Ja, oder was? Ja, ja. Also, tschüss.

Ja, das ist...

Weißt du, das ist ja nicht einfach nur ein bisschen das Liebste, aber auch so Qualität, wo es so eine Situation gibt, wo man auch so ein bisschen überlegt, wie man eigentlich nicht meint. Zumindest so. Aber es ist halt so. Ja. Ja.

Nass. Ich habe dir nass gesagt. Das ist für den Schulrepräsentator gut. Das kann man wirklich brauchen. Wenn du Fidschen holst. Ja. Richtig. Das ist ja gut. Ich verstehe es immer noch. Das kann ich vielleicht erst mal. Ich bin eben auch dabei. Ja.

Fast alle wieder da. Ja, jetzt zwei und drei. Ich glaube, zwei geht wahrscheinlich schnell. Also ich glaube, damals, als wir mit den Daten zum ersten Mal gearbeitet hatten, hatten wir oder hat ihr die meisten von euch Entscheidungsbaum am Ende gewählt. War so, oder? Und da hattet ihr ja auch Folien gemacht.

Für das Fix-It-Management. Das könnt ihr euch ja nochmal anschauen. Also was hier das Ziel ist, einfach nochmal zu gucken, was ihr bei dem Tree zum Beispiel rausgefunden habt an Interpretationen und das zu vergleichen mit dem, was euch diese neuen Widgets jetzt mitteilen. Also wir hatten ja vor der Pause gesehen, was man da alles rausfinden kann. Und dann interessiert mich, ob da irgendwelche Unterschiede existieren und vielleicht auch irgendwas, was man...

mit dem einen erreichen kann und mit dem anderen nicht. Also wenn ich das eine sage, dann meine ich den Tree und das andere dann die Feature Importance, Explain Model und so weiter, neuen Widgets, die wir jetzt benutzt haben. Und wenn man irgendwas nicht erreichen kann, ist das schlimm. Vielleicht, wenn ihr anfangt, darüber nachzudenken, ob diese Fußnote hier noch sind oder auch nicht. Wollen wir erstmal das machen? Hier habe ich geschrieben, logistische Regressionen.

Wisst ihr noch, wie das Widget heißt, mit dem man sich die logistische Regression erklären lassen kann? Genau, Nomogramm. Macht ihr mal ein Tree mit einem Tree-Viewer hintendran oder schaut euch die Folien von damals an und dann so ein Nomogramm und dann guckt ihr mal im Vergleich, was ihr rausfinden könnt und was mich ein bisschen philosophieren wird. Also, mich interessiert vor allem das hier. Ist da irgendwas, das euch in den Sinn gekommen?

Oder habt ihr was entdeckt, was ihr zum Beispiel mit dem Tree könnt, was ihr jetzt mit den neu kennengelernten Widgets nicht könnt? Gibt es da irgendeinen wichtigen Unterschied? Also es fängt hier oben an und dann gehst du runter zu der Vorhersage.

Und was hast du dann, wenn du unten angekommen bist? Was macht er konkret? Ja, ab wann? Also diese...

Diese Diskretisierung von numerischen Atomhofen einzeln, die gefällt dir, oder? Ja, das kann man sagen. Auch wenn die Agenten von diesem Artikel mehr als zwei Jahre Erfahrung haben, dann ist es ein Problem. So in diesem Sinne. Ja, können wir es noch genauer fassen. Was haben wir damals gemacht?

Was habe ich euch damals aufgetragen? Also ich habe jetzt gerade in der Diskussion hier mit Jeremy gemerkt, dass es eigentlich ein super Beispiel ist, noch besser als das, was ich auf der Folie habe. Wir haben damals auch Handlungsempfehlungen abgeleitet. erinnert ihr euch? Also hier gibt es zum

Beispiel, haben wir gesehen, so ein Muster. Wir haben uns ja vor allem für die roten Knoten interessiert und haben gesagt, okay, lass uns mal gucken, wo die Zahlen hier, die absoluten Zahlen auch besonders groß sind, weil da viele Tickets reinfallen.

Und dann hatten wir gesagt, okay, das hier ist interessant, weil da sind sehr viele Tickets drin. Und was ist hier das Muster? Also Problem Category Software aus System, eher weniger erfahrener Agent und dann Request. Also Software System Request, weniger erfahrener Agent. Hat noch jemand im Kopf, erinnert ihr euch noch, was für eine Handlungsempfehlung wir da ausabgeleitet haben? Nikola? Vielleicht den Agenten

Was würden die Probleme am höchsten sind, wenn die Arbeiten mit Agents zu vollen Firmen aus Erfahrung kommen? Ja, also entweder solche Software System Requests nicht den Unerfahrenen geben. Oder, hatten wir gesagt, vielleicht können wir auch die Unerfahrenen spezifisch auf diese Arten von Tickets trainieren. Also das ist ja das Schöne. Wir wissen, die unerfahrenen Agents, die haben keine Probleme mit Access oder Login.

Und nach dem Baum hier auch nicht unbedingt mit Hardware. Hardware ist generell schwierig. Die haben spezifische mit Software und System Probleme im Vergleich zu den erfahreneren. Das heißt, wir könnten sie darauf spezifisch trainieren. Oder eben, wie du sagst, wir geben die den anderen. Können wir das auch rausfinden, diese Handlungsempfehlung, anhand von, wo habe ich das jetzt? Bei mir war es auch sehr langsam, deswegen habe ich das alles gelöscht.

anhand von Feature Importance, X-Plane Model oder ICE und wenn nicht, woran fehlt? Ich glaube, dass ihr das schon eigentlich rausgefunden habt teilweise. Was hat der Tree, Nikola? Was hat der Tree, was diese Widgets mir nicht bieten? Ich hätte das einfach mal bei...

Ja, also ich sehe zum Beispiel Agent Experience ist wichtig, Problem Category ist wichtig. Das waren ja die zwei und Ticket Type. Das waren die drei Features, die in dem

Muster mitgespielt haben. Und ich kann vielleicht auch ein bisschen sehen, so wie es bei der Agent Experience aussieht. Okay. Also je erfahrener, desto weniger Violations. Also worauf ich hinaus will, ist das Wort multivariate Muster. Ja, diese Kombination. Aber es ist wirklich diese Kombination, aus der ich meine Handlungsempfehlung ableite. Also es ist die Kombination, dass ich sage, unerfahrene Agent, also es sind ja die hier drüben sozusagen, ja.

Ja, hier sehe ich, okay, die haben vielleicht tendenziell mehr Probleme, aber ich weiß nicht womit. Und im Tree sehe ich, worauf ich sie trainieren sollte oder welche Tickets ich ihnen wegnehmen muss, je nachdem, was meine Strategie sein soll. Also ich sehe hier, es hat mit Software Systems Requests zu tun und das kann ich aus den anderen Widgets nicht ableiten. Steckt es denn da drin in dem Modell? Was glaubt ihr? Jeremy, wir hatten es schon diskutiert. Ja, oder Sterling, weiß nicht mehr.

Ja, es steckt drin. Es steckt drin. Also man sieht es ja an dem Bild hier, es sind mehrere Bäume, aus denen so ein Gradient-Boosting-Modell besteht. Das heißt, das Gradient-Boosting-Modell hat multivariate Muster. Solche, wie wir sie im Tree jetzt gerade gesehen haben. Es ist nur zu komplex, um die zu zeigen. Und wir können uns hier ein bisschen annähern, was das Modell tut. Aber die multivariaten Muster, die sehen wir nicht. Und das ist schade, weil wir zum Beispiel dann keine Handlungsempfehlungen abgeben können.

Ist klar. Also, ihr könnt natürlich übrigens auch hier ein Tree dranhängen. Oder nehmen wir den da. Jetzt auch gucken. Jetzt weiß ich gar nicht, was passiert, ehrlich gesagt. Aha, jetzt hat das gecappt

zum Gradient Boosting. Ja, also das kann ich auch ein bisschen am Tree selbst sehen. Die Problem Category ist die Wurzel von dem Baum. Aber hier kann ich es mir auch mit den gleichen Features anzeigen lassen. Ja.

Also ich kann natürlich diese Widgets auch für Modelle verwenden, die selbst sowieso schon intrinsisch erklärbar sind, was begrenzt Sinn macht. Okay, habt ihr noch irgendwelche Fragen? Sollen wir uns den letzten Punkt angehen? Na, muss ich doch noch sehen hier, aber er will nicht. Nee, ja, muss ich zumachen. Irgendwas stimmt da nicht. Egal, der letzte Punkt ist die lokale Interpretierbarkeit.

Also hier ist der Tipp, wenn ihr Instanzen, also Tickets auswählen wollt, könnt ihr das in einem Data Table Widget machen. Und die Idee ist jetzt, zwei Tickets auszuwählen. Eins, bei dem es SLA-Violation gegeben hat und eins, bei dem es in Range war. Und dann eine lokale Erklärung für die beiden Tickets, euch in Orange anzeigen zu lassen und die dann in Worten zu beschreiben. Letzte Aufgabe für heute. Seid ihr bereit?

Wenn euer Orange inzwischen sehr langsam ist und dauernd crasht, dann macht vielleicht ein paar Sachen wieder weg oder kappt die Verbindung, so wie ich es auch gerade eben gemacht habe. Um dieses hier zu machen, das sollte nicht so viel Rechenleistung benötigen, weil ihr ja nur jeweils eine Instanz euch erklären lasst. Es sollte einigermaßen schnell gehen und wenn ihr noch testet, scoret und so weiter, dann braucht ihr alles nicht mehr. Kappt ihr einfach die Verbindung, okay? Damit es schneller geht.

Ist klar, was zu tun ist? Es ist erstmal nicht so klar, weil ihr erstmal überlegen müsst, glaube ich, welches Widget ihr nehmt. Also ich kann so viel verraten, es ist eins, was wir noch nicht verwendet haben. Und dann, wir haben noch nicht so viel darüber gesprochen, aber ihr kriegt das hin. Lokale Interpretierbarkeit, vielleicht hilft auch das hier. Ja, so mittelfrott. Vielleicht eher das hier. Okay. So, ich habe so eine schöne Stille.

experimentieren, wie es gehen könnte. Aber jetzt hatte ich ja gesagt, ihr sollt mal mit X-Plane probieren. So, ich habe jetzt mal ein paar Verbindungen wegen der Performance. Also, wie muss ich es machen? Ich habe hier mein Modell. Und wenn ich jetzt ein Ticket selektieren will, dann mache ich jetzt hier mal ein Branch auf und mache dann Data Table hin. Und in dem Data Table...

Kann ich was selektieren? Jetzt machen wir es mal vielleicht ein bisschen smart. Nehmen wir erst mal Violation. Ist egal. Nehmen wir mal den hier. So, wie kann ich jetzt das erklären? Also, was ich brauche, ist dieses Widget, Explain Prediction. Und jetzt ist es, naja, wir machen es mal anders. Schließen wir es mal hier drüber an.

Der hat mir gesehen, der braucht zwei Data-Inputs. Einmal die Background-Data, aus denen er sich sozusagen seine Shepley-Values ausrechnet. Und dann Data steht für die Instanz, die ich erklären will. Das heißt, wenn ich das jetzt anschließe, muss ich es als Data anschließen, nicht Background-Data. Also so. Selected Data ist die eine Zeile, die ich selektiert habe. Und Background-Data, ja, kann ich mir jetzt im Einfachen

Fall von hier holen. Also wieder alle Daten, auf denen das Modell trainiert wurde. Nicht ganz die feine Art, aber wir machen es mal so. Und dann sehe ich auch was. Ich glaube, die meisten von euch haben was gesehen, oder? Sonst sah es vielleicht ungefähr so aus. Ja, jetzt kann ich hier oben wieder wählen, die Target Class und sozusagen, jetzt bin ich gerade ein bisschen verwirrt, vielleicht nehmen wir ein Beispiel, das noch deutlicher ist. Vielleicht irgendein Hardware Request.

Haben wir da einen? Ja. Ich will schon ein Beispiel nehmen, wo es besonders deutlich ist. Das ist doch schön, ne? Also ihr seht jetzt sozusagen, das nennt sich auch Force Plot. Das gibt es auch als Wasserfall. Also das hatte ich auf der Folie. Im Prinzip ist es immer das Gleiche, was da gezeigt wird. Nämlich manche Features, die sozusagen dafür sprechen. Also ihr seht jetzt, die einen zeigen nach links, die anderen nach rechts.

Und hier ist sozusagen nach rechts wird der Score größer, also der Score für die Wahrscheinlichkeit sozusagen, dass es Violation vorhersagt. Und hier geht das eben nach oben sozusagen, das heißt, die roten, die zeigen ja nach oben, die erhöhen die Wahrscheinlichkeit für die Violation. Das heißt, wir wissen ja schon immer, dass Hardware Requests problematisch sind. Das heißt, hier sehe ich, dass der Hauptgrund, warum es hier eine Violation gegeben hat,

Hardware ist und Request hat auch noch was dazu beigetragen. Ein bisschen was hat es geholfen, dass es kein Systems Ticket ist. Ja, hier ist so ein bisschen dieses One-Hot-Encoding für zum etwas seltsamen Effekt. Aber sozusagen hier sehe ich so ein Hardware-Request-Problem und das haben wir ja schon immer gewusst sozusagen, dass die schwierig sind und hier sehe ich eben für ein Ticket, was kann ich jetzt damit machen, wenn ich so ein Widget habe und

Jetzt kann ich das irgendwie bei Fixit verwenden. Also wenn ich das Modell global verstanden habe und die Muster nicht zu komplex sind, dann brauche ich es nicht. Aber was es mir natürlich mitteilt, wenn ich jetzt ein Ticket eröffne und das Gradient Boosting Modell sagt, Vorsicht, es wird eine Violation geben, dann kann ich mir das anschauen und kann verstehen, warum. Also dieses Modell,

Diagramm sagt mir, es wird eine Violation geben, weil es sich um ein Hardware-Request handelt und nicht um eins von den schönen, einfachen Access-Login-Problemen. Jetzt können wir es vielleicht noch. Also hier, ihr könnt auch zwei Branches aufmachen, aber ich finde es am einfachsten, jetzt einfach hier eine andere Zeile auszuwählen. Nehmen wir so ein Access-Login, da ist die Erklärung auch schön einfach. Da seht ihr auch der größte Beitrag, also die Vorhersage ist auch

von dem Modell, dass es in Range sein wird und die Erklärung ist hauptsächlich basiert darauf, dass es eben so ein schön einfaches Access Login ist. Wenn wir was nehmen, was leicht spannender ist. Also da sehen wir

Was war das jetzt? Man sieht gar nicht, welche Kategorie es war, aber es muss wohl Software gewesen sein. Ja, also die problematischen Kategorien, es hilft, dass es nicht eine von den problematischsten Kategorien ist und dass es ein Issue und kein Request ist. Und ja, also man sieht hier, wenn man sieht, dass die zwei Attribute spielen eine Rolle. Also

Problem Category und Ticket Type. Ja, ihr könnt auch noch sicher Erklärungen finden, wo die Agent Experience zum Beispiel auch noch eine Rolle spielt und ihr dann dieses Muster, was wir ja auch entdeckt hatten, dass weniger erfahrene Agents Probleme mit bestimmten Arten von Tickets haben, dann auch in der Erklärung für ein bestimmtes Ticket wiederfindet. Das finde ich jetzt so schnell hier, glaube ich, nicht. Ja, aber der Anwendungsfall ist klar, oder? Ja.

Ihr habt das Modell, das sagt vorher, Achtung, Problem. Und ihr wollt verstehen, warum. Und dann könnt ihr euch so einen Force-Plot anschauen. Oder in der Medizin habt ihr vielleicht ein sehr komplexes Modell, was irgendwie voraussagt, ob jemand eine bestimmte Krankheit hat. Und dann kann man schön sehen, welche Risikofaktoren bei den Patienten eben dafür sprechen und welche dagegen. Okay, jetzt haben wir... Eigentlich passt alles von denen hier durch, von den Widgets. Das hier finde ich persönlich nicht nützlich.

oder habe es noch nicht genügend verstanden, aber ich sehe keinen wirklichen Grund für euch, sich damit zu beschäftigen. Habt ihr noch Fragen? Dazu oder zu eurem Projekt-Assignment? Was auch immer. Nein? Okay, ihr kriegt noch eine E-Mail. Es gibt wieder ein Quiz. Daran erinnere ich euch. Und in die E-Mail schreibe ich auch noch Sachen da rein, damit es hoffentlich klar wird, wie es abläuft.

Okay, gut, dann sehen wir uns doch in jedem Fall, denke ich, nächste Woche. Hoffentlich einfach nach dem Plan für die Coachings. Den veröffentliche ich auch noch und mache den Link in die E-Mail. Bis dann.

Also ich habe jetzt nicht mehr so viel Interesse wie jetzt, weil ich jetzt auf dem Auftrieb und es jetzt keine Probleme gibt.