Gut. Ja, wenn es keine Fragen gibt, dann können wir auch direkt loslegen mit maschinellem Lernen. Vielleicht noch kurz die Frage an der Stelle. Ihr habt ja bei Einführung in die KI schon ein bisschen was über maschinelles Lernen gehört, oder? Also es ist jetzt nicht das erste Mal, dass wir darüber sprechen oder dass ihr davon hört. Eine praktische Frage, das Tool Orange, habt ihr das da auch schon benutzt?

Okay, das war letztes Jahr anders. Gut, dann schauen wir mal, wie es läuft gleich. Wir machen gleich ein kleines Game und da benutzen wir Orange zum ersten Mal und da wird dir ein bisschen kalte Wasser geschmissen, aber das kommt schon gut. Also, was ist maschinelles Lernen? Ich fange mal gerne damit an. Warte hier auf den Bild. Sollen wir das Fenster zumachen? Danke.

Also, was seht ihr? Also keine Angst vor trivialen Antworten? Ein Apfel und eine Birne. Ja, genau. Ein Apfel und eine Birne. Jetzt muss ich dich fragen, wie ist dein Name? Jan. Hast du dieses Bild schon mal gesehen, dieses Spezielle? Nein, das habe ich noch nicht. Hast du noch nicht gesehen. Aber auch euch anderen geht es so, ihr würdet zustimmen, das ist ein Apfel und eine Birne auf dem Bild. Wie ist das denn möglich? Also ihr habt dieses Bild noch nie gesehen, keiner hat es euch verraten, was man da sieht und ihr wisst,

Und ich weiß es auch, da ist ein Apfel und eine Birne drauf. Wie kann das sein? Du hast dir gemerkt, welche Form ein Apfel hat? Genau. Ich begleite es einfach mit meiner Erfahrung, was es sein soll. Oder was ein Apfel ist.

ist dein Schatz ja schon. Also du hast schon viele Äpfel gesehen in deinem Leben. Ja. Also wie nennen wir das im maschinellen Lernen, was du da gesehen hast? Also ihr merkt schon, worauf ich hinaus will, oder? Ihr habt was gelernt. Du hast von den vielen Äpfeln gelernt, die du gesehen hast. Genau. Musst du vergleichen. Ja. Wir haben auch x Äpfel.

Beispiel, also halb bewusst, halb unbewusst im Kopf, wie ein Apfel aussieht, wie eine Birne aussieht. Und wenn man jetzt noch ein nächstes Glas hält, dann gleicht das, gleicht das wie auch. Also Apfel hat mehr so eine Kugelform. Ich weiß nicht, wie ich die Form der Birne beschreiben soll. Vielleicht Birnenthermik. Okay, ist nicht besonders gut. Und ja, Muster, genau. Ja, also am Ende würde ich sagen, das, was wir hier haben,

in unserem Kopf, vielleicht gibt es noch mehr, außer der Form, was wir benutzen zum Vergleich. Das nennst du ein Muster. Das Muster besteht irgendwie, wie heißt das hier, Form? Erinnert ihr euch von dem, was ihr gehört habt über maschinelles Lernen? Ja, Features. Ja, super. Features, also hier gibt es Features. Hat jemand noch eine Idee für ein Feature, das wir nutzen, wenn wir Äpfel und Birnen unterscheiden? Genau, also das hilft uns hier nicht.

Weil Birnen und Äpfel, ich glaube, die können rot, gelb oder grün sein. Die sind selten blau. Aber es hilft uns zum Beispiel vielleicht von Pflaumen zu unterscheiden oder anderen blauen oder violetten Früchten. Aber ich schätze mal, rot ist fast eher eine untypische Farbe für eine Birne. Wir haben es trotzdem erkannt. Wir erkennen sozusagen diese Muster und wir beinhalten auch, dass mal was ein bisschen anders ist als blau.

die ganz typische Birne. Worauf ich hier noch hinaus wollte, Jan, du hast das gesagt, dass du ganz viele Erpfischungen gesehen hast. Jetzt wollte ich noch den korrekten Term dafür haben, wenn wir in Maschinellen lernen, oder lernen überhaupt, wie können wir das nennen, was du da gesehen hast, woraus du das gelernt hast, das Muster? Aus dem Festlauf, das ist das, was

Nicht Testdaten, sondern auch mit T und Daten hinten. Das waren deine Trainingsdaten. Das waren Trainingsbeispiele. Trainingsdaten, Trainingsbeispiele. Nenn es mal Trainingsbeispiel. Also woraus besteht so ein Trainingsbeispiel? Aus dem Apfel, den du siehst. Weiß das? Also wenn man so aufwächst als Kind, dann sieht man viele Äpfel, oder? Und dann irgendwann weiß man, dass die Apfel heißen. Wie funktioniert das?

Also jemand muss ja auch wissen, einer hat ein Label. Ja, genau. Mit Labels. Vielleicht schon der richtige Term. Wahrscheinlich meistens unsere Eltern, die uns sagen, das ist nicht unbedingt explizit, die sagen nicht, das ist ein Apfel, sondern die sagen, du möchtest einen Apfel essen und halten uns einen hin und dann schließen wir daraus, das Ding nennt sich Apfel. Wir sehen es, wir sehen die Form, wir sehen die Farbe, wir sehen die Features und ihre Werte.

und wir sehen das Label, das ist ein Apfel und das ist dann für uns ein Trainingsbeispiel. Wenn wir das oft genug gesehen haben, dann sind wir in der Lage, auch hier den Apfel als Apfel zu erkennen, obwohl wir das Bild noch nicht gesehen haben, weil wir dieses Muster oder nachher werden wir es auch Modell nennen. Also wenn wir bei Menschen bleiben, dann ist es ein mentales Modell sozusagen von einem Apfel, das haben wir in unserem Kopf. Und dieses Modell

Ja, es ist wirklich wie so eine Art Bild eines typischen Apfels und einer typischen Birne. Das heißt, wir vergleichen nicht, also ich glaube nicht, Jan, dass du sozusagen diesen Apfel mit allen Äpfeln vergleichst, die du jemals gesehen hast, sondern du hast sozusagen aus all diesen Trainingsbeispielen, die du gesehen hast, dir dieses Modell in deinem Kopf gebaut, was sieht ein typischer Apfel aus und mit diesem Modell vergleichst du das Bild und dann macht es Klick. Gut, jetzt haben wir schon ein paar wichtige Begriffe.

von Machine Learning, die auf der Date, also Features, wir nennen die auch Attribute oder Merkmale, was einfach das deutsche Wort für Feature ist. Wir werden wahrscheinlich noch andere Namen dafür, uns werden noch andere begegnen, was einfach so ist, ich kann es nicht ändern, die Leute verwenden vielfache, vielfältige Namen dafür. Genau, wir haben gemerkt, dass man sowas wie ein mentales Modell

sich in seinem Kopf aufbaut, was aus Mustern besteht. Und diese Muster sind einfach Kombinationen aus Merkmalen und Features und ihren Werten. Also ich merke mir sozusagen, was ist die typische Form, was ist die typische Farbe von einem Apfel. Genau. Wenn es diese Labels gibt, dann sprechen wir auch von überwachtem maschinellen Lernen. Also wenn ich eins vorspringe, was wir jetzt angeschaut haben mit Äpfeln und Birnen,

Das nennt sich Klassifikation. Das ist eigentlich ganz intuitiv. Wir haben jetzt Obst klassifiziert in Äpfel oder Birnen. Und wir werden natürlich dann auch noch die zweite wichtige Art von überwachtem Lernen kennenlernen, die Regression. Also wenn wir klassifizieren, dann sagen wir etwas vorher, was kategorisch ist, eine kategorische Variable. Und Variablen habt ihr schon mal gesprochen, kategorisch, numerisch. Ich könnte zurückzugeben, wenn ihr schon darüber gesprochen habt.

Ihr widersprecht nicht, also gehe ich davon aus. Genau, wenn es etwas ist, was kategorisch ist, dann ist es Klassifikation. Wenn es eine Zahl ist, die wir vorher sagen, also die numetische Variable, dann ist es Regression. Ihr seht auf der rechten Seite, dass es noch andere Arten von Lärm gibt, unüberwachtes Lärm.

Eine sehr prominente Ausprägung davon ist das Clustering und Reinforcement Learning. Also Reinforcement Learning ist ganz sicher etwas, was der Manuel euch näher bringen wird, den vielleicht

schon mal näher gebracht hat. Clustering wahrscheinlich wird auch er machen, wie es im Moment aussieht. Also was ist das? Was passiert da? Ihr könnt euch vorstellen, dass wenn man Kindern zum Beispiel Bilder von Tieren gibt und ihnen gar nichts sagt,

Dann gibt es eine gute Chance, dass die Kinder anfangen werden, sagen wir mal, es sind Bilder von Hunden und Katzen, dass sie einen Haufen mit Katzenbildern, einen Haufen mit Hundenbildern machen. Oder wenn man ihnen aufträgt, dass sie die Karten irgendwie sortieren sollen, dann ist es relativ wahrscheinlich, dass sie es so machen werden. Das heißt, da gibt es keine Labels und trotzdem werden die Kinder...

Alle Menschen sind sozusagen gewohlt, dass wir Ähnlichkeiten erkennen und Dinge gruppieren, die zusammengehören. Und nichts anderes ist Clustering, also ähnliche Objekte zu gruppieren in einen Cluster. Wenn wir von Business AI reden, dann hat das natürlich auch Anwendungen. Anwendungen hiervon machen wir gleich noch ausführlich. Hier gibt es vor allem die wichtige Anwendung der Anwendung.

der Kundensegmentierung, also im Marketing nutzt man das häufig, um ja, ohne dass man vorher irgendwas festlegt, was es für Kundensegmente gibt, den entsprechenden Cluster-Algorithmus drüber laufen lassen oder man, ja, man macht das häufig zum Beispiel, man führt irgendwelche Umfragen durch und dann guckt man, okay, was gibt es da für Segmente in den Antworten sozusagen. Reinforcement Learning, da geht es um Agenten, die Aktionen ausführen sollen,

und lernen, gute Aktionen zu wählen, um bestimmtes Ziel zu erreichen. Und dieses Ziel ist durch eine sogenannte Reward-Funktion beschrieben. Das heißt, je besser die Aktionen gewählt sind vom Agenten, desto, also je näher er dem Ziel kommt mit der Aktion, desto mehr Reward gibt es und dadurch verstärkt sich, also Reinforcement, diese spezielle Strategie, Aktionen zu wählen. Okay, was wir jetzt eben

Stark in den Vordergrund stellen am Anfang dieses Moduls ist das überwachte Lernen. Also 90 Prozent der Sachen, die wir hier machen, sind überwachtes Lernen. Ich glaube auch in der Welt da draußen ist das so verteilt. Also sicher gibt es auch interessante Anwendungen von überwachten Lernen und sehr interessante Anwendungen von Reinforcement Learning. Da wird euch der Manuel dann noch viel dazu erzählen. Aber gerade Regression und Klassifikation sind sehr, sehr wichtig.

häufig angewendet. Dazu jetzt ein paar Beispiele. Also, was können wir damit machen, wenn wir von Äpfel und Birnen, ich habe das Bild Äpfel und Birnen da unten hingemacht, weil wir, wenn wir jetzt die Beispiele anschauen, uns das eigentlich immer wieder, uns daran erinnern können und sagen können, was sind hier die Äpfel, was sind hier die Birnen. Also können wir mal probieren. Nehmen wir mal Kreditvergabe. Ist ein angeblich Klassifikationsproblem. Also eine Bank will entscheiden, ob ihr von ihr Geld bekommen sollt oder nicht.

Was würdet ihr sagen, was sind da jetzt die Äpfel, was sind die Birnen? Kreditwürdig und kreditunwürdig. Ja, ja. Also wann bist du kreditwürdig? Wenn ich keine Gezahndung habe zum Beispiel. Ja, also das geht jetzt schon in Richtung Features sozusagen. Aber ich hätte jetzt noch gesagt sozusagen, du bist kreditwürdig, wenn wir davon ausgehen, dass du das Geld zurückzahlst. Also ich könnte sozusagen auch direkt vorher sagen,

Die Birnen sozusagen sind die, die das Geld zurückzahlen und die Äpfel oder umgekehrt sind die, die es nicht zurückzahlen. Aber es kommt aufs Gleiche raus. Genau. Und es funktioniert genauso. Also wenn die Maschine lernt, lernt sie im Prinzip genauso wie wir Menschen. In dem Sinne, dass wir der

Maschine Beispiele geben. Wir geben der Maschine Kreditanträge, ein paar Tausend. Und jetzt können wir der Maschine auch mitteilen, zum Beispiel,

wurde der Kredit zurückgezahlt oder nicht und dann fängt die Maschine an, Muster zu lernen. Was die Maschine von uns unterscheidet, zumindest bei dieser Art von Daten, bei manchen Daten ist es anders, bei Bildern zum Beispiel, aber bei solchen, wir sagen, strukturierten Daten, müssen wir der Maschine noch sagen, welche Features sie nutzen soll, beziehungsweise wir geben ihr die Features und die Werte dieser Features für jeden Kreditantrag. Also,

Grob gesprochen, wenn ihr einen Kredit haben wollt, müsst ihr Formulare ausfüllen und das sind schon mal auf jeden Fall Daten, die wir diesem Algorithmus geben. Da sind also verschiedene Sachen abgefragt, ob es Betreibungen gab zum Beispiel, was der Jan gesagt hat. Das heißt, wir haben hier verschiedene Features mit Werten und davon haben wir ein paar tausend Trainingsbeispiele plus jeweils das Label, ob das Geld zurückgekommen ist oder nicht.

Und dann fängt die Maschine an zu lernen, was macht einen typischen, jetzt denken wir mal negativ, also was ist typisch für solche Kreditanträge, bei denen am Ende nicht das Geld zurückgekommen ist. Und wir versuchen das zu lernen, diese Muster, damit wir dann vorhersagen können, wo wir lieber kein Geld oder keinen Kredit vergeben sollen. Betrugserkennung, das ist jetzt wahrscheinlich noch einfacher. Also was sind da Äpfel, was sind da die Birnen?

Sagen wir mal Versicherung. Also wenn jemand etwas verheimlicht zum Beispiel. Ja. Also wenn jemand etwas zum Beispiel an einen Begriff kommt, dass er eine Krankheit hat. Ja, also das wäre dann Betrug. Das heißt Betrugsfälle, also wenn ich einen Schaden einreiche mit betrügerischen Absichten sozusagen oder was verschweigt, dann...

Wobei, du meintest jetzt, bevor es überhaupt zur Polizei kommt, was verschweigt. Ah, okay. Ja, das könnten wir auch Betrugsfall nennen. Sozusagen schon beim Underwriting sozusagen. Ich hatte jetzt mehr gedacht wirklich an Schadensfälle. Ja, wo ich im schlimmsten Fall zünde ich mein Haus an. Ja, das heißt, Betrugsfälle sind die Äpfel und die normalen berechtigten Schadensmeldungen sind die

Die Birnen. Targeted Marketing. Äpfel, Birnen. Was mache ich beim Targeted Marketing? Ich denke, da geht es darum herauszufinden, ob es einen Sinn hat, einem bestimmten Kunden eine bestimmte Bewertung zu zeigen. Ob er darauf anspringt, ob er halt nicht. Also wenn er gewisse Dinge schon gekauft hat, dann ist das eigentlich ein Leben, um Dinge zu verposten oder so. Ja, also ich will was verkaufen, ja.

Und ich habe Kunden X und jetzt ist eigentlich, oder sagen wir mal, die Äpfel sind die Kunden, die darauf anspringen, wie du sagst. Ich will vorher sagen, wer wird darauf anspringen? Das sind meine Äpfel. Und wer wird nicht darauf anspringen? Das sind meine Birnen. Ich versuche, aus alten Kampagnen, die ich gefahren habe und wo ich beobachte, wer wie anspringt, die Muster zu lernen. Also wie sieht der typische Apfel aus? Der typische Kunde, der auf mein Angebot anspringt.

Oder, okay, jetzt lasse ich euch mal in Ruhe, mit Äpfel und Birnen, also ich sage es selber, Kundenabwanderung, also klar, was sind da Äpfel, Birnen? Die Kunden, die bei mir bleiben, hoffentlich die meisten, sind die Äpfel und die Kunden, die mich verlassen, ihren Vertrag auflösen oder mich verlängern, sind die Birnen und wieder versuche ich eben Muster zu lernen, was sind die typischen Merkmale der Leute, die mich verlassen.

verlassen oder ihren Vertrag nicht verlängern und wenn ich diese Muster kenne, dann kann ich vorhersagen, wer aktuell gefährdet ist und dann vielleicht ein spezielles Angebot machen. Also ihr

seht auch jeweils, oder es ist glaube ich leicht zu sehen, dass da ein Wert dahinter steckt. Wenn ich solche Modelle lerne und sowas kann, sowas vorhersagen kann, dann kann ich zum Beispiel Geld sparen, indem ich

Besser, zuverlässiger und auch mit weniger Aufwand erkenne, wo jemand mich betrügen will und dann kann ich mehr Betrugsfälle erkennen und zurückweisen. Da brauche ich Geld, hier kriege ich mehr Geld zurück, was ich verliehen habe. Warum macht man Targeted Marketing? Also einerseits ist es für Kunden ärgerlich, wenn sie zu viel Werbung bekommen, die sie nicht interessiert.

Und andererseits, manchmal kostet es auch was. Also wenn ich zum Beispiel über irgendwie das Telefon oder so oder andere, sagen wir mal, teurere Kanäle Marketing mache, dann ist es da auch natürlich von Vorteil, wenn ich das gezielt machen kann und nicht so breit. Das kostet mich dann auch Geld. Ja, oder Spamfilter ist eigentlich auch Klassifikation. Also die interessanten, relevanten E-Mails sind die Äpfel und die

dies beim email sind die birnen und ich versuche zu erkennen was zeichnet spam aus oder auch irrelevant ist also es muss nicht unbedingt werden es kann auch statt e-mails irgendwelche anderen inhalte sein newsartikel oder so genau dann gibt es auch die regression über die wir noch nicht so richtig viel gesprochen haben ich habe ja gesagt wir werden sehen eigentlich vieles funktioniert sehr ähnlich wenn wir über klassifikation und regression reden

Der Hauptunschätzschied ist wirklich, dass eben das, was vorhergesagt wird, eine Zahl ist. Ich würde sagen, dass wahrscheinlich mit Abstand am häufigsten vorkommendes Szenario für Regression ist, dass man in irgendeiner Form Nachfrage vorhersagt. Also wie viel Stück vom Produkt X werden gekauft? Das ist das, was einem sofort einfällt. Aber Nachfrage kann auch in sehr vielen anderen Formen daherkommen. Also

Wie viele Leute kommen zum Konzert? Also wir werden einige Beispiele haben. Wenn man ein bisschen darüber nachdenkt, kommt man zum Schluss, aha, hier geht es um Nachfrage. Manchmal kann man auch zum Beispiel versuchen vorherzusagen, welchen Preis jemand für etwas zahlen wird oder bis zu welchem Preis man jemandem etwas verkaufen kann. Oder ich kann natürlich auch vorhersagen, was mich etwas kosten wird oder verkaufen.

in welchem Umfang irgendein Risiko eintreten wird, was ein bisschen das Gleiche ist. Oder ich kann auch Erträge vorhersagen, wobei Erträge oft, also Erträge vorherzusagen ist eigentlich oft fast das Gleiche, wie die Nachfrage vorherzusagen, weil Erträge meistens von der Nachfrage irgendwie abhängen. Manchmal kann ich auch, zum Beispiel, wenn ich Marketingkampagnen habe, kann ich vorhersagen, ob sie so viel

Rücklauf haben werden, wie geplant oder so viele Impressions oder Klicks oder was auch immer ich als Erfolg werte. Ja, also auch da, wenn wir jetzt überlegen, was ist der wirtschaftliche Nutzen, den ich ziehen kann, wenn zum Beispiel mir mein Regressionsmodell sagt, wenn du es so lässt, wie es jetzt ist, dann wirst du dein Ziel nicht erreichen, dann muss ich vielleicht noch Budget nachschießen. Also es hat auch da eine direkte Relevanz auf Entscheidungen.

Oder wenn ich sage, ich weiß, wie viele Produkte, wie viele Stück vom Produkt X ich verkaufen werde oder die Kunden kaufen werden, dann muss ich halt eventuell nachbestellen oder eben kann überhaupt entscheiden, wie viel ich bestellen soll. Also auch da hat es wirklich direkte Auswirkungen auf meinen Handel. Ja, wir werden diesen Zyklus hier, der CRISP-DM-Cycle, der steht für

Cross-Industry Standard Procedure for Data Mining. Also Standard Procedure bedeutet, das ist ein Standard, irgendwo definiert, weiß ich nicht. Wenn man den anschaut und mit Leuten redet, die Machine Learning machen in der Praxis, dann nennen die das vielleicht teilweise anders, aber eigentlich folgt das wirklich immer diesen Schritten. Also das ist wirklich das, was in der Praxis passiert.

Und ja, ich denke, es macht Sinn, wenn wir, also ihr werdet sehen, dass sich die Aufteilung der Themen im Modul zumindest die ersten fünf Lektionen an diesem Modell orientieren. Wir fangen heute an mit den zwei obersten Blöcken, Business und Data Understanding.

Ihr seht diese zwei Pfeile dazwischen, irgendwie gehört das zusammen. Also für mich ist das eigentlich quasi eine Box. Business Understanding, wir müssen verstehen, was wir wollen, das ist klar. Und wir müssen das übersetzen können in die Sprache von Machine Learning Tools sozusagen. Also in unserem Fall Orange. Was will Orange haben als Input?

Und wie können wir das, was die Business Stakeholder als Ziel formuliert haben, übersetzen in diesen Input oder in diese Form, die da erwartet wird. Genau, ich nenne das formalisierte Problembeschreibung. Dann kommt das, was, ja, wenn man das in der Praxis betreibt, macht das ungefähr 80 Prozent der Arbeit aus. Das wird uns nächste Woche quälen. Also ich werde euch wirklich ein bisschen quälen. Ihr könnt euch schon

Warm anziehen sozusagen, also in dem Sinne, das wird tough. Wir arbeiten uns durch ein wirklich komplexes Beispiel durch. Wenn ihr mal irgendwo kurz aussteigt, ist das vielleicht nicht schlimm, es wird ja dann jetzt auch aufzeichnen und es wird euch dann wieder begegnen natürlich auch beim Semesterassignment und deswegen machen wir es auch, weil es wirklich eine Hilfe ist.

Hoffentlich dafür. Es ist nicht so wahnsinnig spannend eigentlich. Also oft wird es auch ausgelassen in Vorlesungen über maschinelles Lernen. Also manche Dinge, die in der Praxis passieren, sind auch wahnsinnig langweilig. Einfach Arbeit. Und natürlich gibt es aber schon ein paar Prinzipien. Die Prinzipien sind allerdings relativ vielfältig. Also es ist gar nicht so einfach, das so runterzubrechen, dass man das...

dass man das so in einer kurzen Zeit präsentieren oder einüben kann. Aber wir probieren es. Das ist so etwas, was meistens am intensivsten behandelt wird, wenn man über maschinelles Lernen spricht, die eigentlichen Algorithmen. Also mein Ziel ist nicht, dass ihr solche Algorithmen alle im Detail versteht mit all der Mathematik, die dahinter steckt,

Geschweige denn, dass ihr irgendwelche neuen Algorithmen entwickeln könnt. Also das ist nicht unser Ziel. Und mein Ziel ist eigentlich, dass ihr wirklich in der Lage seid, für das richtige Problem die richtigen Algorithmen auszuwählen. Und wir werden auch sehen, dass diese Algorithmen noch konfigurierbar sind. Die haben so ein paar Parameter. Und da gibt es aber auch so ein paar Gemeinsamkeiten. Also die Parameter, die heißen immer anders und haben was eben mit der Natur dieser Algorithmen zu tun.

Aber es gibt zum Beispiel bei fast jedem Algorithmus ein Parameter, mit dem man die Komplexität des Modells, was da rauskommt, also wie umfangreich ist das, wie kompliziert. Das kann man mit Parametern bei fast allen Algorithmen steuern. Wie gesagt, die heißen immer anders, aber ich versuche euch beizubringen, welche Parameter jeweils bei welchen der wichtigsten Algorithmen für diese Komplexität zuständig ist. Und damit kann man schon viel, wenn man das verstanden hat, kann man schon viel auch

richtig konfigurieren, sage ich mal. Das heißt, das ist das, was wir können wollen am Ende, dass ihr wisst, welcher Algorithmus ist jetzt der richtige für das Problem, was wir hier haben und wie muss ich den konfigurieren. Das ist aber meistens auch so ein Hin und Her zwischen der Evaluation, vielleicht unser wichtigstes Thema, und diesem Modellieren. Das heißt,

Auch wenn ich verstehe, was Parameter bewirken und Änderungen an den Parametern, muss ich trotzdem ausprobieren. Also ich probiere aus. Vorhin wurde schon das Wort Testmenge gesagt. Ich werde dann, wenn ich Evaluation mache, immer so eine Testmenge haben von unbekannten Daten, auf denen das Modell getestet wird. Und das wird nicht am Anfang gleich optimal sein. Da habe ich immer Iterationen und ich iteriere, indem ich zum Beispiel den Algorithmus wechsle oder die Parameter des Algorithmus anders einstelle.

Dass wir verstehen, was die Parameter tun, hilft insofern, als wir dann wissen, in welche Richtung wir gehen müssen. Wenn wir ein Ergebnis sehen, dann sagen wir, aha, vielleicht müssen wir die Komplexität erhöhen oder reduzieren. Genau. Hier gibt es noch einen zweiten Punkt. Wir werden dann, wenn wir hier unten bei Evaluation ankommen, auch noch kurz darüber diskutieren, was es für typische Probleme gibt und wie man die vermeiden kann. Genau, das ist unser Modell. Und jetzt lasst mich nicht lügen.

bin ich schon weitergegangen, als ich eigentlich wollte, aber das macht nichts. Jetzt würde ich gerne mit euch dieses Spiel spielen. Also ich wollte eigentlich schon da aufhören, wo wir das Wort Muster auf der Folie stehen hatten und wo wir erkannt haben, dass es so ein mentales Modell in unserem Kopf gibt von Äpfeln und Birnen und ich wollte, dass wir so ein mentales, nee, nicht mentales, ein maschinengelerntes Modell erstellen mithilfe von Orange. Habt ihr das

erfolgreich installiert. Okay. Vielleicht, bevor ich euch jetzt ins ganz kalte Wasser schmeiße, nee, falsch, zeige ich mal ganz kurz so ein paar absolute Basics in Orange. Also, wenn ihr das aufmacht, dann sieht es bei euch ungefähr so aus. Auf der linken Seite ist bei mir ein bisschen mehr als bei euch, vermutlich, weil ich noch so ein paar Add-ons installiert habe. Manche von denen werden wir im Verlauf, werdet ihr im Verlauf vielleicht auch installieren.

Zum Beispiel das Explain-Addon, aber im Moment brauche ich es noch nicht. Das habt ihr, oder? Muss ich mal gerade gucken. Ich mache euch das Spiel mal verfügbar. Das ist ein Escape-Game. Moment, das mache ich noch nicht. Weil ich weiß, wenn der Link da ist, dann klickt ihr drauf und dann wollt ihr losmachen. Wir machen erst mal die Daten. Ach so. Also gut, melde ich mich noch mal an. Hm.

Ihr könnt euch den Datensatz schon mal runterladen. Das wird auf jeden Fall... Achso, ne, könnt ihr noch nicht, weil ich es noch nicht geschafft habe. Okay, sorry. Muss ich gerade gucken, ob ich ihn auch lokal gerade da habe. Also, wie funktioniert Orange? Orange funktioniert mit sogenannten Workflows und Widgets, die man in diesen Workflows zusammensteckt. Also die Dinger, die ihr hier seht, wenn ihr zum Beispiel sowas hier aufklappt, sind...

Komponenten, die irgendwas mit Daten tun und die nennen wir Widgets in Orange. Und eins der wichtigsten, was ihr eigentlich immer brauchen werdet, ist das File Widget. Das könnt ihr euch einfach anklicken, dann habt ihr es hier und sobald ihr es einmal benutzt habt, wird hier irgendeine Datei ausgewählt sein. Jetzt müssen wir nur die richtige wählen. Hier habt ihr euch runtergeladen. Vielleicht...

Ich glaube, ich bin über. Hat es geklappt? Nein. Also, wenn ihr, wahrscheinlich müsst ihr die Seite nochmal neu laden, wenn ihr es nicht seht. Und dann sollte Escape Game Doppelpunkt Datensatz bei

euch auftauchen. Okay. Okay, ich warte kurz. Redet ihr alle das oder seid ihr noch ein paar Schritte davon entfernt?

Jetzt sieht es noch nicht so aus.

Ich zeige jetzt nicht viel, weil einige Sachen werden euch auch in dem Game einfach dann nahegelegt, was ihr machen sollt und ihr werdet dann schon drauf kommen, wie das funktioniert. Das generelle Prinzip ist,

Dieses Widget hat jetzt die Daten geladen und jetzt könnt ihr sozusagen, hier gibt es diesen Halbkreis sozusagen oder Viertelkreis um das Widget rum und wenn ihr da drauf klickt und die Maus gedrückt lasst, also linke Maustaste und anfangt zu ziehen, dann entsteht hier eine Verbindung zu etwas, was noch nicht da ist, aber jetzt seht ihr, dass...

mir vorschlägt, was ich zum Beispiel mal verbinden könnte. Ich mache zum Beispiel mal ein Data Table. Also ihr könnt natürlich dieses File auch einfach mit eurem Texteditor oder in Excel aufmachen. Aber es gibt zum Beispiel dieses Data Table Widget, da kann ich einfach die Daten mal anschauen. Und ich sage vielleicht kurz, was das für Daten sind, damit ihr auch versteht, was hier die Äpfel und die Birnen sind. Also, das sind alles, vereinfacht gesagt, Bestellungen.

bei einem Online-Shop für, was verkaufen die? Office Supplies, Technology und Furniture. Also irgendwie Büroausstattung. Und jede Bestellung hat ein paar Attribute. Das sind hier die Spaltenüberschriften. Und es gibt hier hinten eine Spalte, die heißt Returned. Und das sind unsere Äpfel und Birnen. Wir wollen vorhersagen, unter welchen Umständen die Kunden ihre

Bestellung wieder zurückschicken. Wir wollen verstehen, was bringt Kunden dazu, ihre Sachen wieder zurückzuschicken. Vielleicht können wir das in Zukunft vermeiden. Das heißt, die Äpfel, die häufigeren hier, sind die, die ihre bestellte Ware behalten. Und die Birnen sind die, die sie zurückschicken. Also wo hier ein Yes steht, zum Beispiel wurde was zurückgeschickt. Und die Frage ist, können wir solche Muster finden? Wenn ja, wie sehen die Muster aus? Okay?

Eben. Also ich kann auch von hier weitere Widgets anhängen. Ihr seht, mir schlägt es hier was vor. Vielleicht kommt das auch im Game vor. Ich sage jetzt weiter nichts, weil ich glaube, das Game leitet euch dann. Wollt ihr eine Pause vorher machen? Ja? Okay. Dann können auf jeden Fall alle sicherstellen, dass sie soweit sind, bis wir wieder weitermachen. Also dieses hier schon mal...

erfolgreich bei sich haben. Und dann würde ich sagen, machen wir im Viertel nach weiter, steigen dann voll ein ins Spiel. Okay. Okay.

Jetzt will ich nicht mehr lange reden. Jetzt geht es los mit dem Spiel. Wenn ihr nochmals die Seite neu ladet, dann solltet ihr jetzt auch das hier sehen. Und ihr solltet es in Gruppen machen. Vielleicht so drei bis vier Leute. Nutzt den ganzen Raum. Da hinten ist noch Platz. Einfach euch mal über Eck zu setzen und zu beraten. Ja, pass auf. Also...

Vielleicht machen wir noch eine Bastelstunde. Aber nach dem Spiel. Einfach draufklicken und gucken, was passiert. Ups, das war das falsche. Entschuldigung. Also, ihr könnt oder ich glaube, ihr müsst sogar euren Namen eingeben. Und bevor ihr dann auf Spielen oder Play drückt oder was da steht, lest, was da steht und kopiert euch eventuell was.

Da steht nämlich gleich was. Ladet die Daten mit einem File-Widget in Orange. Könnt ihr schon abhaken? Habt ihr gemacht. Hängt ein Formular-Widget an. Die würde ich mir kopieren. Okay, und ihr

habt 45 Minuten, um zu entkommen. Sonst werdet ihr gekillt. Genau. Und dann, ja, dann muss ich nichts mehr sagen. Ihr müsst einfach...

euren Instinkten folgen und versuchen, euch von dem Spiel leiten zu lassen und sagt euch, was zu tun ist. Also das sind die ersten Anweisungen, die könnt ihr mal schon durchführen, bevor ihr einsteigt und dann geht's weiter. Ihr findet die Hinweise im Spiel. Also bitte macht Gruppen, okay, und setzt euch irgendwie um die Tische rum, dass ihr wirklich miteinander reden könnt und beraten könnt, wie ihr weitermacht.

Genau, also bei dem Barplot, da seht ihr ja, oder habt ihr alle gefunden, welches die schlechteste Provinz ist? Ah, okay. Also,

Die schlechteste Provinz ist die mit der höchsten Return Rate. Und das Game hat euch ja gefragt, ob das ein Muster ist. Erinnert ihr euch? Und ich habe jetzt auch schon mit der Gruppe da hinten diskutiert. Ein Muster ist es dann, wenn die Return Rate in dieser Provinz sehr viel höher ist als die globale Return Rate. Die war ja 10 Prozent. So weit waren ja alle. Und jetzt...

Das kann ich ja verraten. Ihr werdet es dann auch merken, was genau die schlechteste Provinz ist und was die für eine Returnrate hat. Aber die schlechteste Returnrate in den Provinzen, die ist nicht so wahnsinnig.

Also das könnt ihr jetzt auch, wie ihr schon draußen seid, nochmal verifizieren, wie hoch in dieser schlechtesten Provinz die Return Rate ist. Und die ist nicht viel höher als 10%. Und deswegen würde ich sagen, das ist kein Muster. Also ein Muster ist für mich was, Äpfel Birnen, ihr erinnert euch, sozusagen, wenn ihr eine Kombination von Variablen oder Features und Feature-Werten findet, wo die Return Rate sehr viel höher ist.

Und das tut ja dann der Tree später für uns. Und die Idee hier war, euch einfach zu zeigen, wenn man das selber sucht, also ihr könnt statt der Provinz auch mal die Verpackung nehmen, in der was verschickt wird oder das Kundensegment. Und ihr werdet jeweils merken, dass die Return Rate nicht so wahnsinnig unterschiedlich ist von der höchsten, auch nicht bei den schlechtesten Werten jeweils von diesen Features. Das ist sozusagen...

Hier auch ein Teil der Message, die leicht untergeht bei diesem Spiel, dass das wirklich...

Man kann manuell nach Mustern suchen. Also Orange ist dafür nicht das beste Tool. Da würde man vielleicht andere Tools nehmen, Tableau oder Power BI oder sowas, wo man schnell solche, oder Excel von mir aus, solche Grafiken erzeugen kann, dann sehen kann, okay, Return Rate aufgesplittet nach Provinz, aufgesplittet nach Kundensegment und so weiter. Gibt es da irgendwelche Ausreißer nach oben? Aber ihr werdet sie nicht finden. Aber es gibt eben diese...

Es gibt dieses Muster. Also denkt nochmal drüber nach, was ihr jetzt für ein Muster gefunden habt und warum das mit den Provinzen kein Muster war. Schaut nochmal rein, wie ist da die Return Rate in der schlechtesten Provinz. Okay? So, jetzt...

Also jetzt...

Nehmen wir uns einmal noch kurz die Zeit, dieses Game nochmal durchzugehen, weil es ja wie gesagt einen manchmal dazu verführt, einfach nur rauszuwollen und dann hilft es, wenn man sich nochmal klar macht, was man da jetzt eigentlich gemacht hat und hoffentlich was dabei gelernt hat. Also, das Erste war ja, okay, ihr müsst die, Entschuldigung, die zerschlagen hier.

Das hat noch nichts mit Machine Learning zu tun. Und hier war sozusagen unser erstes Ziel, mal zu verstehen, wie viele Bestellungen insgesamt zurückgeschickt werden und diese neue Variable, die wir definiert haben. Also ihr könnt es auch, wenn ihr das macht, also es geht ja hier über das Formula-Widget, dann habe ich euch noch verraten,

Ihr solltet die numerisch machen. Okay, das hat sich hier schon wieder alles gemerkt. Ich mache es nochmal neu. Also numerisch. Es ist numerisch, weil wir ja wirklich Return Rate, damit wollen wir rechnen. Wir wollen die, also was wir effektiv getan haben mit dieser Formel oder tun, Return Rate, wenn ihr diese Formel eingebt, dann ersetzen wir einfach Ja und Nein durch 1 und 0. Aber 1 und 0 können wir, oder die Einsen können wir auch addieren. Und wenn wir dann hinterher

Also wenn wir ein Average oder ein Mean berechnen, dann bedeutet das ja, wir addieren die Einsen auf und teilen durch die Anzahl Bestellungen. Das heißt, wir haben wirklich sowas wie die Prozentzahl, wie viele Bestellungen zurückgeschickt werden. Was so ein Formula Widget macht, einer von euch hat es gesagt, das ist sowas wie Feature Engineering. Ich glaube, Jeremy, du hast es, oder?

Oder einer von euch hier vorne. Egal. Jedenfalls könnt ihr das auch beobachten. Also wenn ihr hier jetzt wieder ein Data Table anhängt, dann ist die Return Rate als neue Spalte hier. Das heißt, wir haben jetzt sozusagen eine ganze Spalte erzeugt, die mittels dieser Formel berechnet. Also es ist wie wenn ihr in Excel hier eine neue Spalte nehmt, diese Formel eingebt und dann runterzieht. Das ist der gleiche Effekt sozusagen in Orange. Und was das Game aber von euch will, ist, dass ihr hier jetzt die Feature Statistics...

euch anzeigen lasst und dann müsst ihr hier bei Return Rate gucken und dann seht ihr, ja, also was bedeutet das, dass der Mittelwert berechnet wurde? Ja, es wurden die Werte alle aufsummiert, also sprich die Einsen gezählt und dann durch die Anzahl Bestellungen geteilt und das bedeutet, wenn man diese Zahl hier interpretiert, 10% der Bestellungen werden zurückgeschickt oder 10,39 und so weiter und das ist das, was ihr dann hier bei diesem Save eingebt.

10, 3, 9 und dann gibt es neue Instruktionen. Genau. Und jetzt hatte ich ja zwischendrin schon mal gesagt, was heißt es denn, nach einem Muster zu suchen? Also wir suchen jetzt sozusagen nach Konstellationen, wo wir eine besonders hohe Return Rate beobachten. Und jetzt wäre zum Beispiel eine Möglichkeit, dass es was mit dem Ort zu tun hat, wo sich die Kunden befinden. Also wenn sie jetzt aus Provinz X kommen, dann ist vielleicht die Wahrscheinlichkeit, dass sie was zurückschicken, höher.

jetzt gar nicht so eine wahnsinnig plausible Annahme, aber wir können es ja mal ausprobieren und ja, können es da oder da anhängen, vielleicht hier, hat sich gesagt, Group by, ja, und dann sage ich, lass uns mal nach Provinz gruppieren und dann kann ich einen Barplot machen und mir anzeigen lassen, vielleicht noch schöne Farben, nicht, dass das irgendwas bringt, aber

Die Return Rate. Also Achtung, hier oben nochmal die Values richtig auswählen, dass da Return Rate steht. Und klar, die größte Return Rate ist die schlimmste, also die schlimmste Provinz ist Ontario. Aber das fand ich eben wichtig, dass wir nochmal gucken, wie groß ist denn da die Return Rate, die ist bei 13%.

im Vergleich zu 10 Prozent nicht wahnsinnig unterschiedlich, also ist das kein Muster. Und jetzt könnte ich natürlich hingehen und sagen, ja, lass uns mal nach was anderem kopieren. Vielleicht hat es was mit der Verpackung zu tun, also Product Container. Und dann wähle ich hier wieder die Return

Rate und das ist die Container. Das ist die Jumbo Box. Bei der Jumbo Box ist es am schlimmsten, aber da sind es auch nur knapp 16 Prozent Return Rate. Also

Vielleicht findet man das schon eine interessante Abweichung, aber ich würde es jetzt noch nicht ein Muster nennen. So, jetzt kann ich alle von diesen Spalten durchprobieren und gucken, ob es vielleicht daran liegt. Und dann kann ich aber auch noch Kombinationen ausprobieren. Also ich kann gucken, vielleicht hat es was mit der Verpackung und dem Kundensegment zu tun oder mit Verpackung und Produktkategorie zu tun.

Eigentlich habe ich unendlich viele Möglichkeiten. Nein, endlich viele Möglichkeiten, aber sehr, sehr viele Möglichkeiten, die ich durchprobieren müsste. Und deswegen gibt es so ein Tool wie Orange. Orange hilft uns genau dabei, diese Muster zu finden, die wir manuell nur mit sehr, sehr viel Aufwand finden würden. Also natürlich hat man Ideen, irgendwas, was plausibel ist. Ich habe jetzt absichtlich ein Muster in die Daten reingebaut, was nicht plausibel ist, damit man nicht so leicht drauf kommt. Und Orange kommt natürlich aber trotzdem drauf.

Genau, das haben wir gemacht. Dann wissen wir Ontario und dann können wir die Fernbedienung aus der Schublade holen, Ontario eingeben. Nee, muss man richtig schreiben. Und kriegen die nächste Anweisung. Jetzt sozusagen Orange mit seinen Machine Learning Algorithmen als Rettung, um die Muster zu finden. Und eben, was da fehlt in dieser Anweisung ist dieses Select Columns Widget.

mit dem man Orange mitteilen kann, was das sogenannte Target ist, also was es vorhersagen soll. Und vorhersagen soll es ja, ob etwas zurückgeschickt wird oder nicht. Also Returned muss ich hier bei Target reinschieben. Ihr seht, ich habe hier auch noch zwei Features aussortiert. Dazu kommen wir auch gleich noch. Aber das macht nichts, wenn ihr sie drin lasst. Genau. Und dann sagt es, dann sollt ihr ein Tree bauen und der soll nicht so groß sein.

Also ich würde hier radikal sein und die Tiefe des Baumes auf zwei einschränken und dann kann man hinterher den Tree Viewer anhängen und anfangen, diesen Baum zu interpretieren. So sieht es bei euch aus, oder? Jetzt, wie interpretiert man sowas? Ein Baum hat die Vorhersagen immer in den Blattknoten, also auf der untersten Ebene steht, was der Baum vorhersagt.

Und die Bedingung, unter der das vorhergesagt wird, die ergibt sich, indem man hier oben anfängt und den Weg zu diesem Knoten abläuft, sozusagen, also zu diesem Blatt. Und auf der untersten Ebene seht ihr, es gibt die blauen und die roten. Die roten sind Yes und das ist das, was uns interessiert. Also wir wollen ja wissen, wann schickt jemand das zurück. Ihr seht, es ist hier nicht so gut dargestellt. Da hier drüben sowieso alles No ist und uninteressant, kann ich das mal wegklicken und sehe dann...

oder kann jetzt das Muster ablesen. Also das Muster, was ich meine, ist die Kombination aus Customer Segment gleich Small Business und Ship Mode gleich Delivery Trap. Also wenn ein Small Business etwas bestellt und es wird mit dem Lastwagen geliefert, dann schicken sie es zurück. Wie gesagt, das ist kein besonders plausibles Muster. Ich habe es einfach da rein gebastelt. Und das Game fragt dann noch danach,

wie oft das stimmt. Also jetzt müssen wir uns vorstellen, das ist jetzt ein Modell und das sagt vorher, dass alle Bestellungen, die mit einem Delivery Truck zum Small Business kommen, zurückgeschickt werden. Und diese Zahl sagt uns, dass das aber nicht immer stimmt. Also nur in knapp 60 Prozent der

Fälle stimmt das. Und so ist es immer beim Machine Learning. Wir haben Modelle, die sagen was vorher, was nicht immer stimmt.

Also wir werden selten oder eigentlich nie Modelle haben, die immer alles richtig machen. Weil wenn man so ein Modell hat, dann kann man das auch selber meistens als Regel hinschreiben. Dann ist das sozusagen festgeklopft. Wir machen Machine Learning dann, wenn es Muster sind, die eben nicht von vornherein klar sind. Und dann stimmen sie auch nicht immer, sondern eben nur meistens. Und so ist es ja auch. Jetzt kann ich das aber auch noch sichtbar machen, also damit ihr das Muster noch so richtig seht.

Ich kann jetzt mal nach diesen zwei Features, die das Muster umfasst, gruppieren. Also nämlich nach, was war es, Chipmode und Customer Segment. Also ich habe jetzt mit Steuerung gedrückt. Also erst Chipmode und dann mit Steuerung gedrückt halten, das Customer Segment. Jetzt mache ich wieder einen Barplot, wähle hier die Return Rate. Und jetzt kann ich mal noch sagen, also wenn ihr jetzt guckt hier, das sind glaube ich,

Mal schauen. Also wir gruppieren mal nach Chipmode und machen Customer Segment als Farbe. Also jetzt seht ihr hier die Legende, also die Small Business sind immer die Orangenen und hier unten ist Delivery Truck, Express Air und Regular Air. Und jetzt sieht man das Muster wunderbar, oder? Also ihr seht einfach diese riesengroße Säule bei Kombination Delivery Truck und Orange via Small Business.

jetzt habe ich noch hier die legende halb verdeckt so ganz perfekt geht es nicht genau was da auch noch interessant ist wenn man hier ein bisschen so drüber fährt mit der maus na komm ich wollte mir gern noch den count anzeigen lassen orange ist nicht das beste tool für sowas

Es sind aber, ich kann es im Tree nachgucken, da kann ich sehen, wie viele Bestellungen das insgesamt betrifft. Also hier unten sehe ich, dass das 229 Bestellungen sind. Das ist nicht wahnsinnig viel. Also die Daten insgesamt umfassen, glaube ich, ungefähr 8000 Bestellungen.

Das heißt, bei nur 229 schlägt dieses Muster an und dann in 134 von diesen Fällen wird es auch wirklich zurückgeschickt. Wenn ihr das durcheinander teilt, dann kommt 58,5 aus. Oder 0,58. Genau. Aber ich wollte einfach nur nochmal dieses Muster auch sichtbar machen. Da seht ihr sozusagen den Unterschied dann zwischen wie Ontario sich eingeordnet hat und wie hier diese Kombination heraussticht sozusagen.

Okay, habt ihr Fragen? Es waren naturgemäß jetzt ein paar Sachen dabei, die wir später noch besser verstehen werden. Das war jetzt einfach mal so ein Deep Dive in ein paar Sachen, die wir natürlich noch ausführlicher anschauen. Jetzt habt ihr auch mal Orange ein bisschen genutzt. Gehen wir zurück zur Theorie, zur Grauen. Wir waren schon weiter und wollen jetzt auch wieder gleich eine Übung machen und die ist ein bisschen eben

ist ohne Orange, das ist jetzt erstmal auf Papier sozusagen. Und zwar geht es um diesen ersten Schritt von unserem CRISP-DM-Cycle, das Problem zu formalisieren. Eine Sache haben wir jetzt in Orange schon gesehen, die das Formalisieren auch macht, und zwar dieses Target. Was wir am Ende haben wollen. Ihr habt es auch in Orange gesehen als Tabelle.

ist im Prinzip Tabelle mit Spalten und Zeilen und wir haben dafür Namen. Für die verschiedenen Dinge, also für die Zeilen und für die Spalten zum Beispiel. Fangen wir mal mit den Zeilen an. Also was Orange erwartet, ist, dass pro Instanz, für die wir etwas vorhersagen wollen, eine Zeile in der Datei existiert. Also in unserem Beispiel gerade eben,

war jede Bestellung eine Zeile. Das stimmt übrigens nicht, aber egal. Wir gehen mal davon aus. Also eigentlich gibt es mehrere Zeilen pro Bestellung manchmal, wenn mehrere Sachen bestellt wurden. Aber das ignorieren wir mal. Also eine Zeile pro Instanz, für die wir was vorhersagen wollen. Genau. Ich sage Instanz dazu. Der Unterschied zwischen den Instanzen und dem, was nachher die Spalten sein werden, ist,

Das ist sehr implizit. Also es steht nirgends, was die Instanzen sind oder wie die heißen. Das muss man wissen, das muss man im Kopf haben. Das ist vielleicht manchmal auch das Schwierigste, zu definieren, was die Instanzen sind. Also wenn wir zum Beispiel zurückgehen, wir machen es gleich als Übung. Die Spalten sind das, was wir Features nennen oder hier Features genannt haben oder auch, habe ich schon angekündigt, manchmal sagen wir Attribute oder Input Variables oder Independent Variables.

Ah ja, hier habe ich jetzt gesagt, ich entscheide mich für das Wort Attribute. Vielleicht werde ich auch manchmal Features sagen. Ihr müsst das dann beides verstehen. Und dann haben wir das, die letzte Spalte hier. Es muss nicht die letzte Spalte sein. Ihr habt ja gesehen, in Orange, wir können definieren, welches die Zielvariable ist. Oft macht man es in die letzte Spalte.

die das Target, manchmal heißt es Target oder auf Deutsch Zielvariable. Insbesondere bei Regression spricht man eigentlich immer von Zielvariable oder Target. Bei Klassifikation spricht man auch oft von Klassenattribut oder von Label, Class Label, Dependent Variable, Output Variable. Ihr seht es viel Englisch, weil die Abbildung ist auf Englisch. Man könnte das auch noch übersetzen. Also es gibt sehr viele Begriffe. Am besten, ihr kennt die alle, weil dann...

Kann euch keiner verwirren. Und in dieses Format müssen wir es bringen. Jetzt brauchen wir mal ein Beispiel, um es zu verstehen. Also formalisieren heißt, dass wir irgendwie einen Wunsch haben. Jemand wünscht sich, was vorhersagen zu können. Das sind irgendwelche Leute aus dem Business. Irgendwelche Stakeholder, die sagen, ich würde gerne was vorhersagen. Und was wir jetzt machen müssen, ist Daten, die wir irgendwo haben, in so eine Form bringen,

sodass die Dinge, für die wir was vorhersagen wollen, also unsere Instanzen, auf Zeilen dargestellt sind, also für jede Instanz eine Zeile und sodass wir auch wissen, welches ist unser, also was sind unsere Attribute oder Features, für die muss es jeweils eine Spalte geben und eine Spalte werden wir dann in diesem Select Columns Dialog wählen als Klassenattribut oder Zielvariable oder Target, wie es in Orange heißt.

Also letztlich die drei Sachen, die müssen wir definieren. Jetzt das Beispiel. Also wir nehmen mal ein Beispiel aus dem Bereich Marketing. Stellt euch vor, eine Firma, nennen wir sie Swiss Bikes, produziert Fahrräder, beziehungsweise vor allem haben sie auch Shops und in diesen Shops bieten sie auch Reparaturen an für die Fahrräder, die sie verkaufen und insbesondere einen Service jeden Herbst, den sogenannten Wintercheck. Also bevor der Winter beginnt, kann man sein Fahrrad dahin bringen und mal überprüfen lassen,

sind alle Schrauben angezogen, was mit dem Reifendruck und so weiter. Und dafür gibt es jeden November ein Mailing. Das haben die jetzt schon ein paar Jahre gemacht und die haben einfach immer alle angeschrieben, die jemals bei ihnen ein Fahrrad gekauft haben. Also sozusagen ein Mass-Mailing gemacht an alle Bestandskunden. Entschuldigung, nicht ganz alle, sondern nur alle, die schon mal zu einer Reparatur da waren. Und sie haben diesem Angebot einen Gutschein beigelegt mit 10% Rabatt auf den Wintercheck.

Das heißt, sie konnten messen, wenn die dann ihren Gutschein mitgebracht haben, ob die Kunden auf dieses Angebot reagiert haben. Das heißt, wer hat es eingelöst? Und jetzt kommt dieser Wunsch. Was wollen wir vorhersagen? Also in Business-Sprache formuliert, wir würden gerne dieses Jahr nur die Kunden ansprechen, gezielt, die dann auch Interesse haben und ihr Fahrrad zum Wintercheck bringen. Wie kann ich das formalisieren?

Die Instanzen, die werden Kunden sein. Also wir brauchen in dem Datensatz, aus dem wir das lernen wollen, eine Zeile für jeden Kunden oder jede Kundin. Dann werden wir Spalten haben und die Spalten, die müssen irgendwie diese Personen beschreiben und zum Beispiel am besten mit Attributen, die mir erlauben, was vorherzusagen, das Klassenattribut.

Ja, die Reaktion. Also hat die Person den Gutschein eingelöst? Ist mit dem Fahrrad zum Wintercheck gekommen mit dem Gutschein? Und was könnte helfen? Also vielleicht hilft es zu gucken, wie lang es her ist, dass die letzte Reparatur gemacht wurde, wie viele Reparaturen jemand schon hatte. Die Annahme ist natürlich, je mehr Reparaturen jemand machen lässt, desto eher kommt er auch zum Wintercheck. Vielleicht spielt es auch eine Rolle, ob jemand in der Stadt lebt oder auf dem Land oder was für ein Fahrrad er oder sie besitzt.

Da kann man brainstormen. Also wenn es um die Attribute geht, natürlich kann man angucken, was man hat, aber man kann auch noch sehr viel Fantasie reinstecken. Und oft kann man aus den Attributen, die sowieso vorhanden sind, in Daten, die man irgendwo sammelt, noch weitere konstruieren. Und die helfen einem dann vielleicht mehr als die, die schon da sind. Also so kann das aussehen. Ihr seht jetzt hier also zehn Zeilen. Jede Zeile ist ein Kunde oder eine Kundin von Swiss Bikes. Jeder

Personen hat eine Nummer und dann sind die Attribute hier und es gibt in der letzten Spalte diese spezielle, das Klassenattribut Response. Also wir sehen, vier von den zehn haben ihr Fahrrad gebracht und sechs nicht. Ja, okay, das ist unrealistisch hohe Response Rate. So ist es leider in der Realität nicht. Aber ist es klar, wie es funktioniert? Okay.

Gut, bevor wir das jetzt als Übung gleich machen, für Klassifikation und dann auch für Regression, vielleicht noch ein bisschen was zum Auswählen von Attributen. Also ich habe ja gesagt, man kann da so ein bisschen brainstormen und man braucht einfach auch ein bisschen Intuition. Also man muss sozusagen die Mechanismen ein bisschen vorwegnehmen und überlegen, ja, was könnte denn Einfluss haben und seine Erfahrung einsetzen.

Wichtig erstmal vielleicht alles, was helfen könnte, ist erlaubt, wie oft beim Brainstorming. Also bewertet wird später erstmal alles auf den Tisch bringen, was eventuell helfen könnte. Und viele von den Algorithmen, die wir dann später kennenlernen, unter anderem aber eigentlich auch die Entscheidungsbäume, die wir schon kennen, können auch zumindest zu einem gewissen Grad dann selber entscheiden, welche Attribute wichtig sind. Ihr habt es gemerkt. Also wenn wir diesen Baum anschauen,

der gelernt wurde und uns vorher anschauen, vielleicht irgendwo hier nochmal so ein Data Table. Also es gibt neben dem Klassenattribut 1, 2, Entschuldigung, 1, 2, 3, 4, 5, 6, 7, 8 Attribute und der Baum nutzt nur zwei davon. Das heißt, irgendwie hat der Baum entschieden, dass die anderen sechs Attribute nicht wichtig sind.

Also so eine Art Feature Selection, nennt man das, steckt in den Algorithmen oft drin. Und weil das so ist, ist meine Empfehlung immer, nehmt lieber mehr als weniger Attribute. Manchmal wundert man

sich, was dann doch noch hilft und lasst es dann eventuell den Algorithmus am Ende entscheiden, ob es hilft oder nicht. Okay. Es gibt auch...

Ja, genau. Also man kann dann noch prüfen, sozusagen auch Korrelationen sich anschauen zum Beispiel. Also Korrelation ist in Anführungszeichen zu setzen. Gerade wenn es um Klassifikation geht. Korrelation ist ja nur zwischen numerischen Variablen definiert. Erinnert euch so ein bisschen an Mathe? Also was damit gemeint ist, ist, ob sozusagen die Unsicherheit über das Klassenattribut sich stark reduziert, wenn ich den Wert des Attributs kenne.

Oder einfach ausprobieren. Also ausprobieren eben zum Beispiel, nehmt einfach einen Entscheidungsbaum und guckt, was passiert. Attribute, die der komplett weglässt, waren dann wohl nicht so wichtig. Okay, aber es gibt so ein paar Sachen, die sich ganz sicher nicht als Attribut eignen. Da könnt ihr von vornherein ausschließen. Also zum Beispiel IDs. Also hier würde ich die erste Spalte einfach mal weglassen, denn jeder Kunde, jede Kundin hat seine eigene ID. Das heißt,

Wir können nicht hoffen, da irgendein sinnvolles Muster zu lernen. Kleine Klammer auf. Manchmal, wenn man sowas drin lässt und es sieht aus wie eine Zahl, wird Orange vielleicht denken, das ist ein numerisches Attribut. Und es kann sogar passieren, dass es dieses Attribut verwendet und dass es was nützt. Woran könnte es liegen? Da kann man nicht so leicht drauf kommen. Nehmen wir mal an, ich nummeriere meine Kunden aufsteigend durch. Also der erste Kunde, der zu mir kommt, kriegt Nummer 1 und dann 2 und so weiter.

Wieso könnte es jetzt sein, dass das einem Algorithmus hilft, wenn es die ID kennt? Die ID kondiert dann sozusagen mit, wie lange jemand schon dabei ist, also wie loyal jemand ist. Angenommen, dass Kunden, die nicht mehr aktiv sind, irgendwann gelöscht werden, aber in diesem Fall ist das vielleicht nicht so, da bin ich pessimistisch, aber wenn jemand, sagen wir mal,

im Bereich Telekom oder Versicherung oder so dann kündigt, dann würde ich die Kunden ja rausstreichen. Und dann, wenn ich einen mit einer sehr kleinen ID habe, dann weiß ich vielleicht, aha, oder der Algorithmus schließt daraus und lernt Patterns, dass Leute, die schon länger dabei sind, für die irgendwie dies oder jenes gilt. Das heißt, aber das würde ich lieber nicht so machen. Also dann würde ich lieber ein Attribut einführen, Länge der Beziehung oder sowas.

Und dann wirklich messen, wie lange jemand dabei ist und das als Attribut nehmen und nicht darauf hoffen, dass das irgendwie implizit da drin steckt. ID's, die in jeder Zeile einen unterschiedlichen Wert haben, also für die Zeilen dieses Datensatzes unique sind, würde ich weglassen als Attribut. Attribute, die immer den gleichen Wert haben, also wenn in jeder Zeile da eine 1 steht,

dann kann ich es auch weglassen, weil dann kann ich auch keine Muster lernen, mit denen ich unterscheiden kann zwischen verschiedenen Gruppen von Instanzen. Und ganz wichtig auch, das ist manchmal eine Falle, in die man tappen kann, man muss sich auch überlegen, dass manchmal Attribute noch nicht bekannt sind zu dem Zeitpunkt, wo ich die Vorhersage mache. Also jetzt könnte es zum Beispiel sein,

Hier hatte ich jetzt angenommen, oder habe ich irgendwas darüber gesagt, was der Wintercheck kostet? Habe ich nicht. Es könnte also sein, dass ich jetzt noch eine Spalte hinzufüge oder in meinen Daten finde, was der Kunde bezahlt hat. Und das ist natürlich leer, wenn der Kunde das Fahrrad gar nicht zur Reparatur oder zum Wintercheck gebracht hat. Und dann teile ich ja dem

quasi mit, ob diese Wintercheck stattgefunden hat oder nicht. Das heißt, es ist ein Attribut, was unerlaubt ist, weil es eigentlich nicht bekannt ist. Also ich will das Modell ja für neue Kunden anwenden, bei denen ich noch nicht weiß, will ich ja gerade vorhersagen, ob sie hier zum

zum Wintercheck bringen und da werde ich die Kosten, die der Wintercheck verursacht hat, nicht wissen. Und ich werde auch nicht wissen, ob da eben ein Wert steht oder nicht. Ist klar? Ja. Okay, dann habe ich wieder eine Übung für euch. Ja, die schaffen wir noch. Wir machen Klassifikation und dann machen wir vielleicht nochmal fünf Minuten Pause und noch Regression hinterher. Nee, wir können auch beides machen. Wie viele Gruppen hatten wir gerade eben? Fünf, oder? Ja.

Eine Gruppe, zwei Gruppen, drei Gruppen, vier Gruppen. Wir waren fünf Gruppen, oder? Wollen wir gerade wieder die gleichen Gruppen machen? Ich habe nämlich fünf verschiedene Formalisierungsaufgaben. Also bei der Regression, ihr seht ja, ich habe auch keine Folie dazwischen, ist eigentlich alles genau gleich, nur dass es eben nicht das Klassenattribut ist, was ich definieren muss. Also die drei Sachen, die bleiben sich gleich. Ich muss die Instanzenformulierung wählen, ich muss wählen, was die Attribute sind.

Im Fall von Klassifikation wähle ich dann ein Klassenattribut und im Fall von Regression eine numerische Zielvariable. Also irgendeine Zahl, die das Modell vorher sagen soll. Okay? Ich würde sagen, wir machen das jetzt auf einmal. Das heißt, wir verteilen das an die Gruppen. Okay? Ja, das kriegen wir hin. Also, wollt ihr hier... War noch jemand dabei? Ich glaube, Liebindo, ne? Genau, ihr vier. Macht ihr das mit der Bank? Also...

Die Businessfrage ist, könnte nicht die AI die Entscheidung über Kreditvergaben übernehmen? Ihr formuliert das als Klassifikationsaufgabe. Ihr wart zu dritt, ne? Macht den Mobiles vom Ganzen die Tat. Also, wenn wir vorher wüssten, welche Kunden ihren Vertrag nicht verlängern, könnten wir sie mit bestimmten Angeboten halten. Okay, dann gehen wir mal in die zweite Reihe. Da wart ihr zu viert, oder? Ihr macht das mit dem Marketing. Ist ein bisschen mehr zu lesen, lese ich jetzt nicht vor. Aber das kriegt ihr hin.

Und dann in der dritten Reihe, ihr vier macht es mit dem Skigebiet. Okay. Und ihr drei mit der privaten Krankenversicherung. Also Skigebiet, ich plane die Ressourcen, die ich am nächsten Tag brauche. Und das hängt davon ab, wie viele Leute auf den Berg kommen, grob gesprochen. Und bei euch geht es um Versicherungsprämien. Ich würde vorschlagen,

Das wird jetzt nicht nur bis um vier dauern. Ihr werdet ein bisschen mehr Zeit brauchen. Die nehmen wir uns jetzt gerade, also vielleicht so 25 Minuten bis viertel nach. Dann machen wir vielleicht nochmal kurz 10 Minuten Pause und dann sammeln wir alles auf der Tafel zusammen. Ich lege das schon mal an, damit ihr auch seht, was rauskommen soll. Und ihr startet schon mal mit der Diskussion.

Nein, du kannst nur gut sein. Ich habe es nie übergespürt. Nein, ich habe es nicht. Ich habe es nicht. Ich habe es nicht. Ich habe es nicht. Ich habe es nicht. Ich habe es nicht.

Ich glaube, wir sehen uns erst.

Ja, das stimmt.

Ja, ich habe noch eine Kribuzi verzeichnet. Also, ich habe noch eine Kribuzi verzeichnet.

Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja.

Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja. Ja.

Ja, das ist eins. Also, die U-Bahn. Da ist nicht alles so. Da ist ein Auto reingekommen. Da ist ein Lebensmittelstück. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh. Da ist ein Schuh

Sorry, eine Sache noch kurz zu unterbrechen.

Ich würde euch empfehlen, außen anzufangen, also beziehungsweise dem, was an der Tabelle steht. Also überlegt euch, was die Instanzen und was Klassener Attributen beziehungsweise Zielvariable sind und dann fangt ihr an, die Attribute zu brainstormen. Okay? Das ist meistens die beste Reihenfolge. Okay. Ja. Ach so. Ja.

Genau. Welche Bank auch immer. Was sind eure Instanzen? In Instanzen haben wir jeden Kreditantrag. Wir hatten eine kurze Diskussion. Zuerst haben wir Kunden, also dass man Kundeninstanz ist. Aber dann Kunden, die mehrere Anträge machen können.

Wenn es unterschiedliche Beträge gibt, z.B. eine Million, dann kommt doch jeder Kreditkontraut. Könnte sein, ihr gebt mir 1000, aber nicht eine Million. Genau. Okay. Dann lass uns immer auch so erst die Aussen und dann nach innen gehen, auch bei euch. Was habt ihr als Klassenergebot? Als Klassenergebot haben wir Kreditvergabe ja, nein. Vielleicht

Nein, das war nur ein Scherz. Jetzt würde ich gerne die von der Krankenversicherung fragen, ob ihr da irgendwas wiedererkennt. Wir hatten eine Diskussion zu eurer Zielvariante und jetzt...

Vielleicht kann man da was überfragen.

einmal die Überarbeitung, aber das mal konferenzbasiert. Das kann ich nicht erklären. Wir haben das entschieden. Das kann ich nicht sagen. Ich kann nicht sagen, dass die Berichte da groß sind, aber man hat die Berichte, die sagen, das war falsch. Also falsch würde bedeuten, dass was passiert ist, das ist im ganzen Sinne nicht der Fall. Man könnte zum Beispiel berichten,

Das ist eine Art Partie, die mit Bewusstsein gewohnt wird. Dann wäre es eine Regression. Aber so wie es technisch oder so sieht, dass man das manuell nicht macht.

In dem Fall von der Versicherung, okay, kommen wir nachher dazu, aber wir haben auch gesagt, da gibt es was, das kann ich ja tatsächlich beobachten und du hast es eigentlich gesagt. Du hast gesagt, es kommt dann vielleicht zu einem Ausfall. Also ich könnte doch versuchen, den Ausfall vorher zu sagen. Also ich würde den Kredit ja nicht vergeben, wenn es einen Ausfall gibt. Das heißt, wenn ich...

Aber statt auch das vorherzusagen oder auf dem zu trainieren, was damals die Mitarbeitenden entschieden haben, nämlich ob das Geld vergeben werden soll oder nicht, zu gucken, war das eigentlich korrekt. Jetzt könntet ihr sagen, das ist ein bisschen problematisch, weil damals wurden ja auch gewisse Anträge nicht vergeben und da waren vielleicht welche dabei, die hätten es zurückzahlen können oder hätten es zurückgezahlt. Man kann ja berechnen, was die maximale

Ja, also da bist du wahrscheinlich wieder bei einer Zahl. Dann hättest du wieder eine Regression. Also ich habe jetzt hier wirklich mal so eine Ja-Nein-Entscheidung. Und ich glaube, die Ja-Nein-Entscheidung hängt ja innerlich, wenn du dir überlegst, ob du ihn vergeben würdest oder nicht, würdest du es davon abhängig machen, ob du glaubst, dass die Person den zurückzahlt oder nicht. Und dann kann ich eigentlich auch direkt das vorher bleiben. Wie auch immer wir die Trainingsdaten dafür bekommen.

Aber ich mache das mal in Klammern. "Verdietausfall ja/nein als Alternative". Also das ist nicht falsch, das kannst du auch machen. Und das wird auch funktionieren. Das wird dann eben das reproduzieren, was die Menschen gemacht haben. Nur eben wie bei der Versicherung auch: Meine Behauptung, Menschen können sich auch irren. Das könnten wir eigentlich dann versuchen, besser zu sein mit unserem Modell. Okay, erzähl mir ein paar Attribute.

Einkommen, Schulden. Warte mal, ich nehme noch ein paar Fragen. Ich würde nachher raten, was die bedeuten. Schulden. Betrag des Beliefs. Beruf. Familienzustand.

Da wird es uns oder nicht? Nein, wir haben hier einen Stand, glaube ich. Zweck, was du hast. Telefon, also die Ladezeit. Vielleicht Kunde. Eins noch. Ich weiß, ihr habt noch viel mehr, aber wir müssen noch Platz haben für die anderen. Noch ein Highlight. Ja.

Hast du irgendjemanden die Idee, was ich mit den Farben bezwecke? Ah, okay, ich glaube, das ist super. Nee, Schulden müssen wir ja nicht haben. Das könnte ja Nein sein. Ich finde das unbefaltbar. Ah, sehr gut.

Ja, genau, wir nennen die demografisch, also demografische Attribute. Also ich nenne sie mal direkte Instanzattribute, also Instanzen sind ja die, die wir anbringen. Was ist dann grün? Grün ist der Rett, wo ich nicht wusste. Ja, grün.

Ich sage mal, wenn man Menschen beschreibt, dann kann man sie mit demokratischen Verbunden beschreiben oder mit Verhaltensattribut. Wie viele Schulden jemand macht oder ob jemand Sicherheit hat, ist es nicht direkt Verhalten in dem Fall, aber es resultiert aus Verhalten. Ich frage trotzdem mal Verhaltensattribut. Jetzt schauen wir mal, ob sich das bei den anderen noch durchhalten lässt. Aber hier passt es ganz gut. Okay, super. Vielen Dank euch. Dann machen wir weiter mit dem Mobilfunkanbieter.

Das war ihr, oder? Entschuldigung. Entschuldigung. Warte mal. Gibt es noch eine Zahl? Direkte Instanz als Tribute. Okay, dann noch mal. Ja. Also ihr habt gefragt, warum soll man mehrere Probierkundenverträge haben.

Wahrscheinlich haben wir nicht viel, aber gut, machen wir Vertrag. Meistens wird es das Gleiche sein. Ich habe eine Erfüllung genommen. Ich habe eine Erfüllung genommen. Die Klausin hat gesagt, es wäre das Dachgeläufige, der auch hinein geht. Ich habe die Beweisung gefragt.

Welche Farbe? Zusätzliche Schwierigkeit. Schwarz. Also was heißt Bereich sozusagen der monatlichen oder Grundgebühr sozusagen? Ja, genau. Also die Betrauenslaufzeit. Machen wir mit Schwarz weiter. Die Netzwerkgeschwindigkeit, also um das 5G und 4G ist...

Hypothese, wenn jemand 4G hat, wechselt, dann wird vielleicht auch vielleicht so man dafür gebraucht. Auch der Datenverbrauch, das ist noch besonders wichtig. Was, oder? Das ist schon wieder ein demokratischer... Ja. Datenverbrauch. Auch die Revision, also dann kommt das Erwartung, das wäre nicht so gut.

Ja, also ihr habt es gesagt, wenn jemand im Gebirge lebt, dann hat man wahrscheinlich nicht so gutes Netz und dann sind viele Mobilfonds dann wieder nicht unterwegs. Genau. Auch das alte. Also alter des Kunden. Und ja, eins noch. Ja.

ich hätte jetzt ja sowas was ist das ich hätte jetzt fast grün gemacht aber der kunde eigentlich dafür aber vielleicht wird es dann zu vielen beschwerden das wäre dann wieder grün das hätte ich jetzt noch dazu gemacht anzahl beschwerden oder tickets vielleicht macht man tickets auch so ok punkt punkt punkt

Danke euch. Alles klar? Hat noch jemand Fragen, Bemerkungen, bessere Ideen? Gut. Dann kommen wir zu den Kampagnen. Das war das 4. Mal. Ja. Unsere Instanz war der Status der Hälfte der Kampagne. Also Kampagnenstatus. Hälfte der Laufzeit. Muss man dafür noch was erklären für die anderen?

Ja, dann ist das schon... Aber macht man erst vielleicht noch das Klassenattribut, dann kann man es besser zusammen erklären, glaube ich. Klassenattribut, Kurs bei Leib, ja, nein. Also, erklär nochmal für die anderen, was genau man da jetzt vorhersagt. Wir schauen in der Kanzel, in der Kampagne, ob man noch mehr politische Nachschüssen hat, auf Seiten, oder nicht. Und wenn wir das jetzt nicht erreicht haben, auf Seiten, dann einfach nein.

Also erreicht, wir erreichen ein Ziel, was wir uns gesteckt haben, oder? Das ist die Frage. Also wir wollen zum Beispiel so und so viele Klicks generieren mit der Kampagne und die Frage ist, ob wir das erreichen, richtig? Ja. Genau. Und wenn das Modell sagt, zur Hälfte der Laufzeit meiner Kampagne, hey, wenn du so weitermachst, dann wirst du es nicht erreichen, dann schießen wir Budget nach. Wenn das Modell sagt, passt schon, dann lassen wir es. Okay. Dass alle noch das verstanden haben.

Jetzt brauchen wir noch ein paar Attribute. Ja. Das erste wäre die Laufzeit. Also diesmal haben wir ja keine Kunden, deswegen haben wir auch keine demografischen Attribute. Laufzeit, Gesamt der Kampagne. Ja. Verbleibungszeit, Zielvorgabe, die aktuelle Performance. Also Performance in der gleichen Einheit gemessen wie Zielvorgabe, zum Beispiel die Anzahl Klicks oder...

Ja. Was damit gemeint? Oder was ist der Unterschied zu aktuellen Performance? Das ist ja eine Performance, weil der E-Mail-Zone und der Status, weil der E-Mail-Zone und der Status, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob

es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft, ob es gut läuft,

Komm, wir nehmen das nächste Mal. Okay. Was ich noch machen würde, glaube ich, ich würde versuchen, so ein bisschen zu modellieren. Also sozusagen, wenn ich hier die Anzahl Clicks habe und hier habe ich die Zeit, dann ist es ja möglich zum Beispiel, dass es irgendwie so aussieht oder dass es irgendwie so aussieht.

Na ja, runtergehen wird es nicht. Also wenn es kumuliert ist sozusagen, also wie viele Klicks ich insgesamt habe, dann steigt es immer an. Aber es kann vielleicht auch sein, dass es so, also das hier ist vielleicht beruhigender, weil ich den Trend innerlich so fortsetze, dass es zum Ende hin besser wird. Und das hier ist so ein bisschen weniger beruhigend. Da habe ich nicht die Hoffnung, dass noch etwas passiert. Ich würde versuchen, das noch irgendwie zu modellieren. Weiß auch nicht genau wie, aber vielleicht kann man so ein bisschen Etappen bilden und dann gucken, wie war der

Zum Beispiel Anstieg zum Vormonat. Darf ich das noch hinschalten? Ja, das ist das, was wir mit dem Verlauf der Einverkaufsvermögen von einbauen. Also wenn die Kampagne die sechs Monate dauert, und wir die Qualität von den ersten Monaten aufbauen, dann ist das ja auch ein Top-Signal. Dann werden wir das ja nicht vorhalten. Okay.

Ja, also das ist sozusagen ein komplexes Attribut. Da müssen wir uns noch überlegen, wie wir das genau ausprobieren. Das wäre jetzt eine ganz einfache Art, wie man es machen könnte. Wahrscheinlich kann man besser werden, wenn man eben versucht, die Entwurfe so ein bisschen granular nachzubilden. Dass man so ein bisschen Verlauf dem Modell mitgibt und das Modell dann weiß, okay, das ist eher am Ende angestiegen oder eher am Anfang und jetzt stagniert es, dass das so ein bisschen noch als Zeitreihe fortsetzen kann. Okay, super. Fragen irgendwo?

Dann machen wir noch die Regression. Schauen wir, dass wir kurz vor fünf verhelfen und dann kriegt ihr noch einen Zug, falls ihr einen haben wollt. Was waren wir zuerst? Die Viehgebiete. Ja, wir haben bei Franz die Trauung. Wir hatten kurz Diskussion.

Ja, das wurde vorher gesagt. Dann können wir auch gleich sagen, für jeden Tag sagen wir was voraus. Eigentlich geht es darum, um den Skifahrer zu sagen. Ich sage mal Besucher. Also Ihr Ziel ist ja, Ressourcen einzuplanen und es können natürlich auch mal welche mitkommen,

Die nur im Restaurant sitzen. Dann brauchen wir beim Restaurant mehr. Okay, jetzt. Dann haben wir das Wetter, das haben wir aufgeteilt. Die Sichtverhältnisse, die kontraktierten Verhältnisse. Ja, man meint das Sichtverhältnisse. Also das ist halt irgendwie kategorisch so neblig. Ja, das ist aber ein Schiff. Okay, andere haben es wieder vergessen. Sag es nochmal.

Wind, Temperatur, Schnee. Also ob es Schneeball hat oder noch vielmals Regen. Sessau, Medien, Laviergefahr. Okay, warte. Eigentlich hätte ich noch verschiedene Farben benutzen können.

Also Saison wäre sowas wie Sommer oder Winter. Ja. Also es kommen wahrscheinlich im Winter mehr Kiefer im Sommer. Ja. Auf Ferien. Und das ist das Interessante, wir hatten ja darüber diskutiert,

zu einem Tag gehört ein Datum und das Datum selbst ist immer ein schlechtes Attribut. Also generell an alle, wenn ihr ein Datum irgendwo habt, aus dem Datum lassen sich immer gute Attribute machen, aber Datum selbst ist kein gutes Attribut. Es kommt nie wieder.

Wenn ich ein Muster lerne, dass am 1.3.2015 irgendwas passiert, das kommt nie wieder, das wird mir nie wieder was müssen. Aber ich kann lernen, dass das noch Winter war und da waren vielleicht Ferien. Lawinengefahr, Events auf dem Zugebiet, die Zeitung vom Schiffsausflug.

Also angenommen, die sind dynamisch. Wenn ich das Skigebiet betreibe und die sind immer gleich, dann würde ich es wahrscheinlich nicht als Attribut nehmen. Nein, aber die sind ja auch dynamisch. Also wir werden dem später noch begegnen, weil damit machen wir noch eine Übung.

Da sind nicht alle, die ihr aufgezählt habt, dabei. Nicht alle Attribute. Was noch dabei ist, ist die Schneehügel. Wenn ihr überlegt, also wenn Schnee ist oder nicht, habt ihr auch. Also Schneehügel hat eigentlich Schneeverhältnisse. Sorry. Okay. Noch irgendwelche Kommentare? Fragen? Dann machen wir weiter.

Mit der Krankenversicherung? Das hatten wir ja schon andiskutiert vorhin. Genau, wir haben es eigentlich gar nicht andiskutiert. Wir haben es direkt übertragen auf die Kredite. Sag nochmal, also verursacht die Posten, hast du gesagt, oder? Du kannst nochmal sagen, was die Ursprünge gefolgt sind. Ja.

Dann bin ich gekommen und habe gesagt, die Prämie hat ein Mensch festgesetzt und dann habt ihr vielleicht Verlust oder Gewinn gemacht, weil die Abschätzung nicht so gut war. Und wenn wir aber Trainingsdaten bauen, können wir es eigentlich ja so machen, dass wir tatsächlich gucken, wie viel Kosten hat jemand verursacht und versuchen direkt das vorher zu sagen und den menschlichen Fehler rauszuholen. Also es ist eigentlich genau wie bei euch beim Kreditanfragen. Okay, jetzt noch ein paar Worte. Jetzt kann ich wieder Farben nehmen, weil jetzt sind es wieder Leute.

Also eigentlich, Geburtstag wäre wieder ein Datum. Dann würde ich nur das Alter nehmen, was dann aus dem Geburtsdatum abgereicht ist. Farbe? Ja, also Geschlecht, Alter. Farbe? Ja, ist nicht direkt verhalten, aber... Macht ja auch grün, oder? Behandlungen, also Anzahl Behandlungen, oder?

Also das muss die Person uns irgendwie angeben in einem Formular. Also wenn die jetzt noch nicht bei uns, wenn der noch nicht bei uns versichert ist, würden das irgendwie erfragen von irgendwann. Medikamentenkosten. Das wäre es noch. Also wenn man Grün noch nimmt als Verhalten, gibt es noch irgendwelches relevantes Verhalten?

Ja, sowas, oder ob ich Extremsport mache, auch Extremsport, Sport ist ja auch was noch, das ist gut. Uns würden wahrscheinlich noch Sachen einfallen. Okay. Habt ihr noch irgendwelche Fragen oder Bemerkungen? Was ist der Unterschied zwischen Klassenerzibut und der Zielvariante? Also es ist eigentlich das Gleiche, nur dass es eben das eine ist, was kategorisch ist,

Also in unseren Fällen meistens ja oder nein. Und da ist es wirklich eine Zahl, also Anzahl Besucher oder Kosten in den beiden Fällen. Okay, wenn ihr das nochmal üben wollt, ihr kriegt gleich eine E-Mail von mir mit einer kleinen Hausaufgabe. Also ich weise euch einfach nochmal darauf hin, dass hier ein Quiz ist. Mal sitzen. ML-Aufgaben formalisieren, da könnt ihr das auch nochmal üben.

Jeremy, du wolltest noch was sagen? Ja, kurz zum prozentualen Anstieg zum Boden, könnte man auch drei Attribute zum Beispiel nehmen, über zwei Monate? Ja, unbedingt. Den Absolutwert? Ja, so

hätte ich jetzt gedacht. Dass es dann drei Werte hat, die verbleiben können. Ich würde eher, also das soll ja sozusagen über verschiedene Kampagnen hinweg mustern.

erkennen. Und das ist wahrscheinlich leichter, weil die vielleicht auf einer anderen Skala operieren, die Kampagnen. Also eine vielleicht sowieso viel mehr Leute erreichen kann als die andere, wenn du es eher prozentual machst, den Anstieg. Dann ist das Muster leichter erkennbar, weil die absoluten Zahlen vielleicht gar nicht vergleichbar sind. Und ich würde auch immer eine Änderung nehmen, statt der absoluten Zahl, weil

dass Orange zum Beispiel es nicht hinkriegt, da noch Berechnungen oder Vergleiche zwischen, also schon auf eine Art, aber es wird sehr viel schwieriger, als wenn du es direkt mitgibst, wie sozusagen die Entwicklung war jeweils von einem zum nächsten Monat. Okay? Aber mehr als nur den Vormonat zu nehmen, ist auf jeden Fall eine gute Idee. Also hier ein bisschen verschiedene Abschnitte zu machen und jeweils zu messen, wie sich es entwickelt hat.

Dann, wenn es kontinuierlich gestiegen ist, dann ist wahrscheinlich die Vorhersage eher, wir werden es erreichen, als wenn es dann stoppen wird. Okay. Ansonsten sehen wir uns wieder nächsten Mittwoch. Ich werde wahrscheinlich wieder so ganz knapp erst reinkommen, weil ich wieder aus Motens komme. Vielleicht auch zwei Minuten zu spät, also nicht weglaufen. Bis nächste Woche. Danke. Ich sehe viel besser aus. Nein.

Ja, mit Sport und allem schon und so.