Hey, was haben wir letztes Mal gemacht? Hier ist ein leeres Orange. Wir haben über Regressionen gesprochen. Vielleicht erinnert ihr euch. Was haben wir da vorher gesagt? Was ist das noch? Ja, genau. Ich habe mir gedacht, wir gehen es einfach einmal noch mal schnell durch und erinnern uns noch mal an die ein, zwei Sachen, die dabei wichtig waren. Vielleicht auch drei.

Also unser Datensatz. Machen wir mal hier so ein Reset. Sind alle angekommen?

Kann noch dauern. Also, was haben wir als erstes gemacht? Wir haben die Daten geladen, das ist immer das Erste, und haben im Select Columns unsere Spielvariable ausgewählt. Ja, ja, hier hat er sich alles gemerkt, natürlich. Kann noch mal jemand sagen, warum das Attribut Date aussortiert wird?

Wollen wir es wieder zurückbringen oder hatten wir uns geeinigt, dass es da schon richtig ist? Ja, genau. Also die Daten liegen alle in der Vergangenheit und deswegen war meine Message an euch schon immer, vielleicht erinnerst du dich auch daran, Dennis,

dass Datum eigentlich immer ein schlechtes Attribut ist, weil es eben nicht wiederkommt. Man kann aber aus dem Datum gewisse Sachen rausziehen, die gut sind. Monat zum Beispiel hatten wir gesagt, ist sicher ein sehr wichtiges Attribut, um vorher zu sagen, wie viele Leute zum Skifahren kommen. Und dann hatten wir erstmal lineare Regressionen gebaut. So, jetzt sagt mir nochmal jemand, wenn ich jetzt wissen will, was diese lineare Regression tut, was muss ich dann machen? Das muss ich...

Wenn es voll drin ist. Kannst du jetzt noch die Video anzeigen lassen? Also, ich wollte mal schauen, ob wir die Koeffizienten noch lassen. Ja, genau, die wollte ich mir anschauen. Ja, genau. Und dann sage ich, die Koeffizienten sollen dargestellt werden. Das ist gut. Und dann hatten wir das hier. Und da ist uns was aufgefallen. Weiß noch jemand?

Das hat mit dem Monat zu tun, dass das nicht kapituliert in den Leben kommt. Genau, hier ist es numerisch und jetzt fängt die lineare Regression an, damit zu rechnen. Also der Koeffizient mit einem Koeffizienten für Manns und Manns war eine Zahl zwischen 1 und 12. Warum ist es jetzt einerseits gut und andererseits schlecht? Also sagen wir mal so, das Modell tut sein Bestes in dem Sinne, dass der Koeffizient negativ ist.

Und somit die meisten Besucherzahlen für Januar, dann Februar, dann März vorhergesagt werden und dann immer weniger. Also wenn man das mal so multipliziert einfach mit den Monatszahlen, die man ja für manch dann einsetzt, wenn man die Vorhersage macht für einen bestimmten Tag.

Dann sieht man also, dass für Dezember fast 5000 Besucher weniger vorhergesagt werden. Gehen wir mal angenommen, die Bedingungen sind sonst gleich wie für einen Tag im Januar. Und das macht irgendwie keinen Sinn. Also so unterschiedlich sind Dezember und Januar dann auch nicht. Also haben wir was gemacht? Ja, genau. Also immer hier können wir es ja umstellen.

Ihr habt jetzt gesehen, in meinem Select Columns, ich habe den Tag auch noch aussortiert. Der wäre, wenn, dann vermutlich auch keine Zahl, sondern auch, müsste man auch auf kategorial umstellen, aber wir haben auch gesagt, der läuft ja zwischen 1 und 31 und es gibt keinen Grund anzunehmen, warum am 1. des Monats mehr Leute kommen sollen als in der Mitte oder so. Es ist ein, dass die Leute vielleicht am Anfang des Monats mehr Geld haben, aber

aber dann auch irgendwie verworfen und hat auch keinen positiven Effekt, wenn man den Tag mit reinnimmt. Und jetzt fängt es also an, ganz viel Sinn zu machen. Also wir sehen jetzt, dass die meisten Besucher hier für Februar vorhergesagt werden, gefolgt vom Januar. Und dann kommt der, nee, dann

kommt nicht der Dezember. Jetzt bin ich gerade erstaunt, was hier los ist. Aber gut, dann kommt der März und dann kommt aber der Dezember.

mit einem relativ kleinen Koeffizienten jetzt hier und alle anderen sind negativ. Und tatsächlich sind das jetzt erstmal die Wintermonate, die die meisten Besucher haben. Das macht Sinn. Was haben wir noch gemacht? Wisst ihr noch was? Ja, zwischendurch haben wir noch was anderes gemacht. Aber warum haben wir den gemacht?

Also welche Frage hat er uns beantwortet? War es interpretierbar? Ja, also der Gradient Boost ist nicht interpretierbar. Aber wir haben sozusagen, ich habe euch gesagt, man kann den Fehler, also wir hatten Mean Absolute Error als Maß uns angeschaut, als einziges, weil wir das so gut verstanden haben. Und da habe ich euch gesagt, das kann man so bis auf 1000 runterdrücken. Also wenn ich jetzt hier

mir so einen Test- und Score-Widget hole und da so eine lineare Regression anhänge, das muss ich ja immer mir frisch holen. Und ich hatte euch erklärt, darüber reden wir heute dann noch ausführlicher, dass ich, wenn ich hier Random Sampling einstelle, dann, dann nimmt dieses Widget das sozusagen in die Hand, die Daten aufzuteilen in Trainings- und Testmenge. Und wenn ich hier einen Algorithmus als weiteren Input dran tue,

Dann wird dieser Algorithmus mit den Trainingsdaten trainiert und auf den Testdaten dann evaluiert. Und hier bin ich jetzt bei einem Mean Absolute Error von anderthalb tausend Besuchern, um die das Modell im Schnitt falsch liegt. Und dann, Marc, probieren wir mal, ob wir mit Gradient Boosting besser sind. Also wissen wir ja schon. Ja, Dennis, was? Noch kurz eine Frage zu Regulation.

Ja. Also wir schauen da auf jeden Fall nächste Woche drauf. Nächste Woche beschäftigen wir uns mit, wie habe ich es genannt, typischen Problemen. Regularisierung beschäftigt sich mit dem einen dieser typischen Probleme, nämlich dem Overfitting. Und das andere Problem, was wir noch anschauen, ist die unbalancierten Klassen. Also wenn man, ja, ich sage immer Nadeln im Heuhaufen sucht. Also uns

Zum Beispiel sehr viele Leute gibt, die sich nicht für dein Produkt interessieren und nur ganz wenige, die du finden willst. Das stellt einen auch manchmal vor Probleme. Da schauen wir nächste Woche drauf. Wieso? Ja, jetzt. Also hier seht ihr, wir sind noch nicht so richtig weitergekommen. Woran liegt es? Jetzt. Irgendwas läuft hier schief. Sieht jemand, was schief läuft? Ja, es war so um die Tausend.

Erinnere ich mich. Und jetzt frage ich mich gerade, irgendwas mache ich falsch? Ja. Also ich habe jetzt nichts weggenommen, außer dem Tag. Ja. Ja, ich versuche mal jetzt noch so ein Tree. Den Tree hattet ihr ja auch erwähnt.

als etwas, was wir gleich noch interpretieren wollen. So als Test, wo sind wir da? 1100, das war so wie beim letzten Mal, das ist jetzt nicht überraschend. Habe ich hier irgendwas? Learning Rate 0,3 habe ich. Trotzdem gehen. 0,1. Okay.

Also das habe ich jetzt noch selten erlebt, dass die Learning Range so einen Einfluss hat. Aber gut, jetzt sind wir bei ungefähr 1000. Das war das, was wir als so, ja, also knapp unter 1000 konnte man auch schaffen. So sehr viel besser. War schwierig. Warum ist das so? Also was...

Für einen Grund hatten wir uns zurechtgelegt, warum es zum Beispiel hilft, den Gradient Boosting zu nehmen oder in dem Fall auch den Tree gegenüber der linearen Regression. Warum kriegt die lineare

Regression das nicht hin, Richtung Error 1000 zu kommen, sondern bleibt bei anderthalbtausend stecken? Wisst ihr es noch? Warum? Also was für Probleme machen die da, meinst du?

Ah, nein. Abhängigkeit. Also an einem sonnigen Tag und wochenende. Genau, das war, was wir im Tree sehen konnten sozusagen. Ich hatte euch dann gefragt, wann kommen die meisten Leute? Und das kann man sehen, dass eben diese spezielle Kombination, der perfekte Skitag sozusagen, hatte ich euch auch erzählt sozusagen, man könnte das auch nachrechnen sozusagen, indem man die Werte hier einsetzt.

und wird vermutlich auf eine niedrigere Zahl kommen. Vermutlich kann ich das auch mal machen. Also wenn wir uns mal anschauen, was der Tree so vorhersagt. Wir hatten uns hier so ein Tree ausgerechnet. Tiefe 4 klingt gut. Und vielleicht machen wir ihn doch Binary. Ja doch, der ist Binary. Und der sagt sozusagen als erstes Mal...

unterscheide nach Monat und wir können eigentlich den ganzen rechten Teil zuklappen, hatten wir gesagt. Das sind die Monate, die nicht im Winter liegen und hier sieht man schon, der linke Teil des Baums ist Winter. Und die größte Zahl, die der Baum hier vorhersagt, ist 11.500. Ich glaube, das war einigermaßen stabil, also nicht vielleicht genau diese Zahl, aber so um den Dreh. Und jetzt mache ich was Experimentelles.

Wir können mal hergehen und hier Predictions machen mit diesem Modell auf den Originaldaten. Und mal schauen, ob man das hier sortieren kann. Tatsächlich gibt es auch so hohe Werte. Nicht oft. Also wenn man sich überlegt, die Bedingungen, unter denen der Tree jetzt diesen hohen Wert voraussagt, hatten wir gesagt, ist Winter.

Relative Humidity niedrig oder niedriger als 78 und dann hat mir gesagt, mehr als 83 cm Schnee heißt also quasi genug Schnee. Das ist das, was der Baum sagt. Also kann man sich halt überlegen, dass das wahrscheinlich relativ noch einige Tage betrifft und man sieht hier bei, ja, es kommt schon vor. Also hier ist sozusagen die Vorhersage der linearen Regression, habe ich mal nach Größe sortiert und das sind sozusagen die

die größten Zahlen. Man sieht hier vielleicht selbst der Baum kriegt es noch nicht so richtig gut hin, weil die tatsächlichen Zahlen in diesen Tagen teilweise sogar noch viel höher waren. Hier gab es sogar mal einen Tag, wo fast 20.000 Leute gekommen sind. Ein Tag im Februar 2013, Wochenende war es, wo es zwei Meter Schnee, Ferien auch noch, Temperaturen um die 0 Grad. Also genau einer von diesen perfekten Tagen sozusagen.

wo viele Modelle dann Probleme haben, das vorherzusagen und der Tree geht eben in die Richtung, dass er sagt, unter bestimmten perfekten Bedingungen sage ich eine sehr hohe Zahl an Schiebwaren voraus. Okay, ja, also multivariate Muster waren unser Stichwort, ja. Ich wollte kurz noch zurückgehen und zeigen, wie das mit dem Bildchen gemacht wird, aber jetzt nicht.

Mit welcher Prediction? Mit dem... Ach, hier. Ja, ich weiß in dem Fall gar nicht, ob es das hier braucht. Wahrscheinlich nicht. Doch es braucht es. Also du musst ihm Daten geben. Das ist natürlich jetzt... Eigentlich sollte man es so nicht machen, weil ich jetzt auf den Trainingsdaten die Predictions mache. Normalerweise würde ich erst was abspalten und dann Predictions machen. Ich kann es mir auch hier anschauen. Predictions.

Dann wird es noch auf den Testdaten angezeigt. Genau, und hier kann ich es mir auch anschauen, jeweils vorhersagen die Modelle. Also man sieht, dass der Gradient Boosting wahrscheinlich auch

deswegen noch ein Stück besser ist, weil er noch höher einsteigt bei manchen Tagen. Also hier zum Beispiel der Tag, wo fast 17.000 Leute gekommen sind.

Da sind sowohl die lineare Regression als auch der Baum zu konservativ und der Gradient Boosting kommt der Sache irgendwie näher. Also der macht das noch besser, sozusagen diese Multivariantenmuster lernen. So, ist die Erinnerung wieder da? Dennis? Nein, also das müssen eigentlich die Testdaten sein.

Die Trainingsdaten sollten hier nicht gezeigt werden. Also die Daten, auf denen getestet wird, genau. Mit der echten Zahl, also hier die 16.687 und dann die drei Feuersagen von den Modellen, die ich dran gehängt habe und alle anderen Attribute. Ja, okay. Andere Fragen, Zweifel, Kommentare? Wie war es mit dem Quiz?

Schwierig. Sollen wir es anschauen? Echt, war es schwierig. Mal schauen. Ich erinnere mich nicht. Alle Fragen anschauen? Alle. Okay. Die erste war vielleicht deswegen schwierig, weil wir nicht so wahnsinnig lange darüber gesprochen haben, wie Gradient Descent funktioniert. Was ich euch nur gesagt hatte, eigentlich kann man Mathematik, also gerade bei

Zum Beispiel lineare Regressionen hatten wir gesagt, wir wollen ja den Fehler minimieren, dafür machen wir Gradient Descent. Und dann kann man sich erinnern an die Schulzeit und sagen, okay, man kann auch irgendwie die Fehlerfunktion ableiten und die Ableitung null setzen und dann findet man das Minimum. Das ist das, was ich mit analytischer Lösung meine. Und die gibt mir natürlich die genaueste Lösung. Also wenn ich das globale Minimum finden will, dann kann ich das mit der analytischen Lösung minimieren.

Das funktioniert bei Linearregressionen, bei manchen anderen Funktionen, die ich oder deren Parameter ich lernen will, ist es dann schon schwieriger oder funktioniert teilweise nicht. Und Gradient Descent funktioniert halt immer. Also ich ziehe da Lösungen so schrittweise an zu nähern. Und das kann ich euch nochmal zeigen kurz. Wir hatten ja so ein Bild von Gradient Descent und das, was da steht, dass die Schrittweite

mit der ich meine Gewichte anpasse beim Lernen, einen Einfluss auf die Konvergenz hat. Das kann man sehen, wenn man sich das hier anschaut. Also wenn ich hier bin, was man hier sieht sozusagen, ist so ein kleiner Pfeil, um den ich das Gewicht hier verändere. Also das Gewicht ist am Anfang so vielleicht initialisiert, dann berechne ich diesen Gradient

Weil der so monopositiv ist, weiß ich, dass ich mit meinem Gewicht kleiner werden muss, dass ich nach links gehen muss, ein Stückchen. Und man geht aber immer nur ein kleines Stückchen. Na, jetzt kann man sich überlegen, wenn ich jetzt bis hier gehe zum Beispiel, dann lande ich hier. Also mit einer großen Schrittweite, mit so einer Schrittweite lande ich hier. Dann leite ich ab und dann sagt mir mein Gradient, du musst nach rechts. Jetzt gehe ich gleich nach da. Also ich schauke mich dann hoch, statt hier in dem Minimum zu landen.

Das heißt, wenn die Schrittweite zu groß ist, dann verpasse ich mein Minimum und dann komme ich nieder an. Deswegen kann die Schrittweite das sehr wohl beeinflussen. Genau, das erste habe ich nicht eingekreuzt, weil es genauer ist analytisch. Habe ich richtig ruminterpretiert, oder? Lass mich kurz fertig reden, dann komme ich zu dir, Marc. Genau, ich muss trotzdem differenzieren können. Als Fehlerfunktion hätte ich hier schreiben müssen.

Und Gradient-Dicent hilft überhaupt nicht, den richtigen Funktionstyp auszuwählen, das muss ich vorher und alleine machen. Es hilft mir dann nur, die richtigen Werte der Gewichte zu finden. Wenn

du das jetzt wiederholst, dann ist das halt ein bisschen ein Knallzeichen, dass du Gradient-Dicent für die Mehrheit der Tauchen, für die anderen Routen schon willst und das ist ein Zusammenhang mit dem Gradient-Dicent.

Also sozusagen die Fehlerfunktion jetzt bei der linearen Reduktion zum Beispiel für ein Gewicht. Und wir suchen das Minimum des Fehlers, also wo der Fehler am kleinsten ist. Und analytische Lösungen, kennt man so vielleicht aus dem Matheunterricht, kann man finden, indem man die ganze Funktion, die ganze Fehlerfunktion nach diesem b hier ableitet und die Ableitung 0 setzt, weil eben hier im Minimum die Ableitung 0 ist.

Ja, und dann löst man nach W auf und dann weiß man, an welcher Stelle oder also sozusagen für welchen Wert von W der Fehler minimal wird. Das ist die analytische Lösung. Und was wir jetzt hier machen oder was hier aufgezeichnet ist, dass wir immer gucken, wie hoch ist der Fehler, wenn ich W gleich 0,623 wähle. Und dann gucke ich hier, wie ist da die Ableitung. Aha, die ist so, das heißt, die muss noch nach links.

Und ich finde dann auch das Minimum eigentlich ziemlich zuverlässig. Bei linearer Regression sogar sicher. Also da kommt das Gleiche raus bei anderen Funktionen. Gibt es vielleicht mehrere Minima und dann kann es passieren, dass ich nicht das Richtige finde. Weil ich in einem hängen bleibe, was nicht das globale Minimum ist. Also wenn die Fehlerfunktion, vielleicht soll ich das mal hinzeichnen,

Das ist die Mathematik. Das Schöne ist ja, dass...

Tool macht. Was manchmal wichtig zu wissen ist, diese Schrittweite, wenn man sie anpassen kann, also hier bei Orange kann ich bei der linearen Regression nichts daran schrauben, aber wenn man sie anpassen kann, kann noch einen Einfluss haben. Okay? Ja, also eigentlich hatte ich ja letztes Mal, als wir über parametrische und nicht parametrische Funktionen, wo eigentlich der KNN der einzige Vertreter von nicht parametrischen war, die wir wirklich angeschaut haben,

gesehen, dass meistens die parametrischen besser sind. Und ein Argument war, und da hoffe ich, dass ihr euch erinnert, wir hatten das mit den Häusern angeschaut, wenn wir auf ganz vielen Häusern trainieren, die klein sind und daher auch günstiger, dann

lernen wir mit der linearen Regression trotzdem, dass es einen linearen Zusammenhang zwischen Hausgröße und Preis gibt und können dann entsprechend interpolieren. Also wenn wir dann wirklich ein großes Haus haben, dann sagen wir auch entsprechend einen hohen Preis vorher. Der Cane nearest neighbor macht das nicht, weil er nur die kleinen Häuser kennt. Und bei dem großen Haus sind die ähnlichsten Nachbarn dann eben trotzdem kleine Häuser, wenn nur kleine Häuser in der Trainingsmenge waren. Erinnert ihr euch an das Bild und die Diskussion dazu?

Das heißt, nein, die parametrischen Methoden interpolieren besser. Und das haben wir gesehen hier bei den Skifahrern, dass eben nicht alles linear ist, beziehungsweise dass die lineare Regression vor allem eben die Abhängigkeiten zwischen Variablen ignoriert. Und die sind manchmal wichtig. Also die perfekten Skifahrbedingungen, die bestehen aus Kombinationen. Das heißt, diese Interaktion zwischen den Variablen ist wichtig manchmal.

Das heißt, nein, lineare Regression kann nicht immer so eingestellt werden. ENN, ob für Klassifikation oder für Regression, braucht zur Klassifikationszeit immer mehr Zeit als parametrisch verfahren. Da muss ich nur einfach die Werte einsetzen in meine lineare Regressionsformel oder in meinen Baum und habe innerhalb von Millisekunden meine Vorhersage.

Hier muss ich immer mit allen Trainingsdaten vergleichen beim Canaris Neighbor, das heißt, kreuze ich nicht an. Ich glaube, dann müsste es so sein. Ja, Marc. Also,

Man definiert sozusagen, um wie viel man das Gewicht verändern möchte. Das wird sowieso mit dem Gradient multipliziert. Also je steiler der Gradient ist, desto weiter läuft man sozusagen. Gerade deswegen ist es wichtig, also wenn man sehr weit hier drüben, also sehr weit vom Minimum weg ist, dass man nicht zu weit hüpft. Also Schrittweite hat den Einfluss darauf, wie weit ich mein Gewicht verschiebe.

Der Gradient sagt mir, in welche Richtung ich gehen muss, um zum Minimum zu kommen. Und die Schrittweite sagt einfach, wie weit gehe ich. Und ich gehe immer nur ein Stückchen. Weil ich mich viel über rausschießen will, über das Minimum. Das weiß man nicht. Wenn du merkst, dass es nicht zu einem Ergebnis kommt oder völlig viel zu große Koeffizienten dir ausspuckt, sozusagen,

dann kannst du probieren, einfach die Schrittweite zu reduzieren. Beziehungsweise, wenn es nicht konvergiert. Ja, also im Allgemeinen funktioniert es. Nur wenn es eben doch mal schief geht, dann nimm einfach eine kleinere Schrittweite. Okay? Wie seid ihr mit Frage 3 klargekommen? Also die District Number in Frage 3 spielt die gleiche Rolle wie der Monat im Skigebiet.

Das habt ihr dann vielleicht irgendwann gemerkt. Sie sollte nicht numerisch sein. Und trotzdem versucht das Modell, irgendwas damit anzufangen. Also ist euch das klar geworden, warum, wenn ich die Distriktnummer in Wien als numerische Variable behandle, warum dann hier ein negativer Koeffizient, also das ist jetzt schon Teil der Antwort, aber so ist es dann, wenn man das lernt,

Es wird sich ein Negativakoeffizient ergeben. Warum macht der Sinn? Warum kann man wieder sagen, das Modell hat sein Bestes getan? Ich denke jetzt, die Wohnungen in der Stadt sind teuer. Und dann ein klein negatives Betrag mal eins, die weniger verhält einen Vorteilpreisabzug als zum Beispiel 22 mal eine negative Zahl.

Genau, also ich setze ja hier die Nummer ein und wenn ich da die 1 einsetze und das hier ist negativ, dann wird weniger abgezogen, als wenn ich da 22, also was weit vom Zentrum weg ist, einsetze. Wenn das negativ ist, wird sozusagen dann der Preis stark verringert. Insofern werden dann in der Summe im Ergebnis die Häuser im Zentrum höhere Preise machen.

Ja, das macht irgendwie Sinn. Andererseits gibt es, ich weiß jetzt nicht genau, wo hier das Lillenviertel ist, aber das ist ja nicht im Zentrum im Allgemeinen und da sind die Preise dann sicher auch sehr hoch. Also nehmen wir mal an, dieses hier, keine Ahnung, ob das stimmt. Und für dieses sollte ich dann einen anderen Koeffizienten haben, der vielleicht sogar positiv ist. Das heißt, hier macht es auf jeden Fall auch ganz viel Sinn,

District Number als kategorial zu behandeln. Also es wird aktuell als numerisch behandelt, das sieht man, weil es einfach mit einem Koeffizient multipliziert wird. Ja, lass mich schnell noch ankreuzen und so, so hat er gesagt. Ja.

Ja, also wir haben es gesehen vorhin beim Monat, dann wird es sozusagen one hot encoded, dann sieht es genau so aus hier. Also dann steht da District gleich und dann 1 bis 22 und dann hat jeder District seinen eigenen Koeffizienten und dann siehst du, wo es besonders teuer ist und wo es besonders billig ist. Also es

Denke ich, wenn du das dann im Wiener zeigst, dann sagt er ja. Genau. Das, den Distrikt kann sich keiner leisten. Okay, weiter. Ah ja, das können wir machen. Das ist was? Dann machen wir es nicht.

Schade, es wäre so Hands-on gewesen. Aber 5 und 6, war das schwierig? Also das haben wir jedenfalls schnell gemacht.

diskutiert. Da geht es wieder um den K-Nearest Neighbor und der K-Nearest Neighbor funktioniert bei stark nicht, oder sagen wir andersrum. Ich glaube, man muss es andersrum sagen. Der KNN, den habe ich ja immer so gescholten. Aus verschiedenen Gründen, aber die lineare Regression wird natürlich nicht gut funktionieren, wenn der Zusammenhang stark nicht linear ist.

Und deswegen würden wir hier in diesem Fall tatsächlich mal nicht erwarten, dass die lineare Regression besser ist als der Canary's Neighbor, der auch nicht lineare Abhängigkeiten hinkriegen kann. Also ich hatte euch sowas gezeigt. Was war das mit der Sonneneinstrahlung bei den Häusern, wo man sagt, ja, das sieht vielleicht irgendwie so aus. Also zu viel ist schlecht, zu wenig auch. Irgendwo in der Mitte liegt die Wahrheit, also was den Preis für das Haus angeht. Und

Sowas kriegt der KNM ganz gut hin, wenn er genug Trainingsdaten aus allen Bereichen hat und die lineare Regression natürlich gar nicht. Und hier muss man einfach nur einmal nachrechnen. Also der Kreis sagt ja schon, dass wir die fünf nächsten Nachbarn betrachten wollen und für diesen Punkt die Vorhersage machen wollen. Also fünf, habe ich gesagt. Und dann muss man einfach den Mittelwert von diesen fünf betrachten.

Instanzen ausrechnen, also die Zielvariable ist hier jeweils daneben geschrieben, der Wert der Zielvariablen und da kommt glaube ich 4 raus, oder? 3 davon sind eh gleich 4 und dann 2 und 6, da liegt die Mitte auch bei der 4. Okay, also Frage 4 wollte nicht... Okay, muss ich kurz langweilen. Okay.

Machen wir ein neues auf. Wenn man hier ein File-Widget nimmt, was sagt die Aufgabe? Die Aufgabe sagt, wir sollen da Daten wählen, die sozusagen mitgeliefert werden. Browse, Documentation, Data Step und wir nehmen Housing.tab. Mal schauen, ob wir das finden. Habt ihr das auch? Mitgeliefert, wenn ihr Orange installiert. So weit, so gut, oder? Ja.

Okay, und jetzt? Was ist das hier? Also jetzt steht hier, wir sollen eine lineare Regression, beziehungsweise wir sollten vielleicht noch, wahrscheinlich ist das Zielvariable hier schon gewählt, also m e d v ist die Zielvariable, was auch immer das bedeutet,

die vorhergesagt werden soll. Insofern brauchen wir kein Select Columns und machen direkt lineare Regressionen. Und dann steht da, man soll sich die Koeffizienten ausgeben lassen und im ersten Schritt vermutlich Rich Regression. Also, das war jetzt eigentlich das, was ich gesagt habe. Wir schauen da noch näher rein beim nächsten Mal. Insofern ist es eigentlich zu früh jetzt, aber machen wir sozusagen ein Preview. 0,5 stand in der Aufgabe.

Also was Ridge und Latto beide machen, ist die Komplexität des Modells zu reduzieren. Und man sagt, dass die Komplexität von so einer Formel eigentlich gegeben ist durch die Größe der Koeffizienten. Also je kleiner die Koeffizienten sind, desto größer wird dieser Intercept, also der Y-Achsenabschnitt sozusagen. Also hier sage ich jetzt wieder irgendwas mit Häusern voraus. Ich weiß nicht mehr genau, was die Variable bedeutet.

Aber fürs Verständnis können wir auch einfach wieder annehmen, wir sagen den Preis voraus. Dann würde großer Intercept bedeuten sozusagen, ja, wir sagen immer, was weiß ich, ein Haus kostet 200.000 und die Koeffizienten verändern dann nur noch minimal diesen Preis. Und wenn ich, das wäre viel Regularisierung, die das bewirkt sozusagen. Und bei Lasso ist es sogar meistens so, dass

viele Koeffizienten null werden, also so eine Art Feature Selection stattfindet, Attribute aussortiert werden damit und gar keinen Einfluss mehr haben. Und wenn ich es ganz ausschalte, dann werden wir sehen, dann sind die Koeffizienten tendenziell am größten. Also dann ist sozusagen der Preis sehr stark abhängig von den Features. Und was ist der Lastige? Das ist, glaube ich, Combi, wenn man beides, lass mal schauen. Wie soll man dann einen Feature Selection machen?

Das macht eine gute Instellation, um zu schauen, was die Formen mitzuholen werden wollen. Ja, also LASSO hat auch sozusagen meistens auf die Koeffizienten, die übrig bleiben, die werden auch nicht zu groß. Also ich bin eigentlich eher Fan von LASSO. Aber am Ende muss man es ausprobieren, dann wieder mit Test und Score und gucken, was am besten performt. Also Komplexität.

Ist ja was, dein Modell kann Komplexität haben und deine Daten haben eine bestimmte Komplexität. Und ich muss irgendwie finden, was am besten, oder so, dass es zueinander passt. Also das Modell muss so komplex sein, wie die Zusammenhänge in den Daten. Und wenn die Daten sehr einfache Zusammenhänge haben, dann sollte ich auch kein zu komplexes Modell nehmen, weil es dann eben Sachen aus den Daten versucht zu ziehen, die da gar nicht sind und dann Overfitting produziert.

Wenn die Daten sehr komplex sind, dann brauche ich auch ein komplexes Modell, um diese Komplexität irgendwie abzubilden. Und das ist Trial and Error am Ende. Wie melden wir dann jetzt auf jeder Modell? Nächste Woche. Ja, also okay. Du kannst sozusagen...

wenn du Test- und Score-Widget hast, einmal auf den Testdaten evaluieren und dann auf den Trainingsdaten. Und wenn du sozusagen was hast, was sehr stark overfittet, dann wird es auf den Trainingsdaten sehr gut dastehen und auf den Testdaten sehr viel schlechter. Und eins, was wenig Overfitting hat, ist auf den Trainingsdaten schon schlecht und auf den Testdaten. Also so ist es in der Praxis meistens. Aber der Unterschied ist nicht so groß. Das heißt auch nicht zwingend, dass du dich dann für das

entscheidest, was weniger Overfitting produziert, wenn es auf den Testdaten, wenn das komplexere Modell auf den Testdaten immer noch besser ist, vielleicht auch trotzdem da. Zwei Dezimalstellen, also jetzt habe ich Ridge Regression mit 5 gemacht und das Attribut hier soll ich mir anschauen, also minus 4,19, richtig? Moment, was hätte ich jetzt machen sollen? Probiert. Ohne Regularisierung, also jetzt pass auf.

Man muss es schon richtig lesen. Wir machen erst mal ohne. Da ist der Koeffizient für Nox minus 17,77. Richtig. So, jetzt machen wir Ridge. Das haben wir gerade gesehen. Das war minus 4,19. Und jetzt machen wir Lasso. Da wird er dann vermutlich 0, wenn ich die Übung richtig angelegt habe. Also das war das, was ich euch gesagt habe. Lasso führt eben oft dazu, dass so eine Art Feature Selection stattfindet.

Und dann nach 0 werden. Genau, also wenn du das jetzt als Formel hinschreibst, dann steht da sowas wie, also MEGV ist die Zielvariante, ich weiß gerade gar nicht, was das bedeutet, hat auch irgendwas, glaube ich, mit dem Preisniveau in einem bestimmten Neighborhood zu tun oder sowas. Ist dann also 30,26 oder ist es das mittlere Einkommen oder ich weiß nicht mehr.

Und dann schon dieses Krim, steht da für irgendwas wie Kriminalitätsrate oder so, fällt weg. Also ich könnte jetzt hinschreiben, 3026 minus 0 mal Krim, aber ja, kann ich mir auch nicht warnen. Also müssen wir hier noch 0 eintragen. Und dann, so wie ich gesagt habe, führt Lato also zu der Feature Selection. Also zum Beispiel dieses Nox-Attribut dann rausfliegt.

Okay, dann Pause. Okay, machen wir Pause. Bis Viertel nach. Okay, machen wir ein bisschen Input. Nicht zu viel. Wenn wir noch Zeit für andere Sachen haben.

Ich will euch kurz was erzählen. Einerseits, was gibt es für Verfahren, um zu evaluieren und was gibt es für Metriken, also die Zahlen, die dann in der Tabelle im Test- und Score-Widget stehen. Ihr habt ja gesehen, es gibt mehrere Spalten und wir wollen ein bisschen verstehen, was die bedeuten und was wir daraus lesen können. Fangen wir mit den Verfahren an. Also, was wir schon die ganze Zeit besprochen hatten, ich hatte das immer an die Tafel gezeichnet, dass man sagt, man macht

Man nimmt die Daten und man teilt sie. Oft nimmt man so zwei Drittel, ein Drittel und sagt, hier habe ich etwas, was ich zum Trainieren meines Modells benutze und den anderen Teil der Daten halte ich mir zurück. Die sind natürlich auch gelabelt, diese Daten. Das heißt, ich weiß, was zum Beispiel der Wert des Kassenattributes ist für jede Instanz in der Testmenge und kann dann das Modell, was ich hier trainiert habe,

auch die Instanzen anwenden und jeweils vergleichen. Vorhin haben wir ja auch gesehen bei den Predictions. Also ihr habt gesehen, was sagt das Modell und was war eigentlich der richtige Wert. Das nennt man Holdout oder Split oder vielleicht gibt es noch zwei andere Namen. In Orange heißt es. Mal schauen, ob ich das Richtige aufmache. Wobei, ja, genau, ich kann... Ja, jetzt habe ich alles gelöscht. Das war schlecht.

Das war jetzt unsere Skifahrersache. Wir beschäftigen uns jetzt bei den Metriken vor allem mit der Klassifikation wieder. Aber für die Verfahren ist es auch wurscht. Da können wir einfach schauen, was hier auf der linken Seite passiert. Also Random Sampling heißt es hier. Das, was ich Holdout nenne. Und hier seht ihr, man kann gar nicht weniger als zweimal das wiederholen.

Und hier werden zwei Drittel benutzt. Auch da so ein bisschen, wenn ihr jetzt fragt, ist zwei Drittel gut oder sollte man mehr oder weniger als Trainingsmenge nehmen? So ein bisschen kann man sagen, es hängt davon ab, wie viele Daten ich habe. Normalerweise würde ich möglichst viele Daten zum Trainieren benutzen, weil dann das Modell besser wird.

und möchte gleichzeitig aber natürlich eine verlässliche Aussage über die Qualität des Instanzen haben. Wenn ich jetzt eine Million Instanzen habe, dann brauche ich nicht ein Drittel dafür zum Evaluieren. Also ich brauche nicht 300.000 Instanzen zum Evaluieren. Eigentlich typischerweise reichen ein paar hundert, um eine gute Aussage zu haben. Das heißt, da kann ich dann eher so Richtung 97 Prozent oder so zum Trainieren nehmen, um wirklich das Maximal aus den Daten rauszuholen.

Ja, Dennis? Dann ist es eigentlich nur ein Paar und wenn es keine richtige Lösung gibt, kann ich es einfach ausprobieren und schauen, welches Modell die beste Position gibt oder die verschiedenen Methoden, die ich mir ausdrücke, die besten Werte. Aber ich habe ja nie 100% genau die richtige Entscheidung als Ingenieur.

Also gut, das Leben ist sowieso meistens so, dass du dir nie hundertprozentig sicher bist. Also ich weiß jetzt nicht genau, wohin die Frage geht. Aber was du nie machen solltest, ist dein Verfahren. Also zum Beispiel, wir sprechen jetzt gleich Kreuzvalidierung. Und wenn du bei Kreuzvalidierung bessere Ergebnisse kriegst als bei Holdout, dann heißt das nicht, dass du mit Kreuzvalidierung evaluieren solltest oder dass du dadurch ein besseres Modell bekommst oder so.

Du musst dann überlegen, was sagt mir das? Aber eigentlich solltest du dir vorher überlegen, wie du evaluieren willst. Also ich kann diese Faustregel noch weiter spinnen. Das hier, also das, was oben

steht, habe ich gerade erklärt. Kreuzvalidierung bedeutet, wenn man zum Beispiel zehnfache Kreuzvalidierung macht, dass man seine Daten in zehn Teile zerhackt und dann zehn Evaluierungsdurchläufe macht und in jedem Durchlauf, ich habe das hier so schön animiert, spielt also eines dieser

Stückchen vom Gesamtdatensatz die Rolle der Testmenge und man trainiert auf den anderen neun. Naja, jetzt ist es auch wieder gut. Das ist zum Beispiel auch was, was man macht, wenn man wenige Daten hat, weil man dann auch wieder das meiste aus, also dann wird jede Instanz mal zum Trainieren verwendet und man presst sozusagen das meiste aus den Daten raus. Also sozusagen die Faustregel ist, wenn ich sehr viele Daten habe, dann kann ich Holdout machen mit

97% Training und nur ein kleiner Teil Testmenge. Es muss sozusagen einfach eine bestimmte absolute Anzahl an Instanzen zum Evaluieren da sein. Und da würde ich sagen, so ein, zweihundert Stück geben meistens verlässliche Resultate. Und wenn mein Datensatz schon sehr viel kleiner ist und sowas kommt vor, dann ist zum Beispiel auch eine Art, die Verlässlichkeit zu erhöhen, dass ich Kreuzvalidierung nehme. Jetzt habe ich wahrscheinlich deine Frage nicht beantwortet.

Ich versuche es gerade zu überlegen, wie ich es noch besser formulieren kann. In der Programmierung beispielsweise habe ich eine Anforderung, fachliche Seite, und da ist es ziemlich egal, was ich dann für ein Programmieren mache. Die Antwort ist, wir bekommen ein

ein Pfeil an Daten und wir wissen dann ja wie selten herauskommt, können wir diese Daten irgendwie brauchen, um etwas schlussendlich dann auszusorgen. Aber ich kann ja eben verschiedene Aussagen treffen und ich habe ja auch eben verschiedene Modelle, die ich brauchen kann, einige eigentlich mehr, andere weniger. Und dort kann ich dann auch noch zusätzlich Sachen einstellen und

Also, ja, du musst dir erst überlegen, wie will ich evaluieren. Wenn ich dir den Datensatz gebe, dann siehst du, wie viele Instanzen habe ich. Sagen wir mal, du hast 200. Dann würde ich sagen, das ist nicht so viel. Da würde ich eher auf Kreuzvalidierung gehen. Und das entscheidest du aber, bevor du irgendwas anderes machst. Einfach anhand der Menge der Daten, die du hast und deinem Wunsch, dass du verlässliche Aussagen haben willst.

Und dann kommt der zweite Teil der Empirie, dass du verschiedene Algorithmen ausprobierst und guckst, welcher der Algorithmen in der Kreuzvalidierung am besten abschneidet. Auch da hast du vielleicht gewisse, sagen wir mal, Leitplanken, die dich führen oder Ideen. Wir hatten ja so eine Tabelle mit verschiedenen Kriterien und

was dir bei der Auswahl vielleicht hilft, aber letztlich bleibt es Trial and Error, absolut zugegeben. Oder du nimmst AutoML, also systematisierst sozusagen die Suche nach dem besten Algorithmus und den besten Parameterwerten, aber letztlich ist es ein Durchprobieren. Aber die wichtige Aussage ist, wie du evaluierst, entscheidest du vorher. Und du solltest nicht dann hinterher daran noch ändern und denken, du kannst irgendwas verbessern, sondern du musst vorher wissen,

wie du es machen willst. Da gibt es jetzt kein absolutes Richtig-Falsch. Die Faustregel ist eben wenig Daten, eher Kreuzvalidierung. Also übrigens die Extremform der Kreuzvalidierung ist, dass ich die Daten in so viele Teile zerhake, wie ich Instanzen habe. Das nennt sich dann Leave-One-Out, weil man immer eine Instanz auslässt. Das gibt es ja auch. Also das ist, wenn ich ganz wenige Daten habe, dann sozusagen gehe ich da in dieses Extrem.

Und je mehr Daten ich habe, desto eher kann ich so ein Random Sampling machen und auch die Größe der Trainingsmenge erhöhen. Ja, gut. Dann bringen wir das mal. Eine Sache, die manchmal noch interessant ist, jetzt nicht zu fest darauf eingehen,

Wann ist die Lernkurve interessant? Die Lernkurve ist interessant, wenn es teuer ist, Labels zu bekommen. Also zum Beispiel, weil ich es von Hand machen muss, zum Beispiel festlegen, ob ich bestimmte Dokumente relevant finde für meine Arbeit. Dann muss ich das von Hand durchgehen und labeln. Finde ich gut, finde ich nicht gut. Und das kostet Zeit und Geld und Nerven. Und deswegen möchte ich möglichst wenig

Instanzen selber labeln. Oder manchmal in der Medizin zum Beispiel, wenn ich eine Diagnose stellen will, dann gibt es manchmal die Möglichkeit, Gewissheit zu bekommen über ein bildgebendes Verfahren, was teuer ist. Also muss ich vielleicht ein CT machen oder so und möchte ich aber nicht unbedingt, wenn es nicht nötig ist. Das heißt, auch da versuche ich, mit möglichst wenigen gelabelten Daten auszukommen. Und die Lernkurve gibt mir einen Anhaltspunkt, ob ich noch weiter

Labels sammeln muss oder ob ich aufhören kann. Also was macht man da? Kann man auch mal hinzeichnen. Man braucht erstmal eine Testmenge, die man sich labelt. Sagen wir mal 100 oder so. Das muss man auf jeden Fall investieren. Und dann hat man hier, nehmen wir mal das Beispiel aus der Medizin, 500 Patienten, die vielleicht die Krankheit haben.

Und die Frage ist sozusagen, wenn ich ein verlässliches Modell lernen will, bei wieviel muss ich das CT machen? Und dann fange ich mal an. Nehme ich erstmal, also was haben wir hier? Erstmal die ersten zehn Patienten und dann nehme ich nochmal zehn und so weiter und besorge mir die Labels, mit denen ich das CT mache. Nehme erst das und dann evaluiere ich, wie gut ich bin. Dann nehme ich das hier noch dazu und evaluiere damit und so weiter.

und dann gibt es so eine Kurve. Die Kurve steigt meistens am Anfang steil und wird dann flacher. Und das ist dann der Punkt, wo ich mich freue. Meistens sehe ich gar nicht die ganze Kurve, weil ich, wenn ich zum Beispiel hier bin, ja, dann ist es logarithmisch hier unten. Das heißt, da hier bin ich schon dabei, dass ich nochmal hundert und weitere CPs mache. Und dann sehe ich, boah, also da habe ich jetzt, was ist hier, Accuracy, 1% Accuracy gewonnen und dann ist irgendwann der Punkt,

wo ich meinem Chef nicht mehr erklären kann, warum ich jetzt nochmal so viel Geld für CTs brauche, für so wenig Accuracy, die ich dadurch gewinne und dann sagt er, gut, lass uns aufhören. Okay, das waren die Verfahren, jetzt kommen die Metriken. Also sozusagen die Basismetrik ist hier gezeigt, also es ist keine Metrik, sozusagen das ist so die

die Grundzahlen, die man sich manchmal gerne anschaut, die sogenannte Confusion Matrix, wo ich sehe, welche Fehler das Modell macht. Nicht nur wie viele, sondern welche. Also in der Accuracy zähle ich ja, wie viele der Vorhersagen waren korrekt. Aber es gibt ja bei binärer Klassifikation zwei Arten von Fehlern. Ich kann entweder Ja sagen und es war eigentlich Nein, oder ich sage Nein und es war eigentlich Ja. Und oft sind die sehr unterschiedlich gelagert, diese zwei Fälle.

oder haben sehr unterschiedlichen Wert für ein Unternehmen. Da unten seht ihr noch ein paar andere Metriken, die man sich auch ausgeben lassen kann manchmal. Und die Accuracy, oder auch Genauigkeit auf Deutsch, die ist ja einfach zu verstehen. Also einfach gucken in der Testmenge, wie viele Instanzen wurden richtig klassifiziert, prozentual. Oder Fehlerrate wäre 100% minus Accuracy. Und auch einfach. Okay, jetzt...

haben wir bisher immer nur mit Accuracy gearbeitet. Und ich glaube, ich hatte schon mal angedeutet, dass es vielleicht auch einen Grund gibt, warum es noch andere Metriken braucht oder warum man nicht immer die Accuracy verwenden will, auch wenn sie so schön einfach zu verstehen ist. Und jetzt schauen wir uns zusammen was an in Orange. Das sind sehr alte Daten, aber es spielt eigentlich keine Rolle. Ihr werdet merken, dass sie sehr alt sind, weil es da drin Attribute hat, insbesondere Hobbys von Menschen.

Sowas wie PC-Owner oder, ja, wir werden es gleich sehen, oder ob jemand eine Stereo-Anlage hat. Also so Sachen, die man kaum noch kennt. Also die Attribute, mit denen die Menschen, also die Instanzen in dem Datensatz sind Menschen und es gab irgendwo auch mal eine Zielvariable, die besagt hat, wie viel jemand gespendet hat. Also das Modell sollte das vorher sagen. Die habe ich gar nicht mehr, die Zielvariable, die habe ich weggeschmissen. Ist auch egal.

Brauche ich nicht für das, was ich demonstrieren will. Was wir noch haben, sind die Hobbys der Menschen und dann gibt es Attribute. Man kann das hier sehen, zum Beispiel, was für Hobbys sie haben. Also zum Beispiel, ob sie Bibel lesen oder Katalog-Methoden machen oder ob sie Hausbier haben oder ein CD. Also wie gesagt, es ist aus den 90er Jahren des vergangenen Jahrhunderts.

Und diese Variablen hier, die besagen, wie oft, also die sind numerisch, da kann auch mal eine 3 drinstehen zum Beispiel, für bei Collectibles oder Books, nehmen wir Books, dann bedeutet das, wenn da eine 3 steht, dass es drei Gelegenheiten gab, wo die Person kontaktiert wurde, mit dem Angebot Bücher zu kaufen und wir wissen, dass sie darauf reagiert hat. Also wie oft hat jemand reagiert? Ich werde es dann gleich auch binär machen.

um bei einem Attribut und versuchen vorherzusagen als Prozent als Klassifikationsaufgabe sozusagen also ob jemand Interesse an Büchern hat zum Beispiel und dann haben wir noch das Geschlecht jetzt machen wir was verrücktes also lassen wir diese Daten laden in Orang war ich neues auch also ihr müsst einfach nur zugucken ich glaube ich hab die nicht auf Moodle die Daten wir müssen nichts machen einfach nur gucken um

Hobbys, da. Ja, ich muss leider ein bisschen vorverarbeiten, aber das ist auch gut, dann könnt ihr mal sehen, wie man zum Beispiel, doch, wir hatten auch schon mal Formula, ne? Wir machen mal Reset hier. Also genau, wir können mal gucken, mit einem kleinen Data Table. Also ihr seht hier die ganzen Hobbys, Yes oder No, ob jemand einen PC hat oder eine Stereoanlage und solche Sachen oder ein Boot und dann hier die numerischen Attribute,

Also hier zum Beispiel hat jemand einmal auf das Angebot reagiert, ein Family Magazine zu bestellen. Und hier hinten hat es noch das Geschlecht. Ja, beim Geschlecht gibt es natürlich männlich und weiblich, M und F. Es gibt aber auch noch drei weitere Werte. Die nehmen wir mal schnell raus, weil wir das Geschlecht auch gleich vorhersagen wollen. Das hat keinen ökonomischen Nutzen. Ich will daran nur was demonstrieren. Also ich schmeiße mal

Menschen raus, die nicht männlich oder weiblich sind. Das hat es mir auch gleich wieder gemerkt, was ich schon mal gemacht habe. Ich glaube, J hieße, dass es ein Joint Household ist, also dass zwei Leute leben, die unterschiedlichen Geschlechts sind. U heißt, glaube ich, Unknown. C weiß ich nicht mehr. Ist auch egal. Die sind jetzt jedenfalls weg, diese Zeilen. Und was mache ich noch?

Nachher werde ich vorhersagen, ob jemand Interesse an Mail-Magazines hat. Und deswegen mache ich da jetzt eine binäre Variable draus. Also die ist ja numerisch, habe ich gesagt. Also mache ich eine Formel und ich schätze mal, hat sie auch gemerkt. Jawohl. Also sie heißt Mail-Magazines, sie ist

kategorisch, seht ihr hier an dem C. Und die Formel sagt, der Wert von dieser Variable soll Ja sein, wenn...

jemand mehr als null Mal reagiert hat, und sonst eben nein. Logisch. Jetzt machen wir Select Columns. Und ich fange mal damit an. Nicht, ja, das machen wir so. Und das nehme ich weg. Und sag mal das Geschlecht vorher, okay? Wie gesagt, macht keinen Sinn aus ökonomischer Sicht. Jetzt ein Score. Was wollen wir nehmen? Vielleicht ein Tree? Das ist doch schon immer unser Lieblingsalgorithmus gewesen, oder? Ja.

Und wisst ihr noch, was die absolute Baseline war? Na, wir schauen erst mal, wie gut der Tree ist. Also, wir haben ja bisher nur von Accuracy gesprochen. Hier ist die Accuracy von diesem Tree 57,6. Ist das gut? Nein? Sollen wir es mal vergleichen mit unserer dümmsten Baseline, die uns einfällt? Wisst ihr noch, welches das ist? Welchen Algorithmus ich da wählen muss? Genau. Genau.

Was sagt Constant voraus? Was denkt ihr? Wisst ihr noch, wie der Algorithmus funktioniert? Ja, also entweder männlich oder weiblich. Ich glaube, es sind mehr Frauen. Wir können tatsächlich auch hier gucken. Predictions. Constant sagt immer S. Und der Tree, unterschiedliche Sachen. Jetzt ist der Tree irgendwie schlechter geworden. Ah nein, 57,6. Entschuldigung.

Ich habe in die falsche Spalte geguckt. Constant ist bei 56%. Das heißt, der Tree ist ungefähr so gut wie unsere super schlechte Baseline. Also bleibt ihr bei eurer Aussage, dass der Tree kein gutes Modell ist. Ist das korrekt? Okay. Jetzt, also wir können nochmal gucken, wie der Tree aussieht. Nur um zu gucken, dass der nicht trivial ist oder so. Was haben wir hier eingestellt? Tiefe 4, okay.

Dann müssen wir hier auch einen der Tiefe 4 nehmen. Irgendwie so. Ja gut. Benutzt irgendwelche Features. Sagt irgendwas vorher. Ist meistens falsch. Naja, nicht meistens. Jetzt lasst uns mal vorhersagen, ob jemand Interesse an Mail Magazines hat. Und zwar muss ich jetzt die numerische Variante rausnehmen hier. Und die Kategoriale nehme ich jetzt als Target. Will jemanden

Tipp abgeben, wie groß die Accuracy jetzt ist. Ist es jetzt leichter oder schwerer vorherzusagen als das Geschlecht? Ja, wahrscheinlich ist das Spiel einfacher. Sehr viel einfacher? Also du erwartest sehr viel genauere Vorhersagen? Ja. Tatsächlich ja. Also der Tree landet bei 97 und Constant auch. Jetzt muss ich mal gerade gucken, wie der Tree aussieht. Warte mal.

Hm, hab ich das hier richtig gemacht? Ich wollte eigentlich schon keinen trivialen Tree lernen. Sorry. Hatte du gesehen, was der Tree macht? Das ist ein trivialer Tree. Der sieht so aus. Also der hat ja auch genau die gleiche Accuracy wie Constance.

Und der macht auch genau das Gleiche. Also was sagt Constant vorher? Wir können mal gucken, hier hinten. Constant sagt immer No. Und der Tree auch. Also No heißt kein Interesse an Mail Magazines. Warum ist das so gut?

Ja, also wir können sagen, wenn wir hier so eine Feedback-Statistik uns anzeigen lassen, mal gucken. Ich brauche das nicht. Also man sieht, das ist eben nur eine verschwindend kleine

Menge von Leuten, die jemals reagiert haben. Das ist typisch im Marketing. Also egal, was man da verkauft, die meisten Leute reagieren nicht. Und das ist natürlich auch der Fall. Was jetzt doof ist, dass ich den Tree nicht überredet kriege, um flexer zu werden. Ah ja, schaut mal. Okay, also ich muss diesen unteren Haken wegmachen. Dann schauen, dass wir hier ungefähr das Gleiche machen. Tiefe 10.

Ihr habt gesehen, ein alles andere als trivialen Tree. Wo landet der? Warte mal. Habe ich wirklich die gleichen Einstellungen? Ja, müsste eigentlich passen. Aber das ist immer noch gleich. Was heißt denn das jetzt? Ist der jetzt gut oder schlecht, der Tree? Jetzt habe ich es ja auch darauf angelegt, euch zu verwirren. Ist mir wahrscheinlich auch genug.

Ja, Dennis, was denkst du? Die Aussage der Konstanz ist ja eben, es hat 94% der Daten, die reinkommen, werden nein sein. Und dort gibt es ja keine Berücksichtigung von TREEM, wenn man dann oben die anfängt mit einem spezifischen Datensatz, sondern es selber testen würde, würde man mit 97% der Wahrscheinlichkeit auch an einen richtigen Ort gelangen. Ich glaube, es ist ja aussagebefettigend.

Also wenn wir einfach generell sagen, auf ganze Daten, eben 97% dann nein. Das sind zwar schon genau die Kreisen, aber ich glaube... Das ist Zufall jetzt tatsächlich, glaube ich. Ja, lass uns mal Folgendes machen. Ich hänge hier mal so eine Confusion Matrix an, okay? Dass wir auch wirklich sehen, wo landen die Personen aus der Testmenge, in welcher dieser Zellen. Also welche Fehler wurden gemacht jeweils. Aber irgendwas stimmt hier nicht.

Also was habe ich hier? 50, 10, Binary. Das mit der 97 wird auch nicht so bleiben, glaube ich. Irgendwas stimmt da nicht. Sorry. So aus, ja. Man sieht, dass die anderen Zahlen sich verändern. Insofern müsste es eigentlich stimmen. Oh Mann.

Das stimmt, er sagt immer No. Dass wir was rausgenommen haben? Ah, jetzt. Okay, sorry. Jetzt kriege ich es hin. Jetzt wird es. Jetzt wird es.

Okay, also jetzt hat sich auch die Zahl verändert. Ihr seht also, die Accuracy ist jetzt gesunken nochmal. 96 Prozent. Also schlechter als die absolute Baseline. Und jetzt sehen wir hier, was passiert. Also Constant sieht so aus. Wir sehen hier Predicted, No oder Yes und Actual, No oder Yes. Das heißt, Constant sagt immer No. 184 Mal ist es nicht korrekt. Da gab es 184 Kunden, die hätten Interesse gehabt.

Und wenn ich jetzt gucke, was macht der Tree, der sagt auch manchmal Yes, insgesamt 99 Mal, 80 Mal umsonst sozusagen, hat die Person kein Interesse gehabt, 19 Mal kommt eine Antwort. Also hat der Tree jemand Richtiges erwischt sozusagen.

Okay, jetzt muss ich nochmal fragen, was denkt ihr jetzt, wenn man jetzt Constant betrachtet und Tree, welches Modell ist euch lieber? Vielleicht, wenn man mal an eine wirtschaftliche Perspektive denkt. Also ich hätte den Schritt genommen, weil Constant sagt immer Nein. Und wenn er jetzt immer Nein sagt, dann glaube ich, er hat ein Modell dafür.

Ja, also sozusagen gerade bei Fällen, wo das, was mich interessiert, nämlich die Antwort oder es passiert oft auch zum Beispiel bei Fraud Detection, da interessieren mich die Fraud Cases, die sind auch sehr selten, hoffentlich. Das heißt,

Immer dann, wenn die Sachen, die mich interessieren, sehr selten sind, dann ist es sehr leicht, ein Modell zu lernen, was sehr gute Accuracy hat, nämlich unser Constant-Modell. Also wenn es nur ein Prozent Fälle gibt, die von Interesse sind, dann ist es sehr leicht,

dann hat das Constant-Modell 99% Accuracy. Es sagt immer nein und in 99% der Fälle stimmt das. Ist aber völlig wertlos, weil wir die interessanten Fälle dadurch garantiert nicht finden. Das heißt, es ist dann eben doch besser, ein Modell zu nehmen, was auch ab und zu mal ja sagt und vielleicht auch zu

oft und dadurch in der Accuracy schlechter aussieht, so wie dieser Tree mit 96 statt 97% eigentlich schlechter aussieht. Aber immerhin

würden wir 19 Leuten vielleicht ein Abo für so ein Magazin verkaufen. Und mit Constant würden wir überhaupt gar nicht versuchen, irgendjemandem was zu verkaufen. Sorry, war jetzt ein bisschen schwierig gerade. Was ist sozusagen jetzt die Message gewesen?

Nehmt nicht Accuracy. Insbesondere, also ihr könnt es nehmen, wenn die Verteilung der beiden Klassen oder wie viele Klassen es auch sind, einigermaßen ausgewogen ist. Aber wenn es insbesondere die interessante Klasse sozusagen nur sehr, sehr, sehr wenige Beispiele dafür gibt, also zum Beispiel Marketing, die Anzahl der Leute, die reagiert haben, ist typischerweise sehr klein. Wenn das so ist, dann ist Accuracy wirklich

Ein Maß, was einen stark in die Irre führen kann, weil es eben Modelle belohnt, die sehr konservativ sind, sage ich mal, also die eigentlich nie Ja sagen. Und die nützen uns am Ende wenig, wirtschaftlich gesehen. Okay, das war die Aussage. Was gibt es noch? Also wir haben jetzt eigentlich nur Accuracy kennengelernt.

Precision, Recall und das F-Measure. Also, wenn man Precision und Recall verstanden hat, das F-Measure ist nur, dass man die zwei Zahlen Precision und Recall in eine Formel einsetzt, das sogenannte harmonische Mittel, und dann kommt ein Wert raus, der eigentlich nur dann groß wird, wenn sowohl Precision als auch Recall einigermaßen groß sind. Das ist die Idee von der Formel. Das heißt, wir wollen jetzt mal verstehen, was ist Precision und Recall. Diese Folie hier ist relativ, sage ich mal,

Die ist mathematisch formuliert, also hier habt ihr die Formeln, wie man es berechnet. Also das ist hier sozusagen meine Confusion Matrix. Und hier sind die Anzahlen, also A, B, C, D steht für Anzahl von Instanzen, die da jeweils reinfallen, so wie wir es gerade gesehen haben. Ich mag es eigentlich lieber in Worten erklären, das finde ich anschaulicher und einfacher. Also ich habe hier Beispiel Fraud Detection.

wenn bei der Fraud Detection die Precision 84% beträgt. Was bedeutet das? Das bedeutet, dass in 84% der Fälle, in denen das Modell sagt, das hier ist Betrug, das auch stimmt. Also man kann auch andersrum sagen, wenn die Precision sehr niedrig ist, dann ist viel Rauschen dabei bei dem, was es ausgibt oder viel falsche Alarme in diesem Fall. Recall beschreibt

wie viel von dem, was da gewesen wäre, habe ich gefunden. Also wenn bei der Fraud Detection der Recall 28% ist, dann heißt es, ich habe 28% der Betrugsfälle gefunden, die vorhanden waren. Oder andersrum gesagt, 72% der Betrugsfälle habe ich eben leider nicht gefunden. Ist klar? So intuitiv? Also jetzt kann man natürlich auch hier gucken, dass das stimmt. Also wenn ich sage, Precision ist

wie viele der vorhergesagten Betrugsfälle sind auch wirklich Betrugsfälle. Dann sind die vorhergesagten natürlich die Summe aus den beiden hier. Das heißt, ich nehme die echten, geteilt durch alle vorhergesagten. Und bei Recall nehme ich die echten, beziehungsweise die echten gefundenen, geteilt durch alle, die ich hätte finden müssen. Also die alle, die wirklich betrug sind. A plus B. Aber ich finde es einfacher.

einfach nur zu verstehen, was es bedeutet. So, also, wenn ihr auf das Test & Score Widget geht, dann habt ihr hier die letzten, nein, Entschuldigung, die vorletzten drei Spalten, also Precision, Recall und F-Measure. Und normalerweise macht man es so und macht es Sinn, dass man hier auf Yes stellt, also

dass man nur sich Precision und Recall für die, sagen wir mal, interessanten Fälle anschaut, also für die Leute, die

Interesse an Mail-Magazinen haben. Wie viele von denen habe ich gefunden? Also der Tree findet 10% von denen und Constant logischerweise 0%. Und die Precision, also knapp 20% der Fälle, wo das Modell Interesse vorhersagt, war auch wirklich Interesse da. So kann man das interpretieren. Das ist nochmal die erste Auszeit. Genau, das machen wir jetzt.

Also AOC ist ein bisschen kompliziert, was die Mathematik angeht, aber naja, es geht. Also wenn man es ganz high level erklären will oder sagen will, mir egal, wie das funktioniert, ich will einfach nur interpretieren, wann ein Modell gut ist, dann kann man sagen, der Wert ist theoretisch zwischen 0 und 1, aber eigentlich zwischen 0,5 und 1. 0,5 ist ein sehr schlechtes Modell und 1 ist das perfekte Modell.

Und funktioniert das Ganze darüber, dass die meisten oder eigentlich alle Klassifikatoren in der Lage sind, einem irgendwie so eine Art Confidence oder ein Score auszugeben. Der sagt, wie wahrscheinlich, also bleiben wir mal beim Targeted Marketing, wie wahrscheinlich ist es, dass die Person Interesse hat an was auch immer ich verkaufen will. Und die Frage ist, die Area under the Curve, die misst sozusagen,

inwieweit, wenn ich die Instanzen nach diesem Score sortiere, inwieweit wirklich die Leute, die auch wirklich dann reagiert haben und Interesse hatten, nach oben sortiert werden von diesem Score. Und ja, mathematisch kann man sagen, dass AOC die Wahrscheinlichkeit ist, dass ein zufällig gewähltes positives Beispiel höher als ein zufällig gewähltes negatives Beispiel gerankt wird.

Also wenn das 0,5 ist, dann ist es sozusagen eben zufällig. Dann macht mein Ranking nichts, was mir hilft. Das heißt, es ist eine zufällige Sortierung und die ist natürlich nutzlos. Wenn ich sie optimal sortiere, also erst alle Leute, die reagiert haben und dann alle Leute, die nicht reagiert haben, dann ergibt sich eben der Wert 1. Jetzt machen wir das einmal durch. Also

Beispiel Marketing wieder, nehmen wir wieder unseren Wintercheck und die Frage, ob jemand reagiert hat. Das heißt, diese echte Klasse hier, die in der letzten Spalte steht, also das sind Instanzen hier, jede Zeile ist ein Kunde von Swiss Bikes. Plus heißt, hat reagiert, Minus heißt, hat nicht reagiert. Und das ist sozusagen der Score, den der Klassifikator ausgibt. Und nach diesem Score haben wir die jetzt sortiert. Also hier glaubt das Modell, dass dieser Kunde mit 95% Wahrscheinlichkeit reagieren müsste.

was auch gestimmt hat, oder sagen wir mal, er hat reagiert oder sie, und hier denkt das Modell auch, dass es eine hohe Wahrscheinlichkeit ist, dass die Person reagiert, aber da war nicht der Fall. Genau, jetzt erzeugt man eine Kurve, und die Kurve wird erzeugt durch zwei Maße, einmal die True Positive Rate auf der Y-Achse und die False Positive Rate auf der X-Achse, und die True Positive Rate ist der Recall,

Und die false positive rate hat keinen anderen Namen, soviel ich weiß. Schauen wir uns das einfach mal im Beispiel an. Also, ich habe hier zwei Spalten hinzugefügt. Man berechnet sozusagen die true positive rate für jede Instanz. Also man geht diese sortierte Liste durch und nach jeder Instanz hält man an und berechnet die true positive rate. Also der Recall, habe ich gesagt, ist die true positive rate. Also wie viele...

Wenn ihr hier mal zählt, es gibt sechs Kunden, die reagiert haben. Das ist jetzt ein relativ gutes Ergebnis hier. Das heißt, Recall beginnt natürlich mit einem Sechstel. Also die erste Instanz, die erste

Person hier hat reagiert. Das heißt, ich habe jetzt von sechs Interessenten einen gefunden. Also bin ich bei 17% Recall.

False Positive Rate mit sozusagen wie viel, also ist sozusagen das Umgekehrte von Precision, also wie viel Prozent der Sachen, die ich gefunden habe, sind nicht interessierte Kunden, also sollte ich wegschmeißen und das ist Null und bleibt Null, auch wenn ich jetzt oder falsche Alarm steht hier, wenn ich jetzt weitergehe, bleibt Null und der Recall steigt an und bis hier und jetzt kommt also die erste

Instanz, wo ich da diese Wahrscheinlichkeit größer als ein halbes Plus vorher sage, aber es eigentlich Minus war. Jetzt stagniert also der Recall und wir werden es gleich sehen, dann gehe ich sozusagen auf der X-Akte ein bisschen nach rechts. Also jetzt habe ich von allen, die nicht interessiert waren, habe ich jetzt eins leider fälschlicherweise nach oben gerankt. Also am Ende ergibt sich so eine Kurve,

Wir waren jetzt sozusagen die ersten vier Zeilen, haben wir bis hier den Weg gemacht und jetzt biegen wir sozusagen ab. Also wenn ich es sehr schlecht mache oder wenn ich sie zufällig sortiere, dann ergibt sich sowas. Also bei so kleinen Daten, das ist eher so eine Treppe, wie sich dann ergibt. Immer wenn ich

Wenn ich von oben komme, ein Minus-Pell drin bin, gehe ich nach rechts. Wenn ich zum Plus bin, gehe ich nach oben. Da kann ich mir überlegen, das Optimale ist die orange Kurve, dass ich am Anfang nach oben gehe, also alle Positiven nach oben sortiere, bis ich sie alle habe. Dann gehe ich darüber. Wenn ich die Fläche anschaue, unter dieser orangen Kurve, ist der Wert 1. Wenn ich sie zufällig sortiere, dann ergibt sich diese Diagonale. Das könnt ihr auch ausrechnen.

dass die Fläche unter dieser diagonalen Einhalt ist. Also eigentlich muss man das jetzt nicht alles unbedingt verstanden haben. Man muss nur wissen, 0,5 ist wie sozusagen zufällig sortiert und 1 ist perfekt. Also die interessanten Fälle nach oben und nur interessante Fälle nach oben. Okay. Habt ihr noch Kraft für fünf Minuten? Ja.

Eine Bemerkung noch zu dieser Area Under the Curve. Die funktioniert eben auch, wenn es nur wenige von den interessanten Fällen gibt. Also sie ist sehr viel robuster als Accuracy gegenüber dieser Unbalanciertheit. Also wenn es nur ein Prozent Kunden gibt, die reagieren, dann wird Constant, kann man ja ganz schön sehen an dem Beispiel, trotzdem 0,5 als Wert haben.

Und ihr seht, der Tree ist in dem Fall nicht so wahnsinnig nah an der 1, also auch wirklich nicht gut. Aber immerhin etwas besser als zu raten oder zufällig zu sortieren. Also 0,529. Sagt also schon auch, dass es ein schlechtes Modell ist. Aber sozusagen dieses Maß ist sehr viel robuster gegenüber dieser Unbalanciertheit als die Accuracy. Aber jetzt kommt mein Favorite, was ich noch toller finde.

als Area Under the Curve zu benutzen, ist etwas, was man tatsächlich nicht angezeigt bekommt in Orange oder es auch nicht sich so einstellen kann, dass man es angezeigt bekommt, was man aber aus der Confusion Matrix leicht selber ausrechnen kann, wenn man sich denn eine eigene sogenannte Kostenmatrix erstellt hat. Die Kostenmatrix hat die gleiche Form wie die Confusion Matrix,

Und ich erkläre es mal an dem Wintercheck, was der Gedanke ist. Also der Gedanke ist, dass die verschiedenen Fehler, die ein Modell macht, unterschiedlichen finanziellen Impact haben sozusagen. Also wenn ich jemanden, nehmen wir mal an, ich schreibe Briefe, also wenn ich jemanden vom Wintercheck überzeugen will und einen Brief schreibe, der sagen wir mal einen Franken kostet und die Person reagiert nicht, naja, dann habe ich einen Franken verloren. Wenn ich

jemand, der eigentlich interessiert gewesen wäre, einen Brief schreibe, dann lasse ich mir eigentlich eine Chance entgehen, vermutlich mehr als einen Franken zu verdienen. Also hier habe ich es mal so aufgeschlüsselt und gesagt, nehmen wir mal an, der Kunde bezahlt jeweils 100 Franken für diesen Wintercheck. Ich habe intern Kosten von 60, das heißt, ich habe tatsächlich einen Gewinn von 40 Franken. Davon muss ich noch den einen Franken für den Werbebrief abziehen, den ich geschickt habe. Und dann ergibt sich sozusagen die Kostmatrix, die so aussieht wie der Rechner.

Also, wie erkläre ich mir so eine Kostenmatrix? Oder wie stelle ich sie auf? Und das wollen wir nach der Pause dann machen. Zwei Beispiele dürft ihr dann machen. Wie stelle ich sie auf? Eigentlich muss ich mir immer überlegen, was mache ich, wenn das Modell Ja sagt. In dem Fall würde ich dem Modell blind vertrauen und würde einen Brief schicken. Der Brief kostet einen Franken. In dem Fall, dass die Person nicht antwortet, habe ich diesen Franken als Kosten dastehen.

Wenn die Person antwortet, dann mache ich 40 Franken Gewinn minus den Franken für den Brief, also 39 Gewinn. Oder, weil ich immer an Kosten spreche, also minus 39 Kosten. Was mache ich, wenn das Modell No sagt? Dann mache ich gar nichts. Ich schicke keinen Brief und dann habe ich auch keine Kosten. Okay.

Das war es eigentlich schon. Das war die Idee. Und was man dann macht sozusagen, um diese Kostenmatrix anzuwenden, ist, wenn ich jetzt eine Confusion-Matrix habe, in der drinsteht, wie viele Leute hier jeweils hineinfallen aus der Testmenge, dann kann ich das einfach mit diesen Kosten multiplizieren, die Zahlen, die in der Confusion-Matrix stehen, und auch addieren. Also wenn es jetzt...

Ja, ich überlege gerade, können wir das machen? Sollen wir das machen? Hier ist es doch. Da habe ich doch alles auf der Folie. Ich habe auch nicht drauf geschrieben, wie die Kosten sind. Das heißt, ihr könnt es gerade selber ausrechnen. Also, was habe ich jetzt da? Ich habe meine Kostenmatrix, die ich euch gerade erklärt habe, auf der linken Seite. Und ich habe eine Testmenge von, ich glaube, 200 Menschen, also 200 Swissbikes-Kunden und ein Modell, was diese Confusion Matrix produziert, auf der Testmenge.

Also, zwölfmal zum Beispiel sagt das Modell Interesse vorher und es ist tatsächlich welches da. 67 Mal sagt es ja und der Kunde hat aber nicht mehr die Antwort. Könnt ihr mal schnell die Kosten ausdecken? Ja, ja. Ich sage, lass euch mal eine Minute damit ab.

Keiner sich verlässt auf... Wolltest du was fragen? Ja. Kann man auch. Also man kann auch hier eine 1 hinschreiben und da eine 40. Und das ist aber tatsächlich das Gleiche. Also ich kann Zahlen...

in einer Zeile verschieben, in dem ich das Vorzeichen ändere. Ja, ich werde mal zwei Minuten rechnen, oder eine, weil ich kann das mal gerade selber im Kopf versuchen. Habt ihr was raus? Ich sammle mal Vorschläge. Was könnten die Kosten sein? 60, 20. Ah, du meinst, das ist total? Ja, das ist total, genau. Ich will nur eine Zahl. Benny, ne? Das haben Sie zwar im Kopf gerechnet.

Gab es noch andere Ergebnisse? Kannst du sagen, was du gerechnet hast? Also theoretisch multipliziert man immer das, was in der oberen linken Zelle steht, mit dem, was hier in der oberen linken Zelle steht und macht das für jede Zelle und summiert es auf.

Man kann es auch in Worten sagen, was man tut. Also zwölfmal habe ich einen Brief geschickt und die Person hat das Fahrrad zum Wintercheck gebracht. Das heißt, zwölfmal habe ich 39 Franken Gewinn gemacht und 67 Mal habe ich einen Brief geschickt und keiner ist gekommen. Das heißt, ich habe 67 Franken umsonst ausgegeben für Briefe. Und hier 121 Mal sagt das Modell nichts.

oder sagt No und ich habe gar nichts gemacht und hatte keine Post. Und so ergibt sich dann sozusagen der Gewinn von 401 Franken. Das Modell Constant würde immer No sagen, weil es vermutlich auch auf der Trainingsmenge so ist, dass es mehr, warte mal, hier muss ich kurz, also auf der Testmenge gab es, die nicht geantwortet haben. Auf der Trainingsmenge wird es ähnlich sein. Das heißt, das Constant-Modell wird immer No sagen und hat dann, also alles in der Spalte,

und hat dann Kosten von null. Und wir sind aber besser. Also wir haben 400 Franken Gewinn. Es gäbe noch die andere Baseline, die Massenkampagne, dass ich einfach allen Leuten einen Brief schicke. Das könnt ihr ausrechnen. Vermutlich ist die, wenn es dumm läuft, besser als das. Vielleicht sind auch diese Kosten hier ein bisschen zu niedrig angesetzt, denn ich werde ja auch verärgert, wenn ich Briefe bekomme, die ich dann entsorgen muss, weil es mich nicht interessiert.

Okay, aber wir haben verstanden, wie man die Kosten berechnet. Das machen wir gleich als Übung und das machen wir dann am Ende noch. Jetzt machen wir Pause, oder? Ich habe gesehen, manche von euch brauchen Kaffee. Es geht schön los. Es geht schön los.

Jetzt würde ich diese Übung machen, die ist sozusagen eine Trockenübung, also ohne Orange, einfach nur

vier Zahlen sich überlegen. Warum vier Zahlen? Es geht darum, Kostenmatrizen aufzustellen und zwar für zwei Szenarien. Das erste ist relativ klar. Für den zweiten muss ich vielleicht kurz was sagen. In beiden Fällen müsst ihr einfach Annahmen machen. Also ich hatte ja erklärt, wie würde ich dran gehen. Ich würde mir erstmal überlegen, was macht das Unternehmen jeweils, wenn

das Modell das eine oder das andere vorhersagt. Also die Annahme ist immer, wir haben ein Modell und wir wollen wissen, wie gut es ist. Und erstmal muss man sich überlegen, was machen wir bei der einen oder der anderen Vorhersage. Und dann muss man die Fälle unterscheiden, dass die Vorhersage stimmt oder nicht und sich dann jeweils überlegen, was passiert und was kostet das. Und darüber, was die Sachen, die da passieren können, kosten,

Könnt ihr auch oder sollt ihr Annahmen machen? Ihr könnt sie auch X und Y nennen, die Kosten, aber ich finde es noch besser, wenn ihr euch Zahlen dazu ausdenkt, die einigermaßen plausibel sind, dann wird es plastischer. Also das erste Szenario ist das mit der Kreditvergabe in der Bank. Lasst uns annehmen, dass das Modell vorhersagt, ob jemand das Geld zurückzahlen wird oder nicht. Also so ein Modell haben wir gelernt aus

vergangenen Daten und das Modell sagt also entweder zahlt zurück oder zahlt nicht zurück. Mehr muss man dazu eigentlich nicht sagen, habe ich das Gefühl. Bei dem zweiten, es geht um eine Unfallversicherung. Also wenn es dort Betrug gibt, dann ist das meistens sozusagen in Absprache zwischen Arzt und Patient, die gemeinsam diesen Betrug begehen sozusagen.

Warum gibt es diesen Betrug? Hatte ich das schon mal erklärt? Also wenn ihr in der Krankenversicherung versichert seid, dann habt ihr meistens so einen Selbstbehalt und bei der Unfallversicherung nicht. Also wenn ich jetzt bei der Arbeit einen Unfall habe und ich verletze mir den linken Arm, dann muss ich das melden und dann kann ich den behandeln lassen ärztlich und kann die Rechnung eingeben und kriege sie voll erstattet.

Jetzt habe ich vielleicht einen Anreiz, wenn es an meinem rechten Arm auch wehtut, auch wenn es nichts mit dem Unfall zu tun hatte, das auch behandeln zu lassen und auch gleich noch mit auf die Rechnung setzen zu lassen und auch der Unfallversicherung unterzuschieben, die dafür ja eigentlich

nicht zuständig ist. Eigentlich müsste ich es bei der Krankenversicherung einreichen, dann hätte ich meinen Selbstbehalt. Also da ist die Motivation sowas.

wollen die aufdecken. Also könnt ihr euch jetzt wieder vorstellen, es gibt ein Modell, es kriegt eine Rechnung oder vielleicht eher eine Rechnungsposition als Input und sagt dann Betrug, ja oder nein. Und ja, wie gesagt, überlegt euch, was die Versicherung dann jeweils tut und was es kostet. Ich würde jetzt einfach mal so sagen, vielleicht wir schneiden hier so durch und sagen, Fensterseite macht...

Was wollt ihr machen? Vielleicht die Unfallversicherung und die Seite hier die Bank? Okay. Und ich denke, dass jeweils wahrscheinlich sich zwei Untergruppen bilden sollten. So ungefähr. Aber das überlasse ich euch. Okay. Also ich schätze so eine Stunde. Wie gesagt, es sind ja nur vier Zahlen.

Aber Viertelstunde, man sollte sich schon kurz Zeit nehmen zu überlegen. Vielleicht zehn Minuten, Viertelstunde, denke ich. Und dann schreiben wir sie an die Tafel, die zwei Kostenmatrizen. Okay, los geht's. Hallo. Hallo. Hallo. Hallo.

Vielen Dank.

Das ist mega cool. Das sieht auch gut aus.

Vielen Dank.

Okay.

Vielen Dank.

Ich glaube, es gibt am meisten...

Das war eine große Hoffnung. Aber das Leben ist alt.

Vielen Dank.

Ja, das ist gut.

Ja, das ist ja noch zu holen. Ja, das ist ja noch zu holen.

Also, er hat gesagt, dass er das nicht machen soll. Ich wollte nur noch strategisch sagen, dass er da noch zusammenarbeiten muss. Das ist mir auch nicht wichtig. Aber ich kann mich interessieren. Ja, das ist mir auch nicht wichtig. Ich kann mich nicht interessieren.

Vielen Dank.

Was ist denn das?

Vielen Dank.

Ja, das ist ein bisschen schwierig.

Ja.

Also ja, es gibt eine. Ist das so? Also wenn ich das so denke, dann kann ich das auch machen. Ich kann das auch machen.

Das ist wahrscheinlich nicht das beste Stolz.

Ja, also ich meine, wir entscheiden das. Ich habe noch nicht eins. Ja, es ist wie dein Mord, dass du dich auf den Menschen verliebst. Nein, nein. Also es ergeben sich unterschiedliche Karten. Man kann beides so machen. Genau.

Ich bin ganz sicher, dass die Verhütung nicht so gut ist. Ja, vielleicht haben Sie dafür einen gleichen Platz. Aber das ist ja auch kostenlos, oder? Also, wenn es modern heißt, dann ist es eigentlich eine sehr gute Sache.

Also, ja. Also, ja.

...

Also, da ist es so, dass es nur für einen Fall quasi die Kosten, die da losgehen. Zum Beispiel, die Kosten sind auch schnell. Die Mitarbeiter, die bezahlt werden, haben dann irgendwie ein ordentliches Geld. Und dann haben sie noch eine gute Lösung. Also, das ist der Weg.

Zeit. Ja, Zeit. Ja, Zeit.

Ja, das ist schon gut.

Ich weiß nicht, ob ich die Videos sehen kann. Achtung, die Frage ist über. Die zweite Frage. Wir können nicht mit Stunden und dann auch die Zeit. Aber ich glaube, wir können das.

Ja, das ist ein guter Punkt.

Ja, genau.

Ja, Mensch, ich sage es nicht so.

Ja, das ist ein ganz anderes Thema. Wenn es ein Nein-Nein ist, dann ist das ja auch okay. Das würde ich nicht mehr verfolgen. Ja, natürlich. Ja.

...

Erstmal muss er das Ei. Ja, ja. Ja, ja. Ja, ja.

Ja, das ist ein guter Punkt.

Ja, das ist richtig.

...

Ja, ich glaube, das ist eine gute Frage.

Vielen Dank.

Ich weiß nicht, was der Rechner ist. Ich weiß nicht, was der Rechner ist.

Vielen Dank.

Ja, ich habe eine Frage.

Das ist auch ein Teil der Technologie, die wir in der Zeit der Ermittlungszeit haben.

Wenn man faustregelt, ist der Abstand... Ja, es ist wirklich unfaustregelnd.

Denn wenn ich das noch einmal aufwende, dann kann ich das noch einmal aufwenden.

Okay, fangen wir doch mit. Ich glaube, ihr habt sehr ähnliche. Wenn wir anfangen. Okay.

Ja, ja, minus 200. Ja, nein, 10.000. Und minus 10.000. Und speichern, ähm, gestern meinte du...

Okay, ich glaube da gehören jetzt noch ein paar Anlagen dazu, die du vielleicht noch sagen könntest. Also Klick-Tür und sowas, oder? Ja, wir haben einfach gesagt, dass 10.000 einfach so die Einrichtung

auch hat. Insgesamt über 10.000. Das ist über 10.000 Franken. Und dann habt ihr wahrscheinlich bestimmt schon... Also erklär mal, was die Zahlen bedeuten.

Wenn man aus dem Sitzen wird erzielt. Das bedeutet 2%, oder? Ja. Und wenn das Modell Ja sagt, habt ihr gesagt, das ist Geld rausgegeben. Und wenn der Kunde es dann nicht zurückzahlt, dann haben wir die 10.000 Euro. Ja.

Und wenn das Modell Nein sagt, geben wir kein Geld raus. Das ist jetzt nicht. Habt ihr das Gleiche.

Ja, exakt. Minus 210.000. Ist es auch für die andere Seite klar, was die zwei Gruppen da...

Wollt ihr das nochmal erklären? Oder was ist das? Ja, es war ein bisschen unruhig. Also ich sage es nochmal schnell. Wenn das Modell Nein sagt, zahlt die Bank keinen Kredit aus. Sagt Nein, nein, kriegt kein Geld von uns. Dann passiert nichts. So war hier die Erklärung. Wenn das Modell Ja sagt, dann geben wir das Geld raus.

Der Fall hier unten ist, dass gar nicht zurückgezahlt wird, dann haben wir das Geld komplett verloren. Also die Annahme war 10.000 Franken ist die Kredithöhe, die ist hier komplett verloren. Wenn sie zurückgezahlt wird, die Kreditsumme, dann auch inklusive Zinsen, 2%. Die 10.000, die war ja eher unser Geld, das ist kein Beginn in dem Sinne, sondern nur die Zinsen, die 200 Franken, die die Bank dann daran verdient. Okay? Ja. Alle zufrieden? Ja.

Okay, jetzt wird es spannend. Erzähl mal, Unfallversicherung. Also, wenn es vorhergesagte Verflucht ist, dann sagt man, dass es Verflucht ist. Ja, warte mal kurz, Dennis. Mach mal kurz erstmal die Abklärung. Was macht die Versicherung, wenn das Modell Ja sagt? Aha, ja, es ist unser Interim Geschäftsprozess, das ist eine Verflucht, vorhergesagt wird.

finden entweder eine Abteilung statt, also dass wir an einen Software-Reich begreifen, wo wir analysieren, ob der Bedarf plausibel ist und beziehen da, falls nötig, Fachkräfte dazu, um diesen Vorwurf begründen zu können. Schön gesagt. Und was kostet das?

Ja, genau. Das ist schon sinnvoll. Irgendwas so im Schnitt. Durchschnittlich. Durchschnittlich. Ihr seid ein bisschen billiger geworden. Nicht so richtig günstig. Was passiert, wenn das Modell Nein sagt? Kein Betrug? Was macht die Versicherung? Aha, wir...

Wir zahlen total aus. Ja, das ist im Grunde das Geld. Gut, jetzt darfst du die Unterscheidung machen. Wenn wir Ja und Ja machen, dann kostet es ungefähr 500 Franken, obwohl man berücksichtigt kann, ob die Kosten von den Rückstandsverbrechen vom Schuldigen, aber wir sagen einfach mal,

Wir melden es, geben es ab, bis es vor Kost ist. Das ist dann anschließend passiert, dann haben wir nicht angeschaut. Also wir könnten natürlich sagen, lieber Kunde, du wolltest uns betrügen, das kostet die... Genau. Und dann, bei Nein, also bei Joe Megatrack, haben wir 1.500, weil... Also sorry, unten links, bei Joe Megatrack.

Ja, weil wenn es ein Betrug hervorsagt, aber es kein ist, der hat nicht den Kunden und es kann unter Umständen sogar auch eine Gegenwart geben, es können mehr Kosten entstehen. Wir müssen uns vielleicht eine Strafe zahlen, wir müssen uns vor Gericht schützen. So sind unsere Kosten höher und es kostet uns mehr, wenn wir faul stehen.

Ja, also was mich da ein bisschen irritiert, geht also davon aus, dass in einer nicht unerheblichen Zahl der Fälle die Abklärung nicht hilft. Also dass man trotzdem dann noch falsch liegt danach. Also normalerweise, wenn es kein Betrug ist, müsste die Abklärung es doch dann zu Tage fördern. Dann müsste doch der Experte sagen, nein, nein, das passt schon, zahl das mal.

dann würde ich mir nie 1'000 sparen, oder? Ja, deswegen sind das auch nur Klauseln des Durchschnitts. Weil bei so einem Gerät, wenn man die Kosten in der Regel hat, kann die bis 100'000 Franken kommen. Und eben durch das es immer weniger Pälle sind, die entstehen und die so weit vor Gericht getragen werden, ist entsprechend der Durchschnitt signativ angesetzt.

Aber gibt es da noch einen Punkt, wo man manuelle Daten sollten, sag ich mal, 90% der Fälle sollten rausgegeben werden? Auf der letzten Übertragung? Also ich mache das mal so. Ich mache mal die Eins in Klammern und wir sagen, in der perfekten Welt, wo alle Abklärungen erfolgreich sind, sind es nur 500. Und sonst wird es natürlich richtig teuer. Das heißt, auch wenn es nur wenige Fälle sind,

Genau, das Risiko für hohe Kosten. Gut, dann haben wir Nein, Ja, also oben rechts, das ist der Fall, wenn unser Modell Nein sagt, aber das Hauptsache, dass es gut ist. Und dies kann unterschiedlich hoch sein, also wir haben jetzt mal 85.000 als Vorschnitt genommen. Angenommen, wir haben einen Neubau, es gab einen Popalschaden,

Unfallfahrzeug muss ersetzt werden, Schmerzenssicherheit muss umgestellt werden. Aha, es geht eigentlich um medizinische Behandlung. Ist egal. Wie nenne ich das durchschnittlich? Glaubenskosten oder so. Ja. Wir müssen schauen, ob wir die Operationen schnell machen. Die Autowelt ist teuer. Ja.

Das ist der Fall, wo es eigentlich nur der Feindbetrug ist und dort haben wir die Schadenskosten. Also man kann nicht sagen, dass es die Feindsbrot gibt, das ist ja unsere Aufgabe als Versicherungsbüro. Ja, das war das, was ich im Doktor daraus wollte und habe es da nicht überzeugt. Ich habe nicht überlegt, dass es 85.000 Brücken ist. Was sagen wir? Ist es Minus?

Oder Plus? Es ist Plus. Wir haben hier sozusagen, wenn es Plus ist, sind es Kosten. Also schlecht. Und das ist ja auch so. Wir bezahlen hier 500.000, die eigentlich die Krankenversicherung bezahlen müsste. Das heißt, wir haben wirklich Verlust. Wir haben Kosten. Okay? Gut. Alles gut für alle? Ja.

Ja, also ich hätte die Zahlen ein bisschen kleiner gemacht, aber ansonsten ist ein Verstand. Also würden Sie sagen, die Kosten sind höher, wenn es tatsächlich ein Bezug ist und der Bezug nicht vorhergesagt wurde? Weil ich hätte gedacht, die Kosten sind höher, wenn es tatsächlich ein Bezug ist. Nein, wenn es kein Bezug ist, aber man hat vorgesagt, dass es ein Bezug sein wird.

Hat man jemanden, der dann anklagt, als Betrugstechniker? Das war ja das, was wir diskutiert haben. Es passiert hoffentlich nur selten, dass das vor Gericht landet, weil wir ja die Abklärung machen. Und wenn unsere Experten was taugen, dann sagen die schon und Bescheid, dass es eigentlich legitime Kosten sind und dann vermeiden wir die Kosten.

Gerichtskosten und den Reputationsschaden und was da noch alles ist. Das kann natürlich passieren. Aber es ist selten, wenn diese Abklärung nicht gibt. Okay? Gut. Wollen wir nochmal 10 Minuten Pause machen? Und dann habe ich noch eine kleine Übung für euch mit Orange, damit wir nicht ganz ohne Orange sind heute. Also Viertel nach, ja? Vorhin habe ich vergessen, dazu zu sagen. Bis wann?

Da war es ein bisschen länger. Jetzt machen wir es ein bisschen kürzer.

Okay, bereit für die Übung? Beziehungsweise, vielleicht bevor ich es vergesse, ist alles so auf gutem Weg mit eurem Assignment? Oder gibt es irgendwas, was vielleicht alle interessiert, was jemand aufhält, bedrückt?

Wann müsst ihr es abgeben? Gute Frage. Lass mal gucken. Also ich meine, ihr gebt sozusagen eine erste Version ab. Es hat noch keiner abgegeben, komisch. Am 30. Also noch ein bisschen mehr als 10 Tage. Und am 2. April haben wir dann

Erste Coaching-Session. Genau, ich hatte ja letzte Woche schon gesagt, ihr könnt jetzt langsam anfangen und vielleicht solltet ihr jetzt langsam anfangen. So, jetzt habe ich es nochmal gesagt. Gut, dann machen wir jetzt die Übung. Ihr findet sie, Kapitel 5, ein Verzeichnis. Wie so oft gibt es eine CSV-Datei mit Daten drin, die ihr ins Orange ladet und hier gibt es ein bisschen Kontext.

Es geht wieder um Marketing und ihr seht hier, die rufen die an und versuchen sie zu überzeugen, also es ist eine Bank und die rufen die Kunden an, versuchen sie zu überzeugen, in langfristige Einlagen zu investieren und jetzt wollen sie das gezielter machen und haben auch Daten gesammelt von vorher, mit denen sie ihr Modell trainiert haben, beziehungsweise das macht ihr jetzt, Modell trainieren.

Und hier unten gibt es noch ein bisschen Informationen dazu, was die Attribute bedeuten. Also insbesondere ist das hier interessant. Da gibt es die Variable, die heißt Y mit Werten Ja oder Nein. Das ist ja das, was wir vorher sagen wollten, nicht wahr? Genau. Und dann einfach den Anweisungen folgen. Ihr sollt hier diese Algorithmen gegeneinander antreten lassen. Und worum es dann geht, ist,

euch mal all diese ganzen Zahlen im Test & Score Widget anzuschauen und zu überlegen, was sagen die uns jetzt und welches Modell wollt ihr am Ende benutzen von den Vieren. Also Constant wird es wahrscheinlich nicht sein. Also sind es eigentlich noch drei, die zur Auswahl stehen. Und mal schauen, welches euch am besten gefällt. Und dann will ich natürlich auch wissen, warum. Soll wir das getrennt an der Nachpräsentation festmachen? Nein, müsst ihr nicht. Einfach

bereit sein, zu erklären, was ihr gemacht habt und könnt dann auch abstimmen oder wie auch immer, welches Modell euch am besten gefällt. Aber letztlich geht es natürlich darum, dann zu verstehen, worauf gucken wir und warum wählen wir dann ein oder das andere. Okay, macht ihr Gruppen ganz von selbst und wir, denke ich, können nach 20, 25 Minuten oder so mal gucken, wo ihr steht.

Ja.

Lasst mich das schnell reproduzieren. Was war das? Ja, also ich hatte euch ja gesagt, nicht zu viel über die Attribute nachdenken. Customer ID hätte ich jetzt mal rausgeschmissen, weil wir IDs, das sind ja Kundendienstanzen. Hallo.

Gleich vorbei. Dann habt ihr den Test in ScoreWidget gemacht. Und was habe ich gesagt? Also Constant soll da dran. Da gibt es nichts zu konfigurieren. Bei NaiveBase auch nicht. Das ist gut. Dann haben wir auch die gleichen Zahlen, hoffentlich. Beim Tree habe ich nicht gesagt, wie ihr den konfigurieren sollt. Ich mache mal so tiefe 10 vielleicht. Ich glaube, der Tree war eh nicht euer Favorit, oder? Am Ende. Ja.

Ja, weiß nicht. Und dann haben wir noch den logistische Regressierer. Gut, da könnte man ein bisschen was einstellen, aber macht, glaube ich, nicht so einen Unterschied. Und noch den Naive Base. Und dann sollten wir gleich Zahlen sehen. Und jetzt können wir natürlich abstimmen. Lässt es bei euch so aus? Ja.

Also bei mir ist jetzt Random Sampling hier ausgewählt. Bei Cross-Validation sollte jetzt nicht so wahnsinnig viel anderes rauskommen. Können wir mal kurz ausprobieren. Dauert natürlich ein bisschen. Was war schon mit der Cross-Validation? Also 10-fach ist das jetzt. Das heißt, er macht jetzt

10 Durchläufe und immer einen von 10 Teilen der Daten nimmt er jetzt. Ein bisschen anders ist es schon.

Machen wir mal random something. Habt ihr andere Zahlen? Ja. Bei welchen? Marc? Bei Knifebase haben wir nur die gleiche Funktion. Hier bei AOC? Wir haben zuerst...

Ja, sollte keinen Unterschied machen. Also das reicht eigentlich so, wie ich es gemacht habe. Du kannst die noch hier abzweigen und so und sie dir vielleicht anschauen und so, aber das braucht es eigentlich gar nicht mehr jetzt. Aber es ist auch egal, wir müssen auch nicht dieselben Zahlen haben.

Eine Sache, die ich euch alle gefragt habe, war nach dem Dropdown hier. Da steht jetzt Results for Target None, Show Average of All Classes. Das war nicht eure schlussendliche Einstellung wahrscheinlich, oder? Also ich kann hier auch Yes oder No wählen. Was bedeutet das? Und man sieht, wenn ich jetzt zum Beispiel Yes wähle, dass sich manche Zahlen verändern. Und zwar...

Die hinteren Precision Recall vor allem und die vorderen nicht. Und zwar werden die vor allem kleiner. Also sieht schlechter aus. Was habt ihr gewählt? None, No oder Yes? Yes. Warum? Also mit den meisten Gruppen hatte ich auch nochmal die Diskussion, dass man sich nochmal kurz vor Augen führt, was bedeuten eigentlich Precision und Recall?

Insbesondere vielleicht schauen wir mal auf Recall. Hier ist es mit Betrug. Wir müssen das nur übersetzen. Also wenn da steht, dass die logistische Regression, und jetzt lasse ich es mal bei Yes, ein Recall hat von 19%, dann bedeutet das, dass 19% der tatsächlich interessierten Kunden gefunden werden von dem Modell. Und

Der Naive Bayes findet annähernd 60 Prozent der interessierten Kunden. Das ist natürlich massiv besser. Und auf der anderen Seite müssen wir aber vielleicht auch Precision mit einbeziehen. Da sehen wir, dass es genau andersrum ist. Precision heißt, wenn das Modell vorhersagt, dass da Interesse ist, wie oft stimmt das dann? Und wir sehen, logistische Regression ist eher so zielgenauer. In 60 Prozent, knapp 60 Prozent der Fälle stimmt es, wenn es Ja sagt.

Und beim 90%-Basins-Neuen 30%. Also andersrum gesagt, 70% der Fälle rufe ich an und es ist kein Interesse da. Ja, also warum macht es Sinn, auf Yes zu stellen? Wenn ich auf No stelle, dann kriege ich Zahlen, die natürlich viel größer sind.

Insbesondere beim Recall sehe ich zum Beispiel beim Constant habe ich 100 Prozent. Klar, es ist einfach Kunden zu finden, die kein Interesse haben. Insbesondere Constant sagt immer No und findet dadurch alle, die kein Interesse haben. Das ist aber für mich völlig wertlos. Ich rufe niemanden an und verkaufe keine einzige Einlage. Ich will ja was verkaufen, deswegen muss ich auf Yes gehen. Vielleicht kann man es so einfach formulieren. Welches ist denn jetzt euer Favorit?

Ja, hätte ich auch gemacht. Ich glaube, alle haben das gemacht. Letztlich, oder? Fast alle, die ich besucht habe, haben eine Confusion Matrix noch drangehängt.

Und da? Jetzt plötzlich seid ihr billig, das ist sicher auch gesourcet im Ausland.

Also diese 600 meinst du hier, weil er 600 mal Ja sagt. Also genau, ich glaube, dass fast alle intern so eine kleine Kostenmatrix hatten und sich gesagt haben, Recall ist irgendwie wichtiger.

wir wollen lieber mehr finden, dann verkaufen wir mehr und das bringt uns mehr ein, als uns die Anrufe kosten. Also so wie du es jetzt gesagt hast, nächste Woche machen wir mit dem Beispiel weiter und dann stellen wir die Kostenmatrix auch ganz konkret auf. Und eben bei der logistischen

Regression, hier siehst du, ich würde 180 verkaufen, einlagen und hier nur 59. Das ist so das, was dann glaube ich, ich habe glaube ich so gehört.

dass die meisten von euch jetzt Naive Bayes bevorzugen würden am Ende. Ja, also es ging bei der Übung einfach darum, nochmal sich klarzumachen, was diese ganzen Zahlen bedeuten. Bei Precision Recall, gerade bei solchen binären Aufgaben, ist es meistens gut, da oben auf Yes zu stellen beim Dropdown. Also weil Yes meistens die interessanten Fälle sind, ob das jetzt Fraud ist oder Churn oder...

oder Interesse am Kauf, das ist eigentlich egal. Meistens ist es das Yes, was einen interessiert. Dann werden die Zahlen schlecht, weil es davon wenige gibt. Und dann, ja, dann habe ich vielleicht innerlich so eine Kostenmatrix. Und wir können das Ganze natürlich aber auch, haben wir jetzt noch nicht gemacht, machen wir nächste Woche mit einer Kostenmatrix, dann würde ich tatsächlich auch evaluieren und würden vermutlich auch feststellen, wenn es so ist, wie du sagst, Dennis, dass der Naive Base dann auch gewinnt. Können wir dann auch ausrechnen. Okay? Zeit ist um.

Ihr kriegt noch eine E-Mail, es gibt wieder ein Quiz und ansonsten sehen wir uns nächsten Mittwoch.