

Es gibt jetzt erstmal nicht so viel aufzuzeichnen. Wenn wir mit der Diskussion hier einsteigen, dann sage ich nicht mehr viel. Hallo? Ja, genau. Also ich würde es gerne folgendermaßen organisieren. Wie viele sind wir? 2, 4, 6, 8, 10, 12, 13. Ja, wir können mal versuchen, dass wir vier Gruppen machen.

Für jedes Thema, was hier auf den Folien ist, also da wir nicht so viele sind, können wir nicht alle Themen abdecken, macht aber nichts. Für jedes Thema brauchen wir zwei Gruppen. Und zwar brauchen wir immer eine Gruppe, die sagt, ja, es ist okay, das zu machen und sich wirklich da reinfühlt und Argumente sammelt und dann das vertritt und verteidigt. Und wir brauchen die andere Gruppe, die sagt, das ist überhaupt nicht okay, das darf man nicht machen und auch entsprechende Argumente sammelt. Und dann, also wir machen erst mal so zehn Minuten Gruppen,

Brainstorming in den Gruppen. Ihr überlegt euch Argumente und dann machen wir hier vorne so eine Podiumsdiskussion. Dann könnt ihr euch bekämpfen. Also versucht in die Rolle zu schlüpfen jeweils. Ich glaube, dass das erste gut ist. Also es ist eigentlich egal, wir können es mit dem oder mit dem machen. Nehmen wir ruhig die Kreditvergabe. Also das wäre mal das eine.

Ihr seht hier zwei Attribute, die die Bank gerne benutzen würde, um vorherzusagen, ob jemand seinen Kredit zurückzahlen wird oder nicht. Und natürlich auf Basis von dieser Vorhersage dann zu entscheiden, ob das Geld ausgerückt wird oder nicht. Nationalität einmal und Wohnort. Also die einen vertreten quasi die Sicht der Bank, die sagt, ja, ja, das wollen wir unbedingt machen, das ist gut, aus den und den Gründen. Und die anderen sagen, versuchen zu begründen, warum das nicht okay ist. Und dann...

Jetzt bin ich gerade noch spontan am überlegen, ob wir A oder B nehmen. Also wir sollten nicht beides nehmen. Ich glaube, wir nehmen A. Da geht es darum, dass ein Unternehmen ein Modell verwenden will, um auf der Basis von hochgeladenen Dokumenten, zum Beispiel CVs und was da noch alles hochgeladen wird, Zeugnisse und so ein Zeug, anschreiben, eine Vorauswahl von Bewerbern trifft. Also könnt ihr euch auch als einen Klassifikator vorstellen,

Der kann erstmal großzügig sein, das ist ja nur eine Vorauswahl, aber der sagt zumindest bei manchen Bewerbern, Bewerberin, nein, das aussortiert. Okay, ist klar? Erstmal so, das ethische Dilemma jeweils. Dann müssen wir mal gucken und aufteilen. Ja, machen wir doch einfach mal euch drei. Ihr seid die Bank, okay, für die Kreditvergabe. Und vier, wir drei, ja, wahrscheinlich gewisse Bankkunden oder halt die, die dagegen sind.

Ja, dann machen wir es einfach per Reihe, oder? Also die zweite Reihe ist das Unternehmen, was gerne die KI nutzen würde. Und in der letzten Reihe, ihr seid dagegen, dass Bewerbungen per KI vorausgewählt werden. Okay, zehn Minuten Brainstorming und dann machen wir die Bühne auf. Okay, ihr seid bereit, oder? Ja.

jetzt wo ihr richtig gut halten könnt können wir das machen ich hoffe dass da noch was neues dabei ist für euch in dem sinne dass wir jetzt für maschinen learning verwenden bisher habt ihr halten wahrscheinlich noch nicht so viel für maschinen eingesetzt dann ist was neues was will ich machen also wir hatten ja schon mal

Also wir hatten ja beim Thema Klassifikation die Übung gemacht, ich weiß nicht, ob ihr euch erinnert, mit der Firma Fixit und dem Ticket. Ihr erinnert euch, oder? Und dann hatten wir bei Regression die Übung mit den Skifahrern. Und bei der Klassifikationsaufgabe hatte ich ein Notebook bereitgestellt. Also ich habe hier in den Folien auch den...

den Link da drauf, wo das in Python gelöst ist. Im Prinzip genau das, was wir in Orange auch gemacht haben, in Python Code abgebildet. Ich habe da sogar immer die Bildchen von den Orange Widgets in das Notebook reingemacht, damit man es schön wiedererkennt. Und jetzt würde ich es gerne...

übertragen. Also ihr könnt euch an dem anderen Notebook jeweils inspirieren. So wahnsinnig unterschiedlich wird der Code am Ende nicht aussehen. An ein paar Stellen gibt es natürlich Unterschiede, sonst wäre es langweilig. Woher einfach, ja, ihr könnt einfach sozusagen den Code nehmen und anpassen oder ihr erarbeitet euch das selbst oder mit Chat-GPT oder mir ist das eigentlich egal. Ich habe hier mal so ein paar Hinweise gegeben, wie ihr es machen könnt. Also wir können das mal gerade zusammen machen, den ersten Schritt. Ihr geht auf Moodle.

und findet hier das Notebook. Unterster Link beim Kapitel 10. Und wenn ihr da drauf geht, dann kommt ihr zu GitHub. Da ist das Notebook. Da könnt ihr es nicht bearbeiten. Wenn ihr auf diesen Button hier klickt, Open in Colab, dann öffnet es sich. Also, ja, habt ihr einen Google-Account? Das ist, glaube ich, notwendig. Dann öffnet sich das

in Google Colab und ihr habt damit eine private Kopie. Also was auch immer ihr da tut, sieht sonst keiner. Bei mir funktioniert das irgendwie aus einem mir unerfindlichen Grunde nicht, zumindest nicht so richtig im Firefox-Browser. Normalerweise mein Standard-Browser ist, ich würde euch empfehlen, einen anderen zu nehmen, wenn ihr einen habt. Also hier habe ich das jetzt mal in Edge offen und wenn ihr auf diesen Open in Colab-Link klickt,

oder Button geklickt habt, dann solltet ihr sowas hier sehen. Ist das der Fall? Genau. Also ihr könnt jetzt loscoden. Aber natürlich fangen wir jetzt erst mal an, diese Zellen, die schon da sind, die ich für euch schon mal angelegt habe, auszuführen. Und die laden erst mal die Daten, machen ein paar Imports und dann kann es losgehen. Jeremy? Ich bin nicht dran. Anmelden muss man sich nicht. Also man könnte nicht sagen, anmelden, das sollten wir nicht haben. Wahrscheinlich bist du noch angemeldet. Ich starte über rechts anmelden. Okay.

Probier mal. Also ich meine, jetzt fangen wir mal an auszuführen. Also was passiert hier in der ersten Zelle? Da passiert irgendwie so eine Magie, die wir nicht näher anschauen müssen. Na, hallo? Da kommt auch diese Meldung. Ihr klickt einfach auf trotzdem ausführen. Ja, sehr gut. Ich habe schon gedacht, irgendwann wird es passieren. Und dann wird irgendwas ausgeführt und im Prinzip wird die Verbindung zur GitHub hergestellt, da habe ich die Daten nochmal abgelegt. Und ihr müsst euch nicht darum kümmern, wo die Daten sind und wie ihr die in dieses Notebook bekommt. Das habe ich jetzt hier für euch gemacht. Beziehungsweise der eigentliche Code, wo die Daten dann wirklich geladen werden, wo ist denn jetzt meine Maus? Der ist hier, ja? Also hier wird das CSV-File mit diesen Daten aus dem Skigebiet eingelesen, mit diesem Befehl hier. Habt ihr schon mal von DataFrames gehört? Super.

Es wird also in einen Data-Frame gelesen. Genau, hier diese Imports solltet ihr auch ausführen, kann vielleicht zwei, drei Sekunden dauern und dann, wenn ihr das hier ausführt, solltet ihr eine Vorschau auf die ersten fünf Zeilen von dem CSV kommen. Klappt das soweit? Fantastisch. Gut. Ja, also ihr wisst, wie ihr, arbeitet ihr im Python-Modul mit Notebooks? Also jetzt hier könnt ihr neuen Code euch so eine Zelle holen und dann tippen. Wie gesagt, bevor ihr jetzt anfängt zu coden, ich weiß ja, ihr seid jetzt heiß darauf, los zu coden, aber vielleicht noch jetzt die paar Hinweise. Ich habe hier ein Bild gemacht von dem Orange-Workflow, den ich möchte, dass ihr abbildet. Und auf der nächsten Folie, die lasse ich dann einfach offen auf dem

Screen hier vorne steht so die Schritte, die ihr machen sollt und die eigentlich wirklich diesen Workflow abbilden. Wir können uns den hier nochmal in meinem Orange anschauen. Also hier werden die Daten geladen. Vielleicht erinnert ihr euch noch an das Attribut Month. Da hatten wir gesagt, das sollte man vielleicht nicht als numerisch behandeln, vor allem wegen dem Dezember, weil das Modell gerne lernt sonst, wenn es numerisch ist, dass je größer der Month ist, desto weniger Leute kommen und im Dezember stimmt das dann nicht.

Also hatten wir den hier kategorisch gemacht. Das müsst ihr dann irgendwie auch in Python hinkriegen. Das hier müsstet ihr abbilden. Dafür gibt es auch eben in dem anderen Notebook von der Fixit-Übung Code, den ihr ganz leicht anpassen könnt. Ihr müsst auch nicht alle diese Attribute hier nehmen. Macht eure eigene Auswahl. Vielleicht erinnert ihr euch noch, welche Attribute wirklich geholfen haben oder...

Es kommt nicht drauf an, ihr müsst kein super Modell hintunen, sondern einfach mit irgendwelchen Attributen arbeiten, die plausibel erscheinen. Gut, und dann sollte es zwei Sachen geben. Einerseits mal einfach eine lineare Regression trainieren und euch dann hier hinten hatten wir ja die Koeffizienten uns ausgeben lassen, die könnt ihr euch auch in Python ausgeben lassen. Dafür gibt es bei Fixit keine Vorlage, das findet ihr so raus, mithilfe von Google oder ChatGPT.

einfach diese Koeffizienten mal ausgeben lassen. Und das Zweite, was ihr dann lernen sollt zu machen, ist einfach mal so ein trainiertes Modell dann auch zu testen. Also im Prinzip das, was wir in Orange immer mit dem Test- und Score-Widget machen, hatte ich auch bei dem... Also lass mal gerade gucken, wo es ist. Das ist jetzt hier gerade nicht in Colab geöffnet, könnte ich vielleicht aber machen schnell. Das war...

Nein, das ist wieder Regression. Nehmen wir Fix It. Also in den Folien ist der Link oder sonst könnt ihr auch nochmal ins Kapitel 3 auf Moodle navigieren und da findet ihr auch den Link und dann könnt ihr das quasi einfach in einem anderen Tab euch aufmachen und dazwischen hin und her springen, wenn ihr wollt. Da seht ihr, da werden auch erst die Daten gelesen. Da werden dann auch hier sozusagen die Attribute gewählt und das Klassenattribut und dann wird ein Modell trainiert und hier unten

geht es dann los, dass dieses Random Sampling, wie es in Orange heißt, gemacht wird. Also eine Aufteilung in Trainings- und Testmenge. Also ihr sollt jetzt keine Kreuzvalidierung oder sowas machen, sondern ihr macht es so, wie es in dem Orange Workflow ist. Einfach dieses, ihr könnt im Prinzip diesen Code mehr oder weniger übernehmen. Es wird natürlich eine andere Metrik benötigt, weil Accuracy für Regressionen nicht funktioniert.

Da habe ich auf die Folie geschrieben, nehmt Mean Absolute Error, das war auch das, was wir damals immer angesehen hatten. Da müsst ihr nur rausfinden, wie das geht in Python. Und ja, also hier könnt ihr euch bedienen. Was wollte ich noch sagen? Genau, ich war noch hier. Also am Ende solltet ihr dann auch zwei Zahlen haben. Also ihr solltet ein lineares Regressionsmodell trainieren und evaluieren und einmal Gradient Boosting.

trainieren und evaluieren und dann solltet ihr hier so zwei Zahlen haben. Vermutlich werden es nicht die gleichen sein, kommt sicher was anderes raus, auch weil ihr anders sampelt, aber zwei Zahlen solltet ihr haben. Vermutlich sollte die Zahl für XGBoost oder das Gradient Boosting sollte kleiner sein als die von der linearen Regression, also es sollte besser funktionieren. Das schon, glaube ich, sollte so sein. Also, ihr werdet auch sehen, dass manchmal Fehler kommen,

Und dann sollte Colab euch auch anbieten, seine eigene KI einzuschalten und könnt ihr natürlich auch gerne nutzen. Versucht zu verstehen, was es euch vorschlägt und dann könnt ihr jetzt ausprobieren, den Vorschlag anzunehmen und gucken, ob der Fehler dann weggeht. Funktioniert eigentlich ganz gut. Genau. Und hier sind nochmal die Schritte. Also im Prinzip geht das den Workflow entlang. Das hier kann man eigentlich auch umdrehen, wobei eben ihr könnt

Ihr könnt das lineare Regressionsmodell einfach mal nur auf den Trainingsdaten trainieren. Hier haben wir es ja auf allen Daten trainiert. Ich trainiere jetzt mal in Python nur auf den Trainingsdaten. Ist ungefähr Gesundheit klar, was zu tun ist? Fangt mal an. Ich komme sonst rum und helfe, gebe euch Tipps. Aber wichtig ist, keine Angst vor irgendwas. Einfach machen, gucken, was passiert. Okay? Ist immer gut beim Coden.

Auch mit Orange haben wir es ja immer so gemacht, oder? Einfach mal machen. Irgendwas wird schon passieren. Genau, soweit waren wir gekommen. Ja, also ich würde jetzt vielleicht mal gerade so von Hand weiter coden und erläutern, weil es, glaube ich, leichter zu verstehen ist, als wenn ich jetzt einfach eine Lösung zeige.

Ich mache es ohne XGBoost und XGBoost zeige ich dann noch schnell. Ist dann nicht so aufregend. Also wir machen jetzt sozusagen damit weiter. Was haben wir gemacht? Ihr habt es ja beim Abtippen genau verstanden. Jede Zeile Code. Oder wir gucken nochmal. Wo ist meine Maus? Ich mache es mit der Hand. Was haben wir gemacht? Also wir haben jetzt dafür gesorgt, dass fehlende Werte gelöscht wurden.

Dann haben wir diesen Train-Test-Split gemacht. Also das, was im Orange das Test-and-Score-Widget macht, mit Random-Sampling, haben wir jetzt hier in einer Zeile Code. Und das hier ist sozusagen erstmal nichts. Da machen wir sozusagen nur eine Variable. Die ist ein Objekt vom Typ Linear Regression. Und hier in dieser Zeile wird dann dieses Modell trainiert. Das ist bei sehr vielen Modellen möglich.

wird die Funktion fit oder ist die Funktion fit vorhanden und die tut das. Also sie wird auch immer mit diesen zwei Parametern aufgerufen. Also man gibt immer die Trainingsdaten und zwar die Attribute, Komma und dann das Zielvariante, das Klassenattribut. Also in dem Fall Zielvariante. Okay, jetzt ...

Muss ich mal selber gucken, ob das funktioniert. Also was wir jetzt machen wollen, im Orange sieht das so aus. Wir haben jetzt das trainiert und jetzt wollen wir hier sozusagen das hier abbilden, dass wir uns hier so eine schöne Überblick über die Koeffizienten verschaffen können. Muss ich mal selber ausprobieren. Da habe ich das damals, glaube ich, auch gegoogelt. Also es gibt die Möglichkeit zu sagen,

Coefficient ist gleich PD.DataFrame und dann schlägt er mir hier was vor und das könnte, glaube ich, funktionieren. Also, was da passiert ist, er konstruiert mir ein DataFrame und DataFrames sind auch immer schön auszugeben. Ach so, ja. Danke. Soll ich nochmal zurück machen? Ich mach's nochmal zurück. Also, ich hab getippt, Coefficient ist gleich. Weiß nicht, ob das bei euch auch so ist. Ja.

So, jetzt fängt er an, mir vorzuschlagen. Ich weiß nicht, ob das bei euch auch funktioniert. Ich nehme das jetzt einmal mit dem Tab mal an, was er mir da vorschlägt. Also, er konstruiert einen DataFrame mit zwei Spalten. Genauso wie im Orange eine Spalte mit dem Namen des Features. Das sind hier, sagen wir, die Liste der Spalten von meinem Attribut.

und trennt Tabelle sozusagen und eine Spalte, wo die Koeffizienten drinstehen. Und das müsste eigentlich funktionieren. Ich glaube, so funktioniert es nicht. Doch, eigentlich auch. Okay, jetzt

können wir nochmal kurz gucken, ob es plausibel ist, was da rauskommt. Also zum Beispiel bei Monat haben wir jetzt dieses One-Hotting-Solding und wir sehen wiederum, dass der Februar mit einem sehr hohen

Koeffizienten verknüpft ist. Das heißt, da kommen die meisten Besucher gefolgt vom Januar, März, Januar, März und dann müsste Dezember kommen und so weiter. Ist plausibel, oder? Wir sehen auch zum Beispiel, wenn Ferien sind, dann kommen mehr Leute. Also das ist ein hoher positiver Koeffizient, auch am Wochenende. Ich habe jetzt hier, anders als zum Beispiel Katharina, ein

negativen, oder wer hatte das? Weiß ich nicht mehr, negativen Koeffizienten für die Temperatur, was ich plausibel finde. Gerade im Winter kommen wahrscheinlich mehr Leute, wenn es kalt genug ist. So, fehlt uns noch was? Wir könnten noch schnell gucken, wie gut unser Modell ist, oder? Also das, was das Test-and-Score-Widget dann auch noch macht, nämlich und hinten so eine Tabelle rausspucken mit verschiedenen Werten und wir konzentrieren uns mal nur auf diese. Also wir lassen uns jetzt nur mal von der linearen Regression,

den MinAbstruthError ausgeben, der wird wahrscheinlich auch ungefähr in dieser Größenordnung 1500 sein. Ich glaube, er ist ein bisschen niedriger. Warum auch immer. Also, was machen wir? Wir machen, man macht normalerweise so, man lässt sich jetzt die, okay, lass mich überzeugen, mal schauen, ob es Sinn macht. Model.predict ist eine Funktion, die man aufrufen kann auf den

Testattributen und die einem dann die Vorhersagen ausgibt und die speichern wir jetzt hier in so einem Vektor sozusagen. Da steht dann also den Zahlen drin, wie viele Besucher erwartet werden an jedem Tag. Und dann nehme ich diese Vorhersage und vergleiche sie mit den wahren Werten, also wie viele Besucher wirklich gekommen sind. Tatsächlich, man gibt erst die wahren Werte und dann die vorhergesagten Werte ein. Und

Schauen wir mal, ob das so funktioniert. Lass mir dann den Mean Absolute Error ausgeben. Ich habe doch die fehlenden Werte genommen. Vielleicht muss ich mal noch mehr... Ja, ganz schlecht. Noch besser als er war. Ich war vorher bei 1400. Warum auch immer das jetzt so gut rauskommt. Also ihr seht, sozusagen das Modell dann zu evaluieren oder den Mean Absolute Error zu bestimmen, sind dann nicht mehr so viele Zeilenquote.

Jetzt können wir auch kurz meine Lösung angucken, die schalte ich dann nachher noch frei in Moodle. Im Prinzip ist das das, was wir hier gemacht haben. Ein bisschen komplizierter hier der Code, aber es kommen auch Koeffizienten raus. Sollten die gleichen sein. Ich habe eben jetzt hier noch einen kurzen Code-Abschnitt, wo ich

meinen XGBoost trainiere. Hier kann ich noch spielen. Ich habe jetzt 100 Bäume genommen und mit maximaler Tiefe 6, also ein ganz schön komplexes Model. Müsste danach hoffentlich gut sein. Und dann seht ihr hier sozusagen, mache ich jetzt zwei Predictions, eine für lineare Regression, eine für XGBoost. XGBoost, habe hier zwei verschiedene Variablen und dann mache ich auch zweimal die Berechnung des

Min Absolute Error eben für die beiden Vorhersagen und gibt dann die zwei Zahlen aus und ja, hier ist der extra Boost erwartungsgemäß besser um die 1000. Das war ja auch damals das, was wir so als Limit erfahren hatten. Okay, was habt ihr gelernt? Frustration. Ja, also ich glaube, wenn man jetzt, also irgendjemand hat es auch probiert, einfach

Ihr könnt auch, wenn ihr ChatGPT Plus habt zum Beispiel, den ganzen Datensatz da hochladen und dann sagen, bau mir ein lineares Regressionsmodell und evaluiere es und dann schreibt ihr den

Python-Code und führt ihn aus. Und dann kriegt ihr Ergebnisse. Was man dann merkt, ist, dass der Code teilweise ganz schön komplex ist. Sterling, du hast es gemacht. Hat es geklappt? Ja. Okay.

Ja, also es gibt unendliche Möglichkeiten, wie ihr euch helfen lassen könnt. Und ich glaube, vielleicht ist das auch der Schlüssel sozusagen, dass ihr lernen müsst, wo ihr euch Hilfe holt und wie ihr die Hilfe so holt, dass ihr auch versteht, was passiert. Also das soll jetzt nicht sozusagen, wir haben jetzt relativ nah am Code geschafft und haben versucht zu verstehen. Es war aber auch, wenn wir jetzt ein bisschen zurücktreten und reflektieren,

Ein bisschen das Ziel, dass ihr auch mal diese Erfahrung macht, wie das ist, Machine Learning mit Python zu coden. Und da draußen gibt es neben ChatGPT auch noch 100.000 Notebooks, die ihr als Inspiration, also nicht nur fixit, sondern es gibt auch mindestens 10.000 Leute, die schon lineare Regressionen auf irgendwelchen Daten gemacht haben und das Notebook irgendwo verfügbar gemacht haben. Das findet ihr und dann könnt ihr es übernehmen und auf eure Daten anpassen.

Wenn man das jetzt ein, zwei Mal gemacht hat, sieht irgendwie immer gleich aus. Es gibt natürlich dann hier und da noch was zu tunen und alle anderen Algorithmen auszuprobieren, aber das Code-Gerüst bleibt immer gleich. Okay? Wollt ihr nochmal eine kurze Pause machen? Was ist die Meinung der anderen?

Oder machen wir einen Bio-Break bis um vier, nur sieben Minuten als Kompromiss? Ja, also generell stellt ihr euch die Prüfung ähnlich vor wie beim letzten Jahr, im Sinne, dass es ein paar Multiple-Choice-Aufgaben geben wird. Die ähneln vermutlich so ein bisschen den Quizfragen, die ihr kennt. Und dann wird es auch

Ich denke mal, ihr werdet mindestens oder eher ein paar mehr Punkte sammeln können mit den anderen Aufgaben, die dann bei Text sozusagen sind, wo ihr irgendwas beschreibt, begründet, berechnet und so weiter. Genau. Wir hatten gesagt, vielleicht noch die allgemeinen Fragen, ein Cheat-Cheat, hatten wir gesagt, dürft ihr beidseitig beschreiben, von Hand oder, nee, von Hand, hatten wir gesagt. Oder iPad. Okay.

Gut, aber von Hand, letztlich. Und irgendwo kam noch die Frage auf, ob man sowohl bei dem Assignment als auch bei der Klausur eine bestimmte Note erreichen muss. Die Antwort ist nein, vielleicht nächstes Jahr, aber dieses Jahr wird es einfach gemittelt. Also das Mittel aus beiden Noten muss über 3,8 liegen. Also man kann

Gewisse Defizite bei der Klausur ausgleichen mit dem Projekt. Meistens läuft es so rum, nicht andersrum. Aber natürlich geht es auch andersrum. Genau. Also von den Themen, wir haben angefangen, überhaupt erstmal zu verstehen, was ist ein Muster? Multivariates Muster. Seht ihr auch in der alten Prüfung, da war auch nach einem multivariaten Muster gefragt. Also am besten, ihr erinnert euch an die Klassifikationsaufgabe und sagt, okay, unerfahrene Agents,

die System Requests bearbeiten müssen, die führen zu SLA Violations. Das war eins von unseren Beispielen für ein Multivariatsmuster, also Kombination von verschiedenen Attributwerten. Dann hatten wir geübt, ihr könnetet bekommen eine natürlichsprachliche Beschreibung von irgendwas, was vorhergesagt werden soll und solltet es formalisieren. Also müsst ihr sagen, was sind die Instanzen,

Was sind die Features oder Attribute und was ist die Zielvariable oder das Kassenprobe? Genau. Dann haben wir ein paar Daten-Aufbereitungsschritte kennengelernt. Also zum Beispiel sowas wie pivotieren oder Umgang mit fehlenden Werten oder sowas. Faut euch nochmal an. Das war Kapitel 2, die Quälerei Pablo Prep. Ja, ich würde mir da vor allem nochmal die Quizfragen schauen. Okay, ähm.

Wir hatten auch manchmal gesehen, insbesondere bei numerischen Attributen, das heute auch wieder, wir haben es nochmal mit dem Month gesehen, dass man den nicht numerisch lassen möchte, weil dann damit gerechnet wird und es macht nicht so viel Sinn, weil eben Dezember und Januar sich ähnlicher sind als 1 und 12. Wir hatten es auch bei der Klassifikation, hatten wir da auch was. Agent ID zum Beispiel würde man auch nicht als numerisches Attribut lassen. Also, dass ihr da so ein bisschen

vorgewarnt seid, dass manchmal numerische oder Attribute numerisch aussehen, das aber eigentlich nicht Sinn und dass es dann nicht so viel Sinn macht, mit denen zu rechnen oder numerische Vergleiche anzustellen. Dann haben wir angefangen, Modelle zu interpretieren, zum Beispiel Entscheidungsbäume, aber auch andere. Das hau ich vielleicht nochmal an. Wir hatten auch Regellerner,

kommen dann vielleicht ganz viele Regeln raus. Und guckt euch nochmal an, auch bei Trees oder bei den Regelnern, wir hatten so ein paar Prinzipien, wie man es dann runterbricht auf das, was einen wirklich interessiert. Also zum Beispiel, wenn man jetzt sagt, das Management interessiert sich für die Violations, dann gucke ich im Entscheidungsbaum eben nur die Blattknoten an, die Violation vorhersagen. Dann habe ich schon mal deutlich weniger. Und dann...

Gab es da noch Zahlen in diesen Knoten, die sagen, auf wie viele Instanzen lässt sich diese Kombination von Bedingungen anwenden? Und je größer die Zahl, desto interessanter findet das Management das. Also hier habe ich auch hingeschrieben, wo ist es jetzt? Handlungsempfehlungen und gucken, was ist besonders wichtig und häufig. Und da würde man natürlich die Handlung dann priorisieren.

Ja, ich habe jetzt das mit dem Post-Hoc, auch wenn es im Semester sehr viel später kam, da dazu geschrieben, weil es dazu passt. Also zum Thema Interpretieren von Modellen hatten wir dann später noch geschaut. Es geht erstmal für manche Modelle von sich aus, also für logistische Regressionen, für Bäume, für Regellerner. Wenn wir sowas wie XGBoost oder Random Forest haben, dann müssen wir anders vorgehen. Also Shub Summary Plot zum Beispiel solltet ihr interpretieren können, wenn ihr einen vorgelegt bekommt und dann irgendwie...

mir was sagen können, nicht nur, welches sind die wichtigsten Attribute, sondern auch, in welche Richtung wirken die zum Beispiel. Also, habt ihr ja gemacht, auch für euer Assignment. Dann hatten wir so eine Tabelle mit Sternchenwertungen für verschiedene Kriterien und verschiedene Algorithmenfamilien, dass ihr so ein bisschen, ja, könnt ihr euch vielleicht auf euer Chi-Chi so ein paar Highlights raußschreiben, ähm,

Natürlich die Interpretierbarkeit, welche Algorithmen sind da gut, aber auch andere Sachen, welche können gut mit fehlenden oder irrelevanten Attributen umgehen und solche Sachen. Dann haben wir gesehen, okay, Komplexität spielt eine Rolle, also so Richtung Over- oder Underfitting. Ich muss irgendwie das richtige Maß an Komplexität finden für mein Modell, sodass es die Komplexität der Daten gut abbildet. Und wir haben da...

Da gab es auch so ein paar Folien, da könnt ihr nochmal schauen, wie steuere ich die Komplexität bei verschiedenen Algorithmen. Also verschiedene Algorithmen haben von sich aus schon oft unterschiedliche Komplexität. XGBoost hat mehr Komplexität als ein einzelner Baum, logischerweise. Und dann gibt es noch bestimmte Parameter, an denen ich schrauben kann, um die Komplexität noch

weiter zu steuern. Kostenmatrizen, mein Favorite für die Evaluation. Also jetzt sind wir schon beim Thema Evaluation.

Vielleicht gebe ich euch, so wie wir das auch hier im Unterricht in der Übung gemacht hatten, eine Beschreibung von der Situation und dann überlegt ihr euch, wie könnt ihr die Kostenmatrix ableiten. Also wir hatten das für zum Beispiel Loan Approvals, also Kreditvergabe gemacht. Wichtig dabei immer, denkt immer daran, überlegt euch, was mache ich, wenn das Modell Ja sagt. Okay, dann gebe ich dem Kunden das Geld. Und dann überlegt euch,

Wenn ich das gemacht habe und dann war das richtig oder falsch, was resultieren dann für Kosten? Und dann schreibt ihr die Zahlen rein und ist fertig. Genau. Wenn ihr, also ihr solltet wissen, was eine Confusion Matrix ist und die interpretieren können und auch aus den Zahlen, die da drin stehen, zum Beispiel die Accuracy berechnen oder wenn ihr eine Kostenmatrix gegeben habt, die dann anwenden, um mithilfe der Confusion Matrix die Kosten von einem Ergebnis, Evolutionsergebnis zu berechnen.

Ja, ihr solltet natürlich auch wissen, Accuracy hat seine Schwächen. Das wisst ihr noch, insbesondere wenn wir starke Unbalanciertheit haben der Daten, also Nadel im Heuhaufen sozusagen. Modelle, die einfach immer Nein sagen, haben dann eine hohe Accuracy und das bedeutet aber nichts. Bedeutet nicht, dass sie gut sind. Genau, wir hatten auch gelernt, wie man Over- und Underfitting oder vor allem Overfitting diagnostizieren kann. Also denkt dran,

Wenn ein Modell auf den Trainingsdaten sehr, sehr viel bessere Ergebnisse hat als auf den Testdaten, sind Zeichen für Overfitting. Und ja, wenn wir imbalancierte Daten haben, unbalancierte Daten haben, hatten wir als einen Lösungsansatz Over- und Undersampling kennengelernt. Also ihr solltet wissen, warum mache ich das und auch, was ist der Effekt? Also meistens geht ja die Accuracy runter, scheint erstmal schlecht, aber man findet eben meistens mehr Nadeln und das hat dann wirtschaftlichen Nutzen.

Genau, jetzt kommt der Teil von Manuel, da versuche ich es mal. Ihr habt über Dimensionsreduktion diskutiert. Also ihr solltet diesen Curse of Dimensionality kennen und Gegenmaßnahmen erläutern, also die Dimensionsreduktionsmaßnahmen, die ihr diskutiert habt. Und dann ist vor allem wichtig, auch nochmal euch Clustering anzuschauen.

Ich glaube ihr habt K-Means und DB-Scan, ich weiß nicht, ob ihr noch mehr angeschaut habt. Hier rüsch ich das Clustering auch noch. Genau, und da solltet ihr verstehen, wie die ungefähr funktionieren, das erklären können und auch eventuell Vor- und Nachteile für bestimmte Use Cases darlegen können oder erkennen können. Und dann hat der Manuel hier noch hinzugefügt.

kritische Distanz berechnen können, Pythagoras erklären mit Smiling. Ich denke, ihr wisst, was daran lustig ist. Okay. Es scheint so. Ja, habt ihr Fragen? Jetzt könnt ihr Fragen zu dem oder zu der Klausur vom letzten Jahr, wenn ihr euch die angeschaut habt, was auch immer euch noch auf der Seele brennt.

Ja, Marc. Oh je, die war ja nicht von mir, die Aufgabe. Im Zweifelsfall kommt mir in die Lösung. Den Logramme kann ich eigentlich auch lesen, weißt du nicht? Wo ist es? Okay, und was ist deine Frage konkret? Ja,

Also ich meine, erst mal ist es, glaube ich, wichtig, sich dann zu erinnern, wie funktioniert hierarchisches Clustering. Also es verschmilzt ja immer die aktuell zwei ähnlichsten Cluster miteinander. Und das ist eigentlich das, was das Dendogramm abbildet. Also du weißt jetzt hier nicht

genau, wo es angefangen hat, aber du weißt, in irgendeinem Schritt wurden erst DEMI und NASH zusammengenommen und zu diesem Zweiercluster kam dann Catherine dazu. Und das bedeutet...

Das ist zum Beispiel der erste Punkt, der beschreibt ja das, was ich gerade gesagt habe, den würdest du dann als richtig ankreuzen. Was hier, glaube ich, wichtig ist zu wissen, dass beim Dendrogramm sich nie die Linien kreuzen, also man ordnet die immer so an, die Datenpunkte, dass sie sich nicht kreuzen. Der dritte Punkt klingt irgendwie falsch. Also ich hätte jetzt gedacht, dass das Distanzmaß sicher einen Einfluss hat und schneiden an verschiedenen Punkten,

Also hier siehst du ja, wie geschnitten wird sozusagen. Also diese gestrichelte Linie, darüber habt ihr auch gesprochen, oder? Wenn du hier schneidest, dann hast du genau die Cluster C1, C2, C3, C4, C5, C6. Du könntest aber auch hier schneiden, dann hättest du nur noch zwei Cluster. Also sozusagen der Algorithmus hört nie auf, der legt immer die zwei nächsten Cluster zusammen, bis es nur noch einen Cluster gibt am Ende. Und du musst einfach sagen, wo es aufhören soll, nach welchem Schritt. Und dadurch hast du dann deine Cluster definiert.

Jetzt weiß ich nicht, ob ich deine Frage beantwortet habe. Wir gucken mal in die Lösung, ob das so stimmt, was ich gesagt habe. Also das Letzte würde ich jetzt richtig ankreuzen. Ja, warte mal, die Lösung ist ganz unten. Wo ist es? Ja, ich glaube, nein. Ah, doch. Ich habe es jetzt gerade genau andersrum interpretiert. Ganz komisch, aber die Kreuzchen zählen. Dann war es doch so, wie ich es gesagt habe. Ist es klar, Marc?

Ja, was du von Hand so zeichnen kannst. Nein, also der Sinn dieser Maßnahme, dass es handgeschrieben ist, ist natürlich zweierlei. Also einerseits könntest du durch sehr klein kopieren natürlich

wenn du jetzt das druckst, sehr viel mehr drauf bekommen auf dein Cheat-Sheet. Das ist jetzt aber nicht so das Primäre. Und andererseits geht es aber auch darum, dass du nochmal was lernst, wenn du es selber schreibst. Man lernt mehr, wenn man selber... Also auch wenn du es zeichnest, du lernst mehr. Ja, habe ich gesagt. Ich habe doch eine E-Mail dazu geschrieben, oder? Ja.

Ich habe eine E-Mail geschrieben dazu, oder? Okay. Ich glaube, so hatten wir uns geeinigt. Also dann schreibe ich da vielleicht doch nochmal eine E-Mail. Nicht, dass dazu Missverständnisse kommen.

Ja, also es wird auch wieder 60 Punkte haben und ihr habt 60 Minuten. Also ihr könnt grob immer eine Minute pro Punkt rechnen und wenn ihr nach einer halben Stunde erst 10 Punkte abgearbeitet habt, dann mal überlegen, wie es weiter vorgeht. Also sicher nicht in Panik geraten, aber...

Ich meine, man kann ja auch immer erstmal lesen und gucken, wo kann ich es gut, wo gibt es auch viele Punkte und dann priorisieren. Ja, wahrscheinlich. Ich glaube, vielleicht zwei weniger. Also es kann sein, weil letztes Jahr war im hinteren Teil viel Multiple Choice. Also das ist jetzt, war damals der Andreas, jetzt ist der Manuel und ich glaube, Manuel wird vielleicht nicht so viel Multiple Choice machen, dann sind es zwei weniger, aber der Umfang ist ähnlich.

Ja, genau. Also es kann sich jetzt leicht verschieben, dass es vielleicht ein bisschen weniger Multiple-Choice ist, oder ein bisschen mehr. Ich kann es nicht so genau sagen. Nein. Es kommt dann auf einmal. Also ihr kriegt auch ein detailliertes Feedback, aber erst dann nach der Prüfung. Ja, ja.

Und die ganze Theorie, das gehen wir eins auf die Folie. Damit meinst du jetzt, alles, was nicht auf den Folien steht, kann nicht gefragt werden? Genau. Darauf geht es hinaus. Das ist jetzt natürlich schwierig, da würde ich mich jetzt nicht drauf festlegen. Weil was für andere Länder... Natürlich. Ist

das so? Ja. Aber ich meine, natürlich haben wir auch ab und zu Sachen gesagt hier oder an die Tafel geschrieben, die...

die eventuell nicht so eins zu eins, Folien sind ja immer nicht ausformulierter Text, das heißt, da wird was drumherum gesprochen und das hat auch Relevanz. Jetzt nur den Text auf den Folien zu lesen, ist wahrscheinlich gut, aber nicht zu absolut 100 Prozent. Dafür gibt es auch noch die Aufzeichnung von den Vorlesungen. Marc, wolltest du noch was fragen? Habt ihr noch was?

Habt ihr es schon angeschaut, die Prüfung vom letzten Mal? Tja, also wenn noch Fragen kommen, ich überlege gerade, wir haben ja kein Teams, oder, für den Kurs. Wir sind ein bisschen altmodisch hier, aber wir könnten natürlich hier noch ein Forum machen. Also was ich immer gut finde, wenn Fragen sind, von denen ihr denkt, dass sie alle betreffen, dass man sie vielleicht irgendwo postet, wo dann auch alle die Antwort sehen.

Ich mache hier noch so ein Forum hin, dann kommt das noch in die E-Mail. Ich meine, wenn ihr denkt, die Frage, die ist mir jetzt aber peinlich, da frage ich lieber per E-Mail, dann ist das natürlich auch okay. Aber wenn ihr denkt, das wäre für alle interessant, dann das mache ich noch gleich. Dieses hier ist, glaube ich, das Einzige, oder? Ja, ich denke, das solltet ihr von vorne bis hinten gelesen haben. Ja.

Gut, dass du fragst. Nein, das ist optional. Sonst eben, ich hatte noch Blick geworfen auf euer Feedback. Vielen Dank dafür noch an die, die mitgemacht haben. Ich hatte mir hier noch zwei, drei Sachen rausgeschrieben. Also manches war in verbalem Feedback. Die drei unteren Punkte habe ich aus der Linie. Die kann ich auch gleich noch zeigen.

Aber vielleicht fangen wir mit dem an. Das sind sozusagen Einzelmeinungen und dann finde ich es immer ganz interessant, wenn man, oder manchmal ist es interessant, in die Runde zu fragen. Also ein Feedback war mehr Mathematik. Es war wenig Mathematik. Also auch was die Algorithmen angeht, bei Dimensionsreduktion wurde es, glaube ich, in dem Kommentar erwähnt, sollte es mehr sein, schon gut. Also vielleicht suchen wir ein Buch raus.

Wo man das nachlesen kann, oder? Ich mache es mal so, für die, die wollen. So, das scheint mir jetzt, sollen wir abstimmen? Ich habe so das Gefühl, also wer würde gerne mehr Mathematik haben? Ja, kommt, Rauter, es ist ja nichts Verwerfliches, Mathematik ist was Schönes. Also, wo jetzt nochmal richtig die Hände hoch.

Eins, zwei, drei, vier. Siehst du, es lohnt sich doch. Und wer ist dagegen? Wer möchte nicht mehr Mathematik? Diese Enthaltung? Na ja, mal schauen. So, ja. So, oder weiteres. Ich meine, wo hapert es noch an Mathematik? Auch bei den Algorithmen jetzt?

Lineare Regression zum Beispiel oder sowas. Ja, also jetzt jenseits von Mathematik. Ich meine damit, ich bin gerade überfordert von meinem, also die Algorithmen näher kennenlernen ist ja nicht unbedingt zwingend Mathematik. Ja. Ja.

Jaja, bei manchen schon. Okay, gut. Schauen wir mal, was wir da machen können. Direkteres Feedback in den Quizzes kam, glaube ich, auch mehrfach. Also kann ich gut verstehen. Okay, können wir daran arbeiten. Genau. Und die habe ich jetzt mal rausgegriffen. Also hier gibt es ja mal diese Linie. So sieht die da aus. Immer wenn es rechts von diesem hier ist, bin ich eigentlich zufrieden. Und dann gab es eben drei Punkte, wo es links davon war. Roter Faden. Ja.

Ja, also vielleicht sucht ihr euch mal aus, wozu ihr was sagen wollt. Wobei, was mich besonders interessiert, insgesamt, ob ihr noch da Feedback habt, weil dazu habt ihr gar nichts geschrieben zum

Assignment und das war ja jetzt auch ein wesentlicher Teil des Selbststudiums, sage ich mal, zusammen mit den Hausaufgaben, den Quizzes. Was jetzt unter dem Durchschnitt der anderen Bewertungen lag, gab es da irgendwas, was euch gefehlt hat oder?

Ich habe jetzt im verbalen Feedback dazu nichts gefunden, aber die Bewertung war nicht so. Wie bei den anderen Sachen. Nichts direkt. Unterrichtsmethoden vor Ort. Was ihr lieber gemacht hättest. Weniger Folien, mehr Folien. Warum grinst du so? Also ich meine, es gibt ja auch sogenannte Flipped Classroom. Ich weiß nicht, ob euch das was sagt.

wo man dann umfangreiche Vorbereitungsaufträge bekommt. Mache ich auch gerne im Master zum Beispiel, wo dann die Sachen, die hier jetzt auf den Folien dargeboten wurden, vorher zum Lesen sind. Quizzes macht man dann auch vorher. Und dann kann man wirklich die Zeit hier nutzen für Übungen und Interaktion sozusagen. Funktioniert dann nicht so gut, wenn nicht alle sich vorbereiten. Was denkt ihr? Habt ihr sowas schon mal erlebt? Das...

Das, was ich jetzt hier mit den Folien gemacht habe, Manuel auch, dass das eher Vorbereitung zu Hause war. Habt ihr auch nicht so? Dann ist es auch schwer zu beurteilen. Wollt ihr noch zu irgendeinem der Punkte was sagen? Ich kann euch ja nichts mehr. Okay. Ich glaube, wenn ihr nichts mehr habt, dann habe ich auch nichts mehr. Dann wünsche ich euch ganz viel Erfolg bei der Klausur. Ich glaube, viele von euch werde ich nächstes Semester im Wissensmanagementmodul wiedersehen.

Also wir werden uns begegnen. Dann, wie gesagt, viel Erfolg.