Vorlesung 1

Summary
Der Vortrag beginnt mit einer Einführung in maschinelles Lernen, wobei anhand eines Beispiels mit Äpfeln und Birnen die Konzepte von Trainingsdaten, Features (Merkmalen) und Labels erläutert werden. Es wird erklärt, dass maschinelles Lernen Muster erkennt und daraus Modelle erstellt, die Vorhersagen treffen können. Überwachtes Lernen, insbesondere Klassifikation und Regression, wird als Hauptfokus genannt, daneben unüberwachtes Lernen (Clustering) und Reinforcement Learning. Anwendungsbeispiele aus dem Business-Bereich wie Kreditvergabe, Betrugserkennung, Targeted Marketing und Kundenabwanderung werden vorgestellt, um die praktische Relevanz zu verdeutlichen. Das Tool Orange wird eingeführt, mit dem Daten geladen, visualisiert und Modelle erstellt werden können. Ein Spiel dient dazu, Muster in Daten zu erkennen und zu verstehen, wie Entscheidungsbäume funktionieren. Der CRISP-DM-Zyklus wird als Standardprozess für Data Mining und maschinelles Lernen erläutert, mit den Phasen Business Understanding, Data Understanding, Modeling und Evaluation. Es wird betont, dass die Auswahl und Formalisierung von Instanzen, Attributen und Zielvariablen entscheidend ist. Beispiele aus Marketing, Kreditvergabe, Skigebietsplanung und Krankenversicherung illustrieren die Problemformulierung und Attributauswahl. Abschließend werden praktische Tipps zur Attributwahl gegeben, etwa das Vermeiden von IDs oder konstanten Attributen, und die Bedeutung von Feature Selection durch Algorithmen hervorgehoben. Insgesamt vermittelt der Text ein umfassendes Verständnis für die Grundlagen, Anwendungen und den praktischen Umgang mit maschinellem Lernen im Business-Kontext.

KeyPoints
Was versteht man unter maschinellem Lernen laut dem Vortrag?
Maschinelles Lernen ist ein Prozess, bei dem ein Modell aus Trainingsdaten Muster lernt, um neue, unbekannte Daten zu erkennen oder vorherzusagen, ähnlich wie Menschen aus Erfahrungen lernen.
Was sind Features im Kontext des maschinellen Lernens?
Features sind Merkmale oder Attribute, die zur Beschreibung von Datenobjekten verwendet werden, zum Beispiel Form oder Farbe bei Früchten.
Was ist der Unterschied zwischen Klassifikation und Regression?
Klassifikation sagt eine kategorische Variable vorher, zum Beispiel Apfel oder Birne, während Regression eine numerische Variable vorhersagt, wie zum Beispiel die Nachfrage oder Kosten.Was ist überwachtes Lernen? Überwachtes Lernen ist eine Art des maschinellen Lernens, bei der Trainingsdaten mit Labels (Zielvariablen) vorliegen, anhand derer das Modell Muster lernt. Welche praktischen Anwendungen von maschinellem Lernen wurden genannt? Beispiele sind Kreditvergabe (Klassifikation in kreditwürdig oder nicht), Betrugserkennung, Targeted Marketing, Kundenabwanderung und Spamfilter.
Wie funktioniert das Tool Orange im Kontext des maschinellen Lernens?
Orange arbeitet mit Workflows und Widgets, die man zusammensteckt, um Daten zu laden, zu analysieren und Modelle zu erstellen. Es unterstützt das Finden von Mustern und das Trainieren von Modellen. Was ist der CRISP-DM-Zyklus? CRISP-DM ist ein Standardprozess für Data Mining, der Schritte wie Business Understanding, Data Understanding, Modellierung und Evaluation umfasst. Warum ist die Auswahl der richtigen Attribute wichtig? Weil nur sinnvolle Attribute helfen, Muster zu erkennen. Ungeeignete Attribute wie IDs oder solche, die erst nach der Vorhersage bekannt sind, sollten ausgeschlossen werden. Wie wird ein maschinelles Lernproblem formalisiert?
Man definiert Instanzen (z.B. Kunden), Attribute (Features) und eine Zielvariable (Label oder Zielwert), die das Modell vorhersagen soll. Was ist ein Beispiel für eine Klassifikationsaufgabe aus dem Vortrag? Die Vorhersage, ob ein Kunde einen Kredit zurückzahlt (kreditwürdig oder nicht), oder ob ein Kunde einen Gutschein für einen Wintercheck einlöst (ja/nein).

Vorlesung 2

Summary
In der Lehrveranstaltung wurde der Zyklus der Datenvorbereitung für maschinelles Lernen anhand eines Beispiels der Telekom-Firma Teleflow detailliert erläutert. Dabei wurden Vertragsdaten in Jahresperioden aufgeteilt, um Instanzen für eine Klassifikationsaufgabe (Kündigung ja/nein) zu erzeugen. Es wurde gezeigt, wie man mit Tableau Prep neue Zeilen generiert, berechnete Felder erstellt und Daten aus mehreren Dateien (Tickets, Services) mittels Joins zusammenführt. Die Aggregation von Tickets pro Vertragsjahr und die Pivotierung von Services in Spalten (One-Hot-Encoding) wurden demonstriert. Zudem wurden Herausforderungen wie fehlende Werte, Ausreißer, falsche Datentypen und Normalisierung behandelt. Die Bedeutung der sorgfältigen Datenaufbereitung für die Modellqualität wurde hervorgehoben, ebenso wie praktische Tipps zum Umgang mit fehlenden Daten und zur Feature-Konstruktion. Abschließend wurde der Export der aufbereiteten Daten für die weitere Analyse in Orange gezeigt.

Key Points
1. Was ist der erste Schritt im Zyklus der Datenanalyse, wie im Text beschrieben?
Der erste Schritt ist die Formalisierung des Problems, also das Problem zu definieren und als Klassifikations- oder Regressionsaufgabe zu formulieren.

2. Was versteht man unter dem Klassenattribut in einer Klassifikationsaufgabe?
Das Klassenattribut ist die Zielvariable, die vorhergesagt werden soll, zum Beispiel 'Betrug' ja oder nein oder 'Kündigung' ja oder nein.

3. Wie werden Instanzen in dem Beispiel mit Vertragsdaten definiert?
Instanzen sind einzelne Vertragsjahre eines Kundenvertrags. Aus einer Zeile mit Vertragsdaten werden mehrere Instanzen erzeugt, indem man den Vertrag in einzelne Jahresperioden aufteilt.

4. Welche Funktion wird in Tableau Prep verwendet, um neue Zeilen zwischen zwei Datumswerten zu generieren?
Die Funktion heißt 'Neue Zeilen' und erzeugt Zeilen zwischen zwei Datumsfeldern, zum Beispiel zwischen Vertragsstart und Vertragsende, mit einer definierten Schrittweite (z.B. 12 Monate).

5. Wie wird das Enddatum eines Vertrags in Tableau Prep berechnet, wenn kein Kündigungsdatum vorhanden ist?
Wenn kein Kündigungsdatum vorhanden ist (Wert 0), wird als Enddatum ein fiktives Datum, z.B. der 31.12.2024, gesetzt, um die Analysezeit zu begrenzen.

6. Wie wird das Klassenattribut 'Cancellation' für die einzelnen Vertragsjahre definiert?
Wenn das Period End dem Vertragsende entspricht, wird 'yes' gesetzt, sonst 'no', also ob in diesem Vertragsjahr gekündigt wurde oder nicht.

7. Was ist der Zweck eines Left Join bei der Verknüpfung von Vertragsdaten mit Tickets?
Der Left Join sorgt dafür, dass alle Vertragsjahre erhalten bleiben, auch wenn es keine zugehörigen Tickets gibt, um keine Instanzen zu verlieren.

8. Wie wird in Tableau Prep die Anzahl der Tickets pro Vertragsjahr berechnet?
Man erstellt ein berechnetes Feld, das für jede Zeile 1 setzt, wenn ein Ticketdatum vorhanden ist, sonst 0, und aggregiert dann mit Summe gruppiert nach Vertragsjahr.

9. Was bedeutet Pivotieren in Tableau Prep im Kontext der Services-Daten?
Pivotieren bedeutet, die Zeilen mit verschiedenen Service-Typen in einzelne Spalten umzuwandeln, sodass für jeden Service eine eigene Spalte mit 0 oder 1 steht, ob der Service gebucht wurde.

10. Welche Strategien werden im Text für den Umgang mit fehlenden Werten beschrieben?
Man kann fehlende Werte ignorieren, Zeilen mit fehlenden Werten löschen, Spalten mit vielen fehlenden Werten löschen oder fehlende Werte durch Mittelwert, Median, Modus oder Werte ähnlicher Instanzen ersetzen.

Vorlesung 3

Summary
Der Vortrag behandelt die Datenaufbereitung und Modellierung anhand eines Beispiels mit Telekom-Verträgen, bei dem Vertragsjahre als Instanzen genutzt werden. Es wird gezeigt, wie man Daten mit Tableau Prep aufbereitet, Zeilen splittet, berechnete Felder erstellt, Joins und Pivotierungen anwendet. Anschließend werden Klassifikationsalgorithmen vorgestellt: Einfache Baseline-Modelle, Entscheidungsbäume mit Fehlerberechnung und Pruning, K-Nearest Neighbor mit Distanzmessung und Gewichtung, logistische Regression mit Koeffizienteninterpretation sowie Gradient Boosting als Meta-Algorithmus. Die Vor- und Nachteile der Algorithmen werden diskutiert, ebenso deren Interpretierbarkeit und Parameterhandhabung. Praktische Übungen erfolgen mit Orange, inklusive Modelltraining, Test & Score und Präsentation der Ergebnisse. Eine Übung mit Daten aus dem Gesundheitswesen zielt darauf ab, SLA-Verstöße bei Tickets vorherzusagen. Abschließend werden Präsentationen der Ergebnisse besprochen und weitere Analysen geplant.

Key Points
1. Was war der Ausgangspunkt für die Datenaufbereitung im Teleflow-Case?
Der Ausgangspunkt waren Verträge von Telekom-Kunden sowie Informationen über Tickets und gebuchte zusätzliche Services.

2. Wie wurden die Instanzen für die Analyse definiert?
Als Instanzen wurden jeweils einzelne Vertragsjahre definiert, um zu prüfen, ob der Vertrag am Ende des Jahres gekündigt wurde oder nicht.

3. Was bedeutet 'berechnetes Feld erstellen' in Tableau Prep?
Es bedeutet, eine neue Spalte zu erstellen, der man einen Namen gibt und eine Formel definiert, ähnlich wie in Excel eine Formel in einer Spalte nach unten gezogen wird.

4. Wie wurde mit mehreren Services oder Tickets pro Vertrag umgegangen?
Man konnte aggregieren, indem man zum Beispiel die Anzahl der Tickets pro Jahr zählte, oder man pivotierte die Daten, um für jeden Service eine eigene Spalte mit 1 oder 0 zu erstellen.

5. Warum ist es problematisch, Zeilen mit fehlenden Werten einfach zu löschen?
Weil dabei oft die seltenen, aber wertvollen Fälle verloren gehen, wie zum Beispiel Schadensfälle, bei denen ein Experte involviert war, was die Modellqualität beeinträchtigen kann.

6. Was ist Diskretisierung und wie wurde sie im Beispiel angewandt?
Diskretisierung bedeutet, numerische Variablen in Intervalle zu unterteilen. Im Beispiel wurde das Alter in Intervalle eingeteilt, um es als kategoriale Variable zu verwenden.

7. Warum sollte man Diagnose-Codes nicht als numerische Werte behandeln?
Weil Diagnose-Codes zwar Zahlen sind, aber keine numerische Bedeutung haben, und man nicht mit ihnen rechnen sollte, sondern sie als kategoriale Werte behandeln sollte.

8. Was ist One-Hot-Kodierung und wann wird sie verwendet?
One-Hot-Kodierung wandelt kategoriale Variablen in mehrere binäre Spalten um, wobei pro Instanz nur eine Spalte den Wert 1 hat. Sie wird verwendet, wenn Algorithmen keine kategorialen Attribute direkt verarbeiten können.
9. Warum wurde im Beispiel jedes Vertragsjahr als eigene Instanz betrachtet?
Weil Kunden jährlich kündigen können und so jedes Jahr eine Entscheidung getroffen wird, was für das Lernen des Modells sinnvoller ist als ein gesamter Vertrag über mehrere Jahre.
10. Was macht der Constant-Algorithmus als Baseline?
Er sagt für jede neue Instanz immer die Klasse vorher, die in den Trainingsdaten am häufigsten vorkommt, also zum Beispiel immer 'No'.
11. Wie funktioniert der One-Rule-Algorithmus (OneR) grob?
Er wählt für jedes Attribut Regeln, die auf den häufigsten Klassenwerten basieren, berechnet die Fehler und wählt das Attribut mit dem geringsten Fehler als Modell.
12. Was ist der Unterschied zwischen einem Entscheidungsbaum und dem One-Rule-Algorithmus?
Der Entscheidungsbaum baut mehrere Ebenen von Regeln auf und kann komplexere Muster lernen, während OneR nur ein Attribut zur Klassifikation verwendet.
13. Wie funktioniert der K-Nearest Neighbor Algorithmus?
Er speichert die Trainingsdaten und klassifiziert neue Instanzen anhand der Klassen der k nächsten Nachbarn im Attributraum, basierend auf einem Distanzmaß.
14. Was ist der Unterschied zwischen gewichtetem und ungewichtetem K-Nearest Neighbor?
Beim gewichtetem KNN zählen Stimmen der Nachbarn je nach Nähe unterschiedlich stark, beim ungewichteten KNN haben alle Nachbarn gleiches Gewicht.
15. Welche Nachteile hat der K-Nearest Neighbor Algorithmus?
Er benötigt viel Rechenzeit bei großen Datenmengen, ist empfindlich gegenüber Ausreißern und hat keine interpretierbare Modellstruktur.
16. Was ist eine wichtige Annahme der logistischen Regression?
Dass die Attribute voneinander unabhängig sind.
17. Wie interpretiert man die Koeffizienten in der logistischen Regression?
Jeder Koeffizient gibt an, wie stark ein Attribut den Logit der Wahrscheinlichkeit beeinflusst; positive Koeffizienten erhöhen die Wahrscheinlichkeit, negative senken sie.
18. Was ist der Vorteil der logistischen Regression bezüglich der Ausgabe?
Sie liefert Wahrscheinlichkeiten für die Klassen, die zwischen 0 und 1 liegen und somit interpretierbar sind.
19. Was ist Gradient Boosting und wie funktioniert es grob?
Gradient Boosting ist ein Meta-Algorithmus, der mehrere schwache Modelle (meist Entscheidungsbäume) sequenziell trainiert, wobei jedes neue Modell die Fehler der vorherigen korrigiert.
20. Welche Kriterien helfen bei der Auswahl eines Klassifikationsalgorithmus?
Performance, Geschwindigkeit beim Lernen und Klassifizieren, Robustheit gegenüber fehlenden oder irrelevanten Attributen, Interpretierbarkeit und Handhabung der Modellparameter.

Vorlesung 5

Summary
In der Lehrveranstaltung wurde zunächst an die Grundlagen der linearen Regression erinnert, insbesondere die Bedeutung der Datenvorbereitung wie das Entfernen oder Umwandeln von Datumsattributen und die Interpretation von Koeffizienten. Es wurde gezeigt, dass lineare Regressionen oft nicht alle komplexen Zusammenhänge erfassen, weshalb Methoden wie Gradient Boosting und Entscheidungsbäume eingesetzt werden, die multivariate Muster besser lernen können. Die Bedeutung der Schrittweite beim Gradient Descent wurde erläutert, da sie die Konvergenz beeinflusst. Weiterhin wurden Regularisierungsmethoden wie Ridge und Lasso vorgestellt, die die Modellkomplexität reduzieren und Feature Selection ermöglichen. Zur Modellbewertung wurden verschiedene Verfahren wie Holdout und Kreuzvalidierung erklärt, wobei die Wahl vom Datenumfang abhängt. Die Metriken Accuracy, Precision, Recall und F-Measure wurden vorgestellt, wobei insbesondere bei unbalancierten Klassen Accuracy irreführend sein kann. Die Area Under the Curve (AUC) wurde als robusteres Maß eingeführt. Zudem wurde die Bedeutung von Kostenmatrizen diskutiert, um den finanziellen Impact von Fehlentscheidungen in Anwendungen wie Kreditvergabe und Unfallversicherung zu quantifizieren. Praktische Übungen mit Orange zeigten, wie verschiedene Algorithmen (Tree, Naive Bayes, logistische Regression) im Marketingkontext bewertet werden können. Dabei wurde deutlich, dass Modelle mit höherem Recall oft wirtschaftlich sinnvoller sind, auch wenn die Accuracy geringer ist. Abschließend wurde auf die Bedeutung der Kostenmatrix bei der Modellwahl hingewiesen und eine weitere Übung angekündigt.

Key Points
1. Was wurde im letzten Mal in der Schulung behandelt?
Es wurde über Regressionen gesprochen, insbesondere lineare Regressionen und deren Anwendung auf Datensätze wie Besucherzahlen im Skigebiet.

2. Warum wird das Attribut 'Date' bei der Datenvorbereitung oft aussortiert?
Weil die Daten alle in der Vergangenheit liegen und das Datum als Attribut nicht wiederkehrt. Allerdings kann man aus dem Datum wichtige Informationen wie den Monat extrahieren.

3. Was ist das Problem bei der Verwendung des Monats als numerisches Attribut in der linearen Regression?
Die lineare Regression behandelt den Monat numerisch und nimmt an, dass die Besucherzahlen linear mit dem Monatswert steigen oder fallen, was zu unrealistischen Vorhersagen führt, z.B. dass im Dezember viel weniger Besucher sind als im Januar.

4. Wie wurde das Problem mit dem Monat als numerisches Attribut gelöst?
Der Monat wurde als kategoriales Attribut behandelt, also one-hot-encoded, sodass jeder Monat einen eigenen Koeffizienten bekommt und die lineare Regression sinnvollere Vorhersagen macht.

5. Warum ist der Gradient Boosting Algorithmus oft besser als die lineare Regression bei komplexen Daten?

Weil Gradient Boosting multivariate Muster und nicht-lineare Abhängigkeiten besser lernen kann, während die lineare Regression nur lineare Zusammenhänge modelliert und Interaktionen zwischen Variablen ignoriert.

6. Was bedeutet Mean Absolute Error (MAE) und wie wurde er in der Übung verwendet?
MAE misst den durchschnittlichen absoluten Fehler zwischen vorhergesagten und tatsächlichen Werten. In der Übung wurde MAE genutzt, um die Genauigkeit der Modelle zu bewerten, wobei Gradient Boosting einen niedrigeren MAE als die lineare Regression erreichte.

7. Was ist Regularisierung und warum ist sie wichtig?
Regularisierung ist eine Methode, um die Komplexität eines Modells zu reduzieren und Overfitting zu vermeiden. Sie sorgt dafür, dass die Koeffizienten kleiner werden und das Modell generalisierbarer wird.

8. Was ist der Unterschied zwischen Ridge und Lasso Regression?
Ridge Regression reduziert die Koeffizienten, ohne sie auf Null zu setzen, während Lasso Regression viele Koeffizienten auf Null setzt und somit eine Art Feature Selection durchführt.

9. Was versteht man unter Holdout und Kreuzvalidierung bei der Modellbewertung?
Holdout teilt die Daten in Trainings- und Testmenge auf, um das Modell zu trainieren und zu evaluieren. Kreuzvalidierung teilt die Daten in mehrere Teile und trainiert und testet das Modell mehrfach, um verlässlichere Ergebnisse zu erhalten, besonders bei kleinen Datensätzen.

10. Warum ist Accuracy bei unausgewogenen Klassen oft ein schlechtes Maß?
Weil bei stark unausgewogenen Klassen ein Modell, das immer die Mehrheitsklasse vorhersagt, eine hohe Accuracy erreichen kann, obwohl es für die Minderheitsklasse völlig nutzlos ist.

11. Was sind Precision und Recall und wie unterscheiden sie sich?
Precision ist der Anteil der korrekt vorhergesagten positiven Fälle an allen als positiv vorhergesagten Fällen. Recall ist der Anteil der korrekt gefundenen positiven Fälle an allen tatsächlichen positiven Fällen.

12. Was misst die Area Under the Curve (AUC) und warum ist sie nützlich?
Die AUC misst die Fähigkeit eines Modells, positive von negativen Fällen zu unterscheiden, unabhängig von der Klassengleichverteilung. Sie ist robuster als Accuracy bei unausgewogenen Daten.

13. Wie kann eine Kostenmatrix bei der Modellbewertung helfen?
Eine Kostenmatrix berücksichtigt unterschiedliche Kosten oder Gewinne für richtige und falsche Vorhersagen und ermöglicht so eine wirtschaftlich orientierte Bewertung von Modellen.

14. Was war das Ziel der Übung mit dem Orange-Tool?

Das Ziel war, verschiedene Klassifikationsalgorithmen auf einem Marketing-Datensatz zu vergleichen, die Metriken wie Precision, Recall und Accuracy zu verstehen und das beste Modell auszuwählen.

15. Warum sollte man bei binären Klassifikationsproblemen oft die Metriken für die positive Klasse (Yes) betrachten?

Weil meist die positive Klasse die interessantere ist (z.B. Betrug, Kaufinteresse) und die Metriken für diese Klasse aussagekräftiger für die Modellbewertung sind.

Vorlesung 6

Summary

Der Vortrag beginnt mit einer Wiederholung verschiedener Klassifikationsalgorithmen (Constant, logistische Regression, Naive Bayes, Entscheidungsbaum) im Kontext einer Bankkunden-Übung. Es wird die Bedeutung von Precision und Recall diskutiert, insbesondere im Szenario eines Callcenters, das Kunden anruft. Die Kostenmatrix wird als wichtiges Werkzeug zur Bewertung von Modellen vorgestellt, da sie die wirtschaftlichen Auswirkungen von Fehlentscheidungen berücksichtigt. Die Kommunikation mit Stakeholdern wird erleichtert, wenn man Modelle anhand von Kosten und Gewinn bewertet statt nur mit abstrakten Metriken wie AUC.

Es folgt eine Erklärung zu Overfitting und Underfitting: Overfitting entsteht durch zu komplexe Modelle, die Trainingsdaten zu genau abbilden und auf neuen Daten schlechter performen, während Underfitting durch zu einfache Modelle entsteht, die Muster nicht erfassen. Methoden zur Steuerung der Modellkomplexität werden erläutert, z.B. Regularisierung bei logistischer Regression oder Begrenzung der Baumtiefe.

Das Problem der Klassenungleichheit (Class Imbalance) wird anhand von Beispielen wie Marketing, Fraud Detection und Medizin erläutert. Strategien wie Undersampling (Reduktion der Mehrheitsklasse) und Oversampling (Vervielfachung der Minderheitsklasse) werden vorgestellt, um Modelle sensibler für seltene Ereignisse zu machen. Wichtig ist, dass Rebalancing nur auf Trainingsdaten angewandt wird, nicht auf Testdaten.

Die Bedeutung von Fehleranalyse wird betont: Man kann Modelle trainieren, die Fehler des ersten Modells vorhersagen, um typische Fehlermuster zu erkennen und gezielt zu verbessern. Ein praktisches Beispiel mit Bankdaten zeigt, wie durch Erhöhung der Baumtiefe False Negatives reduziert werden können.

Abschließend wird ein Escape Game vorgestellt, in dem Oversampling praktisch umgesetzt wird. Dabei wird eine Kostenmatrix definiert, die Kosten für falsche positive Anrufe und Gewinne durch erfolgreiche Abschlüsse berücksichtigt. Das Spiel zeigt, dass durch Oversampling die Anzahl der positiven Vorhersagen steigt, die Kosten sinken und der Gewinn steigt, obwohl Accuracy und Precision sinken und Recall steigt.

Organisatorische Hinweise zum Semesterprojekt und Coaching-Slots werden gegeben. Insgesamt vermittelt der Text praxisnahes Wissen zu Modellbewertung, Umgang mit Datenungleichgewicht, Modellkomplexität und Fehleranalyse im maschinellen Lernen.

Key Points

1. Welche Algorithmen wurden im Beispiel mit den Bankkunden ausprobiert?
   Es wurden vier verschiedene Algorithmen ausprobiert: Constant (Baseline), logistische Regression, Naive Bayes und Entscheidungsbaum (Tree) mit einer Tiefe von 5.

2. Was ist der Unterschied zwischen Precision und Recall in diesem Kontext?

Precision gibt an, wie viele der als positiv klassifizierten Fälle tatsächlich positiv sind (Trefferquote unter den Vorhersagen), während Recall angibt, wie viele der tatsächlich positiven Fälle vom Modell gefunden wurden (Erkennungsrate).

3. Warum ist es wichtig, eine Kostenmatrix bei der Modellbewertung zu verwenden?
Eine Kostenmatrix hilft, die wirtschaftlichen Auswirkungen von Fehlklassifikationen zu quantifizieren und ermöglicht eine bessere Entscheidungsfindung, da sie Gewinn und Kosten verschiedener Vorhersagen berücksichtigt. So kann man Modelle auswählen, die den höchsten Gewinn oder die geringsten Kosten verursachen.

4. Was versteht man unter Overfitting und Underfitting?
Overfitting bedeutet, dass ein Modell zu komplex ist und sich zu stark an die Trainingsdaten anpasst, wodurch es auf neuen Daten schlechter generalisiert. Underfitting bedeutet, dass ein Modell zu einfach ist und die zugrundeliegenden Muster der Daten nicht ausreichend erfasst.

5. Wie kann man Overfitting erkennen?
Overfitting erkennt man, wenn das Modell auf den Trainingsdaten sehr gute Ergebnisse erzielt, aber auf den Testdaten deutlich schlechter abschneidet. Ein großer Unterschied zwischen Trainings- und Testfehler deutet auf Overfitting hin.

6. Welche Methoden gibt es, um mit Klassenungleichgewicht (Class Imbalance) umzugehen?
Man kann Undersampling der Mehrheitsklasse (Heuhaufen) oder Oversampling der Minderheitsklasse (Nadeln) verwenden. Beim Oversampling können auch synthetische Beispiele erzeugt werden (z.B. SMOTE). Wichtig ist, dass diese Methoden nur auf den Trainingsdaten angewendet werden.

7. Warum sollte man nur eine Metrik zur Optimierung verwenden und andere als Randbedingungen?
Weil die gleichzeitige Optimierung mehrerer Metriken oft nicht möglich ist und zu Verwirrung führt. Es ist einfacher und klarer, eine Hauptmetrik zu optimieren und andere Anforderungen als Nebenbedingungen zu definieren, um ein praktikables Modell zu erhalten.

8. Was bedeutet es, wenn ein Modell eine hohe Precision, aber niedrigen Recall hat?
Das Modell trifft wenige falsche positive Vorhersagen (hohe Genauigkeit bei den positiven Vorhersagen), findet aber nur einen kleinen Teil der tatsächlich positiven Fälle (niedrige Erkennungsrate). Im Callcenter bedeutet das, dass viele potenzielle Kunden nicht kontaktiert werden.

9. Wie kann man die Komplexität eines Entscheidungsbaums steuern?
Man kann die maximale Tiefe des Baumes begrenzen oder andere Parameter wie die Anzahl der Bäume bei Ensemble-Methoden (z.B. Gradient Boosting) anpassen, um Overfitting zu vermeiden oder Underfitting zu beheben.

10. Was ist der Zweck des Escape Games im Kontext des maschinellen Lernens?

Das Escape Game dient dazu, das Gelernte praktisch anzuwenden, insbesondere das Oversampling bei unbalancierten Daten, das Aufstellen einer Kostenmatrix und die Optimierung eines Modells anhand realer Bankkundendaten in Orange.

Vorlesung 9

Summary
Die Präsentation findet privat mit Manuel, der Gruppe und ohne weiteres Publikum statt, ähnlich einer mündlichen Prüfung mit 25 Minuten pro Gruppe. Die ersten zwei Minuten sind für eine kurze Vorstellung für das MyVC-Management vorgesehen, danach folgen etwa zehn Minuten für eine high-level technische Erklärung der Modelle, Features und Evaluationsstrategien. Die restliche Zeit wird für Fragen genutzt. Die Abgabe umfasst Präsentationsfolien und technische Deliverables wie Tableau Prep Workflows, Code und Input-Dateien, die offen und vorbereitet sein müssen. Es wird kein Bericht, sondern nur die Präsentation erwartet, die in einen offiziellen und einen Backup-Teil mit technischen Details gegliedert ist. Interpretierbarkeit von Modellen wird als wichtiges Thema behandelt, um Vertrauen zu schaffen und unerwünschte Muster oder Fehler zu erkennen. Beispiele aus Medizin und Bildverarbeitung illustrieren die Bedeutung. Intrinsisch interpretierbare Modelle wie lineare Modelle und Entscheidungsbäume werden vorgestellt, ebenso post-hoc Erklärmethoden wie Permutation Feature Importance, Partial Dependence Plots und SHAP, die modellagnostisch sind und in Orange genutzt werden können. Die praktische Anwendung in Orange wird gezeigt, inklusive der Nutzung von Widgets zur globalen und lokalen Interpretation von Modellen. Eine Übung fordert die Studierenden auf, ein Gradient-Boosting-Modell zu trainieren, wichtige Features zu identifizieren und das Modellverhalten zu beschreiben, insbesondere um Target Leakage zu erkennen. Abschließend wird die lokale Interpretation einzelner Instanzen mit Explain Prediction Widgets demonstriert, um nachvollziehbare Erklärungen für Vorhersagen zu erhalten. Die Bedeutung multivariater Muster und deren Erkennung durch Entscheidungsbäume wird hervorgehoben, ebenso die Limitationen der neuen Widgets im Vergleich zu Bäumen. Die Sitzung endet mit einer Zusammenfassung, weiteren Aufgaben und einem Ausblick auf kommende Coachings.

Key Points
1. Wie findet die Abschlusspräsentation statt und wer ist das Publikum?
   Die Abschlusspräsentation findet privat statt, nur Manuel, ich und die jeweilige Gruppe sind anwesend. Es ist vergleichbar mit einer mündlichen Prüfung, bei der jede Gruppe etwa 25 Minuten Zeit hat.

2. Wie ist die Zeit der Präsentation aufgeteilt?
   Die ersten zwei Minuten sind für eine Vorstellung und eine Erklärung für das Management von MyVC vorgesehen, danach zehn Minuten für die eigentliche Präsentation und etwa 13 Minuten für Fragen und Diskussion.

3. Was soll in der Präsentation gezeigt werden?
   Die Präsentation soll in zwei Teile gegliedert sein: Ein Präsentationsteil mit einer High-Level-Übersicht, der die wichtigsten Aspekte und Entscheidungen erklärt, und ein Backup-Teil mit technischen Details, die bei Bedarf gezeigt werden können.

4. Welche Deliverables müssen abgegeben werden?
   Es müssen Dateien abgegeben werden, die die Datenaufbereitung nachvollziehbar machen, wie Tableau Prep Workflows, Code für Datenverarbeitung, Orange-Workflows und die Input-Dateien für Orange. Außerdem die Präsentationsdatei.

5. Soll man auch Features zeigen, die nicht verwendet wurden?
Man soll sich auf die Features konzentrieren, die tatsächlich verwendet wurden und funktionieren. Es ist aber interessant zu wissen, welche Gedanken man sich gemacht hat, auch wenn manche Features nicht genutzt wurden.

6. Warum ist Interpretierbarkeit bei Modellen wichtig?
Interpretierbarkeit schafft Vertrauen, hilft beim Debugging, erkennt Verzerrungen und unerwünschte Muster, und ist besonders wichtig bei kritischen Anwendungen wie Medizin oder Justiz, wo Entscheidungen nachvollziehbar sein müssen.

7. Was sind intrinsisch interpretierbare Modelle?
Das sind Modelle wie lineare Modelle, logistische Regressionen, Entscheidungsbäume und Regelbasierte Modelle, die von sich aus verständlich sind, da man ihre Entscheidungsregeln oder Koeffizienten direkt interpretieren kann.

8. Was sind Post-Hoc-Erklärungen?
Post-Hoc-Erklärungen sind Methoden, die auf bereits trainierte, oft komplexe und nicht interpretierbare Modelle angewendet werden, um deren Verhalten zu erklären, ohne das Modell selbst zu verändern.

9. Was ist Permutation Feature Importance?
Eine globale Methode, die misst, wie wichtig ein Feature ist, indem man die Werte dieses Features permutiert und beobachtet, wie stark sich die Modellleistung verschlechtert.

10. Was sind Partial Dependence Plots (PDP) und Individual Conditional Expectation (ICE) Plots?
PDP zeigen den durchschnittlichen Einfluss eines Features auf die Modellvorhersage, während ICE die Vorhersageverläufe für einzelne Instanzen darstellen, um die Variabilität zu verdeutlichen.

11. Was sind SHAP-Werte und wie funktionieren sie?
SHAP-Werte basieren auf Shapley-Werten aus der Spieltheorie und erklären lokal, wie viel jedes Feature zu einer einzelnen Vorhersage beiträgt, indem alle möglichen Feature-Kombinationen betrachtet werden.

12. Wie kann man lokale Interpretierbarkeit in Orange umsetzen?
Man kann das Widget "Explain Prediction" verwenden, das eine einzelne Instanz erklärt, indem es die SHAP-Werte für die Features anzeigt und visualisiert, welche Features die Vorhersage beeinflussen.

13. Was sind die Herausforderungen bei der Interpretation von komplexen Modellen wie Gradient Boosting?
Solche Modelle erfassen multivariate Muster, die schwer zu visualisieren sind. Globale Erklärungen zeigen oft nur einzelne Feature-Einflüsse, während komplexe Interaktionen verborgen bleiben.

14. Wie kann man Target Leakage erkennen?
Durch Interpretierbarkeit und Analyse der Feature Importance kann man feststellen, ob ein Modell Informationen verwendet, die zum Vorhersagezeitpunkt nicht verfügbar sein sollten, wie z.B. das Attribut "Days Open" im Beispiel.

15. Welche Vorteile bietet die Interpretation von Modellen für Unternehmen?
Sie ermöglicht es, Muster zu erkennen, die zu besseren Entscheidungen führen, z.B. welche Kunden auf Marketing reagieren, und unterstützt die Ableitung von Handlungsempfehlungen.

Vorlesung 10

Summary
In der Unterrichtseinheit wurden zunächst ethische Diskussionen zu KI-Anwendungen in der Kreditvergabe und Bewerberauswahl in Gruppen geführt, wobei jeweils Pro- und Contra-Argumente gesammelt wurden. Danach erfolgte eine praktische Einführung in maschinelles Lernen mit Python-Notebooks über Google Colab, inklusive Datenimport, Modelltraining (lineare Regression und Gradient Boosting), Evaluation mittels Mean Absolute Error und Interpretation der Koeffizienten. Es wurde betont, dass der Code anpassbar ist und Hilfsmittel wie ChatGPT genutzt werden können. Weiterhin wurden zentrale Kursinhalte zusammengefasst: Mustererkennung, Datenaufbereitung, Modellinterpretation (Entscheidungsbäume, Shapley-Werte), Komplexitätssteuerung, Kostenmatrizen, Evaluation (Confusion Matrix, Accuracy), Umgang mit unbalancierten Daten, Clustering-Methoden (K-Means, DB-Scan, hierarchisches Clustering) und Dimensionsreduktion. Abschließend wurden Prüfungsmodalitäten erläutert, darunter Multiple-Choice- und Textaufgaben, sowie Feedback der Studierenden zu Mathematikanteil, Unterrichtsmethoden und Prüfungsformat diskutiert. Ein Forum für Fragen wurde eingerichtet, und es wurde auf die Bedeutung des selbstständigen Lernens und Verstehens hingewiesen.

Key Points
1. Wie wurde die Diskussion zur ethischen Bewertung von KI-Anwendungen organisiert?
Die Teilnehmer wurden in vier Gruppen aufgeteilt, wobei jeweils zwei Gruppen pro Thema gebildet wurden: eine Gruppe, die die Anwendung befürwortet und Argumente dafür sammelt, und eine Gruppe, die dagegen ist und Gegenargumente sammelt. Anschließend fand eine Podiumsdiskussion statt, in der die Gruppen ihre Positionen vertreten und verteidigen sollten.

2. Welche zwei Attribute wurden von der Bank für die Kreditvergabe als Vorhersagekriterien genannt?
Die Bank wollte die Attribute Nationalität und Wohnort verwenden, um vorherzusagen, ob jemand seinen Kredit zurückzahlen wird oder nicht.

3. Was ist das ethische Dilemma bei der Verwendung von KI zur Vorauswahl von Bewerbern?
Das Unternehmen möchte ein Modell verwenden, das auf Basis hochgeladener Dokumente wie Lebensläufen und Zeugnissen eine Vorauswahl von Bewerbern trifft und einige Bewerber aussortiert. Die ethische Frage ist, ob es in Ordnung ist, Bewerbungen per KI vorauszuwählen und damit möglicherweise Bewerber auszuschließen.

4. Wie können die bereitgestellten Python-Notebooks genutzt werden?
Die Notebooks sind auf Moodle verlinkt und können über GitHub in Google Colab geöffnet werden. Dort erhält jeder Nutzer eine private Kopie, die er bearbeiten und ausführen kann. Es ist ein Google-Account notwendig, und es wird empfohlen, einen Browser wie Edge zu verwenden, falls Firefox Probleme macht.

5. Welche Schritte wurden im Python-Notebook zur Datenverarbeitung und Modellierung durchgeführt?

Zunächst wurden die Daten geladen und fehlende Werte gelöscht. Dann wurde ein Train-Test-Split durchgeführt. Anschließend wurde ein lineares Regressionsmodell trainiert und die Koeffizienten ausgegeben. Danach wurde das Modell mit dem Mean Absolute Error evaluiert. Zusätzlich wurde ein Gradient Boosting Modell (XGBoost) trainiert und ebenfalls evaluiert.

6. Was ist der Unterschied zwischen der linearen Regression und dem Gradient Boosting Modell in Bezug auf die Fehlerwerte?

Das Gradient Boosting Modell (XGBoost) erzielt in der Regel einen kleineren Mean Absolute Error als die lineare Regression, was bedeutet, dass es besser funktioniert und genauere Vorhersagen liefert.

7. Welche Tipps wurden für den Umgang mit Fehlern im Google Colab Notebook gegeben?

Wenn Fehler auftreten, bietet Google Colab an, eine eigene KI zur Fehlerbehebung einzuschalten. Es wird empfohlen, die Vorschläge zu verstehen und auszuprobieren, ob der Fehler dadurch behoben wird.

8. Welche Lernziele wurden für die Prüfungsvorbereitung genannt?

Die Lernziele umfassen das Verständnis von Mustern, multivariaten Mustern, Formalisierung von Vorhersageproblemen, Datenaufbereitung, Modellinterpretation, Umgang mit Over- und Underfitting, Kostenmatrizen, Evaluation mit Confusion Matrix, Umgang mit unbalancierten Daten, Dimensionsreduktion, Clustering-Methoden wie K-Means und DB-Scan, sowie das Verständnis von Distanzmaßen.

9. Wie wird die Prüfung voraussichtlich aufgebaut sein?

Die Prüfung wird ähnlich wie im Vorjahr sein, mit Multiple-Choice-Aufgaben und weiteren Aufgaben, bei denen man etwas beschreiben, begründen oder berechnen muss. Es wird ein Cheat-Sheet geben, das von Hand oder auf dem iPad beschrieben werden darf. Die Note setzt sich aus dem Mittel von Klausur und Assignment zusammen.

10. Welche Empfehlungen gab es zur Vorbereitung auf die Prüfung?

Es wird empfohlen, die Folien gründlich zu lesen, die Quizfragen zu wiederholen, die Aufzeichnungen der Vorlesungen anzuschauen und sich mit den behandelten Themen intensiv auseinanderzusetzen. Außerdem wurde ein Forum für Fragen eingerichtet, um den Austausch unter den Studierenden zu fördern.