DAV

DEUTSCHE
AKTUARVEREINIGUNG e.V.

Überarbeitete Version vom 04.09.2022.

Schriftliche Prüfung im Spezialwissen

# Actuarial Data Science Advanced

gemäß Prüfungsordnung 4.1
der Deutschen Aktuarvereinigung e. V.

am 22.10.2021

*Hinweise:*

- Die Gesamtpunktzahl beträgt 180 Punkte. Die Klausur ist bestanden, wenn mindestens 90 Punkte erreicht werden.
- Bitte prüfen Sie die Ihnen vorliegende Prüfungsklausur auf Vollständigkeit. Die Klausur besteht aus 11 Seiten.
- Alle Antworten sind zu begründen und bei Rechenaufgaben muss der Lösungsweg ersichtlich sein.

*Mitglieder der Prüfungskommission:*

Axel Kiermaier, Dr. René Külheim, Dr. Jonas Offtermatt,
Tobias Renner, Dr. Felix Spangenberg

**Aufgabe 1** [6.1 Datenmanagement, 6.3. Cloud Computing, 5.2 Datenschutz, 8.1 Data Mining 2] [42 Punkte]

Ihr Vorstand ist der Meinung, dass Ihr Unternehmen besser für Data Science Anwendungen aufgestellt werden soll. Sie wurden von Ihrem Fachbereich als Projektmitarbeiter auserkoren und sollen nun gemeinsam mit einem interdisziplinären Team Ihr Unternehmen „*Data Science Ready*" machen.

a) [6 Punkte] In einem ersten Meeting sollen die verschiedenen Datenquellen im Unternehmen konsolidiert und zusammengeführt werden. Es geht allerdings wild durcheinander. Erklären Sie in einem ersten Schritt den Teilnehmern den Unterschied zwischen dispositiven und operativen Datenhaltungen. Nennen Sie hierfür jeweils 3 Eigenschaften der Datenhaltungen.

b) [3 Punkte] Die IT möchte unbedingt, dass das schon bestehende DWH verwendet wird. Sie plädieren dagegen für einen neuen Data Lake. Erläutern Sie das Konzept eines Data Lakes.

c) [6 Punkte] Bisher hat sich im Projektteam keiner Gedanken über das Thema Datenschutz gemacht. Nennen Sie Ihren Kollegen die zentrale Rechtsquelle für Datenschutz in der EU und erläutern Sie Ihren Kollegen, warum eine der Aufgaben des Projektes die Erarbeitung einer Datenschutzfolgenabschätzung sein muss.

d) [10 Punkte] Der Vorstand hat in der aktuellen Computerwoche gelesen, dass in der Cloud alles billiger und besser ist. Deswegen soll das Projektteam nun eine Cloud-Strategie erarbeiten. Leider kennen sich Ihre Kollegen mit dem Thema kaum aus. Erläutern Sie Ihren Kollegen tabellarisch die verschiedenen Service-Modelle, welche in der Cloud typischerweise verwendet werden (Cloud-Typen). Erläutern Sie zusätzlich die Bereitstellungsmodelle und ihre Unterschiede (Cloud-Services). Geben Sie eine begründete Empfehlung (max. 4 Sätze) für einen Cloud-Typen und ein Bereitstellungsmodell für versicherungsrelevante Fragestellungen ab.

e) [17 Punkte] Für die Datenbeladung des neuen Data Lakes schlägt die IT drei verschiedene Methoden vor. Um diese bezüglich ihrer Schnelligkeit zu vergleichen wurden mehrere Testbeladungen durchgeführt. Die jeweiligen Ladezeiten finden Sie in der Tabelle unten. Nach Blick auf die Tabelle schwanken Ihre Kollegen zwischen Methode 1 und Methode 2. Sie sehen gleich, dass bei Methode 3 der Mittelwert zwar größer, die Streuung aber geringer ist. Zeichnen Sie eine geeignete Visualisierung der Daten um Ihren Kollegen diese Vorteile von Methode 3 zu veranschaulichen. Folgende Dinge sollte aus dieser ersichtlich sein: Maximale + Minimale Ladezeit der verschiedenen

Methoden, Mittelwert der verschiedenen Methoden, Streuung der Messwerte, Median der Methoden. Achten Sie bitte auch auf die grundlegenden Anforderungen an Diagramme (bspw. Achsenbeschriftung u.ä.).

| | **Methoden** | | |
|---|---|---|---|
| **Testlauf** | 1 | 2 | 3 |
| | Zeit in Millisekunden | | |
| **1** | 5626 | 6175 | 5440 |
| **2** | 1096 | 8700 | 5406 |
| **3** | 6884 | 8682 | 5903 |
| **4** | 8238 | 8519 | 5399 |
| **5** | 796 | 8832 | 6128 |
| **6** | 3465 | 8687 | 5159 |
| **7** | 3493 | 3285 | 5329 |
| **8** | 6438 | 2326 | 5244 |
| **9** | 3515 | 4197 | 5794 |
| **10** | 1993 | 6578 | 5681 |
| **11** | 9246 | 2012 | 5062 |
| **12** | 6244 | 1947 | 5535 |
| **13** | 7304 | 3052 | 5716 |
| **14** | 3872 | 5843 | 5556 |
| **15** | 6550 | 5579 | 5079 |
| **16** | 5626 | 3051 | 5602 |
| **17** | 2561 | 4220 | 5808 |
| **18** | 3278 | 4649 | 5204 |
| **19** | 2456 | 4893 | 5779 |
| **20** | 9794 | 5744 | 5121 |
| Mittelwert | **4923,8** | **5348,6** | **5497,3** |
| Median | 4749 | 5236 | 5487,5 |
| Minimum | 796 | 1947 | 5062 |
| Maximum | 9794 | 8832 | 6128 |

**Lösungsvorschlag:**

a) Als richtige Lösung gelten auch tabellarische Gegenüberstellungen mit Stichworten oder kurze erklärende Sätze. Wichtig ist, dass die fett gedruckten Stichworte genannt werden.

**Operative Datenhaltung**:

Operative Datenhaltungen sind für einen **schnellen und fehlerfreien Zugriff der operativen Systeme optimiert** und zeichnen sich mitunter durch folgende Eigenschaften aus:

- **hohe Zugriffsfrequenz,**
- **dynamische transaktionsbezogene Aktualisierung der Daten,**
- **keine Redundanz der Datenspeicherung und referentielle Integrität, ermöglicht durch die dritte Normalform.**

Die letzte Eigenschaft (keine Redundanzen, dritte Normalform) hat zur Folge, dass die Informationen welche in den Daten stecken auf sehr viele Tabellen verteilt sind. Operative Datenbanken sind daher für Datenauswertungen, wie sie für Reporting Fragestellungen oder Data Analytics Projekte benötigt werden, eher ungeeignet.

**Dispositive Datenhaltung:**

Um analytische und planerische Managementaufgaben zu unterstützen werden *dispositive Informationssysteme* benötigt. Diese sollen die Entscheidungsgrundlage für das Management liefern und die **Datenbasis für internes oder externes Reporting** (z.B. für Solvency II oder IFRS) sowie für moderne Data Science Anwendungen bilden.

Dispositive Informationsverarbeitung **basiert auf Daten, die aus operativen Anwendungssystemen und externen Quellen extrahiert wurden**. Typischerweise erfolgt diese Extraktion mit Hilfe eines *Data Warehouse (DWH).* Mitunter werden Daten **bewusst redundant** gespeichert.

b) Die grundlegende Idee eines Data Lakes ist die Bereitstellung eines zentralen und riesigen Datenspeichers, der mit **möglichst vielen dem Konzern zur Verfügung stehenden Daten** – die potentiell für Analysen relevant sein könnten – versorgt wird. Zudem werden verschiedenste **Analytics Tools angebunden**, so dass der Data Lake eine **konzernweite Analytics Plattform** mit zentraler Datenhaltung darstellt.

Die Daten können **in verschiedensten Formaten** vorliegen:

- strukturiert (also tabellarisch),
- semistrukturiert (z.B. im Json Format),
- unstrukturiert (z.B. Texte, Bilder oder Videos).

Ein wesentlicher Aspekt ist die ***zentrale* Bereitstellung der Daten**. Das heißt, es gibt im Idealfall *einen* Datentopf, auf den alle Konzernmitarbeiter, die Datenanalysen durchführen, zugreifen können. Beim Data Lake **werden die Daten typischerweise in ihrem ursprünglichen Format abgelegt (*Schema on Read*)** (z.B. Text, Json, Bild, Audio). Zu einem späteren Zeitpunkt, wenn die Daten für Analysezwecke ausgelesen werden, müssen sie dann in der Regel in tabellarische Form geparsed werden. *(Pro fett gedruckten Stichworten 0,5 Punkt)*

c) Zentrale Rechtsquelle ist natürlich die EU/DSGVO (1 Punkt).

Gründe für Folgenabschätzung. In **Artikel 35** steht: Hat eine Form der Verarbeitung, insbesondere bei **Verwendung neuer Technologien**, aufgrund der Art, des Umfangs, der Umstände und der Zwecke der Verarbeitung **voraussichtlich ein hohes Risiko** für die **Rechte und Freiheiten natürlicher Personen** zur Folge, so **führt der Verantwortliche vorab eine Abschätzung** der Folgen […] für den Schutz personenbezogener Daten durch. *(Pro fett gedruckten Stichworten 1 weiterer Punkt)*

*d) Cloud-Modelle:*

| **Public Cloud** | **Private Cloud** | **Hybrid Cloud** |
|---|---|---|
| Externer Betreiber, viele Nutzer; Sämtliche Hard- und Software-Ressourcen gehören dem Betreiber und werden von ihm verwaltet<br><br>Zugriff über Internet | Exklusive Nutzung durch ein Unternehmen<br><br>Zugriff im Intranet | Kombination aus Public und Private Cloud<br><br>Daten können zwischen Private und Public verschoben werden |
| Amazon AWS, Microsoft Azure, Google Cloud Platform | | |

(3 Punkte, wenn alle Stichworte genannt. Beispiele sind optional)

*Bereitstellungsmodelle:*

| **IaaS – Infracstructure as a Service** | **PaaS – Platform as a Service** | **Saas – Software as a Service** |
|---|---|---|
| Infrastruktur<br>Server, Netzwerk, Speicher | Infrastruktur<br>Server, Netzwerk, Speicher | Infrastruktur<br>Server, Netzwerk, Speicher |
| | Plattform<br>OS, Entwickler-, Admin-Software | Plattform<br>OS, Entwickler-, Admin-Software |
| | | Anwendungen<br>Mobile, IoT, Office, Data Science ... |

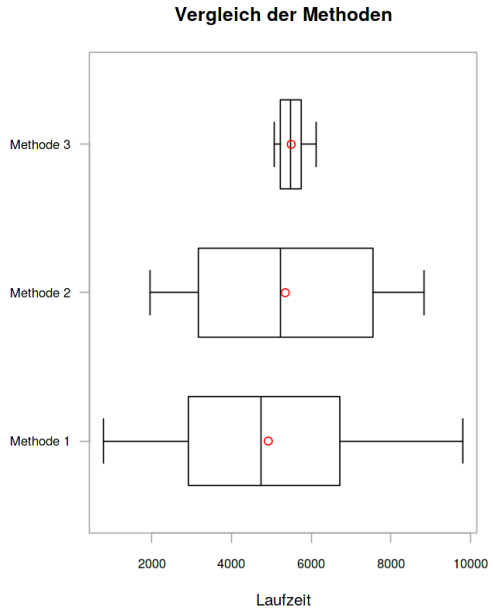
(3 Punkte, wenn alle Stichworte in den Tabellen genannt sind. Beispiele sind optional.)

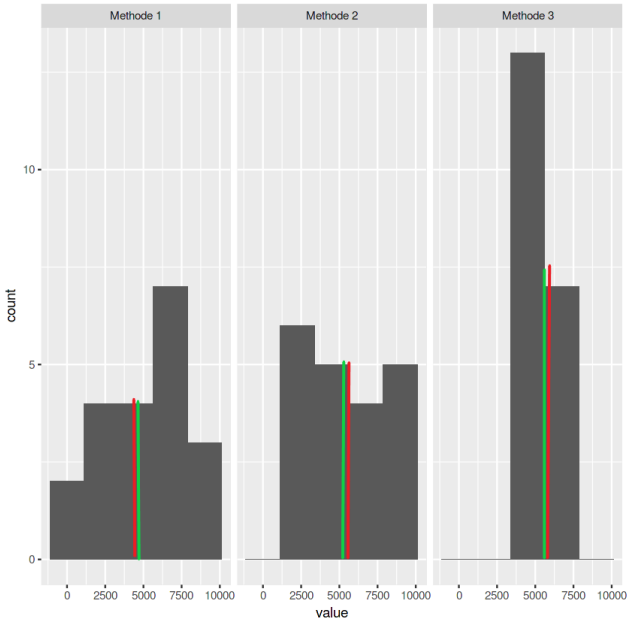
Im Prinzip ist jede Empfehlung richtig. Je nach Anwendungsfall könnte durchaus auch eine public cloud und software as a service genutzt werden. Bei personenbezogenen Daten (bspw. Diagnosen o.ä.) ist eine privat cloud und so wenig Service wie möglich zu empfehlen.

(4 Punkte für eine Empfehlung.)

e) Wichtig für volle Punktzahl: Die genannten Anforderungen (Maximum, Minimum, Mittelwert, Median, Streuung) müssen enthalten sein. Die Basics müssen da sein, also Achsenbeschriftung, Titel, Legende, etc.

Beispiel für eine gültige Visualisierung (nur der Boxplot erfüllt die Basics):



**Aufgabe 2** [6.1 Datenmanagement 2, 6.2 Datenverarbeitungstechnologien 2, 8.2 Analytics 2, 8.3 Innovative Produkte 2] [42 Punkte]

In der Entwicklung eines neuen Telematik-Tarifs in Ihrem Versicherungsunternehmen sollen die Standortdaten und die geografischen Orte berücksichtigt werden, an dem die Versicherungsnehmer gefahren sind. In Abhängig davon, wie hoch die Schadenhäufigkeiten der Orte sind, soll die Prämie der Versicherungsnehmer angepasst werden.

a) [6 Punkte] Die Berücksichtigung der tatsächlich gefahrenen Wegstrecken im Telematik-Tarif ist eine innovative Weiterentwicklung zu den klassischen Kfz-Tarifen. Beschreiben Sie jeweils zwei mögliche Vor- und Nachteile des Telematik-Tarifs für ihr Versicherungsunternehmen im Vergleich zu klassischen Kfz-Tarifen.

Bei der Verarbeitung der Telematikdaten wird der geografische Ort sekundengenau erfasst. Hierdurch entstehen große Datenmengen, die parallel verarbeitet werden sollen.

Ein Kollege von Ihnen schlägt vor, eine Datenbank aufzubauen, in der zu jedem Versicherungsnehmer (VN) und Ort ein Zähler vorhanden ist (relationale Tabelle „VN_ORT" mit den Spalten „VN_ID", „Ort" und „Zaehler"). Initial haben alle Zähler den Wert 0.

Bei jeder Datenerfassung (der Telematikdaten) soll die folgende Funktion aufgerufen werden und hierdurch der Zähler um den Wert eins erhöht werden:

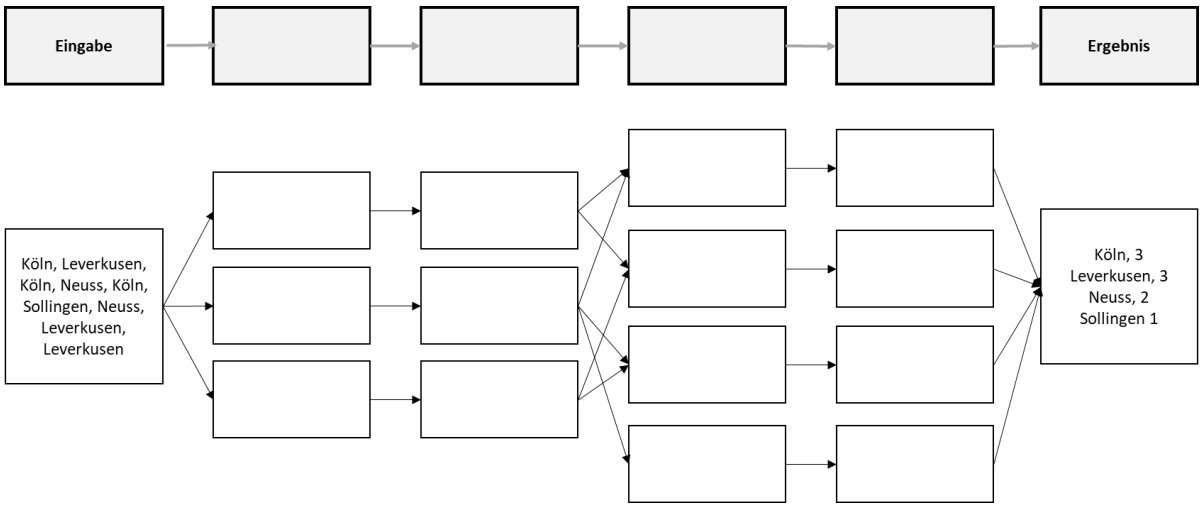
```
def Increase_VN_Ort (VN_ID, Ort):

    # Datenbankabfrage zur Ermittlung der Anzahl der besuchten Ort für einen Versicherungsnehmer (VN)
    # Speicherung der Anzahl der Besuche / des Zaehlers in der Variable temp_Zaehler
    sql_anweisung = """ select Zaehler from VN_ORT where VN_ID = ? and Ort = ? """
    zeiger.execute(sql_anweisung, [VN_ID, Ort])
    temp_Zaehler = zeiger.fetchall()[0][0]

    # Datenbankupdate zur Erhöhung der Anzahl der Besuche / des Zaehlers um den Wert 1
    sql_anweisung = """ update VN_ORT SET Zaehler = ? where VN_ID = ? and Ort = ? """
    zeiger.execute(sql_anweisung, [temp_Zaehler+1, VN_ID, Ort])

    Anzahl_Besuche = temp_Zaehler+1;

    return Anzahl_Besuche
```

b) [6 Punkte] Beschreiben Sie, warum diese Funktion nicht den Anforderungen der funktionalen Programmierung entspricht und für die parallele Datenverarbeitung nicht geeignet ist.

Für die Datenverarbeitung der Telematik soll (anstelle der oben genannten Funktion) das Map/Filter/Reduce-Konzept angewandt werden.

c) [16 Punkte] Vervollständigen Sie die folgende Grafik des Map/Filter/Reduce-Konzeptes. Hierbei sind die Verarbeitungsschritte (obere Teil der Grafik) zu benennen und die Verarbeitung für die neun exemplarischen Datensätzen (unter Teil der Grafik) zu befüllen.

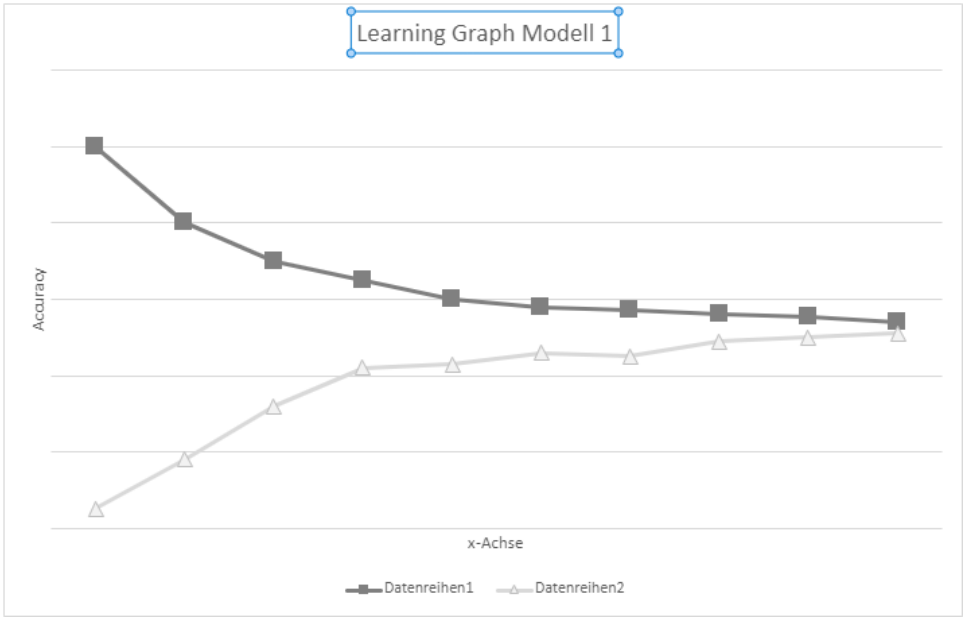
Erläutern und beschreiben Sie hierbei die Verarbeitungsschritte für den Anwendungsfall.

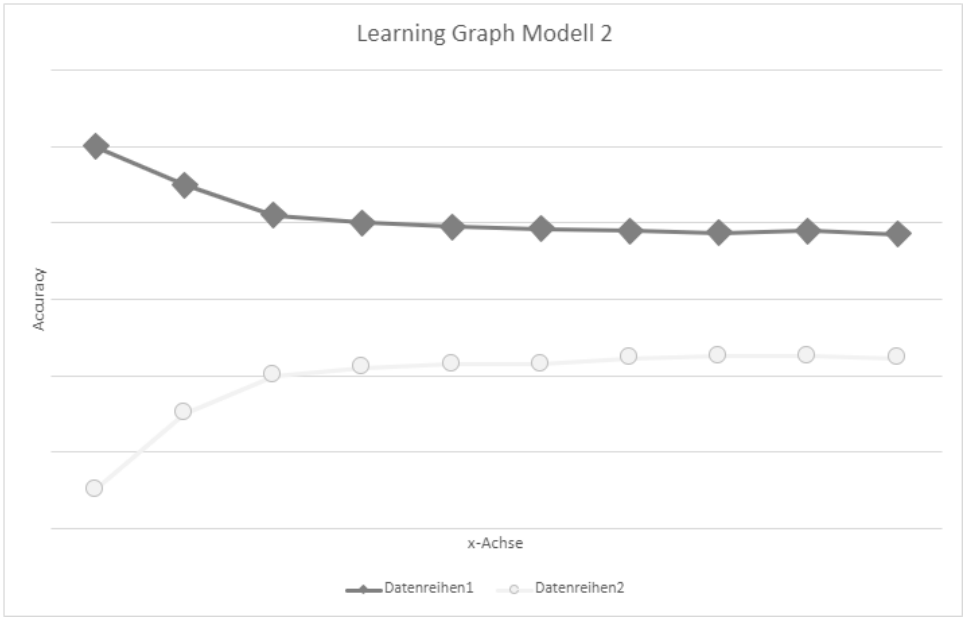



Zur Verwaltung und Analyse der Telematikdaten soll eine Datenhaltung aufgebaut werden. Die Telematikdaten werden hierzu durch einen externen Dienstleister geliefert. Hierbei sind die Inhalte der Datenlieferung und der Schnittstelle klar definiert und festgelegt.

d) [7 Punkte] Für den Aufbau der Datenhaltung steht das „schema on read" und „schema on write" Konzept zur Auswahl. Wählen Sie eines der Konzepte für die Verwaltung der Telematikdaten aus. Begründen Sie hierbei die Auswahl des Konzeptes und benennen Sie zwei Vorteile von dem ausgewählten Konzept.

Nach Bereitstellung und Aufbereitung der Daten soll ein Prognosemodell der Schadenhäufigkeit erstellt werden. Hierbei sollen zwei Modelle, die mit unterschiedlichen Methoden erstellt wurden, verglichen werden. Die Bewertung der Modelle soll über Learning Graphen erfolgen. Im Folgenden sind die Learning Graphen der beiden Modelle dargestellt, wobei die Bezeichnungen in den Diagrammen unvollständig sind:




e) [7 Punkte] Beschreiben Sie (kurz) die Funktionsweise von Learning Graphen und benennen Sie sinnvolle Bezeichnungen von den folgenden Bestandteilen/Elementen der dargestellten Learning Graphen:

- x-Achse
- Datenreihe1
- Datenreihe2

Welches Modell soll als Prognosemodell für die Schadenhäufigkeit verwendet werden? Begründen Sie die Auswahl.

Hinweis: Gehen Sie von der gleichen Skalierung in den beiden Learning Graphen aus.

**Lösungsvorschlag:**

(a) Folgend sind Vor- und Nachteile des Telematik-Tarifs aufgelistet.

Vorteile des Telematik-Tarifs sind:

- Möglichkeit zur besseren und genaueren Risikodifferenzierung
- Bessere Erfassung und Controlling des tatsächlichen Fahrtverhalten der Versicherungsnehmer

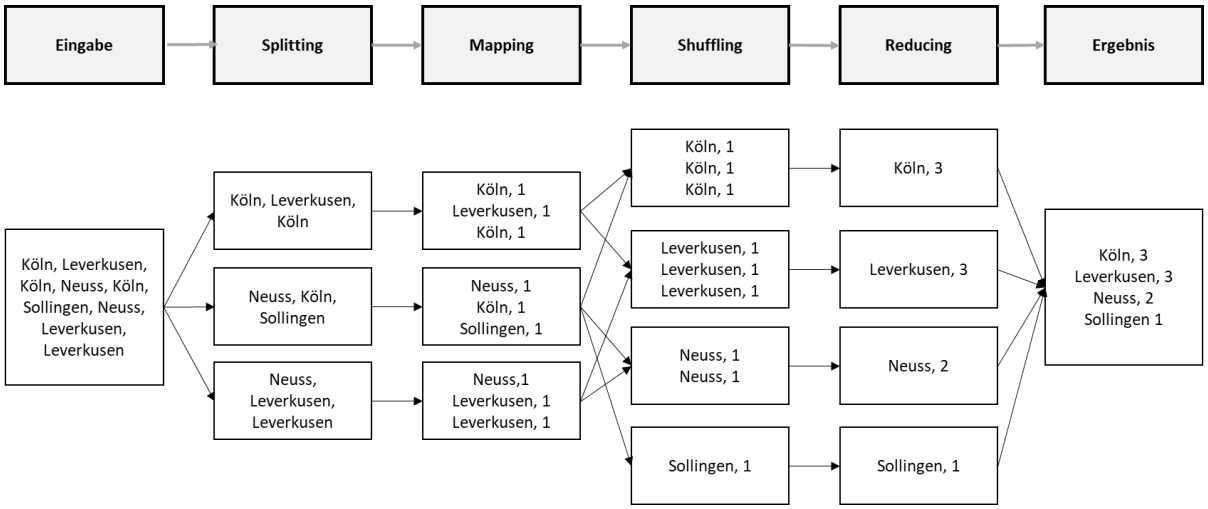
Nachteile des Telematik-Tarifs sind:

- Größere Anforderungen in der Datenverarbeitung und -auswertung
- Besondere Anforderungen zur Einhaltung der Datenschutzrichtlinien
- Gegenfalls geringere Kundenakzeptanz auf Grund von Skepsis gegenüber Telematik-Tarifen

Alternative Vor- und Nachteile sind möglich.

(b) Eine wesentliche Anforderung an eine funktionale Programmierung ist, dass durch den Aufruf von Funktionen keine Daten geändert werden, d.h. keine Seiteneffekte vorliegen. In der dargestellten Funktion werden Daten aus einer Datenbank verändert. Dies ist ein Seiteneffekt und widerspricht hierdurch den Anforderungen einer funktionalen Programmierung. Hierdurch ist eine Parallelisierung der Datenverarbeitung nicht möglich.

(c) In der folgenden Grafik sind die Verarbeitungsschritte des Map/Filter/Reduce-Konzeptes mit den neun exemplarischen Datensätzen dargestellt:




Die Verarbeitungsschritte für den Anwendungsfall sind hierbei:

- Splitting: In dem Splitting-Schritt werden die Datensätze (aus der Eingabe) in verschiedene gleichgroße Blöcke aufgeteilt.

- Mapping: In dem Mapping-Schritt wird jedem Wert / jeder Stadt der Wert 1 zugeordnet.
- Shuffling: In dem Shuffling-Schritt werden die Datensätze zu den gleichen Werten / Städten zusammengefasst.
- Reducing: In dem Reducing-Schritt wird die Summe für jeden Wert / jeder Stadt ermittelt.
- Ergebnis: Im letzten Schritt werden die Ergebnisse aus dem Reducing-Schritt zusammengefasst und hierdurch das Ergebnis ermittelt.

(d) Die Schnittstelle der Telematikdaten ist klar definiert. Daher ist davon auszugehen, dass die Inhalte der Datenlieferung klar festgelegt sind und zukünftige Änderungen nicht unerwartet auftreten. Die Vorteile des „Schema on Read" Konzepts - d.h. Umgang mit unstrukturierten, teilweise unbekannten und sich veränderten Daten - sind für diese Daten daher nicht zwingend gegeben. Da die Schnittstelle klar definiert ist, ist das „Schema on Write" Konzept zu wählen. Die Vorteile des „Schema on Write" Konzepts sind:

- Performancevorteil: Da die Struktur der Daten bekannt ist, ist die Abfrage und Weiterverarbeitung der Daten mit einer guten / besseren Performance möglich.
- Benutzerfreundlichkeit: Der Umgang und die Analyse der Daten sind einfacher möglich, da die Struktur der Daten festgelegt und klar definiert ist.
- Fehlerrobustheit: Die Ablage / Speicherung der Daten erfolgt in klaren Strukturen. Hierdurch ist die Weiterverarbeitung und Analyse weniger fehleranfällig auf Grund von fehlerhalten oder falschen Daten.

Alternative Argrumente sind möglich.

(e) Ein Learning Graph ist eine Darstellung der Modellperformance auf Trainings- und Testdaten in Abhängigkeit der Größe der Testdaten.

Folgend sind die Bestandteile des Learning Graphen aufgelisted:

- x-Achse: Menge der Trainingsdaten
- Datenreihe1: Trainingsdaten
- Datenreihe2: Testdaten

Modell 1 ist zur Prognose der Schadenhäufigkeit zu wählen. Aus dem Learning Graph des Modells 1 ist ersichtlich, dass mit steigender Anzahl von Trainingsdaten die Prognosegüte (Accuracy) von Trainings- und Testdaten annähernd gleich ist und hierdurch das Modell 1 verallgemeinert. In dem Learning Graph von Modell 2 ist dies nicht der Fall und ein Overfitting liegt vermutlich vor.

Weiterhin ist die Prognosegüte (Accuracy) auf den Testdaten bei Modell 1 höher als bei Modell 2.

**Aufgabe 3** [6.1 Datenmanagement 2, 6.2 Datenverarbeitungstechnologien 2, 7.1 Regressions- und Clustermethoden 2] [40 Punkte]

(a) [15 Punkte] In Ihrem Bestandssystem bestehen folgende Verknüpfungen für die Zuordnung von Vertreterdaten zu Kundendaten:

Ein Vertreter besitzt keine oder mehrere Vorschläge, wobei jeder Vorschlag zu genau einem Vertreter gehört. Aus jedem Vorschlag kann ein Vertrag entstehen, jeder Vertrag gehört zu genau einem Vorschlag. Jedem Vertrag ist genau einem Kunden zugeordnet, ein Kunde besitzt mindestens einen Vertrag. Ein Kunde besitzt mindestens eine Kontaktmöglichkeit und jede Kontaktmöglichkeit gehört zu genau einem Kunden.

Für die Vertreter sollen geeignete KPI-Werte ermittelt werden. Diese werden vereinfacht kumuliert erfasst. Dazu gehört die Anzahl der erstellten Vorschläge (AEV), die Anzahl der abgeschlossenen Verträge (AAV) sowie die Anzahl der erfassten Kontaktmöglichkeiten (AEK).

Stellen Sie zunächst in einer geeigneten Weise (Grafik) die beschriebene Situation dar. Überführen Sie dann diese Darstellung in ein geeignetes Datenmodell für ein Data Warehouse, wobei genau zwei Dimensionstabellen verwendet werden sollen. Benennen Sie einen Vorteil und einen Nachteil dieses Vorgehens.

(b) [10 Punkte] Bei Ihren Analysen im Data Warehouse haben Sie den im Bild (s.u.) dargestellten Datensatz vor sich. Für eine Modellierung verwenden Sie zwei verschiedene Methoden. Zunächst eine polynomiale Modellierung:

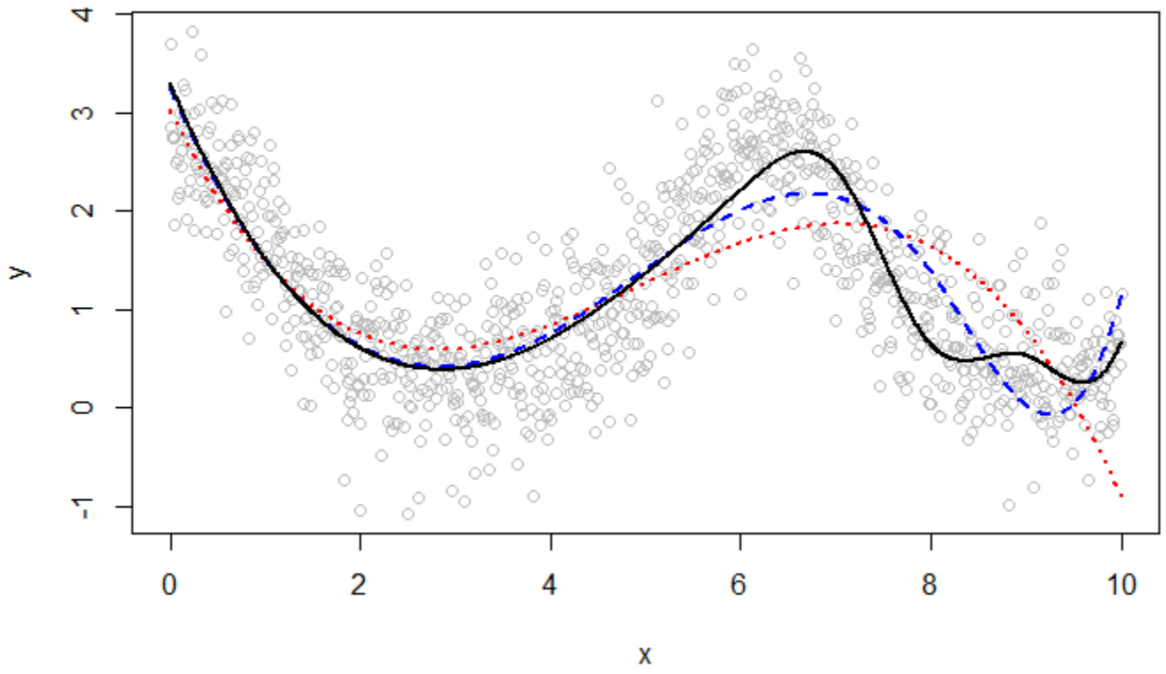
```
fit_poly <- lm(y ~ poly(x, 3), data=df)
pred_poly <- predict(fit_poly, newdata=list(x=x.grid), se=TRUE)
```

Zusätzlich verwenden Sie eine Modellierung mit Splines, wobei Sie verschiedene Knotenanzahlen verwenden (einmal drei und einmal fünf Knoten). Die Modellierung erfolgt nach dem folgenden Schema:

```
fitsp1 <-  lm(y~bs(x, knots=c(...)), data=df)
predsp1 <- predict(fitsp1, newdata=list(x=x.grid), se=TRUE)
```

Die Datenpunkte und die Ergebnisse der Modellierung sind in der

folgenden Abbildung dargestellt:




Leider ist die Zuordnung, welche Kurve zu welchem Modell gehört, nicht mehr vorhanden. Orden Sie den Kurven (schwarz durchgezogen, blau gestrichelt und rot gepunktet) mit einer Begründung einem entsprechenden Modell zu. Welche Splines werden zur Modellierung verwendet? Gehen Sie auch darauf ein, wo Sie (begründet!) die Lage der Knoten vermuten!

(c) [7 Punkte] Sie stellen sich die Frage, wie viele Knoten bei der Modellierung in Teil (b) berücksichtigt werden sollten. Sie haben mehrere Knotenmengen (bezeichnet mit $K_1$, $K_2$, …, $K_n$) zur Verfügung. Wie können Sie sich entscheiden? Ist die alleinige Betrachtung der Residuenquadratsumme eine gute Idee (Begründung)? Wie kann man alternativ vorgehen?

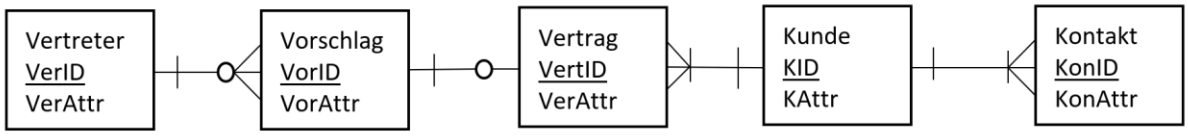
(d) [8 Punkte] Bei einer Modellierung betrachten Sie die folgende Funktion:

$$s(x) = \begin{cases} (x+1) + (x+1)^3, -1 \leq x < 0 \\ 4 + (x-1) + (x-1)^3, 0 < x \leq 1 \end{cases}$$

Welche Eigenschaften eines natürlichen kubischen Splines bzgl. den Knoten $x_0 = a = -1$, $x_1 = 0$ und $x_2 = b = 1$ liegen vor bzw. liegen nicht vor? Fertigen Sie eine Skizze an.

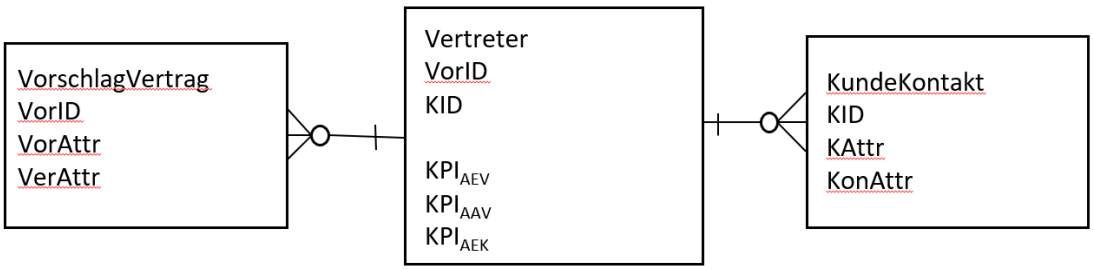
**Lösungsvorschlag:**

(a) Die beschriebene Situation kann z.B. wie folgt dargestellt werden:




Zunächst können geeignete Primärschlüssel angelegt werden. Die Namen der Tabellen sind in der ersten Zeile und die Primärschlüssel in der zweiten Zeile dargestellt. Ein Eintrag für stellvertretende Attribute (z.B. VerAttr) ist in der dritten Zeile dargestellt.

Bei der Überführung in ein Datenmodell für ein Data Warehouse wird bei der Faktentabelle der Vertreter mit seinen Kennzahlen abgebildet. Da nur zwei Dimensionstabellen verwendet werden sollen, werden auf der einen Seite die Informationen zum Vorschlag und Vertrag (Tabelle VorschlagVertrag), und auf der anderen Seite die Informationen zum Kunden und die Kontaktdaten (KundeKontakt) zusammengefasst. Die jeweiligen Attribute sind in diesen neuen Tabellen zusammenzufassen. Das kann wie folgt dargestellt werden:




Ein Vorteil ist dabei, dass ein Zusammenspielen von benötigten Informationen mit einfacheren Select-Queries erfolgen kann. Das Datenmodell wird zudem einfacher zu lesen und intuitiver.

Nachteilig ist dabei, dass redundante Informationen gespeichert werden. Bei der Zusammenfassung der Kommunikationsdaten und Kundendaten werden die Kundendaten (falls mehrere Kommunikationsdaten zum Kunden vorliegen) mehrfach gespeichert. Das kann zu einer Verletzung der Datenintegrität führen.

(b) Bei dem polynomialen Fit handelt es sich um ein Polynom vom Grad drei. In der Graphik ist nur die in rot (gepunktet) eingezeichnete Kurve zu einem Polynom vom Grad drei ähnlich. Damit handelt es sich bei der schwarzen (durchgezogen) und blauen (gestrichelten) Kurve um Splineschätzungen. R verwendet hier standardmäßig kubische Polynome. Beide Kurven haben im Intervall von 0 bis 6 einen ähnlichen Verlauf. Unterschiede gibt es im Intervall von 6 bis 10. Die schwarze Kurve hat hier einen deutlich unruhigeren Verlauf, was darauf hindeutet, dass sich in diesem Bereich mehrere Knoten befinden, die diesen Verlauf verantworten.

Anmerkung: Bei diesen simulierten Daten wurden für die blaue Kurve nur die Knotenpunkte 0, 4 und 8 verwendet. Für die schwarze Kurve hingegen 6, 7, 8, 9 und 10.

(c) Die alleinige Betrachtung von RSS ist keine gute Idee, da man hierbei nur von einem gemessenen Wert ausgeht. Besser wäre folgendes Vorgehen: Einteilung des Datensatzes in Trainings- und Validierungsteil. Training des Modells auf dem Trainingsteil, Anwendung des Modells auf dem Validierungsteil. Dieses Verfahren wiederholen und RSS als Durchschnittswert berechnen. Das kann für die verschiedenen Knotenmengen $K_1$, …, $K_n$ durchgeführt werden. Die Menge K mit dem kleinsten durchschnittlichsten RSS wird schließlich ausgewählt („Kreuzvalidierung").

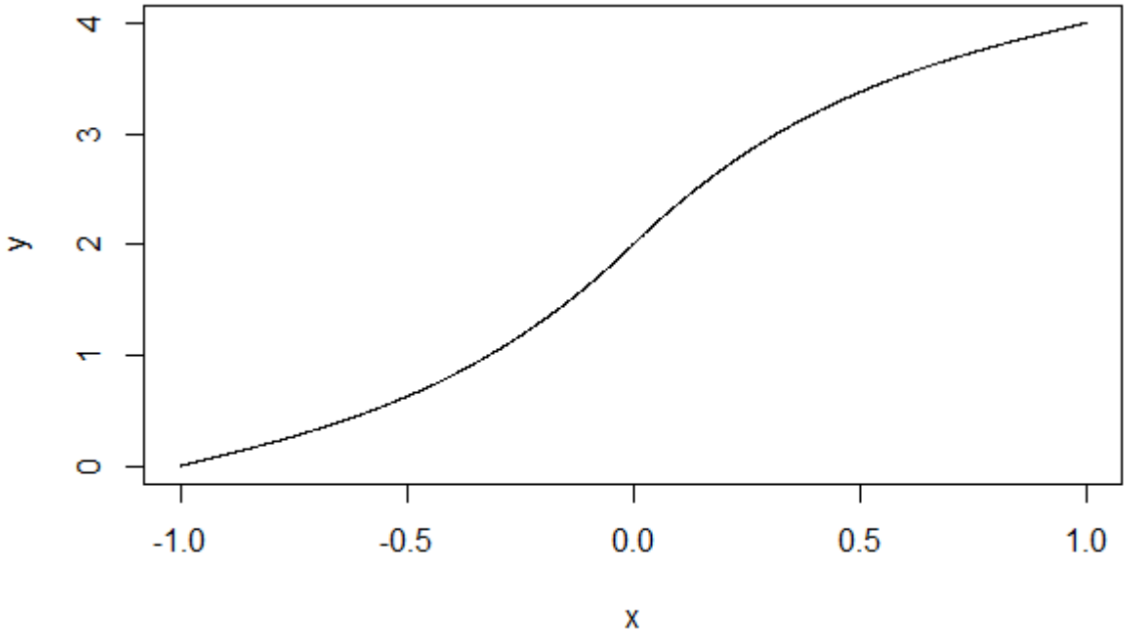
(d) Bezeichne mit $f_1(x) = (x+1) + (x+1)^3$ und $f_2(x) = 4 + (x-1) + (x-1)^3$. Daraus ergibt sich $f_1'(x) = 1 + 3(x+1)^2, f_1''(x) = 6(x+1)$ und $f_2'(x) = 1 + 3(x-1)^2, f_2''(x) = 6(x-1)$.

Damit folgt:

- $f_1(0) = f_2(0) = 2$
- $f_1'(0) = 4 = f_2'(0)$
- $f_1''(0) = 6 \neq f_2''(0) = -6$
- $f_1''(-1) = 0 = f_2''(1)$

Auf den jeweiligen Intervallen liegt eine Funktion vom Grad 3 vor. Jedoch ist eine Glattheitsbedingung verletzt.

Skizze der Funktion:

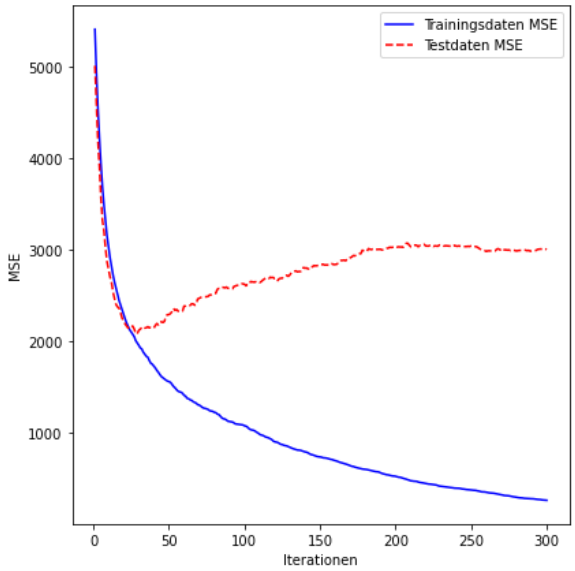


**Aufgabe 4** [7.1 Regressions- und Clustermethoden 2, 8.2 Analytics 2] [24 Punkte]

a) [6 Punkte] Beschreiben Sie das Boosting Verfahren.

b) [6 Punkte] Die folgende Grafik zeigt, wie sich der MSE auf den Trainings- und Testdaten in Abhängigkeit von der Anzahl der Iterationen des benutzten Boosting-Verfahrens entwickelt.

Welches Phänomen sehen Sie? Wie würden Sie aufgrund der Grafik die Anzahl der Bäume bei dem Boosting-Verfahren festlegen?




c) [6 Punkte] Wie unterscheiden sich Bagging und Boosting? Kann es bei Entscheidungsbäumen in Verbindung mit Bagging zu Overfitting kommen? Begründen Sie bitte Ihre Antwort.

d) [6 Punkte] Warum ist es sinnvoll, neben einem Trainings- und einem Testdatensatz noch einen Validierungsdatensatz zu erzeugen?

**Lösungsvorschlag:**

a) Es wird ein additives Modell erstellt. Dazu werden rekursiv Entscheidungsbäume gelernt. Der nächste Entscheidungsbaum wird basierend auf den Residuen des bereits trainierten Modells, d.h. auf der Differenz zwischen Vorhersage und tatsächlichen Werten, trainiert. Die Summe der Entscheidungsbäume wird zur Vorhersage genutzt.

b) Hier ist deutlich ein Overfitting zu sehen. Bis zur ungefähr 25. Iteration sinken die MSEs auf den Trainings- und Testdaten und sind fast gleich. Ab dann sinkt der MSE auf den Trainingsdaten weiter stark, während der MSE auf den Testdaten wieder ansteigt. Aufgrund der Grafik sollte man die Anzahl Bäume auf ungefähr 25 festlegen.

c) Beim Boosting werden die einzelnen Lerner rekursiv ermittelt und diese dann additiv kombiniert. Beim Bagging werden stochastisch unabhängig mehrere Lerner ermittelt und diese dann durch Mittelung (bei Regression) oder Mehrheitsentscheidung (bei Klassifikation) kombiniert. Beim Bagging kommt es in der Regel nicht zum Overfitting, da die einzelnen Lerner in der Theorie alle gleich stark und stochastisch unabhängig sind, sodass die Vorhersagequalität mit der Anzahl der benutzten Bäume steigt.

d) Trainingsdaten werden benutzt, um das Modell zu trainieren. Validierungsdaten werden benutzt, um die richtigen Parameter zu bestimmen. Testdaten werden benutzt, um nach Ende der Modellerstellung das Modell zu evaluieren. Die Testdaten in dieser Aufgabe sind also eigentlich Validierungsdaten.

**Aufgabe 5** [8.1 Data Mining 2, 8.2 Analytics 2] [32 Punkte]

Als Data Scientist sollen Sie in einem Team ein Versicherungsunternehmen zum Thema Stornoanalyse unterstützen.

(a) [16 Punkte] In einer Diskussion mit dem Auftraggeber fallen die folgenden beiden Aussagen: „Es gibt inzwischen Data Mining Tools, mit denen man automatisiertes Machine Learning durchführen kann. Diese Tools könnten wir doch einfach auf unser Data Warehouse ansetzen, um unser Stornoproblem zu lösen." – „Der Data Mining Prozess erfordert kaum menschliche Überwachung."

Setzen Sie sich mit diesen Aussagen auseinander. Wie äußern Sie sich dazu (in ganzen Sätzen)? Nennen Sie für die Auseinandersetzung mit der zweiten Aussage sämtliche Phasen des CRISP-DM-Modells, wählen und benennen Sie aus jeder Phase (mindestens) eine Aufgabe und nutzen Sie diese zur Argumentation für eine menschliche Überwachung des Prozesses.

(b) [9 Punkte] Nach eingehender Datenbereinigung und Voranalysen entwickeln Sie ein Modell zur Klassifikation der Kunden in die Kategorien „storniert" und „storniert nicht".

Ihre Trainingsdatenmenge umfasst 100.000 Kunden. Davon haben 12.000 Kunden gekündigt. Ihr Modell klassifiziert davon 10.800 Kunden als „storniert". Von den Kunden, die nicht gekündigt haben, klassifiziert das Modell 70.400 als „storniert nicht".

Fassen Sie die genannten Daten zu einer Confusion Matrix zusammen. Erläutern Sie dabei kurz, welche Größen allgemein in einer Confusion Matrix eingetragen werden.

Bestimmen Sie die Accuracy, die Sensitivität und die Spezifität auf den Trainingsdaten.

Bei Ihren Testdaten (bestehend aus 20.000 Kunden) haben 17.600 nicht gekündigt, was das Modell in 16.600 Fällen erkennt. 300 Kunden, die gekündigt haben, klassifiziert das Modell als „storniert nicht".

Somit haben Sie auf den Testdaten eine True Positive Rate (TPR) von 87,5 % und eine False Positive Rate (FPR) von 5,7 %. Was vermuten Sie aufgrund der Trainings- und Testergebnisse?

(c) [7 Punkte] Ihr Modell in (b) klassifiziert zwischen „storniert" und „storniert nicht" mittels eines bestimmten Schwellenwerts. Für andere Schwellenwerte erhalten Sie mit Ihren Trainingsdaten folgende neun weitere Ergebnisse:

| Sensitivität | Spezifität |
|---|---|
| 0,92 | 0,30 |
| 0,30 | 0,98 |
| 0,70 | 0,95 |
| 0,95 | 0,10 |
| 0,89 | 0,85 |
| 0,98 | 0,05 |
| 0,80 | 0,90 |
| 0,91 | 0,50 |
| 0,85 | 0,89 |

Integrieren Sie die Ergebnisse aus der vorigen Teilaufgabe in diese Tabelle und skizzieren Sie auf dieser Basis eine ROC-Kurve. Erläutern Sie dabei, wie Sie von den Zahlen zur Grafik kommen.

Was ist typisch oder untypisch an der Grafik?

**Lösungsvorschlag:**

(a) *[2 Punkte]* „Automatisiertes ML": Auch sogenannte AutoML-Tools können Probleme nicht selbstständig lösen, ohne von einem Fachmann mit entsprechendem Domänenwissen angewandt und überwacht zu werden. Vielmehr ist Data Mining ein iterativer und interaktiver Prozess.

*[2 Punkte]* „Prozess ohne Überwachung": Ein Blick auf die einzelnen Phasen des CRISP-DM-Prozesses zeigt, dass in jeder Phase die menschliche Steuerung unerlässlich ist. Generell gilt: Jede von einem System automatisch getroffene Entscheidung kann – wie natürlich auch jede menschliche Fehlentscheidung – zu ungenauen, falschen oder im Extremfall genau entgegengesetzten Ergebnissen führen.

*Anm.: Die im Folgenden ausgewählten Aufgaben sind nur beispielhaft; grundsätzlich können alle Aufgaben des CRISP-DM-Prozesses für die Argumentation herangezogen werden. Für die volle Punktzahl muss aber ein klarer logischer Zusammenhang zwischen der Argumentation und einer Aufgabe erkennbar sein. Die Ausführungen müssen nicht so umfangreich sein, wie hier dargestellt, aber dennoch nachvollziehbar.*

*[2 Punkte]* Phase 1 – Geschäftsverständnis, Aufgabe „Bestimmen von Geschäftszielen" etc.: Hier sind die Geschäftsziele zu definieren, die Ausgangssituation zu analysieren und die konkrete Data Mining Aufgabe zu beschreiben und geeignete Erfolgskriterien festzulegen – alles Tätigkeiten, die nicht an eine Software delegiert werden können, sondern bestenfalls in einzelnen Punkten maschinell unterstützt werden können. In der Aufgabe „Bestimmen von Data-Mining-Zielen" sind Erfolgskriterien für das Data Mining festzulegen. Ob hier eine Treffergenauigkeit von 90 % der Vorhersagen oder eher eine erwartete jährliche Ersparnis von 2 % des Umsatzes das Ziel ist, kann nicht automatisiert entschieden werden.

*[2 Punkte]* Phase 2 – Datenverständnis, Aufgabe „Sammeln ursprünglicher Daten": Hier müssen Entscheidungen getroffen werden, ob die vorliegenden Daten ausreichend sind oder von vornherein zu sehen ist, dass entscheidende Daten fehlen. Die Software kann dies nicht ad-hoc wissen. Selbst eine Software, die bereits für einen ähnlichen Zweck eingesetzt wurde und in dem Zusammenhang „gelernt" hat, kann nicht alle Besonderheiten des jeweiligen Geschäftshintergrunds kennen und berücksichtigen – ähnlich wie eine Checkliste, die ein Berater aus einem anderen Projekt mitgebracht hat, die aber nicht exakt auf die aktuelle Situation passt.

*[2 Punkte]* Phase 3 – Datenvorbereitung, Aufgabe „Bereinigen von Daten": Wenngleich es vielfältige Verfahren zur Imputation fehlender Daten gibt, lässt sich die Auswahl des Verfahrens nicht treffen ohne Domänenwissen und nicht ohne eingehende Beschäftigung mit den vorhandenen Daten und der möglichen Ursache des Fehlens. Die Analyse kann auch zu der Erkenntnis führen, dass man ohne nachträgliche Erfassung der fehlenden Daten nicht weiterkommt.

*[2 Punkte]* Phase 4 – Modellierung, Aufgabe „Bewerten des Modells": Beispielsweise sollte bei einem erstellten Entscheidungsbaum geprüft werden, ob die enthaltenen Regeln dem gesunden Menschenverstand entsprechen bzw. wenn nicht, ob und weshalb sie bei genauerer Betrachtung dennoch valide sind – eine naturgemäß dem Menschen vorbehaltene Aufgabe.

*[2 Punkte]* Phase 5 – Evaluierung, Aufgabe „Überprüfungsprozess": Hier ist der gesamte Data Mining Prozess einer Prüfung zu unterziehen dahingehend, ob unnötige Irrwege eingeschlagen wurden und wo Optimierungspotential vorhanden ist. Vielleicht war auch ein Irrweg hilfreich, um die schließlich als optimal eingestufte Lösung überhaupt zu finden. Diese Analyse kann wieder maschinell unterstützt, aber nicht automatisiert durchgeführt werden.

*[2 Punkte]* Phase 6 – Bereitstellung / Anwendung, Aufgabe „Planen der Bereitstellung": Hier ist zu überlegen, wie die gefundenen Ergebnisse und / oder die entwickelten Modelle in den regulären Arbeitsablauf integriert bzw. transformiert werden können. Mithin ist hier die Kenntnis des gesamten Unternehmens erforderlich, die nicht oder bestenfalls in einer stark abstrahierten Form maschinenlesbar vorhanden ist. Die erforderlichen Entscheidungen will man nicht an die Maschine delegieren.

(b) *[2 Punkte]* Generell hat die Confusion Matrix für eine binäre Klassifikation folgende Gestalt:

| | *Pred. Class* | | |
|---|---|---|---|
| *True Class* | Pos | Neg | |
| Pos | TP | FN | P |
| Neg | FP | TN | N |
| | .P | .N | Total |

Hier werden gegenübergestellt die Anzahlen der True Positives (TP), True Negatives (TN), False Positives (FP) und False Negatives (FN). Zusätzlich

werden häufig die Zeilen- und Spaltensummen eingetragen, die bei der Berechnung verschiedener Kenngrößen ebenfalls einfließen.

*[2 Punkte]* Für die Trainingsdaten ergibt sich folgende Matrix.

| | **Trainingsdaten** | | |
|---|---|---|---|
| | *Pred. Class* | | |
| *True Class* | Pos | Neg | |
| Pos | 10.800 | 1.200 | 12.000 |
| Neg | 17.600 | 70.400 | 88.000 |
| | 28.400 | 71.600 | 100.000 |

*[2 Punkte]* Man erhält für die Trainingsdaten

Accuracy = (TP + TN) / Total = (10.800 + 70.400) / 100.000 = 81,2 %

Sensitivität = TP / P = 10.800 / 12.000 = 90 %

Spezifität = TN / N = 70.400 / 88.000 = 80 %

*[1 Punkt]* Allgemein gilt

TPR = TP / P = Sensitivität

FPR = FP / N = FP / (FP + TN) = 1 – TN / (FP + TN) = 1 – Spezifität

Somit ergibt sich folgende Gegenüberstellung:

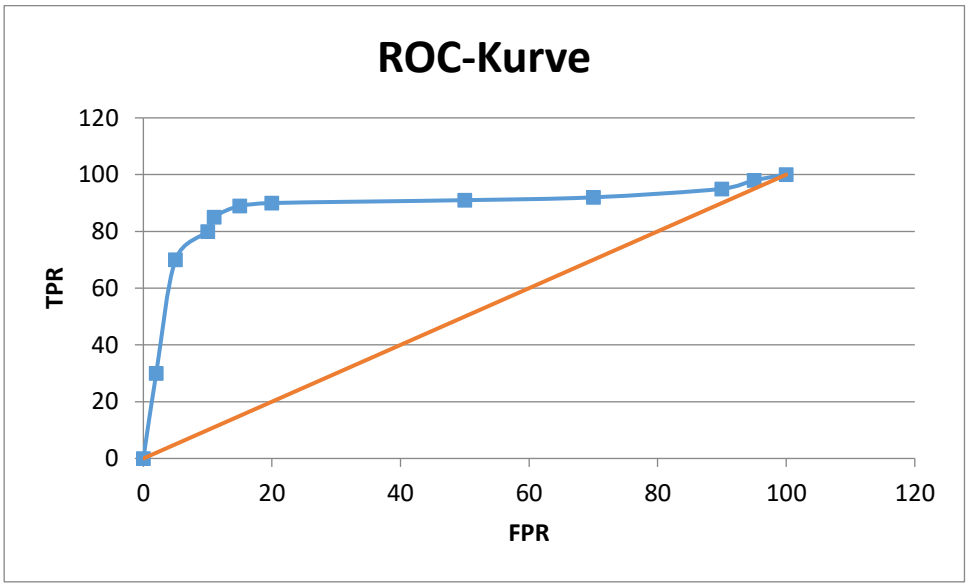
| | TPR | FPR |
|---|---|---|
| Trainingsdaten | 90,0% | 20,0% |
| Testdaten | 87,5% | 5,7% |

*[2 Punkte]* Normalerweise gibt es auf den Testdaten (idealerweise geringfügig) schlechtere Werte, d.h. TPR(Training) > TPR(Test) und FPR(Training) < FPR(Test). Die extrem bessere FPR (und auch die entsprechend deutlich bessere Accuracy von 93,5 %) lässt vermuten, dass die Auswahl der Testdaten nicht zufällig erfolgt ist, sondern einen (aus diesen Zahlen allein nicht zu erklärenden) Bias aufweist. Hier ist eine genauere Untersuchung der Auswahl der Testdaten erforderlich.

(c) *[2 Punkte]* Die ROC-Kurve trägt die True Positive Rate (TPR) auf der y-Achse gegen die False Positive Rate (FPR) auf der x-Achse auf. Es gilt TPR = Sensitivität und FPR = 1 – Spezifität. Zusätzlich gilt stets TPR=0 für FPR=0 und TPR=1 für FPR=1. Somit hat man folgende (sortierte) Daten:

| FPR | TPR |
|---|---|
| 0% | 0% |
| 2% | 30% |
| 5% | 70% |
| 10% | 80% |
| 11% | 85% |
| 15% | 89% |
| 20% | 90% |
| 50% | 91% |
| 70% | 92% |
| 90% | 95% |
| 95% | 98% |
| 100% | 100% |

*[2 Punkte]* Hieraus ergibt sich folgende Grafik:




*[3 Punkte]* Typischerweise bewegt sich die Kurve oberhalb der Diagonalen, da die Diagonale die ROC-Kurve eines Modells darstellt, das zufällig in positiv und negativ klassifiziert. Ideal wäre ein Verlauf, der vom Nullpunkt aus unmittelbar zum Wert 100 % ansteigt. Die vorliegende Kurve liegt dazwischen, ist insofern typisch.

Bei stabilen Modellen ist der Verlauf der Kurve gleichmäßiger als das hier der Fall ist; insofern ist diese Kurve untypisch. Hier bestehen also noch Optimierungsmöglichkeiten.