DAV

DEUTSCHE
AKTUARVEREINIGUNG e.V.

Überarbeitete Version vom 04.09.2022.

Schriftliche Prüfung im Spezialwissen

Actuarial Data Science Basic

gemäß Prüfungsordnung 4
der Deutschen Aktuarvereinigung e. V.

am 22.05.2021

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 [Digitalisierung (1.2), Gesellschaftliches Umfeld & Ethik (1.3), Datenschutz (1.4), Informationsverarbeitung in Versicherungen (2.3), Innovative Produkte (4.3)] [38 Punkte]

Sie sind als Actuarial Data Scientist bei der Pfefferminzia a.G. angestellt und sollen ein Projekt zur Einführung eines neuen datengetriebenen Tarifs leiten. Bei dem Tarif sollen personenbezogene Daten in Echtzeit zur Berechnung der Prämien verwendet werden und alle Geschäftsvorfälle vom Kunden online durchgeführt werden.

a) [10 Punkte] Wie sich schnell herausstellt, ist der Vertrieb von dem Produkt wenig begeistert. Sie erhalten von dort keinerlei Unterstützung. Kurz nach Projektbeginn ergibt sich DIE Gelegenheit, Sie fahren mit dem Vertriebsvorstand gemeinsam Aufzug. Formulieren Sie einen „Elevator Pitch" (max. 10 Sätze) mit dem Sie den Vertriebsvorstand von ihrem Projekt und dem Produkt überzeugen können. Verweisen Sie dabei auf die Wettbewerbsvorteile durch Digitalisierung und die Vorteile durch innovative Produkte.

b) [10 Punkte] Nach dem Sie den Vertriebsvorstand überzeugt haben, erreicht Sie eine E-Mail des Datenschutzbeauftragten des Unternehmens. Er weist daraufhin, dass seit Mai 2015 die EU-Datenschutzgrundverordnung gelte und er davon ausgehe, dass die im Tarif angedachte Verwendung von personenbezogenen Daten nach Artikel 6 EU/DSGVO grundsätzlich verboten wäre. Bitte formulieren Sie eine kurze Antwort-E-Mail (max. 8 Sätze), in welcher Sie drei konkrete Ausnahmen von diesem Verbot nennen und machen Sie einen konkreten Vorschlag, welche davon man in diesem Fall verwenden könnte. Begründen Sie Ihren Vorschlag.

c) [6 Punkte] So schnell gibt der Datenschützer nicht auf. Er verweist nun auf Artikel 20 und verlangt von Ihnen ein konkretes Löschkonzept aller erhobenen Daten. Sie veranstalten darum einen Workshop mit den IT-Mitarbeitern des Projektes, um alle IT-Systeme zu ermitteln, in denen im Rahmen Ihres neuen Tarifs Daten gespeichert werden könnten. Zu Beginn präsentiert Ihnen ein Kollege der IT folgendes Schaubild Ihrer Unternehmens-Anwendungslandschaft:

Ihnen fällt schnell auf, dass dies keineswegs die vollständige IT-Landschaft sein kann. Nennen Sie 3 weitere Anwendungssysteme einer typischen Versicherungssystemlandschaft.

d) [12 Punkte] Einzig die Kollegen in der Marketing-Abteilung sind begeistert von dem neuen Produkt. In der Cafeteria treffen Sie einen befreundeten Kollegen aus der Marketing-Abteilung, welcher euphorisch über das neue Produkt redet und begeistert die ganze Cafeteria über die Vorteile des neuen Verkaufsschlagers unterrichtet. Bremsen Sie Ihren Freund ein und nennen Sie ihm die mit dem Produkt verbundenen Reputationsrisiken, sowie die mit der Verwendung von individualisierten personenbezogenen Daten verbundenen ethischen Probleme. (Max. 10 Sätze)

Lösungsvorschlag:

a) Der „Elevator Pitch" sollte ausformuliert sein. Dabei sollten für eine vollständige Lösung der Aufgabe mind. 5 der folgenden Stichworte genannt werden (pro Stichpunkt 2 Punkte, egal ob Digitalisierung oder Produktentwicklung):

Wettbewerbsvorteile durch Digitalisierung

- Höhere Kundenzufriedenheit
- Niedrigere Verwaltungskosten
- Weniger Aufwand durch Standardisierung
- Datenhandel
- Bewertungsmöglichkeiten für neue Risiken und sich verändernde Risiken

Vorteile innovativer Produktentwicklung:

1. Aufbau von detaillierten Datengrundlagen zur Risikobewertung
2. Genauere Differenzierung und Selektion von Risiken
3. Verbesserte Möglichkeit zur versicherungstechnischen Bewertung und Analyse von Risiken (u.a. Betrugserkennung und Schadenvermeidung)
4. Aufbau von detaillierten Datengrundlagen zur Bewertung der Kundenbeziehung
5. Verbesserte Analyse und Modellierung der Kundenbeziehung
6. Optimierung von Kundengruppenmanagement und Database Marketing

b) Es müssen 3 Ausnahmen von §6 genannt werden (pro Ausnahme 2 Punkte). Die Ausnahmen sind:
   - Zustimmung der betroffenen Person

- Verarbeitung zur Erfüllung eines Vertrages notwendig
- Erfüllung einer rechtlichen Verpflichtung
- Lebenswichtige Interessen der betroffenen Person schützen
- Erfüllung einer Aufgabe, die im öffentlichen Interesse liegt
- Wahrung berechtigter Interessen des Verantwortlichen oder eines Dritten

Ein Vorschlag für eine Ausnahme muss gemacht werden (2 Punkte). Es muss eine Begründung gegeben werden (2 Punkte). Eigentlich macht nur 1. oder 2. Sinn. Die anderen Punkte sind nicht erfüllt.

c) Es müssen mindestens 3 der untenstehenden Anwendungssysteme genannt werden (pro genanntem System 2 Punkte). Synonyme oder ähnlich klingende Vorschläge zählen auch:

Bestandsführung, Inkasso, Exkasso, Partnersystem, Rückversicherungssystem, Dokumentenarchiv, CoC-Komponente, Hochrechnung-/Simulationssystem, Antragsübernahme, Provisionssystem, Meldesystem (GDV, ZfA, RV o.ä.), Zulagenverwaltung, Schadenverwaltung, Widerspruchsverwaltung, Rechte-Management, Web&Mobil, Arbeitssteuerung/Workflow-System, Leistungssystem, bAV, Produktentwicklung, Finanz- u. Rechnungswesen Komponente (SAP), Reporting, Personal, ….

d) Hier ist leider ein kleiner Aufsatz nötig. In den maximal 10 Sätzen sollten sich mind. 5 der folgenden Stichpunkte (2 Punkte pro Stichpunkt) so oder so ähnlich wiederfinden:

- *Risiko, dass Pfefferminzia als Datenkrake wahrgenommen wird*
- *Risiko des Datenverlustes kann zu empfindlichen Strafzahlungen führen*
- *Risiko des Datenverlusts kann das Neugeschäft empfindlich einbremsen*
- *Risiko, dass aufgebauter Markenname nicht mehr mit „Vertrauen", sondern mit „Kunden ausspähen" in Verbindung gebracht wird*
- *Risiko, dass Daten zweckentfremdet werden und missbräuchliche Datennutzung zu Rufschädigung führt*
- *Individualisierte Datenerhebung kann zu sehr kleinen Kollektiven führen, welche den angedachten Ausgleich im Kollektiv nicht mehr möglich machen*
- *Die Grundidee der Versicherung, dass in der Gemeinschaft Sicherheit besteht, kann durch Individualisierung ausgehebelt werden*
- *Durch genaue personenbezogene Datenerhebung können durch Querverbindungen Informationen über den Einzelnen gewonnen werden, die weit über den ursprünglichen Zweck hinausgehen und*

somit weitere Rückschlüsse auf den Kunden zulassen (bspw. Erkennung von Krankheiten, die zu Risikoausschluss führen)

- *Die personenbezogenen Daten können zweckentfremdet verwendet werden, um bspw. Kunden auszuspähen / stalken / erpressen / …*

Aufgabe 2 [Datenmanagement (2.1), Regressions- und Clustermethoden (3.1) Datenaufbereitung zur Modellerstellung (3.4), Modellselektion und Regularisierung (3.5) & Analytics (4.2)] [43 Punkte]

Im Rahmen der Tarifierung für die Sparte Kfz in Ihrer Versicherung soll die Datengrundlage für die Modellierung erzeugt werden. Datenquellen sind die Daten aus zwei disjunkten Bestandssystemen und einem Schadensystem in Ihrem Versicherungsunternehmen. Folgend sind exemplarisch jeweils 10 Datensätze aus den drei Systemen dargestellt.

Tabelle: *Bestandsdaten_1*

VNR	VN_Alter	Geschlecht	Vertrags-beginn	Mit-arbeiter
100001	35	m	01.01.2020	nein
100002	21	m	01.08.2019	nein
100003	29	w	01.08.2000	nein
100004	18	w	01.07.2020	nein
100005	81	m	01.01.1982	nein
100006	22	w	01.07.2019	ja
100007	44	m	01.01.2013	nein
100008	62	m	01.01.1999	ja
100009	29	w	01.01.2013	nein
100010	77	w	01.01.1973	nein

Tabelle: *Bestandsdaten_2*

VNR	VN_Alter	VN_ Geschlect	Berufs-schlüssel	Vertrags-beginn	Mit-arbeiter
920002	55	m	1	01.01.1999	nein
920003	32	m	6	01.01.2018	nein
920004	19	w		01.06.2020	ja
920005	55	w	2	01.01.1980	nein
920006	74	m	2	01.01.1977	
920007		w	1	01.08.2000	nein
920008	46	m	1	01.08.2001	nein
920009	39	m	5	01.01.2010	ja
920010		w	2	01.07.1990	
920011	60	w		01.01.1988	nein

Tabelle: *Schadendaten*

SNR	VNR	Schaden-höhe in Tsd. €	Schaden-Datum
20110002	100002	2,0 €	01.07.2020
20110003	920004	2,0 €	03.07.2020
20110004	100003	3,0 €	05.07.2020
20110005	920006	3,0 €	05.07.2020
20110006	920011	7,5 €	05.07.2020
20110007	100027	2,0 €	08.07.2020
20110008	100568		09.07.2020
20110009	920565	2,2 €	10.07.2020
20110010	920451	5,4 €	10.07.2020
20110011	100557	4,8 €	10.07.2020

a) [7 Punkte] Erstellen Sie zwei SQL-Statements, um die Bestandsdaten mit den Schadendaten anzureichern. Hierbei sollen keine Informationen verloren gehen. Als Ergebnis soll eine Tabelle entstehen, in der alle Bestandsdaten vorhanden sind und (falls vorhanden) die Schadendaten zu dem Bestand angereichert sind.

Hinweis: Erstellen Sie im ersten Schritt eine Zwischentabelle, in der die Daten aus den beiden Bestandssystemen zusammengeführt werden und reichern Sie anschließend die Schadendaten an. Gehen Sie davon aus, dass zu einem Vertrag maximal ein Schaden vorliegt.

b) [7 Punkte] Welche Datenauffälligkeit - aus der Datenbasis aus Aufgabe a.) - gibt es für die Variable „VN_Alter"? Benennen und erläutern Sie kurz vier Möglichkeiten, wie mit dieser Datenauffälligkeit als Vorbereitung zur Modellierung umgegangen werden kann?

Voraussetzung für die weiteren Analysen ist es, dass alle stetigen Variablen normalverteilt sind. Hierzu soll die Normalverteilung bei den stetigen Variablen mit einem Probability Plot (Q-Q Plot) überprüft werden und ggf. eine Transformation durchgeführt werden.

c) [7 Punkte] Beschreiben Sie das Vorgehen zur Erstellung eines Probability Plot (Q-Q Plot) zur Überprüfung der Normalverteilung.

Zur Analyse der Schadenaufwendungen soll aufbauend auf der Datengrundlage ein Prognosemodell erstellt werden. Hierzu soll ein Regression Tree erzeugt werden, in dem der Schadenaufwand vorhergesagt werden soll.

d) [15 Punkte] Erstellen Sie für die nachfolgend dargestellten zehn Datensätze den ersten Split des Regression Tree. Bestimmen Sie hierzu die Zielvariable des Regression Tree und benennen Sie eine geeignete Verlustfunktion für den Aufbau des Regression Tree. Hier soll ein binärer Split verwendet werden, bei dem der Wert der Verlustfunktion minimiert wird. Rechnen Sie mit zwei Nachkommastellen.

VNR	VN_Alter	Geschlecht	Schaden vorhanden	Schaden-aufwand in Tsd. €
100001	35	m	nein	- €
100002	21	m	ja	3,00 €
100003	29	w	ja	3,00 €
100004	18	w	nein	- €
100005	81	m	nein	- €
920002	55	m	nein	- €
920003	32	m	nein	- €
920004	19	w	ja	3,00 €
920005	55	w	nein	- €
920006	74	m	ja	1,00 €

Hinweise:

Die Verlustfunktion hat im Root-Knoten des Entscheidungsbaums den Wert 16 und bei einem Split durch die Variable Geschlecht den Wert 8.

Als Bestandteil der Verlustfunktion kann der „Mittlere Quadratischer Fehler" (MSE) verwendet werden.

Es gilt $(3-2{,}25)^2 = 0{,}56$; $(0-2{,}25)^2 = 5{,}06$; $(1-0{,}17)^2 = 0{,}69$; $(0-0{,}17)^2 = 0{,}03$ (gerundet auf zwei Nachkommastellen)

Bei der Überprüfung der Verallgemeinerung des Modells hat sich ein Overfitting gezeigt. Zur Vermeidung des Overfitting soll eine Regularisierung durchgeführt werden.

e) [7 Punkte] Erläutern Sie, was Overfitting bedeutet und benennen und beschreiben Sie vier Parameter bei der Erstellung von Entscheidungsbäumen, um ein Overfitting zu vermeiden.

Lösungsvorschlag:

(a) Folgende zwei SQL-Statements können verwendet werden, um die Datengrundlage zu erstellen:

SQL 1:

create table Bestandsdaten as

select VNR, VN_Alter, Geschlecht as VN_Geschlecht, Vertragsbeginn, Mitarbeiter, ´´ as Berufsschlüssel from Bestandsdaten_1

UNION

select VNR, VN_Alter, VN_Geschlecht, Vertragsbeginn, Mitarbeiter, Berufsschlüssel from Bestandsdaten_2

SQL 2:

select B., S.**

from Bestandsdaten B LEFT JOIN Schadendaten S

ON B.VNR = S.VNR

(b) Die Variable „VN_Alter" enthält fehlende Werte / NULL Values. Im Vorfeld zu der Modellierung gibt es folgende Möglichkeiten, mit der Datenauffälligkeit umzugehen:

- Variable ausschließen. Die Variable, bei der Datensätze mit fehlenden Werten vorliegen, wird für die weiteren Analysen und die Modellierung nicht verwendet.

- Datensätze ausschließen. Die Datensätze bei denen die Variable fehlende Werte hat, werden für die weiteren Analysen und die Modellierung nicht verwendet.
- Fehlende Werte ersetzen (Imputation). Die fehlenden Werte werden durch eine statistische Kennzahl (z.B. Mittelwert oder Median) aus den Datensätzen der Variable mit vorhandenen Werten ersetzt.
- Vorhersage der fehlenden Werte. Durch ein Regressionsmodell werden die Werte bei den Datensätzen ohne Werte vorhergesagt und die fehlenden Werte entsprechend ersetzt (Lernziel ADS Advanced).
- Keine Anpassungen der Datengrundlage. An der Datengrundlage werden keine Anpassungen durchgeführt und in der Modellierung werden Methoden verwendet, die mit fehlenden Datensätzen umgehen können (z.B. Entscheidungsbäume).

Weitere Antworten zum Umgang mit fehlenden Werten sind möglich.

(c) Die Überprüfung, ob die Datensätze einer Variablen X normalverteilt sind, ist unter Anwendung eines Probability Plot (Q-Q Plot) in folgenden Schritten möglich:

- Die n Datensätze der Variable X werden aufsteigend sortiert: $x_1, \dots, x_j \ \dots, x_n$
- Zu den aufsteigend sortierten n Datensätzen werden jeweils die j-ten Quantile der Normalverteilung ermittelt $\hat{x}_1, \dots, \hat{x}_j \ \dots, \hat{x}_n$
- Die n Datensätze der Variable X werden gegen die n Quantile der Normalverteilung geplottet
- Befinden sich die Datenpunkte $(x_1, \hat{x}_1), \dots, (x_j, \hat{x}_j) \dots, (x_n, \hat{x}_n)$ auf der Diagonalen des Diagramms (mit Abweichungen innerhalb eines Toleranzbereichs), dann haben die Daten der Variable X eine Normalverteilung. Grobe Abweichungen von der Diagonalen sprechen gegen eine Normalverteilungsannahme.

(d) Ziel der Analyse ist es die Schadenaufwendungen zu untersuchen und vorherzusagen. Die Zielvariable des Regression Tree‘s ist daher die Variable „Schadenaufwand“.

Da die Variable „Schaden vorhanden“ nicht als beschreibende Variable für das Prognosemodell verwendet werden kann und der Wert der Verlustfunktion für den Split durch die Variable „Geschlecht“ bekannt ist, so muss der Wert der Verlustfunktion bei dem Split der Variable „VN Alter“ bestimmt werden.

Die Verlustfunktion für den Regression Tree mit einem binären Split ist:

$$Verlustfunktion = \frac{m_1}{m} * MSE_1 + \frac{m_2}{m} * MSE_2$$

Hier ist gilt:

- m = Anzahl der Datensätze im Ausgangsknoten
- $\mathrm{m_i}$ = Anzahl der Datensätze im Knoten i unterhalb des Ausgangsknotens
- $\mathrm{MSE_i}$ = Mittlerer Quatratischer Fehler im Knote i unter des Ausgangsknotens, mit

  $\mathrm{MSE_i} = \sum_{\mathrm{j} \in \mathrm{Knoten\ i}} \left(\widehat{y_\mathrm{i}} - \mathrm{y^j}\right)^2$ und

  j = Anzahl der Datensätze im Knoten i

  $\mathrm{y^j}$ = Wert der Zielvariable für den $\mathrm{j}-$ ten Datensatz

  $\widehat{y_\mathrm{i}}$ = Prognosewert für die Zielvariable im Knoten i

(Weitere Varianten der Verlustfunktion sind möglich.)

Aus den Datensätzen ist ersichtlich, dass Schäden verstärkt bei jungen Versicherungsnehmern auftreten. Als Grenze des Splits wird daher das Alter 31 ausgewählt. Der Wert der Verlustfunktion berechnet sich hierbei wie folgt:

- Split 1 (Alter VN <= 31):

  $\widehat{y_1} = \frac{3+3+0+3}{4} = 2{,}25$

  $MSE_1 = 3 * (3-2{,}25)^2 + (0-2{,}25)^2 = 6{,}75$

- Split 2 (Alter VN > 31):

  $\widehat{y_2} = \frac{0+0+0+0+0+1}{6} = 0{,}17$

  $MSE_2 = 5 * (0-0{,}17)^2 + (1-0{,}17)^2 = 0{,}83$

  Der Wert der Verlustfunktion berechnet sich wie folgt:

  $$\frac{4}{10} * 6{,}75 + \frac{6}{10} * 0{,}83 = \ 3{,}2$$

Da der Wert der Verlustfunktion bei dem binären Split der Variable „VN Alter" geringer ist als bei der Variablen „Geschlecht", soll der erste Split im Regression Tree bei dem VN Alter <= 31 erfolgen.

(e) Bei einem Overfitting passt sich das Modell zu stark an die Daten an, auf denen es trainiert wurde. Das Overfitting beeinflusst negativ die Fähigkeit des Modells zur korrekten Vorhersage von neuen Daten.

Parameter zur Regularisierung bei der Erstellung von Entscheidungsbäumen sind:

- Maximale Tiefe des Baums: Festlegung der maximalen Anzahl von Ebenen in einem Baum.
- Minimale Anzahl von Datensätzen pro Split: Anzahl der Datensätze, die in einem inneren Knoten mindestens vorhanden sein müssen, um weitere Split durchzuführen.
- Minimale Anzahl von Datensätzen in einem Blatt: Anzahl der Datensätze die in jedem Blatt / Leaf Node mindestens vorhanden sein müssen.
- Maximale Anzahl von Blättern / Leaf Nodes: Anzahl der Blätter / Leaf Nodes, die maximal in einem Baum vorhanden sein dürfen.

Weitere Antworten zu den Parametern des Entscheidungsbaums sind möglich.

Aufgabe 3 [Regressions- und Clustermethoden (3.1.2)] [25 Punkte]

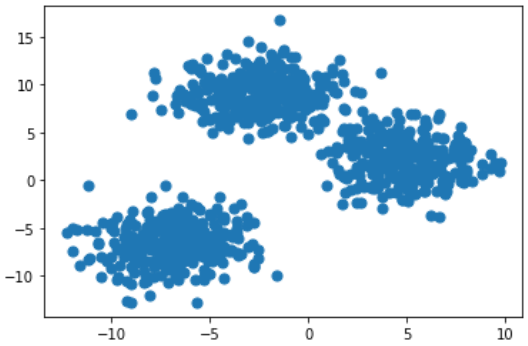
a) [6 Punkte] Beschreiben Sie den k-Means-Algorithmus in wenigen Sätzen.

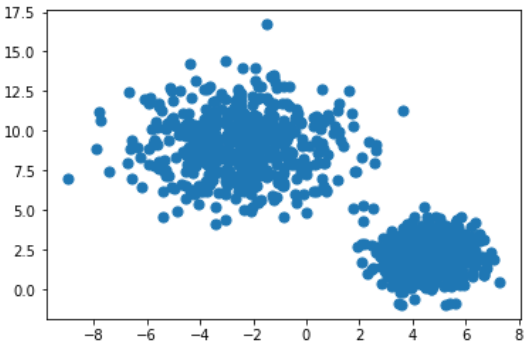
b) [6 Punkte] Geben Sie zwei Beispiele aus dem Versicherungsumfeld an, bei denen man Clusteringverfahren anwenden kann, und den jeweiligen Nutzen.

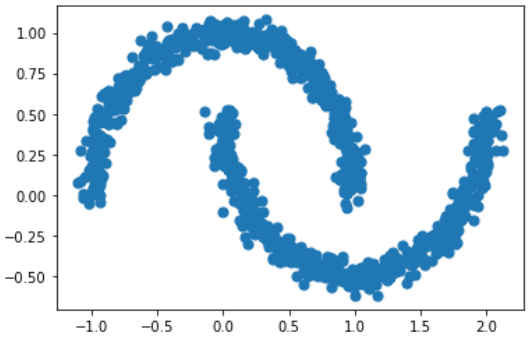
c) [9 Punkte] Entscheiden Sie bei den folgenden Plots zunächst, wie viele Cluster k jeweils vorliegen. Skizzieren Sie auf der Basis dieses von Ihnen gewählten k jeweils ein mögliches Ergebnis des k-Means-Algorithmus.

Bei welchem Plot liefert der Algorithmus kein sinnvolles Ergebnis?

Nennen Sie eine Eigenschaft, die Cluster haben sollten, damit der Algorithmus sinnvolle Ergebnisse liefert.

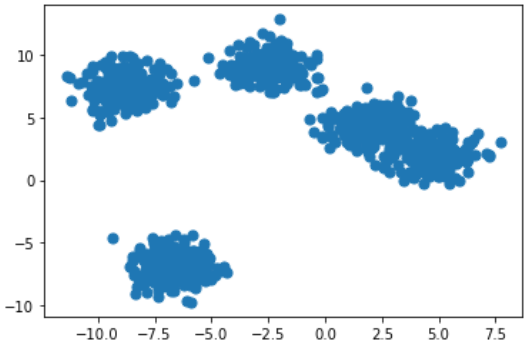




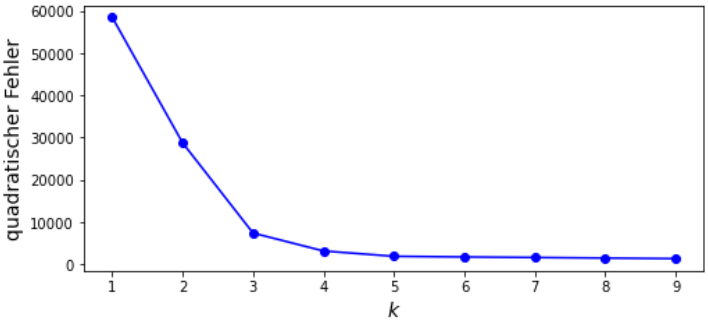


d) [4 Punkte]

Entscheiden Sie, analog zu Teil c), wie viele Cluster Sie bei diesem Datensatz vermuten:



Sie führen für diesen Datensatz den k-Means-Algorithmus für ein bis neun Cluster durch. Anschließend plotten Sie die jeweiligen Summen der quadratischen Abweichungen und bekommen den folgenden Verlauf:



Wie viele Cluster würden Sie anhand dieses Plots auswählen (mit Begründung)?

Lösungsvorschlag:

a)

Die Datenpunkte werden zunächst zufällig einem der k Cluster zugeordnet. (1 Punkt)

Solange bis sich die Cluster nicht mehr ändern, werden folgende Schritte ausgeführt: (4 Punkte)

- Berechnung der Clusterzentren als (koordinatenweise) Durchschnittswerte aller Punkte im jeweiligen Cluster
- Zuordnung der Datenpunkte zu dem Cluster, bei dem das Clusterzentrum den geringsten euklidischen Abstand aufweist.

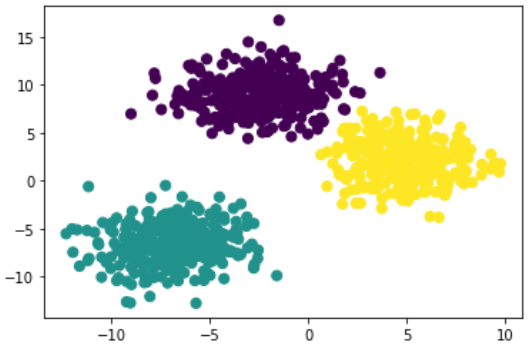
b) Beispiel 1: Clusteringverfahren können zur Segmentierung der Versicherungsnehmer benutzt werden. Die Segmentierung kann dann für maßgeschneiderte Bestandsaktionen benutzt werden.

Beispiel 2: Für Bestandsprojektionen können Clusteringverfahren benutzt werden, um den Bestand zu verdichten und die Laufzeit der Projektion zu reduzieren.

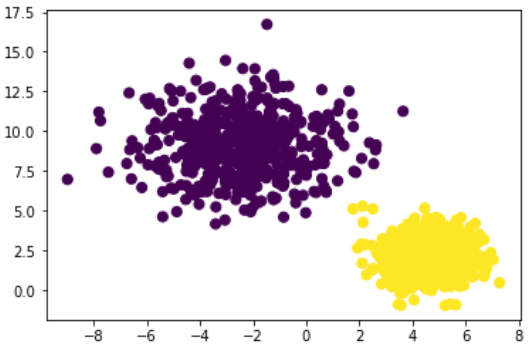
Beispiel 3: Bei Bestandsmigrationen können die Verträge geclustert werden, um eine repräsentative Auswahl als Testbestand festzulegen.

c)

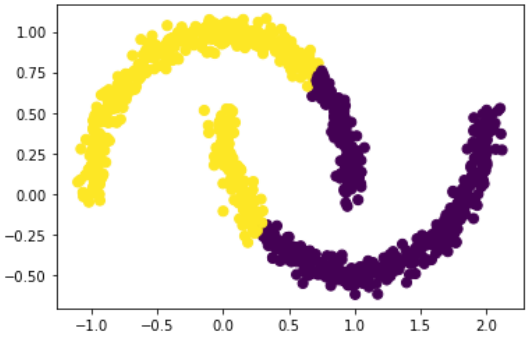
Beim ersten Plot liegen drei Cluster vor. k-Means könnte folgendes Clustering liefern: (2 Punkte)



Beim zweiten Plot liegen zwei Cluster vor. k-Means könnte folgendes Clustering liefern: (2 Punkte)



Beim dritten Plot liegen zwei Cluster vor. K-Means könnte folgendes Clustering liefern: (2 Punkte)



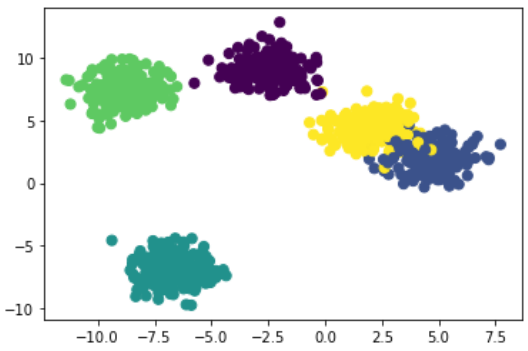
Offensichtlich liefert k-Means beim dritten Plot kein sinnvolles Ergebnis. (1 Punkt) Damit k-Means sinnvolle Ergebnisse liefert, sollten die einzelnen Cluster näherungsweise kreisförmig (bzw. kugelförmig in höheren Dimensionen) sein. Diese Eigenschaft ist beim dritten Plot verletzt. (2 Punkte)

d)

Es sind vier Cluster zu erkennen. Tatsächlich wurden für das Beispiel fünf kreisförmige Cluster erzeugt. Richtige Antworten sind daher vier oder fünf. (2 Punkte)

Anhand des Plots der quadratischen Fehler wählen wir auch vier Cluster. Der quadratische Fehler sinkt bis zu vier Clustern deutlich. Bei mehr als vier Clustern sinkt der quadratische Fehler nur noch unwesentlich.

Hinweis: Der zugrundeliegende Datensatz sieht wie folgt aus (nicht Teil der Lösung):


10
5
0
−5
−10
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5

Aufgabe 4 [Datenmanagement 2.1] [42 Punkte]

Die Versicherung SunnySide AG ist dabei, den bisherigen Beratungsprozess (von Papier) in eine digitale Form zu überführen. Dabei ist eine Webapplikation geplant, die die bisherigen Schritte von der Vorstellung des Vertreters bis zur Vorschlagserstellung abbildet. Als Actuarial Data Scientist sind Sie Mitglied des Projektteams.

(a) [15 Punkte] Zunächst überlegen Sie mit an der Gestaltung der Objekttypen und deren Beziehungen. Folgende Informationen liegen Ihnen vor:

Ein Vertreter betreut keine oder mehrere Personen, und eine Person ist mindestens einem Vertreter zugeordnet. Für jede Person gibt es keine oder mehrere Beratungen, und jede Beratung ist genau einer Person zugeordnet (hier werden Beratungsgruppen zunächst nicht weiter betrachtet). Aus Beratungen können Vorschläge resultieren, müssen aber nicht. Jeder Vorschlag ist genau einer Beratung zuordenbar.

Für Vertreter werden die Vertreternummer, Vorname, Name und Agentur gespeichert. Bei Personen sind hier nur eine PersonenID, Vorname, Name, Geburtstag und Adresse (bestehend aus Straße, Ort und Postleitzahl) relevant. Bei Beratungen wird eine BeratungsID, Datum der Beratung, die ausgewählten Themen (hier betrachten wir nur die Themen „Haftpflichtversicherung", „Kfz-Versicherung" und „Unfallversicherung") sowie Notizen (die themenbezogen erfasst werden) gespeichert. Vorschläge haben (sehr vereinfacht) eine VorschlagsID, Sparte, ein Feld für einen Produktnamen und einen Preis.

Erstellen Sie auf Basis dieser Informationen ein Entity-Relationship-Diagramm (in der „Krähenfuß-Notation").

(b) [7 Punkte] Nach der Erstellung des Diagramms diskutieren Sie dieses in ihrem Projektteam. Eine Mitarbeiterin hat eine Anmerkung zum Objekttyp „Beratung". Was ist aufgefallen? Führen Sie, falls nötig, notwendige Korrekturen durch.

(c) [12 Punkte] Nach mehreren weiteren Abstimmungsrunden entscheiden Sie sich dazu, aus dem finalisierten Diagramm ein Datenbank-Modell abzuleiten. Dabei haben Sie sich auf eine relationale Datenbank geeinigt. Geben Sie die Typbeschreibung für alle Objekttypen aus (b) in Tabellenform an. Berücksichtigen Sie dabei, falls nötig, Koppeltabellen sowie Primär- und Fremdschlüssel.

(d) [8 Punkte] Die durchgeführten Beratungen können über ein Drittsystem gespeichert werden. Umgekehrt können Beratungen aus diesem System wieder geladen werden. Dabei ist geplant, die gespeicherten Beratungen

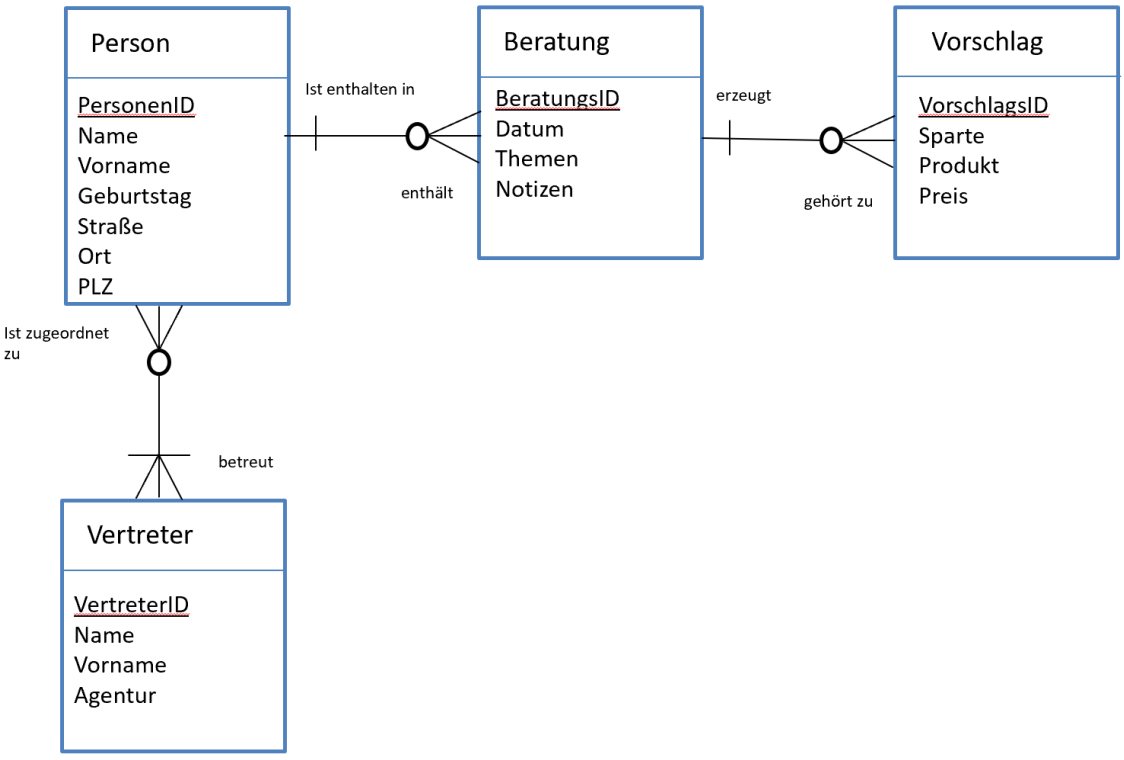
über ein JSON-Format zuzuschicken. Konstruieren Sie anhand des folgenden Beispiels eine entsprechende Struktur, in der Sie die rekursive Eigenschaft des Datentyps ausnutzen:

Die Person Erika Mustermann (PID 1717) geboren am 12.08.1995, wohnhaft in der Musterstraße 1 in 10115 Berlin, hatte zwei Beratungsgespräche. Im ersten Gespräch (ID 1) am 19.01.2020 wurde das Thema „Unfallversicherung" (ID 3) besprochen. Es wurde keine Notiz erfasst.

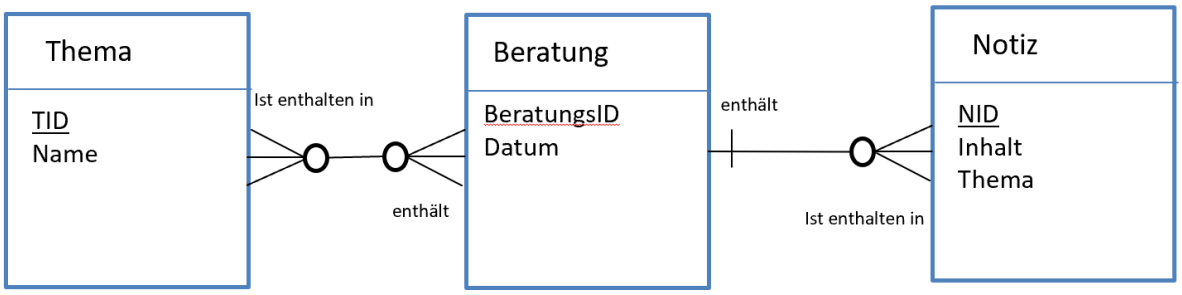
Im zweiten Gespräch (ID 2) am 06.03.2020 wurden die Themen „Kfz-Versicherung" (ID 1) und „Haftpflichtversicherung" (ID 2) besprochen. Bei der „Kfz-Versicherung" wurde die Notiz „Vorschlag ausstehend" (ID 2) erfasst. Bei der „Haftpflichtversicherung" wurde die Notiz „Vorschlag besprochen" (ID 3) erfasst. Zudem wurde ein Vorschlag (ID 1) mit der Sparte „Sach", dem Produkt „Haftpflicht" und dem Preis (€ 11,93) erfasst.

Lösungsvorschlag:

(a) Aus der Aufgabenstellung ergibt sich das folgende Diagramm:



(b) Das Modell verstößt gegen die erste Normalenform, da multiple Eigenschaften (Themen und Notizen) vorhanden sind. Diese müssen aufgelöst werden:



Hinweis: Hier wird eine Möglichkeit der Modellierung gezeigt.

(c) Aus dem Diagramm ergeben sich die folgenden Typbeschreibungen:

Vertreter(**VertreterID,** Name, Vorname, Agentur)

Vertreter/Person**(⇑VertreterID⇑+⇑PersonenID⇑)**

Person(**PersonenID**, Name, Vorname, Geburtstag, Straße, Ort, PLZ)

Beratung(**BeratungsID**, Datum, ⇑PersonenID⇑)

Hinweis: Der Fremdschlüssel ist nicht unikal jedoch eingabepflichtig.

Notiz(**NID**, Inhalt, Thema, ⇑BeratungsID⇑)

Beratung/Thema(⇑**BeratungsID⇑ + ⇑TID⇑**)

Thema(**TID**, Name)

Vorschlag(**VorschlagsID**, Sparte, Produkt, Preis, ⇑BeratungsID⇑)

(d) Die folgende JSON Struktur bildet die angegebenen Informationen ab:

```
{
  "PersonenID": 1717,
  "name" : "Mustermann",
  "vorname" : "Erika",
  "geburtstag" : "12.08.1995",
  "strasse" : "Musterstraße 1",
  "Ort": "Berlin",
```

```
    "PLZ": 10115,
    "beratung": [
        {"BeratungsID": 1, "Datum" : "19.01.2020",
            "thema": [
                {"TID" : 3, "name": "Unfallversicherung"}]},
        {"BeratungsID": 2, "Datum": "06.03.2020",
            "thema": [
                {"TID": 1, "name" : "Kfz-Versicherung"},
                {"TID": 2, "name": "Haftpflichtversicherung"}],
            "notiz": [
                {"NID": 2, "inhalt": "Vorschlag ausstehend",
                    "thema" : "Kfz-Versicherung"},
                {"NID": 3, "inhalt": "Vorschlag besprochen"}],
            "vorschlag": [{
                    "vorschlagsID": 1, "sparte": "Sach",
                    "produkt": "Haftpfl.vers.",
                    "preis": 11,93}]}
            ]
}
```

Aufgabe 5 [Data Mining (4.1)] *[32 Punkte]*

Sie arbeiten in einem internationalen Team und wirken als Actuarial Data Scientist bei der Datenvorbereitung mit. In diesem Rahmen bekommen Sie Fondsdaten aus zwei Quellen, die Sie zusammenführen sollen. Ihre Tätigkeit sollen Sie in einem Datenintegrationsbericht dokumentieren.

Die Daten sind hier auszugsweise exemplarisch aufgeführt:

Quelle1:

Fondsname	Fondskurs	Datum	Währung
DKF_VM	12.7	01.02.2021	SchwFr
DWX_BP	13.21	01.02.2021	Euro
RFX_AB	2.38	01.02.2021	US-Dollar

Quelle2:

Date	Curr	Name	Rate
1.2.21	EUR	DWXBP	13,21
1.2.21	SFR	ABCDE	1.937,21
1.2.21	EUR	S1PRD	257,38

(a) *[17 Punkte]* Erstellen Sie die Datenstruktur für eine neue Tabelle FundRates, welche die Daten aus den beiden Quellen aufnehmen soll (Attributnamen, Reihenfolge der Attribute, Skalenniveaus und Datentypen), und beschreiben Sie das Mapping in tabellarischer Darstellung.

Erläutern und begründen Sie jede Ihrer bei der Definition der Datenstruktur getroffenen Entscheidungen in ganzen Sätzen.

(b) *[8 Punkte]* Füllen Sie die neue Tabelle FundRates basierend auf den oben aufgeführten exemplarischen Inhalten aus. Nehmen Sie dabei alle Ihnen sinnvoll erscheinenden Anpassungen der Inhalte vor.

Erläutern und begründen Sie jede Ihrer bei der Befüllung getroffenen Entscheidungen in ganzen Sätzen.

(c) *[7 Punkte]* Betrachten Sie nochmals die exemplarische Befüllung. Sehen Sie angesichts eines bestimmten Datensatzes den Bedarf, vor einer Verwendung der Daten weitere Recherchen anzustellen? Welche Vermutung haben Sie und weshalb? Wie verifizieren Sie Ihre Vermutung? Was unternehmen Sie, wenn sich Ihre Vermutung bestätigen sollte?

Lösungsvorschlag:

(a) *[1 Punkt]* Die Attribute der neuen Tabelle werden wegen der internationalen Teamzusammensetzung *englisch* benannt (konsistent mit dem neuen Tabellennamen).

[5 Punkte] Folgende Tabelle beschreibt das *Mapping* und die Skalenniveaus und Datentypen:

Attr.Name Quelle1	Attr.Name Quelle2	Attr.Name in Tabelle FundRates	Reihenfolge	Skalenniveau	Datentyp
Fondsname	Name	FundName	1	Nominalskala	Character
Fondskurs	Rate	FundRate	4	Verhältnisskala	Numerisch
Datum	Date	RateDate	2	Intervallskala	Character / Date
Währung	Curr	Currency	3	Nominalskala	Character / Factor

Die Entsprechungen der Attribute von Quelle 1 und 2 sind in diesem Fall offensichtlich.

[2 Punkte] Um der Klarheit und Kontextunabhängigkeit willen wurden die englischen *Bezeichnungen* aus Quelle 2 verlängert. Curr könnte z.B. auch für Current stehen. Ein Attributname Date könnte je nach Entwicklungsumgebung Probleme mit vorhandenen allgemeinen Datentypen oder Funktionen namens Date verursachen; daher die Entscheidung für RateDate.

[2 Punkte] Bei der *Reihenfolge* der Attribute ist zu beachten, dass die Schlüsselattribute FundName, RateDate und Currency zuerst kommen sollten (also FundRate am Schluss). Über die Reihenfolge der Schlüsselattribute lässt sich diskutieren; üblicherweise wird der FundName als führend angesehen.

[2 Punkte] Die Attribute *FundName* und *Currency* liegen in den Originalquellen als Zeichenfolgen vor; daher ist der Datentyp Character. Da die (lexikalische) Reihenfolge in beiden Fällen keine Rolle spielt, handelt es sich um nominalskalierte Merkmale.

[1 Punkt] Das Attribut *Currency* lässt sich wegen des begrenzten und wohldefinierten Ausprägungsumfangs in R als Factor-Variable behandeln (kategorielle Variable).

[2 Punkte] Das *RateDate* ist streng genommen eine Zeichenfolge, die ein Datum enthält, und wäre damit ebenfalls nominalskaliert. Das (technische) Format hängt hier von der Entwicklungsumgebung ab. Fachlich ist

es aber möglich, Differenzen zwischen zwei Werten zu berechnen; daher ist das Attribut intervallskaliert.

[2 Punkte] Die *FundRate* ist offenbar numerisch. Da es einen natürlichen Nullpunkt gibt, handelt es sich um eine verhältnisskalierte Größe.

Anmerkung: Andere Lösungen sind möglich, eine entsprechende differenzierte Begründung vorausgesetzt.

(b) *[4 Punkte]* Die Befüllung könnte wie folgt aussehen:

FundRates:

FundName	RateDate	Currency	FundRate
DWXBP	01.02.2021	EUR	13,21
ABCDE	01.02.2021	CHF	1937,21
S1PRD	01.02.2021	EUR	257,38
DKF_VM	01.02.2021	CHF	12,70
DWX_BP	01.02.2021	EUR	13,21
RFX_AB	01.02.2021	USD	2,38

Generell ist bei der Befüllung eine Orientierung an allgemeinen Standards sinnvoll.

[1 Punkt] Der FundName kann zunächst 1:1 übernommen werden. Er enthält in beiden Quellen nicht die WKN (Wertpapierkennnummer) oder die 12-stellige ISIN (International Securities Identification Number), sondern offenbar quellenspezifische Namen. Daher ist ggf. eine Vereinheitlichung erforderlich. Diese sollte aber in einem gesonderten Schritt erfolgen (siehe Teilaufgabe (c)).

[1 Punkt] Beim RateDate sollte möglichst ein eindeutiges, kontextfreies Datumsformat gewählt werden (z.B. Jahresangabe 4-stellig). Wegen der Internationalität wäre hier auch „2021-02-01" denkbar.

[1 Punkt] Currency sollte – entsprechend Quelle 2 – mit den gemäß ISO genormten Währungs-Abkürzungen belegt werden.

[1 Punkt] Das Zahlenformat der FundRate ist anzugleichen (in der Musterlösung: Dezimalkomma, kein Tausender-Punkt; international denkbar wäre auch „0000.00"). Hinweis: Es gibt auch Fondskurse unterhalb der Cent-Schwelle; ob solche im System verwendet werden sollen, wäre vor der technischen Umsetzung abzuklären.

(c) *[1 Punkt]* Vermutlich handelt es sich bei DWXBP und DWX_BP um den gleichen Fonds. Dafür sprechen ähnliche Namen und gleiche Kurse.

[2 Punkte] Wenn möglich sollten in solchen Fällen immer die Lieferanten der Datenquellen bzw. eine Dokumentation hinzugezogen werden (wenn nicht schon vor dem ersten Blick auf die Daten geschehen). Ansonsten

wäre die Beobachtung im ersten Schritt anhand von Einträgen zu anderen Terminen zu verifizieren.

[4 Punkte] Wenn es sich offenbar um den gleichen Fonds handelt: Ist die Benennung innerhalb der Originalquellen konsistent (also Quelle 1 immer DWXBP und Quelle 2 immer DWX_BP) oder liegt in einem Fall ein Tippfehler vor?

- Ein Tippfehler müsste korrigiert werden.
- Im Konsistenzfall dagegen wäre zu klären, welcher Name in der neuen Tabelle zu verwenden ist.

  Insbesondere wäre zu klären, ob aus den beiden Quellen weitere Tabellen übernommen werden, in denen ebenfalls Namen für Fonds verwendet werden; in einem solchen Fall sind diese Bezüge ebenfalls auf den vereinheitlichten Namen anzupassen.