Schriftliche Prüfung im Spezialwissen

Actuarial Data Science Basic

gemäß Prüfungsordnung 4.1
der Deutschen Aktuarvereinigung e. V.

am 21.05.2022

Hinweise:

- Als Hilfsmittel ist ein Taschenrechner zugelassen.
- 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 13 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 [1.1 Grundlagen, 1.2 Digitalisierung, 1.3 Gesellschaftliches Umfeld & Ethik, 1.4 Datenschutz 1] [31 Punkte]

Sie sind auf der firmeneigenen Weihnachtsfeier und unterhalten sich. Sie erzählen, dass Sie die Ausbildung zum Actuarial Data Scientist angefangen haben. Ein*e Kolleg*in ist überrascht und meint, das sei nichts anderes als die Arbeit als Aktuar*in.

a) [4 Punkte] Sie widersprechen mit den anderen Anforderungen an einen Actuarial Data Scientist. Nennen Sie hierfür vier Tätigkeiten eines Actuarial Data Scientist.

b) [7 Punkte] Sie konnten noch nicht vollständig überzeugen, darum argumentieren Sie mit der voranschreitenden Digitalisierung. Nennen Sie vier Wettbewerbsvorteile der Digitalisierung für Unternehmen und nennen Sie drei Auswirkungen hieraus auf die Aufgaben eines Aktuars.

c) [3 Punkte] So langsam kommen Sie voran. Nennen Sie drei Beispiele für Datenquellen aus dem aktuariellen Kontext, welche für Data Science-Anwendungen relevant sein können.

d) [6 Punkte] Jetzt haben Sie überzeugt. Ihre Kollegen möchten sofort die *bestehende* Prämienkalkulation anpassen. Hierfür sollen die Kunden durch einen 5 € Gutschein auf die Unternehmenshomepage gelockt werden und anschließend durch Cookies ihr Surfverhalten analysiert werden. Dieses soll dann zur Tarifierung verwendet werden. Nennen Sie den Kollegen drei Gründe aus dem Code of Conduct der Versicherungsbranche, warum dies nicht möglich ist.

e) [11 Punkte] Die Kollegen sind enttäuscht und wollen die Einwände durch Einführung eines neuen innovativen Produktes umgehen. Das neue Produkt soll als Zusatzversicherung im Bereich Krankenversicherung angeboten werden. Hierbei sollen die Kunden einen Rabatt auf ihre Prämie bekommen, wenn sie eine gesunde Lebensführung einhalten. Zur Überwachung der Lebensführung müssen die Kunden regelmäßig Gesundheitsdaten (Werte aus Fitnesstrackern, Essgewohnheiten, sportliche Aktivitäten, ...) auf einer Website eintragen. Aus diesen Daten wird nach einem von Ihrem Unternehmen festgelegten Algorithmus ein Lebensführungs-Score berechnet. Damit die Kunden einen höheren Ansporn haben, werden die erreichten Scores auf der Website veröffentlicht. Nennen und erläutern Sie Ihren Kollegen fünf ethische/gesellschaftliche Bedenken, welche bei so einem Produkt bedacht werden sollten. (Bitte nennen Sie die Bedenken kurz und knapp. Maximal 2 Sätze pro Bedenken.)

Lösungsvorschlag:

a) Es müssen vier der folgenden Stichpunkte genannt werden (1 Punkt pro Stichpunkt):

- Schnittstelle zwischen Management und IT (Übersetzer in beide Richtungen)
- Umgang mit großen Datenmengen
- Auswertung der Daten und Erstellung von Modellen
- Berücksichtigung von Datenschutz und Datensicherheit
- Konzeption von technischen Lösungen
- Kommunikation mit vielen unterschiedlichen Bereichen (IT, Fachabteilung, Management, …)
- …

b) Es müssen vier der folgenden Wettbewerbsvorteile genannt werden (1 Punkt pro Vorteil):

- Höhere Kundenzufriedenheit
- Niedrigere Verwaltungskosten durch weniger Mitarbeiter, weniger Sachkosten
- Weniger Aufwand durch Standardisierung
- Datenhandel
- Bewertungsmöglichkeiten für neue Risiken und sich verändernde Risiken
- …

Es müssen drei der folgenden Auswirkungen genannt werden (1 Punkt pro Auswirkung):

- Berichte werden automatisiert erzeugt
- Es müssen deutlich mehr Datenquellen berücksichtigt werden

- Es müssen unstrukturierte Daten aus verschiedensten Quellen berücksichtigt werden
- Das heißt mehr Fähigkeiten eines Data Scientist nötig
- Mehr Zeit, um Zahlen zu interpretieren (Berichtsqualität erhöhen)
- Mehr Zusammenarbeit mit IT nötig (oder IT selber machen)

c) Es müssen drei Beispiele genannt werden. (1 Punkt pro Beispiel). Mögliche Beispiele sind: Bestandsführungssysteme, DWH-Systeme, Informationen aus Wearables, Daten aus Apps, Daten aus der Homepagenutzung, Kundendaten (über klassische Bestandsdaten hinaus), Wetterdaten, Unfalldaten, Daten zu Einbrüchen, Daten aus Unfallmeldesystemen, usw. usf.

d) Im CoC heißt es: Artikel 2 (1): Die Erhebung, Verarbeitung oder Nutzung personenbezogener Daten erfolgt grundsätzlich nur, soweit dies zur Begründung, Durchführung oder Beendigung eines Versicherungsverhältnisses erforderlich ist. Oder auch (ebenso Artikel 2 (1)) Die personenbezogenen Daten werden grundsätzlich im Rahmen der den Betroffenen bekannten Zweckbestimmung verarbeitet und genutzt. Daneben gilt noch Artikel 6: Die Verarbeitung besonderer Art personenbezogener Daten (zum Beispiel Gesundheitsdaten) bedürfen der gesonderten expliziten Zustimmung des Betroffenen. Und schließlich könnte man noch Artikel 7 anführen: Personenbezogene Daten werden grundsätzlich bei den Betroffenen selbst erhoben. Drei dieser Gründe sollten genannt werden (2 Punkte pro Grund).

e) Hier kann ein kleiner Aufsatz folgen, es sind aber auch Stichpunkte ausreichend (2 Punkte pro Stichpunkt). Folgendes sollte aus ethischer Sicht bedacht werden:

- Die Veröffentlichung der Scores (ggf. mit Namen, das ist im Aufgabentext nicht ausgeschlossen) kann zu einem großen Druck bei den Versicherten führen.
- Die Festlegung, was eine gesunde Lebensführung ist, wird von einem Versicherungsunternehmen zwar sicher nach bestem Wissen und Gewissen und auf wissenschaftlicher Basis festgelegt. Dies sollte aber nicht von einem Unternehmen festgelegt werden.
- Personen mit ungesundem Lebensstil könnten diskriminiert werden.
- Die Angabe von Gesundheitsdaten kann zur Überwachung der Versicherten führen.

- Das Solidaritätssystem verliert ggf. seine Wirkung, da ein Ausgleich im Kollektiv nicht mehr möglich ist.

- …

Aufgabe 2 [2.1 Datenmanagement 1] [32 Punkte]

a) [15 Punkte] In einer relationalen Datenbank liegen die folgenden Tabellen vor:

id	num	value
1	1	A
2	1	B
3	2	D
4	3	C

table 1

id	num	value
1	2	R
2	2	S
3	3	T
4	3	U

table 2

id	num	value
1	2	V
2	2	W
3	3	X
4	5	Y
5	8	Z

table 3

Geben Sie das Ergebnis in Tabellenform für folgende SQL-Statements an:

i. SELECT a.value as aval, b.value as bval
FROM table1 as a INNER JOIN table2 as b
ON a.num = b.num;

ii. SELECT a.value as aval, b.value as bval
FROM table1 as a LEFT JOIN table2 as b
ON a.num = b.num;

iii. SELECT b.id as bid, b.value as bval, c.value as cval, a.value as aval
FROM (table3 as c LEFT JOIN table2 as b
ON b.num = c.num)
RIGHT JOIN table1 as a
ON c.num = a.num;

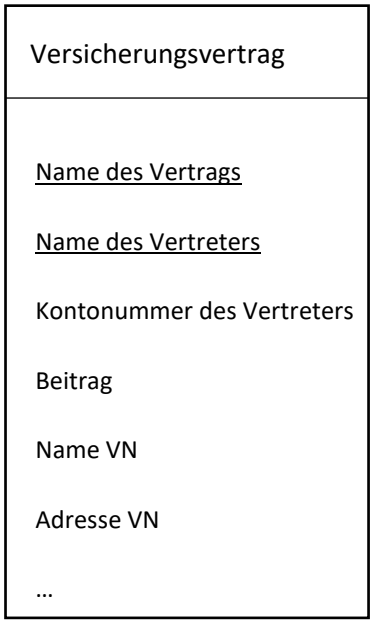
b) [17 Punkte] Ein Datenmodell liegt in der zweiten Normalform vor, wenn

   a. es sich in der ersten Normalform befindet und

   b. jede beschreibende Eigenschaft eines Objekttyps zwar vom Gesamtschlüssel, aber nicht bereits von einem Teilschlüssel dieses Objekttyps funktional abhängig ist.

   Dabei ist innerhalb eines Objekttyps eine Eigenschaft B dann von einer Kombination von Elementen (Eigenschaften und/oder Beziehungstyp-Richtung) E1, …, En funktional abhängig, wenn sich für jedes konkrete Objekt dieses Objekttypen aus der Kombination der Werte dieser Elemente direkt auf den Wert der Eigenschaft B schließen lässt.

   Sie betrachten den folgenden Objekttypen:



   a. Begründen Sie, warum dieser Objekttyp nicht in der zweiten Normalform vorliegt.

   b. Warum ist die Herbeiführung der zweiten Normalform sinnvoll?

   c. Führen Sie bei dem Objekttyp „Versicherungsvertrag" die zweite Normalform herbei. Begründen Sie Ihre Entscheidungen.

Lösungsvorschlag:

a) Nachfolgend sind die Ergebnistabellen abgebildet

i. Ergebnis des Statements:

	aval	bval
▶	D	R
	D	S
	C	T
	C	U

ii. Ergebnis des Statements:

	aval	bval
▶	A	NULL
	B	NULL
	D	S
	D	R
	C	U
	C	T

iii. Ergebnis des Statements:

	bid	bval	cval	aval
▶	NULL	NULL	NULL	A
	NULL	NULL	NULL	B
	1	R	W	D
	2	S	W	D
	1	R	V	D
	2	S	V	D
	3	T	X	C
	4	U	X	C

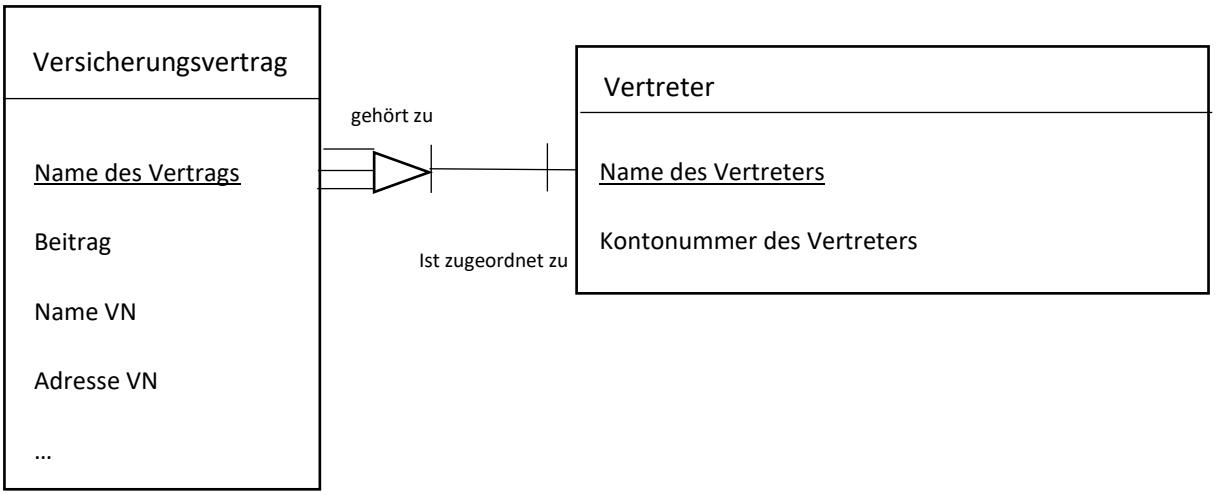
b) Für den Objekttyp ist folgendes zu beachten:

a. Der Objekttyp liegt in der ersten Normalform vor (keine multiplen Eigenschaften). Zu beachten ist, dass der Objekttyp einen zusammengesetzten Schlüssel aufweist. Der Gesamtschlüssel besteht hierbei aus den beiden teilidentifizierenden Eigenschaften „Name des Vertrags" und „Name des Vertreters". Die beschreibende Eigen-

schaft „Kontonummer des Vertreters" ist vom Teilschlüssel „Name des Vertreters" funktional abhängig.

b. Das Ziel bei der Herbeiführung der zweiten Normalform liegt in der Beseitigung von Redundanzen bei der Datenspeicherung.

c. Für die Herbeiführung der zweiten Normalform muss die teilidentifizierende Eigenschaft „Name des Vertreters" und die von ihr funktional abhängige beschreibende Eigenschaft „Kontonummer des Vertreters" aus dem Objekttyp „Versicherungsvertrag" herausgelöst werden. Sie bilden einen neuen Objekttyp „Vertreter":



Aufgabe 3 [Datenmanagement 1 (2.1), Datenverarbeitungstechnologien 1 (2.2), Data Mining 1 (4.1), Analytics 1 (4.2), Regressions- und Clustermethoden 1 (3.1)] [42 Punkte]

Zur Optimierung der Vertriebsaktivitäten in Ihrem Unternehmen sollen gezielt Bestandskunden kontaktiert werden, die eine hohe Wahrscheinlichkeit für den Abschluss einer Zusatzversicherung haben. Vom Vorstand in Ihrem Unternehmen haben Sie die Aufgabe bekommen, hierzu ein Prognosemodell zu erstellen.

Grundlage für die Analysen sind die Vertriebsdaten in Form von JSON-Dateien. Diese enthalten die Information, ob im Vorjahr eine Zusatzversicherung abgeschlossen wurde. Folgend sind zwei exemplarische JSON-Dateien dargestellt:

Json 1

```
{
 "Kunden-Name": „Müller“,
 "Kunden-Vorname": „Max“,
 "Kunden-ID": 123456,
 "Alter": 53,
 "Geschlecht": „m“,
 "Adressen": [
   {
    "PLZ": 80796,
    "Ort": „München“,
    "Straße": „Mittermayrstraße 2“,
   },
{
    "PLZ": 10719,
    "Ort": „Berlin“,
    "Straße": „Kurfürstendamm 8“,
   }
  ],
  " Verträge": [
   {
    "Vertragart": „Hauptversicherung“,
    "Vertragsabschluss": 2010,
    " Jahresbeitrag“: 220.20
   },
{
    "Vertragart": „Zusatzversicherung“,
    "Vertragsabschluss": 2020,
    " Jahresbeitrag“:
   }
  ]
}
```

Json 2

```
{
 "Kunden-Name": „Schmidt“,
 "Kunden-Vorname": „Monika“,
 "Kunden-ID": 123457,
 "Alter": „07.11.1980“,
 "Geschlecht": false,
 "Adressen": [
   {
    "PLZ": 1127,
    "Ort": „Dresden“,
    "Straße": „Leipziger Str. 11“,
   }
  ],
  " Verträge": [
   {
    "Vertragart": „Hauptversicherung“,
    "Vertragsabschluss": 2001,
    "Jahresbeitrag": 90.30
   }
  ]
}
```

a) [10 Punkte] Beschreiben Sie die notwendigen Schritte zur Verarbeitung der JSON-Dateien und die Datengrundlage für die Modellierung des Prognosemodells. Gehen Sie hierbei auf folgende Punkte ein:

   a. Benennen Sie die abhängige Variable und drei unabhängige Variablen.

   b. Benennen Sie Typ und Wertebereiche der abhängigen und unabhängigen Variablen.

   c. Benennen Sie drei Datenauffälligkeiten und jeweils eine Möglichkeit damit umzugehen.

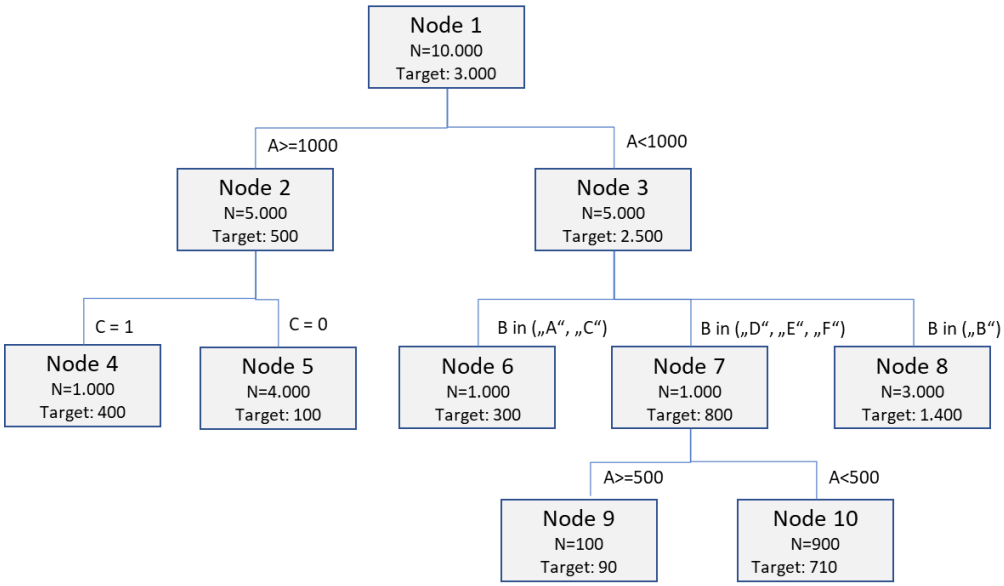
   d. Erläutern Sie kurz die notwendigen Datentransformationen zur Ermittlung der abhängigen Variablen.

Ein Kollege macht den Vorschlag, vor der Modellierung zur Datenreduzierung eine Dimensionsreduzierung durchzuführen.

b) [6 Punkte] Benennen Sie jeweils drei Vor- und Nachteile von einer Dimensionsreduzierung.

Zur Erstellung des Prognosemodells soll ein Entscheidungsbaum und eine logistische Regression erstellt werden und die Modelle verglichen werden.

In der nachfolgenden Grafik ist ein vereinfachter Entscheidungsbaum dargestellt (mit den erklärenden Variablen A, B und C). Hierbei beschreibt „N" die Anzahl der Datensätze und „Target" die Anzahl der Daten mit Abschluss einer Zusatzversicherung.



c) [11 Punkte] Bestimmen Sie für die nachfolgenden Datensätze die Prognosewerte und bewerten Sie die Prognosegüte mit einem geeigneten Gütemaß. Treffen Sie hierzu geeignete Annahmen.

Datensatz	Erklärende Variablen	Prognosewert	beobachtete Zielvariable
ID_1	A=100, B="A", C=1		1
ID_2	A=2000, B="E", C=1		0
ID_3	A=100, B="B", C=0		0
ID_4	A=50, B="F", C=0		1
ID_5	A=900, B="E", C=1		1
ID_6	A=1200, B="A", C=0		0

Nach der Erstellung der Prognosemodelle ergeben sich für 10 exemplarische Datensätze die folgenden Werte:

Datensatz	Prognostizierte Wahrscheinlichkeit		beobachtete Zielvariable
	logistische Regression	Entscheidungsbaum	
ID_1	0,57	0,57	0
ID_2	0,93	0,54	1
ID_3	0,26	0,18	0
ID_4	0,41	0,48	0
ID_5	0,6	0,56	0
ID_6	0,39	0,88	0
ID_7	0,84	0,64	1
ID_8	0,32	0,32	0
ID_9	0,74	0,91	1
ID_10	0,58	0,16	0

d) [15 Punkte] Erstellen Sie für die Datensätze einen Liftplot mit den Werten der beiden Prognosemodelle und interpretieren Sie das Ergebnis. Hinweis: Verwenden Sie die Informationen aus dem Entscheidungsbaum aus Teil c.)

(Anmerkung: Die Aufgabe soll die Funktionsweise von Liftplots abfragen / darstellen. Für die praktische Erstellung und Interpretation von Liftplots ist eine deutlich größere Datenmenge notwendig.)

Lösungsvorschlag:

a) Ausgehend von den beiden JSON-Dateien ergibt sich folgende Datengrundlage und die dazu notwendigen Verarbeitungsschritte:

- Abhängige Variable: Die abhängige Variable „Abschluss Zusatzversicherung im Vorjahr" enthält die Information, ob eine Zusatzversicherung im Vorjahr abgeschlossen wurde. Die Variable ist vom Typ Boolean mit den Werten 0 und 1. Zur Ermittlung der abhängigen Variable ist unter „Verträge" zu prüfen, ob ein Eintrag mit Vertragsart „Zusatzversicherung" und „Vertragsabschluss" 2021 (d.h. im Vorjahr) vorliegt. Ist dies der Fall, dann hat die abhängige Variable den Wert 1, ansonsten den Wert 0.
- Mögliche unabhängige Variable sind:
  - Variable „Alter" vom Typ Integer und dem Wertebereich 0 bis 200
  - Variable „Geschlecht" vom Typ Character mit den Werten „m", „w" oder „d"
  - Variable „Jahresbeitrag" vom Typ Decimal mit dem Wertebereich 0 bis 1.000.000
  - Variabel „Anzahl Verträge" vom Type Integer und dem Wertebereich 0 bis 20

  Hinweis: Die Benennung anderer Variablen, Datentypen und Wertebereiche sind möglich.
- Datenauffälligkeiten in der JSON-Files:
  - Im JSON 2 hat das Alter einen ungültigen Eintrag. Bei der Verarbeitung ist zu prüfen, ob unter Alter ein Datum gespeichert ist. Sollte ein Datum gespeichert sein, dann ist das Alter zu ermitteln und in der weiteren Datenverarbeitung zu verwenden.
  - Im JSON 2 hat das Geschlecht einen ungültigen Eintrag. Bei der Verarbeitung ist zu prüfen, ob das Geschlecht einen ungültigen Wert hat. Ist dies der Fall, dann sollte das JSON-File / der Datensatz ausgeschlossen werden.
  - Im JSON 1 fehlt der Wert zum Jahresbeitrag. Der fehlende Wert kann durch einen geeigneten Wert (z.B. durch einen Mittelwert) ersetzt werden.

  Weitere Varianten zum Umgang mit den Datenauffälligkeiten sind möglich.

b) Vorteile einer Dimensionsreduzierung:

- Identifikation und Fokussierung auf die wichtigsten Einflussfaktoren
- Reduzierung von redundanten Einflussfaktoren / Informationen
- Verringerung der notwendigen Zeit zum Training der Modelle
- Verringerung des benötigten Plattenspeichers

Nachteile einer Dimensionsreduzierung:

- Verlust der Interpretation der erklärenden Variablen
- Aufwand und Komplexität in der Datenaufbereitung (im Vorfeld der Modellierung)
- Möglicherweise Verlust von Informationen und hierdurch Verschlechterung der Modellperformance

Weitere Vor- und Nachteile sind möglich.

c) Aus dem dargestellten Entscheidungsbaum können in den Leaf-Knoten (durch Berechnung: Anzahl „Target" dividiert durch die Anzahl „N") die Prognosewerte ermittelt werden, wie folgend (ohne Nachkommastellen) dargestellt:

- Node 4: 40 % (= 400/1.000)
- Node 5: 3 % (= 100/4.000)
- Node 6: 30 % (= 300/1.000)
- Node 8: 47 % (= 1.400/3.000)
- Node 9: 90 % (= 90/ 100)
- Node 10: 79 % (=710/900)

Aus den Werten A, B und C können für die gegebenen Datensätze die Leaf-Knoten ermitteln werden. Hieraus ergeben sich folgende Prognosewerte:

Datensatz	Erlärende Variablen	**Prognosewert**	beobachtete Zielvariable
ID_1	A=100, B="A", C=1	30%	1
ID_2	A=2000, B="E", C=1	40%	0
ID_3	A=100, B="B", C=0	47%	0
ID_4	A=50, B="F", C=0	79%	1
ID_5	A=900, B="E", C=1	90%	1
ID_6	A=1200, B="A", C=0	3%	0

Ein mögliches Gütemaß ist die Misclassification Rate. Diese beschreibt den Anteil der falsch vorhergesagten Datensätze. Hierzu ist festzulegen, ab

welcher Wahrscheinlichkeit (=Cutoff) die Zielvariable den Wert 1 zugewiesen bekommt. Bei einem Cutoff von 50 % ergeben sich folgende Werte (andere Werte sind möglich):

Datensatz	Erlärende Variablen	**Prognosewert**	beobachtete Zielvariable	**Prognose Zielvariable**	Vorhersage korrekt
ID_1	A=100, B="A", C=1	30%	1	0	nein
ID_2	A=2000, B="E", C=1	40%	0	0	ja
ID_3	A=100, B="B", C=0	47%	0	0	ja
ID_4	A=50, B="F", C=0	79%	1	1	ja
ID_5	A=900, B="E", C=1	90%	1	1	ja
ID_6	A=1200, B="A", C=0	3%	0	0	ja

Die Misclassification-Rate ergibt sich hiermit zu 17 %.

Weitere Gütemaße sind möglich.

d) Die Erstellung des Liftplots erfolgt pro Prognosemodell in folgenden Schritten:

- Sortierung der Datensätze absteigend nach den prognostizierten Wahrscheinlichkeiten (Scorewerten):

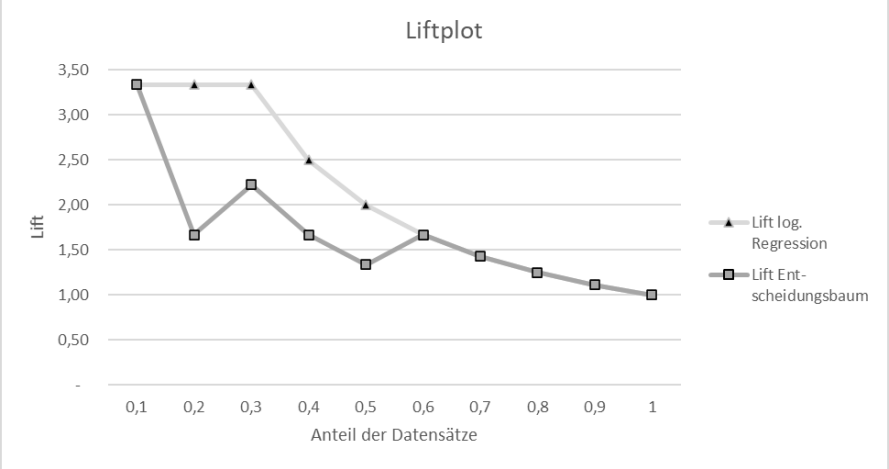
logistische Regression		Entscheidungsbaum	
Prognose	beobachtete Zielvariable	Prognose	beobachtete Zielvariable
0,93	1	0,91	1
0,84	1	0,88	0
0,74	1	0,64	1
0,6	0	0,57	0
0,58	0	0,56	0
0,57	0	0,54	1
0,41	0	0,48	0
0,39	0	0,32	0
0,32	0	0,18	0
0,26	0	0,16	0

- Ermittlung des Lifts pro Datensatz. Hierbei beschreibt der Lift den Faktor, um den ein Prognosemodell besser ist (zur Identifikation von Kunden zum Abschluss einer Zusatzversicherung) im Vergleich zur zufälligen Auswahl.

  Aus dem dargestellten Entscheidungsbaum ist ersichtlich, dass für die Grundgesamtheit eine Wahrscheinlichkeit von 30 % für den Abschluss einer Zusatzversicherung besteht . Hieraus ergeben sich folgende Lift-Werte:

	logitische Regression				Entscheidungsbaum			
Erwartete kummulierte Anzahl (Zufall)	Prognose	beobachtete Zielvariable	Erwartete kummulierte Anzahl (Prognose)	**Lift**	Prognose	beobachtete Zielvariable	Erwartete kummulierte Anzahl (Prognose)	**Lift**
0,30	0,93	1	1	**3,33**	0,91	1	1	**3,33**
0,60	0,84	1	2	**3,33**	0,88	0	1	**1,67**
0,90	0,74	1	3	**3,33**	0,64	1	2	**2,22**
1,20	0,60	0	3	**2,50**	0,57	0	2	**1,67**
1,50	0,58	0	3	**2,00**	0,56	0	2	**1,33**
1,80	0,57	0	3	**1,67**	0,54	1	3	**1,67**
2,10	0,41	0	3	**1,43**	0,48	0	3	**1,43**
2,40	0,39	0	3	**1,25**	0,32	0	3	**1,25**
2,70	0,32	0	3	**1,11**	0,18	0	3	**1,11**
3,00	0,26	0	3	**1,00**	0,16	0	3	**1,00**

- Die ermittelten Lift-Werte werden in einem Diagramm dargestellt:



Interpretation des Liftplots: Aus dem Liftplot ist ersichtlich, dass die logistische Regression für einen großen Teil der Datensätze einen höheren Lift-Wert hat im Vergleich zum Entscheidungsbaum (vor allem für die ersten drei Datensätze). Dies bedeutet, dass die logistische Regression besser geeignet ist zur Identifikation und Auswahl von Kunden, die eine hohe Wahrscheinlichkeit haben, eine Zusatzversicherung abzuschließen. Aus diesem Grund sollte die logistische Regression als Prognosemodell ausgewählt werden.

Aufgabe 4 [3.1 Regressions- und Clustermethoden] [40 Punkte]

a) [16 Punkte] Sie möchten eine lineare Regression durchführen. Sie nehmen dafür an, dass

$$Y = \beta_0 + \beta_1 \cdot X_1 = X \cdot \beta$$

mit $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$ gilt.

Sie beobachten folgenden Zusammenhang zwischen erklärender Variable X_1 und Zielvariable Y:

Erklärende Variable X_1	1	2	3	4
Zielvariable Y	3	2	3	6

Geben Sie die Designmatrix X an. (2 Punkte)

Schätzen Sie die Regressionskoeffizienten β_0 und β_1 mittels der Methode der kleinsten Quadrate. (8 Punkte)

Bestimmen Sie die Residuen $\hat{\varepsilon}$. (2 Punkte)

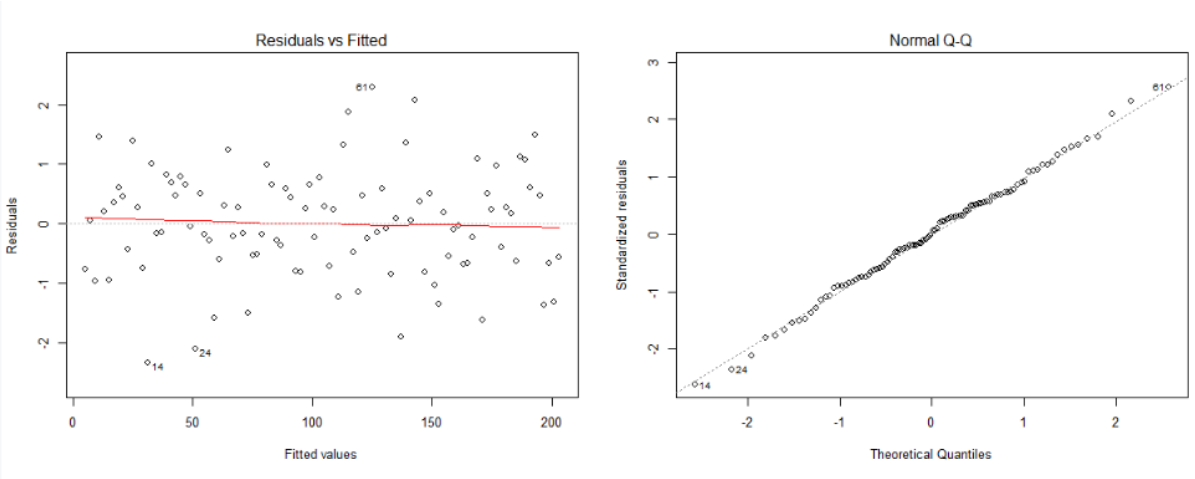
Bestimmen Sie die Standardfehler von $\widehat{\beta_0}$ und $\widehat{\beta_1}$. (4 Punkte)

Hinweis 1: Es gilt $\begin{pmatrix} 4 & 10 \\ 10 & 30 \end{pmatrix}^{-1} = \begin{pmatrix} 1{,}5 & -0{,}5 \\ -0{,}5 & 0{,}2 \end{pmatrix}$

Hinweis 2: Die Ergebnisse der Schätzer $\widehat{\beta_0}$ und $\widehat{\beta_1}$ sind für dieses Beispiel ganzzahlig.

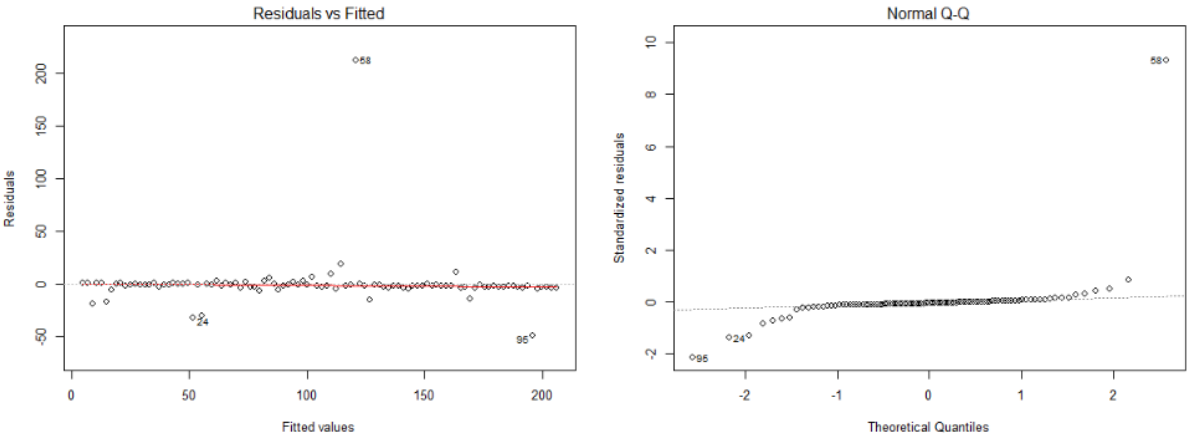
b) [4 Punkte] Sie haben einen Datensatz mit 100 Werten und führen mit R eine lineare Regression unter der Annahme, dass ein linearer Zusammenhang besteht und dass die Störterme unkorreliert und normalverteilt sind, durch.

Beurteilen Sie anhand der folgenden Plots, ob die Modellannahmen erfüllt sind. Schlagen Sie ggfs. einen sinnvollen Ansatz vor, um das Modell zu verbessern.



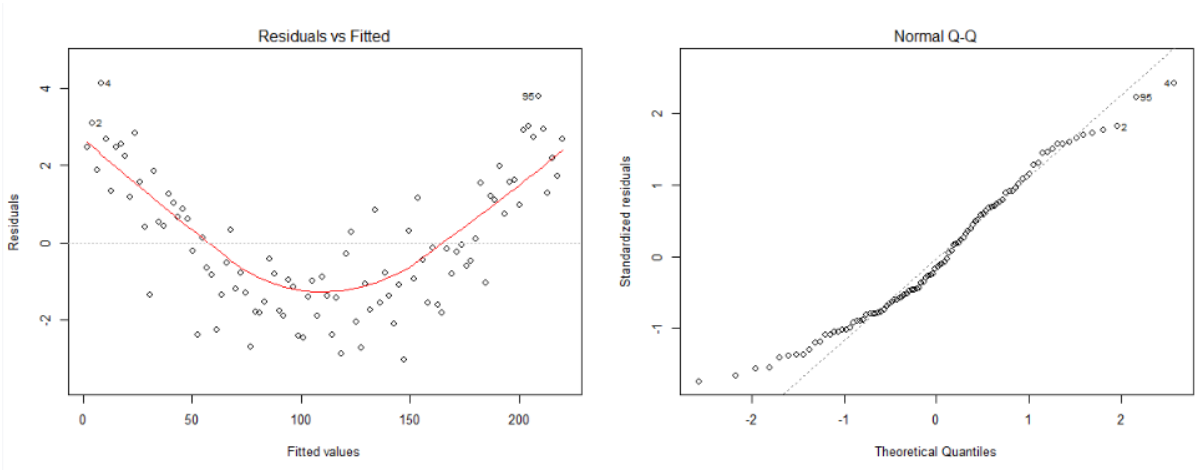
c) [4 Punkte] Sie haben einen zweiten Datensatz mit 100 Werten und führen mit R eine lineare Regression unter der Annahme, dass ein linearer Zusammenhang besteht und dass die Störterme unkorreliert und normalverteilt sind, durch.

Beurteilen Sie anhand der folgenden Plots, ob die Modellannahmen erfüllt sind. Schlagen Sie ggfs. einen sinnvollen Ansatz vor, um das Modell zu verbessern.



d) [4 Punkte] Sie haben einen dritten Datensatz mit 100 Werten und führen mit R eine lineare Regression unter der Annahme, dass ein linearer Zusammenhang besteht und dass die Störterme unkorreliert und normalverteilt sind, durch.

Beurteilen Sie anhand der folgenden Plots, ob die Modellannahmen erfüllt sind. Schlagen Sie ggfs. einen sinnvollen Ansatz vor, um das Modell zu verbessern.



e) [12 Punkte] Ihnen liegen die Ergebnisse von zwei Modellierungen aus R vor. Informationen zu Modell 1:

```
Call:
lm(formula = y ~ a + b + c + d + e, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-11.5632  -2.8029   0.1062   2.3034  11.0830

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.09762    1.46406  -0.067    0.947
a            5.01900    0.07336  68.419  < 2e-16 ***
b            0.87334    0.15106   5.782 2.92e-08 ***
c            2.03561    0.13690  14.870  < 2e-16 ***
d            0.68415    0.06133  11.156  < 2e-16 ***
e            0.06182    0.08573   0.721    0.472
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.007 on 194 degrees of freedom
Multiple R-squared:  0.9617,    Adjusted R-squared:  0.9607
F-statistic: 974.3 on 5 and 194 DF,  p-value: < 2.2e-16
```

Informationen zu Modell 2:

```
Call:
lm(formula = y ~ a + b + c + d, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-11.4769  -2.9301   0.2202   2.3182  11.5448

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02873    1.45174   0.020    0.984
a            5.02057    0.07323  68.555  < 2e-16 ***
b            0.86620    0.15054   5.754 3.34e-08 ***
c            2.03458    0.13672  14.881  < 2e-16 ***
d            0.68613    0.06119  11.213  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.002 on 195 degrees of freedom
Multiple R-squared:  0.9616,    Adjusted R-squared:  0.9608
F-statistic:  1221 on 4 and 195 DF,  p-value: < 2.2e-16
```

Berechnen Sie für beide Modelle sowohl AIC als auch BIC. Für welches Modell würden Sie sich demnach entscheiden? (Hinweis: Der Residual Standard Error berechnet sich wie folgt: $\sqrt{\frac{SS_{Res}}{df_{Res}}}$)

Lösungsvorschlag:

a)

Die Designmatrix hat die folgende Form:

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix}$$

Es gilt

$$X^t X = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix} = \begin{pmatrix} 4 & 10 \\ 10 & 30 \end{pmatrix}$$

Es gilt

$$\begin{pmatrix} 4 & 10 \\ 10 & 30 \end{pmatrix}^{-1} = \frac{1}{30 \cdot 4 - 10 \cdot 10} \cdot \begin{pmatrix} 30 & -10 \\ -10 & 4 \end{pmatrix} = \begin{pmatrix} 1{,}5 & -0{,}5 \\ -0{,}5 & 0{,}2 \end{pmatrix}$$

Damit gilt

$$(X^t X)^{-1} X^t = \begin{pmatrix} 1{,}5 & -0{,}5 \\ -0{,}5 & 0{,}2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 0{,}5 & 0 & -0{,}5 \\ -0{,}3 & -0{,}1 & 0{,}1 & 0{,}3 \end{pmatrix}$$

$$\hat{\beta} = (X^t X)^{-1} X^t y = \begin{pmatrix} 1 & 0{,}5 & 0 & -0{,}5 \\ -0{,}3 & -0{,}1 & 0{,}1 & 0{,}3 \end{pmatrix} \begin{pmatrix} 3 \\ 2 \\ 3 \\ 6 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Es gilt $\hat{y} = X\beta = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}$ und damit $\hat{\varepsilon} = \begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \end{pmatrix}$.

Es gilt

$$\hat{\sigma}^2 = \frac{1}{n-(p+1)} \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{1}{n-(p+1)} \cdot \sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2} = \frac{1}{2} \cdot 4 = 2$$

Damit gilt

$$se(\widehat{\beta_0}) = \sqrt{2} \cdot \sqrt{1{,}5} = \sqrt{3} \approx 1{,}7321$$

$$se(\widehat{\beta_1}) = \sqrt{2} \cdot \sqrt{0{,}2} = \sqrt{0{,}4} \approx 0{,}6325$$

b)

Der QQ-Plot zeigt, dass die beobachteten Residuen normalverteilt sind. Die Residuen scheinen nicht vom gefitteten Wert abzuhängen. Die Modellannahmen sind erfüllt. Anhand dieser beiden Plots ist keine notwendige Anpassung des Modells erkennbar.

c)

Der QQ-Plot zeigt, dass es einige Ausreißer gibt. Die Annahme, dass die Residuen normalverteilt sind, ist nicht erfüllt. Hier bietet sich z.B. eine logarithmische Transformation an.

Die Residuen scheinen nicht vom gefitteten Wert abzuhängen.

d)Der QQ-Plot zeigt, dass die beobachteten Residuen normalverteilt sind. Die Residuen scheinen aber vom gefitteten Wert abzuhängen. Es bietet sich an, einen zusätzlichen quadratischen Term zu betrachten.

e)

In beiden Fällen handelt es sich um einfache lineare Modelle mit n = 200 Beobachtungen.

Berechnung für Modell 1: Residual standard error $\sqrt{\frac{SS_{Res}}{df_{Res}}}$, und damit ergibt sich $SS_{Res} = 3114{,}88$ und mit $d = 7$ berechnet sich

$$AIC = n + n \cdot \ln(2\pi) + n \cdot \ln\left(\frac{SS_{Res}}{n}\right) + 2d = 1130$$

$$BIC = n + n \cdot \ln(2\pi) + n \cdot \ln\left(\frac{SS_{Res}}{n}\right) + \ln(n) \cdot d = 1153$$

Eine entsprechende Berechnung für Modell 2 ergibt

$$AIC = 1129, \qquad BIC = 1149$$

Demnach wird Modell 2 von beiden Kriterien bevorzugt.

Aufgabe 5 [4.1 Data Mining 1, 4.2 Analytics 1, 3.3 Datenvisualisierung] [35 Punkte]

a) [7,5 Punkte] Erläutern Sie zunächst den Unterschied zwischen den 3 wichtigsten Analysearten (deskriptiv, …) sowie zwischen den 2 wichtigsten Lernmethoden (überwacht, …).

b) [14 Punkte] In Ihrer Bestandsführung ist aufgefallen, dass es seit einiger Zeit zu vermehrtem Storno kommt. Daher ist die Idee entstanden, vor der Entwicklung von geeigneten Gegenmaßnahmen zur Unterstützung verschiedene Methoden aus Statistik und/oder Data Science heranzuziehen.

   Nennen Sie 4 einfache, aber möglichst verschiedene Ansätze hierfür und erläutern Sie diese ggf. mit jeweils einem Satz.

   Schlagen Sie für jeden der von Ihnen genannten Ansätze vor, mittels welcher/n Anwendungsklasse(n) (Klassifikation, Regression, etc.) Sie an das jeweilige Thema herangehen würden. Geben Sie an, zu welcher Analyseart und ggf. zu welcher Lernmethode im Sinne von Teilaufgabe a) Ihr Vorgehen gehört.

   Nutzen Sie für Ihre Antwort folgendes Schema:

      Ansatz 1: …

         Anwendungsklasse: …

         Analyseart: …

         Lernmethode: …

   Anm.: Es geht hier nicht um die Nennung konkreter Algorithmen.

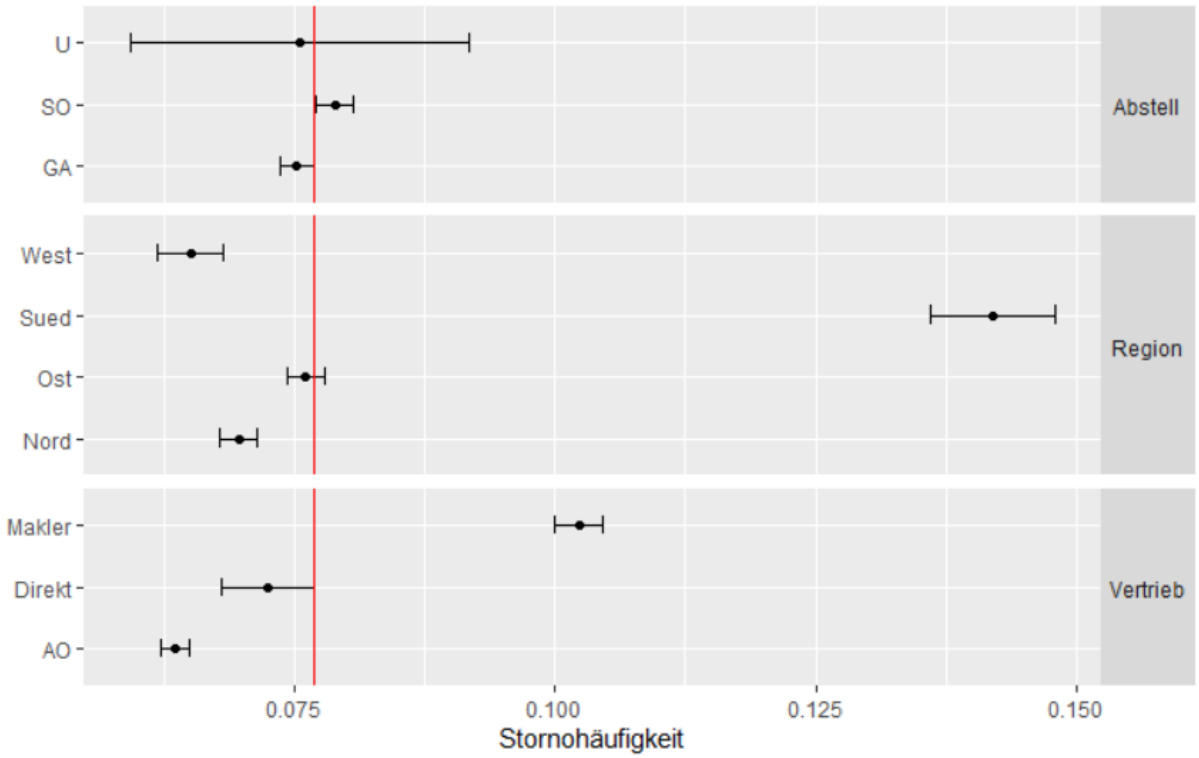
c) [7 Punkte] Für regionale Unterschiede in Ihren Daten haben Sie ein Merkmal *Region*, das die Ausprägungen Nord, Süd, Ost, West annehmen kann.

   Welchen Merkmalstyp hat dieses Merkmal? Wie würden Sie dieses Merkmal aufbereiten, wenn Sie Ihre Daten beispielsweise mittels logistischer Regression klassifizieren wollen?

   Zeigen Sie diese Aufbereitung an Beispiel-Daten mit mehreren Datensätzen (Beobachtungen).

   Wie nennt man und wozu dient ein solches Vorgehen?

d) [6,5 Punkte] Im Rahmen der Stornoanalyse wurde die folgende Grafik erstellt.



Was zeigt diese Grafik? Erläutern Sie die Grafik, wobei Sie plausible Interpretationen für unzureichend bezeichnete Grafikelemente vornehmen. Welche Informationen können Sie ablesen? Welche Auffälligkeiten erkennen Sie und welche Auswirkungen haben diese Erkenntnisse auf Ihre weiteren Analysen?

Lösungsvorschlag:

a) Deskriptive Analysen: Bei deskriptiven Analysen erfolgt eine Beschreibung und Interpretation von vorhandenen Daten. Hierdurch können Ergebnisse und Erkenntnisse aus der Vergangenheit gewonnen werden.

Prädiktive Analysen: Ausgehend von Daten aus der Vergangenheit werden durch prädiktive Analysen Vorhersagen über mögliche zukünftige Ergebnisse abgeleitet. Hierbei erfolgt eine Vorhersage, was in der Zukunft mit einer gewissen Wahrscheinlichkeit eintreten wird.

Präskriptive Analysen: Durch präskriptive Analysen wird der Effekt von unterschiedlichen Handlungen auf zukünftige Ereignisse quantifiziert. Hierdurch werden Empfehlungen für Entscheidungen und Handlungen erzeugt.

Überwachtes Lernen: Beim überwachten Lernen (Supervised Learning) wird der Lernalgorithmus an Beispielen mit bekanntem Ergebnis (Zielmerkmal) trainiert und dann bei neuen Daten zur Vorhersage verwendet.

Unüberwachtes Lernen: Beim unüberwachten Lernen (Unsupervised Learning) gibt es keine vorherzusagende Zielgröße. Das Ziel ist die Beschreibung von Zusammenhängen und Mustern zwischen den beschreibenden Merkmalen.

b) Folgende Ansätze sind (neben anderen) denkbar:

Ansatz 1: Quantifizierung des Stornoverhaltens der jüngeren Vergangenheit / Ermittlung der „Stornoquote"

Anwendungsklasse: deskriptive Statistik

Analyseart: deskriptiv

Lernmethode: - (kein maschinelles Lernen)

Ansatz 2: Vorhersage des Stornoverhaltens für die kommenden fünf Jahre

Anwendungsklasse: Regression / Zeitreihenanalyse

Analyseart: prädiktiv

Lernmethode: überwacht

Ansatz 3: Vorhersage des Stornoverhaltens einzelner Versicherungsnehmer

Anwendungsklasse: Klassifikation

Analyseart: prädiktiv

Lernmethode: überwacht

Ansatz 4: Klärung möglicher Gründe für das Storno
z.B. durch Mustererkennung hinsichtlich der Vertriebswege, Produktklassen, etc. oder durch Analyse von Korrespondenz-Häufigkeit und –Inhalten

Anwendungsklasse: Mustererkennung bzw. Textmining

Analyseart: deskriptiv

Lernmethode: unüberwacht

c) Bei *Region* handelt es sich um ein kategorielles Merkmal.

Für die logistische Regression wie für viele andere Machine-Learning-Verfahren müssen kategorielle Merkmale zunächst geeignet in numerische Merkmale transformiert werden. Ein verbreitetes Transformationsverfahren ist die sog. Dummy-Kodierung, auch One-Hot-Encoding genannt. Hierbei wird für jede Ausprägung *a* des kategoriellen Merkmals *M* ein neues binäres Merkmal *Ma* eingeführt. Dieses Merkmal *Ma* zeigt mit den Ausprägungen 1 bzw. 0 an, ob in dem jeweiligen Datensatz das Merkmal *M* die Ausprägung *a* hat oder nicht.

Das folgende Beispiel demonstriert dieses Vorgehen anhand eines Auszugs aus fiktiven Daten.

Originaldaten:

id	Region	…
1	Nord	…
2	West	…
3	Nord	…
4	Ost	…
5	Süd	…
6	West	…
…		

Transformierte Daten:

id	RegionNord	RegionOst	RegionSued	RegionWest	…
1	1	0	0	0	…
2	0	0	0	1	…
3	1	0	0	0	…
4	0	1	0	0	…

5	0	0	1	0	…
6	0	0	0	1	…
…					

d) Die Grafik zeigt die Stornohäufigkeit in Abhängigkeit von den Ausprägungen kategorieller Merkmale. Eingezeichnet ist als rote Linie die durchschnittliche Stornohäufigkeit insgesamt (also sämtlicher Beobachtungen). Für jede mögliche Ausprägung der verschiedenen Merkmale sind die durchschnittliche Stornohäufigkeit (schwarzer Punkt) sowie die Standardabweichung (waagerechte schwarze Linie) eingezeichnet.

Hierdurch lässt sich ablesen, welche Ausprägungen der einzelnen Merkmale mit über- oder unterdurchschnittlichen Stornohäufigkeiten korrelieren.

Folgende Auffälligkeiten sind hervorzuheben:

- Die *Region* Süd zeigt eine wesentlich höhere Stornohäufigkeit und sollte deshalb genauer untersucht werden.
- Ähnliches gilt, wenngleich nicht im selben Ausmaß, für den *Vertrieb* Makler.
- Das Merkmal *Abstell* lässt keinen Zusammenhang zur Stornohäufigkeit erkennen. Allerdings zeigt sich hier bei der Ausprägung U eine sehr große Standardabweichung.