DAV

DEUTSCHE
AKTUARVEREINIGUNG e.V.

Schriftliche Prüfung im Spezialwissen

Actuarial Data Science Basic

gemäß Prüfungsordnung 4
der Deutschen Aktuarvereinigung e. V.

am 12.06.2020

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 11 Seiten.
- Alle Antworten sind zu begründen und bei Rechenaufgaben muss der Lösungsweg ersichtlich sein.

Mitglieder der Prüfungskommission:

Axel Kiermaier, Dr. René Külheim, Dr. Stefan Nörtemann, Tobias Renner, Dr. Martin Seehafer, Mareike Welter

Aufgabe 1. *[*Modul *ADS Grundlagen & Umfeld] [30 Punkte]*

Sie sind Mitarbeiter eines mittelständischen Versicherungsunternehmens (VU). Dies ist Ihre erste Stelle nach dem Studium. Ihre Chefin spricht Sie auf das Thema Actuarial Data Science an.

(a) *[6 Punkte]* Erläutern Sie Ihrer Chefin drei zentrale Begriffe im Umfeld Data Science. Erklären Sie zudem die Bedeutung von Actuarial Data Science für ein VU.

(b) *[2 Punkte]* Wo liegt der Unterschied zwischen Actuarial Data Science und Data Science?

(c) *[2 Punkte]* Wo im VU könnte ein Actuarial Data Scientist arbeiten? Nennen Sie zwei Bereiche.

  (i) *[4 Punkte]* Begründen Sie, wieso der Einsatz eines Actuarial Data Scientist in diesen Bereichen sinnvoll ist.

  (ii) *[3 Punkte]* Welche Eigenschaften sollte ein Actuarial Data Scientist mitbringen? Nennen Sie mindestens drei Eigenschaften.

(d) *[6 Punkte]* Welche Datenquellen sind für einen Actuarial Data Scientist im aktuariellen Umfeld relevant? Nennen Sie mindestens drei Datenquellen. Nennen Sie zudem mindestens drei mögliche Anwendungen im aktuariellen Kontext.

(e) *[4 Punkte]* Erläutern Sie die Begriffe strukturierte, semistrukturierte und unstrukturierte Daten.

  (i) *[1 Punkt]* Welcher der genannten Begriffe ist Ihrer Ansicht nach im Versicherungsumfeld besonders relevant?

  (ii) *[2 Punkte]* Begründen Sie Ihre Wahl.

Lösungsvorschlag:

a.)

Data Engineering: Der Fokus liegt im Vergleich zu Data Science stärker auf der IT und weniger auf dem Fach-Knowhow (und auch weniger auf Mathematik). Dadurch kann der Data Engineer den Data Scientist bei den technisch „kniffligeren" Aufgabenstellungen unterstützen.

Data Analysis / Datenanalyse: Bedient sich einfacher Methoden von Data Science (z.B. nicht unbedingt Machine Learning / KI)

Data Mining: Der Prozess der Generierung von Erkenntnissen aus Daten

Big data:

Nennung der 6 Vs:
1. Volume Menge an zu verarbeitenden Daten
2. Velocity Geschwindigkeit, mit der die Daten anfallen
3. Variety unterschiedliche Datenformate und -strukturierung
4. Variability Dateninkonsistenz und unterschiedliche Aktualität
5. Veracity eingeschränkte Vertrauenswürdigkeit der Daten(herkunft)
6. Validity Datenqualität: Genauigkeit und Korrektheit

künstliche Intelligenz (k.I.) / Artificial Intelligence (AI): "is the study of how to make computers do things which, at the moment, people do better."

Digitalisierung:

Digitalisierung umschreibt den Prozess und die Auswirkungen der wachsenden Verbreitung digitaler Technik.

oder

Der Begriff Digitalisierung bezeichnet allgemein die Veränderungen von Prozessen, Objekten und Ereignissen, die bei einer zunehmenden Nutzung digitaler Geräte erfolgt.

Im ursprünglichen und engeren Sinne ist dies die Erstellung digitaler Repräsentationen von physischen Objekten, Ereignissen oder analogen Medien. *Im weiteren (und heute meist üblichen) Sinn steht der Begriff insgesamt für den Wandel hin zu digitalen Prozessen mittels Informations- und Kommunikationstechnik.*

… weitere Begriffe möglich

Actuarial Data Science wird für VU zunehmend an Bedeutung gewinnen. Hier werden die wissenschaftl. Methoden zum Umgang mit großen Datenmengen im aktuariellen Kontext betrieben. Diese werden immer wichtiger um im Zeitalter der Digitalisierung moderne und zeitgemäße Versicherungsprodukte anzubieten und um den Umgang mit den Daten im VU (das Gold unserer Zeit) zu optimieren.

oder

Lösung über die Tätigkeiten eines ADS:
- 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.)

data science und actuarial data science unterscheiden sich im Bereich aktuarielles / versicherungsspezifisches Fachwissen.

c.) (i)

Ein actuarial data scientist könnte z.B. in der Produktentwicklung arbeiten. Hier würde er Produkte entwickeln, deren Grundlage der Einsatz großer Datenmengen sind. Z.B. eine KFZ-Versicherung pay as you drive.

oder

Ein actuarial data scientist könnte z.B. in der DV-Koordination eines VUs arbeiten. Hier ist er in der Schnittstelle zwischen Fachbereich und IT angesiedelt und konzipiert technische Lösungen. Hier könnte er sein Fachwissen bei der Erstellung von technischen Lösungen einbringen.

oder

… weitere Einsatzmöglichkeiten möglich

c.) (ii)

- Übersetzer zwischen Fachbereich und IT
- 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, …)
- …

d.)

Datenquellen:
- Informationen aus den Bestandsführungssystemen, DWH-Systemen, …
- Informationen aus Wearables, Daten aus der Nutzung von Apps, Homepage, …
- Kundendaten (neben den klassischen Bestandsdaten auch aus der Korrespondenz mit dem Kunden), z.B. Infos über Familienstand und Familienmitglieder, …
- Wetterdaten
- Unfalldaten, Daten zu Einbrüchen, …
- Daten aus Unfallmeldesystemen beim Auto
- …

Anwendungen
- pay as you drive Tarife
- bessere Auswertung der eigenen Tarife z.B. Rückschlüsse zu Zusammenhängen aus der Risikoprüfung zu späteren Schadenfällen
- bessere Vorhersagen von Unwettern, Erdbeben, …
- …

weitere Lösungen möglich

e.)

strukturierte Daten
- haben ein vorgegebenes Format, in das sich alle Informationen einordnen lassen: normalisierte Form
- innerhalb einer relationalen Datenbank haben sie eine Zeilen- und Spaltenposition
- sind leicht zu finden und zu bearbeiten
- man kann sie z.B. mit SQL auslesen

semistrukturierte Daten
- haben keine allgemeine Struktur, tragen aber einen Teil der Strukturinformationen in sich
- sind nicht vollständig strukturiert und nicht vollständig unstrukturiert
- Beispiel: E-Mail: Grundstruktur mit Absender, Empfänger, Betreff und weitere Informationen des Nachrichtenkopfes. Der eigentliche Inhalt ist ein strukturloser Text.

unstrukturierte Daten:
- Informationen, die in einer nicht identifizierbaren und nicht normalisierten Datenstruktur vorliegen
- kein Datenmodell vorhanden
- schwer zu analysieren und zu verarbeiten
- Beispiele: Texte, Bilder, Audio- und Videodateien, Präsentationen

e.) (i) & (ii)

<u>mögliche Lösung:</u>

Besonders relevant sind unstrukturierte Daten

Die in einem VU während der normalen Geschäftstätigkeit anfallenden Daten sind in der Regel unstrukturiert. Es handelt sich dabei um große Datenmengen, von denen sich ein Teil in eine strukturierte Form bringen lässt. Ein Großteil bleibt jedoch nach wie vor unstrukturiert. Die Verwaltung, Speicherung und Verarbeitung der unstrukturierten Daten stellt die VU vor große Herausforderungen, da die herkömmlichen Verarbeitungsprogramme und Datenbanken hierfür nicht nutzbar sind. Hier können (actuarial) data scientisten einen wichtigen Beitrag leisten.

Aber auch denkbar:

Besonders relevant sind strukturierte Daten

In einem VU liegen in den Bestandsführungs- und Angebotssystemen große Datenmengen in strukturierter Form vor. Diese spielen im klassischen Versicherungsgeschäft bis heute eine zentrale Rolle.

Wichtig für die Punktevergabe ist also die Begründung unter (ii) und nicht eine eindeutige Antwort unter (i).

Aufgabe 2. *[*Modul *ADS Grundlagen & Umfeld*: Baustein *Datenschutz] [30 Punkte]*

Die (fiktive) international tätige *Modern Times Insurance Ltd.* (MTI) mit Sitz in Zürich (in der Schweiz) plant, ausschließlich für den deutschen Markt, die Einführung eines neuen innovativen Produktes in der Sparte „Leben".

Die Produktentwicklungsabteilung hat dazu einen Vorschlag für ein „pay-per-use-Produkt" erarbeitet. Nach der Registrierung über die Homepage der MTI kann der Kunde für sich über eine App beliebig viele jeweils 24 Stunden laufende Unfallversicherungen kaufen.

Die gesamte Abwicklung des Geschäfts, insbesondere Vertrieb und Bestandsverwaltung, soll über eine cloud-Lösung direkt in der Firmenzentrale in Zürich erfolgen.

(a) *[6 Punkte]* Erläutern Sie die Rechtssystematik hinsichtlich des Datenschutzes, die in diesem Fallbespiel anzuwenden ist und benennen Sie die Rechtsquellen, die für dieses Fallbeispiel einschlägig sind.

(b) *[6 Punkte]* Ist eine Verarbeitung personenbezogener Daten in diesem Fall erlaubt? Begründen Sie Ihre Antwort und belegen Sie sie mit einer Rechtsquelle.

(c) *[6 Punkte]* Erläutern Sie (kurz) die zentralen Datenschutzgrundsätze, die in diesem Fallbeispiel anzuwenden sind.

Nach erfolgreicher Einführung möchte das Aktuariat ein Profittesting durchführen, um zu prüfen, ob das neue Angebot profitabel ist. Für die Analyse sollen die erhobenen personenbezogenen Daten der Kunden verwendet werden.

(d) *[6 Punkte]* Nehmen Sie Stellung zur Rechtmäßigkeit des Vorhabens mit Blick auf den Datenschutz.

Da das Angebot im deutschen Markt sehr erfolgreich ist und bereits im ersten Jahr weit mehr Kunden gewonnen werden konnten als erwartet, möchte das Unternehmen Maßnahmen zur Betrugserkennung ergreifen. Sie beauftragt daher die auf Betrugserkennung spezialisierte *Artificial Intelligence Agency* AG (AIA) in Köln mit der Analyse des Kundendatenbestands.

(e) *[6 Punkte]* Nehmen Sie Stellung zur Rechtmäßigkeit der Beauftragung mit Blick auf den Datenschutz. Begründen Sie Ihre Position und formulieren Sie eine kurze Empfehlung dazu, was bei der Beauftragung auf beiden Seiten zu beachten ist.

Lösungsvorschlag:

a.) Die Schweiz ist nicht Mitglied der EU und in der Schweiz gilt die Datenschutzgrundverordnung DSGVO nicht. Dennoch unterliegt die Datenverarbeitung (obwohl in der Schweiz durchgeführt) der DSGVO, da hier Daten von Bürgern der EU verarbeitet werden. („Marktortprinzip"!) Damit ist auch das Bundesdatenschutzgesetz zu beachten. Daneben sind die Datenschutzbestimmungen der Schweiz zu beachten. Ob zudem der Code-of-Conduct Datenschutz (CoC) des GDV zu beachten ist, hängt davon ab, ob die MIT dem CoC beigetreten ist.

b.) Gemäß Artikel 6 DSGVO ist die Verarbeitung personenbezogener Daten, die zur Erfüllung eines Vertrages oder zur Durchführung vorvertraglichen Maßnahmen erforderlich sind, erlaubt. Daneben kann die MIT die explizite Einwilligung der Verarbeitung von der betroffenen Person erheben.

c.) In diesem Fallbeispiel sind die sieben Grundsätze für die Verarbeitung personenbezogener Daten gemäß Artikel 5 DSGVO zu beachten. Im Einzelnen: „Rechtmäßigkeit, Verarbeitung nach Treu und Glauben", „Transparenz", „Zweckbindung", „Datenminimierung", „Richtigkeit", „Speicherbegrenzung" sowie „Integrität und Vertraulichkeit".

d.) Gemäß Artikel 6 f DSGVO ist die Verarbeitung personenbezogener Daten auch dann, wenn die Verarbeitung ist zur Wahrung der berechtigten Interessen des Verantwortlichen oder eines Dritten erforderlich ist. Die Verarbeitung der Daten für eine interne Analyse wie das Profittesting kann man als berechtigtes Interesse der MTI ansehen. Zum Schutz der betroffenen Personen sind geeignete Maßnahmen, wie Anonymisierung oder Pseudonymisierung der Daten in Betracht zu ziehen.

e.) Zunächst ist zu klären, ob eine Verarbeitung personenbezogener Daten in diesem Fall überhaupt erlaubt ist. Dies kann man jedoch – wie in d.) – mit den berechtigten Interessen des Verarbeiters (der MTI) begründen.

Die Beauftragung der AIA zur Analyse des Datenbestands der MTI mit Blick auf die Betrugserkennung ist eine Auftragsverarbeitung im Sinne des Artikels 28 DSGVO.

Gemäß Artikel 28 (1) muss die MTI vorab prüfen und sicherstellen, dass die AIA hinreichend Garantien dafür bietet, dass geeignete technische und organisatorische Maßnahmen so durchgeführt werden, dass die Verarbeitung im Einklang mit den Anforderungen der DSGVO erfolgt und den Schutz der Rechte der betroffenen Person gewährleistet. Die Verarbeitung durch einen Auftragsverarbeiter erfolgt auf der Grundlage eines Vertrags oder eines anderen Rechtsinstruments nach dem Unionsrecht oder dem Recht der Mitgliedstaaten.

Die Verantwortung für die ordnungsgemäße Verarbeitung der Daten kann nicht auf den Auftragsverarbeiter übertragen werden.

Aufgabe 3. *[*Module *Informationstechnologie* & *Insurance Analytics] [30 Punkte]*

Zur Steigerung des Bestandswachstums in der Sparte „Hausrat" in Ihrer Versicherung sollen durch Data Science Verfahren die Vertriebsaktivitäten optimiert werden. Hierzu sollen Bestandskunden mit einem hohen Cross-Selling Potenzial identifiziert werden.

(a) [*Datengrund und Bestandsverwaltung*] Grundlage der Analyse sind die Daten aus dem Bestandsverwaltungssystem in Ihrer Versicherung, in dem zu allen Kunden die Vertragsinformationen abgelegt sind.

(i) *[3 Punkte]* In dem Bestandsverwaltungssystem werden die Daten zu den „Produkten" und „Verträgen" abgelegt. Erläutern Sie den Begriff „Vertrag" und grenzen Sie diesen von dem Begriff „Produkt" ab.

(ii) *[3 Punkte]* Für die Analyse sind noch weitere Daten notwendig. Benennen sie ein weiteres Anwendungssystem, das als Datenquelle für die Analyse verwendet werden kann. Begründen Sie ihre Antwort.

(b) [*Bestandsverwaltung und relationale Datenbanken*] Das unvollständige Entity-Relationship-Diagramm, welches die Grundlage der relationalen Datenbank des Bestandsverwaltungssystems ist, ist folgend dargestellt:

(i) *[3 Punkte]* Erklären Sie das Konzept von „Primary Key".

(ii) *[2 Punkte]* Vervollständigen Sie das Diagramm mit einer sinnvollen Kennzeichnung der Attribute mit „PK" für „Primary Key" und „FK" für „Foreign Key".

(iii) *[3 Punkte]* Vervollständigen Sie das Diagramm um die Relation zwischen den Entitäten „Vertrag" und „Kunde". Hierbei soll folgende Beziehung bestehen: „Ein Kunde kann mehrere Verträge besitzen, jedoch mindestens einen Vertrag und ein Vertrag gehört genau zu einem Kunden".

(iv) *[4 Punkte]* Zur Datenaufbereitung sollen die Daten per SQL aus der Datenbank ermittelt werden. Erstellen Sie ein SQL-Statement zur Abfrage und Verknüpfung der Daten aus den Entitäten „Vertrag" und „Kunde".

Hinweis: Für die Erstellung sollen folgende Einschränkungen gelten:

- Auswahl der Attribute „Vertrags_NR", „Vertragsbeginn" und „Vertragsende" aus der Entität „Vertrag"
- Auswahl der Attribute „Kunden_ID" und „Name" aus der Entität „Kunde"
- Selektion der Kunden aus dem Land „Deutschland"

(c) [*Daten und Variablen*] Nach Bereitstellung und Verknüpfung der Daten aus den verschiedenen Anwendungssystemen wurde ein Datenabzug erzeugt, in dem alle Kunden enthalten sind, bei denen in der Vergangenheit eine Vertriebsaktivität für Hausratsversicherungen erfolgte. Folgend ist ein Ausschnitt aus den Daten dargestellt:

	Kunden-ID	Name	Alter	Geschlecht	Berufs-schlüssel	VIP-Kunde	KFZ-Vertrag vorhanden	Wohngebäude-Vertrag vorhanden	Lebens-versicherung vorhanden	Hausrat Vertriebs-aktivität	Abschluss Hausrat nach Vertriebsaktivität
	2631001	Maier	20	m	12012	1	1	1	0	1	0
	2631002	Schmidt	-	w	2150	0	0	0	0	1	0
	2631003	Müller	55	w	15501	0	1	0	1	1	1
	2631005	Maier	77	w	2150	0	0	0	1	1	1
	2631006	Schneider	45	m	2108	0	0	0	0	1	0
	2631007	Fischer	-	m	2182	1	0	0	0	1	0
	2631008	Weber	48	w	2150	1	0	0	0	1	0
	...										
Datentyp											

(i) *[4 Punkte]* Vergeben Sie für die dargestellten Variablen sinnvolle Datentypen.

(ii) *[2 Punkte]* Ist die Variable „Berufsschlüssel" ordinal oder nominal skaliert? Begründen Sie Ihre Antwort.

(iii) *[2 Punkte]* Welche der dargestellten Variablen ist die abhängige Variable?

(iv) *[4 Punkte]* Die Variable „Alter" enthält fehlende Werte („missing values"). Benennen Sie eine Möglichkeit zur Bereinigung der Daten und berechnen Sie für die dargestellten (sieben) Datensätze die fehlenden Werte der Variable „Alter".

Lösungsvorschlag:

a.)

(i) Ein Vertrag ist eine Vereinbarung zwischen dem Versicherungsunternehmen und dem Versicherungsnehmer zur Dokumentation des Versicherungsschutzes. In einem „Vertrag" werden die Eigenschaften und beschreibenden Attribute, wie z.B. der Vertragsbeginn und die Vertragslaufzeit, zu einem Versicherungsvertrag dokumentiert. Der Versicherungsnehmer erhält den „Vertrag" als Dokument. Ein „Vertrag" kann ein oder mehrere „Produkte" beinhalten, in dem der Umfang des Versicherungsschutzes festgelegt ist.

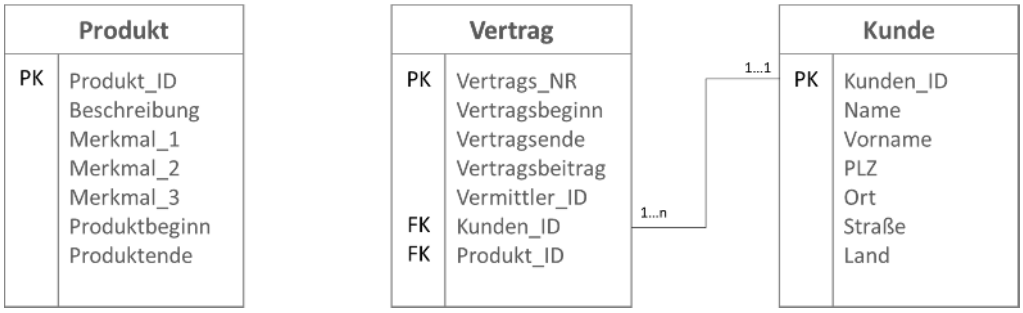
(ii) Zur Anwendung von Data Science Verfahren zur Optimierung der Vertriebsaktivitäten sind Informationen zu den Akquiseaktivitäten der Vergangenheit erforderlich. Diese Informationen sind (in der Regel) in den Vertriebssystemen der Versicherung abgelegt. Daher sind die Vertriebssysteme, neben dem Bestandsverwaltungssystem, relevante Anwendungssysteme und somit relevante Datenquellen für die Analyse.

b.)

(i) Der „Primary Key" ist eine eindeutige Kennzeichnung jedes Datensatz in einer Entität und besteht aus einem oder mehreren Attributen der Entität. Durch den „Primary Key" wird jeder Eintrag in der Entität / Tabelle eindeutig identifiziert.

(ii) Die Kennzeichnung der „Primary Keys" (PK) und „Foreign Keys" (FK) ist in dem nachfolgenden Entity-Relationshhip-Diagram dargestellt.

(iii) Die Relation zwischen den Enitäten „Vertrag" und „Kunde" ist in dem nachfolgenden Entity-Relationshhip-Diagram dargestellt:



(iv) Das SQL-Statement ist folgend dargestellt:

```
SELECT V.Vertrags_NR, V. Vertragsbeginn, V.Vertragsende,
        K.Kunde_ID, K.Name
FROM Vertrag V INNER JOIN Kunde K
ON V.Kunden_ID = K.Kunden_ID
WHERE K.Land = `Deutschland`
```

c.)

(i) In der folgenden Tabelle sind die Datentypen der Variablen dargestellt:

	Kunden-ID	Name	Alter	Geschlecht	Berufs-schlüssel	VIP-Kunde	KFZ-Vertrag vorhanden	Wohngebäude-Vertrag vorhanden	Lebens-versicherung vorhanden	Hausrat Vertriebs-aktivität	Abschluss Hausrat nach Vertriebsaktivität
Datentyp	**Integer**	**Character**	**Integer**	**Character**	**Integer**	**Boolean**	**Boolean**	**Boolean**	**Boolean**	**Boolean**	**Boolean**

(ii) Die Variable „Berufsschlüssel" ist nominal skaliert, da zwischen den Ausprägungen keine Anordnung existiert.

(iii) Ziel der Analyse ist es, den Abschluss der Hausratversicherung in Abhängigkeit von unabhängigen Variablen vorherzusagen. In den Daten enthält die Variable „Abschluss Hausrat nach Vertriebsaktivität" diese Information und ist somit die Zielvariable für die Modellerstellung.

(iv) Eine Möglichkeit zur Bereinigung der Daten ist es, die fehlende Werte durch den Mittelwert zu ersetzen. Hierbei wird der Mittelwert über alle Datensätze der Variable ermittelt, bei denen ein Wert vorhanden ist. Für den dargestellten Datensatz ergibt sich folgender Mittelwert: (20 + 55 + 77 + 45 + 48) / 5

= 49. Für die Datensätze ohne Angabe eines Wert in der Variable „Alter" werden die fehlenden Werte durch 49 ersetzt.

Aufgabe 4 [Module *Insurance Analytics* & *Mathematik*] *[35 Punkte]*

(a) [*Selektion von informativen Variablen*] Im Rahmen einer Modellerstellung sollen informative Variablen ausgewählt werden. In Abhängigkeit des Informationsgewinns soll die Variable „Geschlecht" oder „VIP-Kunde" selektiert werden. Für die Analyse sind folgende Werte gegeben und bekannt:

- In der Gesamtmenge enthält die Zielvariable 50% positive und 50 % negative Ausprägungen
- Die Entropie der Gesamtmenge ist 1,0
- Der Informationsgewinn der Variable „VIP-Kunde" ist 0,1
- 60 % der Kunden sind weiblich und 40 % der Kunden sind männlich
- Bei weiblichen Kunden ist die Zielvariable in 2/3 der Fälle positiv und in 1/3 der Fälle negativ
- Bei männlichen Kunden ist die Zielvariable in 1/4 der Fälle positiv und in 3/4 der Fälle negativ

Hinweis: Es gilt:

$$\log_2\left(\frac{2}{3}\right) = -0{,}5850;\ \log_2\left(\frac{1}{3}\right) = -1{,}5850;\ \log_2\left(\frac{1}{4}\right) = -0{,}4150;\ \log_2\left(\frac{3}{4}\right) = -2$$

(i) *[6 Punkte]* Berechnen Sie die Entropie für die Ausprägungen „Geschlecht = weiblich" und „Geschlecht = männlich".

(ii) *[5 Punkte]* Berechnen Sie den Informationsgewinn der Variable „Geschlecht".

(iii) *[10 Punkte]* Der Informationsgewinn beim Split eines Knotens *v* in einem Entscheidungsbaum mit insgesamt *N* zugeordneten Datensätzen durch die Aufteilung nach den *k* Attributwerten eines bestimmten Merkmals wird definiert durch

$$\Delta = I(v) - \sum_{j=1}^{k} \frac{N_j}{N} I(v_j),$$

Zeigen Sie, dass $\Delta \geq 0$.

Bezeichnungen:

I(.) die Entropie,
v_j die neu erzeugten Unterknoten und

N_j deren Datensatzanzahl

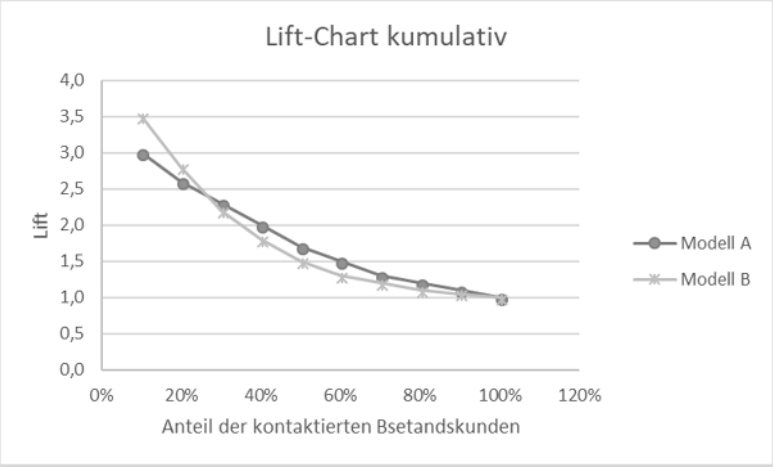
Hinweis: Die Logarithmusfunktion ist konkav, d.h. für $x_1, ..., x_n > 0$ und $w_1, ..., w_n$ mit $0 \leq w_j \leq 1$ und $w_1+...+w_n = 1$ gilt

$$\sum_{j=1}^{n} w_j log_2(x_j) \leq log_2\left(\sum_{j=1}^{n} w_j x_j\right).$$

(iv) *[2 Punkte]* Wählen Sie eine der beiden Variablen „Geschlecht" oder „VIP-Kunde" aus, die im Rahmen der Modellerstellung selektiert werden soll. Begründen Sie Ihre Antwort.

(v) *[3 Punkte]* In einer weiteren Analyse soll der Informationsgewinn bezüglich einer kontinuierlichen Variablen berechnet werden. Begründen Sie, warum die Berechnung des Informationsgewinns unter der Verwendung der Entropie nicht möglich ist.

(b) [*Model-Fitting und Modellauswahl*] Unter Anwendung von zwei Modellierungsmethoden wurden zwei Prognosemodelle erstellt. Das kumulative Lift-Chart der beiden Modelle ist in der folgenden Grafik dargestellt:



(i) *[3 Punkte]* Das Lift-Chart wurde unter Verwendung der Testdaten erzeugt. Erläutern Sie, warum das Lift-Chart nicht auf den Trainingsdaten, sondern auf den Testdaten erzeugt wurde.

(ii) *[3 Punkte]* Für zukünftige Vertriebsaktivitäten sollen 50 % der Bestandskunden kontaktiert werden, bei denen noch keine Vertriebsaktivitäten stattgefunden hat. Welches der beiden Modelle sollte für die Prognose verwenden werden? Begründen Sie Ihre Antwort.

(iii) *[3 Punkte]* Das Modell B hat einen Lift von 3,5 bei 10 % der kontaktierten Bestandskunden. Beschreiben Sie die Berechnung des Lift-Werts.

Lösungsvorschlag:

a.)
(i) Die Berechnung der Entropie erfolgt wie folgt:

$$\begin{aligned} Entropie\ (Geschecht = weiblich) &= -\frac{2}{3} * \log_2\left(\frac{2}{3}\right) - \frac{1}{3} * \log_{0,2}\left(\frac{1}{3}\right) \\ &= \frac{2}{3} * 0{,}5850 + -\frac{1}{3} * 1{,}5850 \\ &= 0{,}9183 \end{aligned}$$

$Entropie\ (Geschecht = männlich) = -\frac{1}{4} * \log_2\left(\frac{1}{4}\right) - \frac{3}{4} * \log_{0,2}\left(\frac{3}{4}\right) = \frac{1}{4} * 0{,}4150 + -\frac{3}{4} * 2 = 0{,}8113$

(ii) Die Berechnung des Informationsgewinns erfolgt wie folgt:

$$\begin{aligned} &Informationsgewinn\ (Geschlecht) \\ &\quad = Entropie\ (Gesamtmenge) - (0{,}6 \\ &\quad * Entropie(Geschlecht = weiblich) + 0{,}4 * Entropie(Geschlecht = männlich)) \\ &\quad = 1{,}0 - (0{,}6 * 0{,}9183 + \ 0{,}4 * 0{,}8113) \\ &\quad = 1{,}0 - \ 0{,}8755 \\ &\quad = 0{,}1245 \end{aligned}$$

Der Informationsgewinn der Variable „Geschlecht" beträgt 0,1245.

Musterlösung Aufgabe 4.) a.) (iii)

Wir betrachten

$$\Delta = I(v) - \sum_{j=1}^{k} \frac{N_j}{N} I(v_j) =: (I) - (II)$$

wobei I die Entropie, k die Anzahl der Attributwerte und N_j die Anzahl der Elemente im j-ten Unterknoten bezeichne. Nach Definition ist

$$I(v) = -\sum_{i \in C} p(i|v) \log_2 p(i|v)$$

wobei C die Menge der Klassen und $p(i|v)$ den Anteil der Datenpunkte mit Klasse i am Knoten v bezeichne. Hierbei sei $0 * log_2 0 = 0$.

Nun gilt

$$\begin{aligned}
(II) &= \sum_{j=1}^{k} \frac{N_j}{N} I(v_j) \\
&= -\sum_{j=1}^{k} \frac{N_j}{N} \sum_{i \in C} p(i|v_j) \log_2 p(i|v_j) \\
&= \sum_{i \in C} \sum_{j=1}^{k} \frac{N_j}{N} p(i|v_j) \left(-\log_2 p(i|v_j)\right) \\
&= \sum_{i \in C} p(i|v_j) \sum_{j=1}^{k} \frac{N_j p(i|v_j)}{N p(i|v)} \left(\log_2 \frac{1}{p(i|v_j)}\right)
\end{aligned}$$

Die Anzahl der in Klasse i fallenden Datensätze am Knoten v lässt sich schreiben als

$$N p(i|v) = \sum_{j=1}^{k} N_j p(i|v_j),$$

was gleichbedeutend ist mit

$$\sum_{j=1}^{k} \frac{N_j p(i|v_j)}{N p(i|v)} = 1.$$

Damit gilt offenbar auch

$$0 \leq \frac{N_j p(i|v_j)}{N p(i|v)} \leq 1$$

und mit der Konkavität des Logarithmus ergibt sich

$$\begin{aligned}
(II) &\leq \sum_{i \in C} p(i|v_j) \log_2 \left(\sum_{j=1}^{k} \frac{N_j p(i|v_j)}{N p(i|v)} \frac{1}{p(i|v_j)} \right) \\
&= \sum_{i \in C} p(i|v_j) \log_2 \left(\frac{1}{p(i|v)} \right) \\
&= I(v)
\end{aligned}$$

Es folgt $\Delta = I(v) - (II) \geq 0$.

1

(iv) Für die Modellerstellung soll die Variable „Geschlecht" verwendet werden, da der Informationsgewinn von 0,1245 höher ist als der Informationsgewinn der Variabel „VIP-Kunde" von 0,1. Durch die Selektion der Variable „Geschlecht" kann die Impurity (Unreinheit) in einem höheren Maß reduziert werden.

(v) Für die Berechnung des Informationsgewinns für kontinuierliche Zielvariablen ist die Verwendung der Entropie nicht geeignet. Anstelle der Entropie kann der Informationsgewinn unter Verwendung der Reduzierung der Varianz bestimmt werden.

b.)

(i) In Rahmen der Modellerstellung werden Training Daten verwendet, um das Modell und die Modellparameter zu fitten / zu bestimmen. Zur neutralen Bewertung der Modellqualität und somit der Bewertung, wie gut das Modell neue Daten vorhersagen kann, ist eine unabhängige Stickprobenmenge erforderlich. Die Test Daten sind zu den Trainings Daten eine unabhängige Datenmenge. Bei der Verwendung der Training Daten zur Modellbewertung besteht die Gefahr, dass ein Overfitting nicht erkannt wird.

(ii) Der Lift ist ein Maß zur Bewertung der Prognosegüte des Modells. Bei 50 % der Daten hat das Modell A einen höhen Lift-Wert und kann somit mit einer höheren Wahrscheinlichkeit der Zielvariable vorsagen. Das Modell A sollte daher für zukünftige Vertriebsaktivitäten verwendet werden. Dies gilt unter der Annahme, dass die Testdaten die gleiche Verteilung haben, wie die Verteilung auf dem das Modell angewandt wird.

(iii) Zur Berechnung des Lift-Werte werden die Testdaten (d.h. Bestandskunden) nach ihrer Abschlusswahrscheinlichkeit laut Modell sortiert. Für die 10% der Bestandskunden mit der höchsten Abschlusswahrscheinlichkeit laut Modell wird die beobachtete Abschlussrate p_{model} ermittelt. Der Wert wird in das Verhältnis zur Abschlussrate der Gesamtheit p_{random} gesetzt. Der Lift-Wert bei 10 % der kontaktieren Bestandskunden ergibt sich somit wie folgt:

$$Lift\_{10\%} = p\_{model} / p\_{random}$$

Aufgabe 5. [Module *Mathematik* und *Informationstechnologie*] *[25 Punkte]*

(a) Grundbegriffe:

(i) *[5 Punkte]* Grenzen Sie die Begriffe überwachtes und unüberwachtes Lernen voneinander ab. Nennen Sie beispielhaft zwei Methoden aus dem Bereich überwachtes Lernen und zwei Methoden aus dem Bereich unüberwachtes Lernen.

(ii) *[5 Punkte]* Beschreiben Sie die Funktionsweise der *k*-fachen Kreuzvalidierung. Gehen Sie dabei davon aus, dass die Response-Variable des Modells numerisch ist und die Anpassungsgüte mit dem MSE (Mean Squared Error) berechnet werden soll. Wie berechnet sich der Schätzer für den Testfehler basierend auf der *k*-fachen Kreuzvalidierung?

(b) Logistische Regression

(i) *[5 Punkte]* Für die Klassifikation einer binären Zielvariablen *y* (kodiert mit 0 und 1) wollen Sie die logistische Regression verwenden. Dafür steht Ihnen in einem Datensatz neben y_i die Beobachtungen x_{i1}, x_{i2} und x_{i3} zur Verfügung. Geben Sie für diesen Fall das allgemeine Modell der logistischen Regression an.

(ii) *[5 Punkte]* Angenommen, der Wert für x_{i1} in ihrem Modell aus a) erhöht sich auf $x_{i1} + 1$. Was gilt dann für das Verhältnis der Chancen? Welche Änderungen können sich ergeben?

(iii) *[5 Punkte]* In einem Datensatz sind die erklärenden Variablen X_1, X_2 und X_3 sowie die Zielvariable Y enthalten:

Beobachtung	X_1	X_2	X_3	Y
1	0	0	3	0
2	-1	1	0	1
3	0	2	0	0
4	1	1	0	0
5	1	0	1	1

Berechnen Sie für den Referenzpunkt $X_1 = 0$, $X_2 = 0$ und $X_3 = 0$ die Klassifizierung mittels k-nächster Nachbar mit k = 3 unter Verwendung der euklidischen Metrik.

Lösungsvorschlag:

a.)

(i) Bei dem überwachten Lernen stehen neben den Kovariaten x_i sogenannte Response-Variablen ("Labels") y_i zur Verfügung. Ziel ist es, ein Modell zu finden mit dem eine Vorhersage der y_i basierend auf den Informationen x_i möglichst akkurat möglich ist. Typische Vertreter sind Regressionsmethoden(z.B. die Generalized Linear Models (GLMs) oder Baumverfahren) aber auch Klassifikationsverfahren (Logistische Regression oder Support Vector Machines (SVMs))

Beim unüberwachten Lernen stehen keine Response-Variablen y_i zur Verfügung. Hier wird versucht, Beziehungen zwischen Variablen bzw. zwischen den Beobachtungen zu ermitteln. Typische Beispiele sind z.B. k-Means oder aber Principal Component Analysis (PCA).

(ii) Bei der k-fachen Kreuzvalidierung wird der Datensatz in k Gruppen unterteilt (mit ungefähr gleicher Anzahl). Die erste Gruppe stellt den Validierungsdatensatz V_1 dar. Die Bestimmung des Modells erfolgt auf den restlichen k-1 Gruppen. Nach erfolgter Anpassung wird MSE_1 auf dem Datensatz V_1 berechnet. Dieses Verfahren wird nun k-Mal wiederholt. In jedem Durchlauf wird eine andere Gruppe für den Validierungsdatensatz V_1 gewählt. Damit ergeben sich insgesamt *k* Schätzer für den Testfehler, MSE_1, …, MSE_k.

Daraus errechnet sich der Testfehler wie folgt:

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} MSE_i$$

b.)

(i) Das allgemeine Modell der logistischen Regression:

$$ln\left(\frac{\pi_i}{\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

wobei:

$$\pi_i = P(y_i = 1 \mid x_{i1}, x_{i2}, x_{i3})$$

(ii) Das Verhältnis der Chancen berechnet sich zu

$$\frac{P(y_i = 1 \mid x_{i1}, x_{i2}, x_{i3})}{P(y_i = 0 \mid x_{i1}, x_{i2}, x_{i3})} \Bigg/ \frac{P(y_i = 1 \mid x_{i1} + 1, x_{i2}, x_{i3})}{P(y_i = 0 | x_{i1} + 1, x_{i2}, x_{i3})} = \exp(\beta_1)$$

Damit ergeben sich die folgenden Szenarien für die Chance $\frac{P(y_i=1)}{P(y_i=0)}$:

$\beta_1 > 0$: Die Chance wird größer.

$\beta_1 < 0$: Die Chance wird kleiner.

$\beta_1 = 0$: Die Chance bleibt gleich.

(iii) Die Beobachtungen 2, 4 und 5 haben nach der euklidschen Metrik den geringsten Abstand zum Referenzpunkt $(\sqrt{2})$. Damit ergibt sich $P(Y = 1 \mid X = 0) = \frac{1}{3} * 2$, $P(Y = 0 \mid X = 0) = \frac{1}{3} * 1$. Damit wird der Referenzpunkt mit 1 klassifiziert.

Aufgabe 6. *[*Modul *Informationstechnologie] [30 Punkte]*

Der *SunnySide VVaG* ist ein Schaden/Unfallversicherer und möchte seinen Kunden neue digitale Services zur Unterstützung bei Urlaubsreisen anbieten. Es wurde bereits ein Gerüst und die benötigte Infrastruktur für eine Kunden-App entwickelt, die verschiedene Dienste anbieten können soll.

Eine derzeit diskutierte Idee besteht darin, dass die App dem Kunden auf Basis von Wetter- und Ortungsdaten Vorschläge hinsichtlich der Freizeitgestaltung am Urlaubsort macht.

(a) *[6 Punkte]* Führen Sie mindestens zwei Beispiele aus, wie sich diese Idee mit den Versicherungsprodukten und letztlich deren Verkauf verknüpfen ließe. Wodurch könnte dem Unternehmen hierdurch ein ökonomischer Vorteil entstehen?

Für die Wetterdaten evaluiert man u. a. die Services von *openweathermap.org*, die aktuelle Wetterdaten und Vorhersagen über eine Web-API anbieten. Für die Abfrage der aktuellen Wetterdaten über die API des Anbieters liegt folgendes Beispiel vor:

API-call: api.openweathermap.org/data/2.5/weather?lat=35&lon=139

Sample-response:


```
{"coord":{"lon":139,"lat":35},
"sys":{"country":"JP","sunrise":1369769524,"sun-
set":1369821049},
"weather":[{"id":804,"main":"clouds","description":"overcast
clouds","icon":"04n"}],
"main":{"temp":289.5,"humidity":89,"pres-
sure":1013,"temp_min":287.04,"temp_max":292.04},
"wind":{"speed":7.31,"deg":187.002},
"rain":{"3h":0},
"clouds":{"all":92},
"dt":1369824698,
"id":1851632,
"name":"Shuzenji",
"cod":200}
```


(b) *[4 Punkte]* Erklären Sie, welche Informationen in der Anfrage an den Anbieter übermittelt wurden. Kann man davon ausgehen, dass die benötigten Daten für die App auf einem Mobile verfügbar bzw. leicht zu ermitteln sind?

Welche Probleme sind bei der Datenerfassung auf dem Endgerät zu erwarten und wie kann man sie adressieren?

(c) *[6 Punkte]* Interpretieren Sie die Antwort des Servers:

(i) Welches Austauschformat wird benutzt? Nennen Sie drei Vorzüge des Formats.

(ii) Für welchen Ort in welchem Land gilt der Datensatz?

(iii) Was ist die aktuelle Temperatur? Welche Temperatureinheit vermuten Sie?

Zur schnellen Evaluierung parsen Sie den Datensatz in R. Sie haben den String in der Variable *openweatherResponse* gespeichert und folgendes Script entworfen:

```
parsed <- jsonlite::fromJSON(openweatherResponse)
toGMTDate <- function(arg) {
  as.POSIXct(as.numeric(arg), origin = '1970-01-01', tz = 'GMT')
}
print(toGMTDate(parsed["dt"]))
```

Die Ausgabe lautet:

```
[1] "2013-05-29 10:51:38 GMT"
```

(d) *[6 Punkte]* Das Datenfeld *„dt*" mit Wert 1369824698 repräsentiert den Zeitpunkt der Abfragenstellung und ist ein UNIX-UTC-Timestamp (entspricht der Anzahl der Sekunden seit dem 1.1.1970, 0:00 Uhr). Erklären Sie die einzelnen Schritte im Skript oben, um den Ausgabewert zu bestimmen.

Hinweis: Der Parameter tz=‘GMT‘ definiert die Zeitzone, in der das Ergebnis ausgegeben werden soll; es gilt GMT="Greenwich Mean Time"=UTC).

(e) *[3 Punkte]* Wie viele Stunden ist zum Abfragezeitpunkt der Sonnenaufgang am Ort des Geschehens her?

Sie wollen das Ergebnis jedes API-Calls in einer eigenen Datenbank abspeichern.

(f) *[3 Punkte]* Welche zusätzlichen Informationen würden Sie neben den API-Daten noch abspeichern, um ein sinnvolles Monitoring des Service zu ermöglichen und was ist dabei ggf. zu beachten?

(g) *[2 Punkte]* Nennen und erklären Sie einen weiteren Grund (neben der Evaluation des Service), warum eine Speicherung der API-Ergebnisse in einer eigenen Datenbank (z.B. finanziell) sinnvoll sein kann.

Lösungsvorschlag:

a.)

1. Absatz: Direkte Angebote zur Absicherung bei bestimmten Events, für die ggf. zusätzlicher Versicherungsschutz erworben werden kann, wie Risikosportarten.

2. Prävention/Schadenverhinderung: Warnungen vor bestimmten Aktivitäten wegen Unwetter/Lawinen (z.B. Berg-/Skitour) um Gefahren zu reduzieren

3. Schadenreduktion: Es könnten Hinweise wie „denken Sie an warme Kleidung/Schneeketten/genügend Getränke" gegeben werden, die ggf. Schadenaufwendungen reduzieren

Andere Lösungen möglich.

b.)

Die Parameter im API call sind (lat=35, lon=139). Hierbei handelt es sich um Koordinaten auf der Erdoberfläche. Diese Daten sollten normalerweise im Smartphone über das Betriebssystem abfragbar sein, falls der App entsprechende Berechtigungen eingeräumt wurden. Für den Fall, dass dies nicht funktioniert könnte der Nutzer aufgefordert werden, die Ortungsdienste zu aktivieren oder seinen Aufenthaltsort direkt anzugeben.]

c.)

(i) Das Austauschformat ist JavaScript-Object-Notation (JSON). Es ermöglicht eine kompakte, menschenlesbare Darstellung und ist leicht maschinell verarbeitbar.

(ii) Shuzenji in Japan.

(iii) "temp":289.5, die Angabe ist in Kelvin, das entspricht in etwa 16°C.]

d.)

Zunächst wird der Datensatz in eine R-Liste geparst. Das übernimmt die Funktion *fromJSON* aus dem package *jsonlite*.

Anschließend wird eine Funktion definiert, die einen UNIX-UTC-Timestamp in einen Datums-String transformiert. Die Aufgabe wird an die R-Funktion as.POSIXct delegiert, die aber ein numerisches Argument erwartet, welches durch die Transformation mit as.numeric erstellt wird.

Zuletzt wird diese Funktion auf dem Feld „dt" des Datensatzes aufgerufen und das Ergebnis an der Konsole ausgegeben.]

e.)

(1369824698 - 1369769524)/(60*60)=15.3h

f.)

Wichtig ist, welcher Kunde angefragt hat und welche Freizeitaktivitäten ihm letztlich empfohlen wurden. Es ist aber zu beachten, dass eine Einwilligung des Kunden zur Verarbeitung der Daten vorliegt und dass die Daten anonymisiert gespeichert werden.

g.)

Caching – bei Abfragen, die zeitlich/örtlich nahe bei bereits getätigten Abfragen liegen, kann das Ergebnis ggf. ohne expliziten API-Call direkt aus der Datenbank ausgeliefert werden – das spart Geld, da jeder API-Call beim Anbieter kostet und man kann auch approximative Antworten liefern, wenn der Openweathermap-Server nicht erreichbar ist.