DAV

DEUTSCHE
AKTUARVEREINIGUNG e.V.

Schriftliche Prüfung im Spezialwissen

Actuarial Data Science Advanced

gemäß Prüfungsordnung 4
der Deutschen Aktuarvereinigung e. V.

am 21.10.2022

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 30 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 [5.1 Gesellschaftliches Umfeld & Ethik 2, 5.2 Datenschutz] [36 Punkte]

Sie lesen im Zug auf dem Weg zum nächsten DAV-Seminar in einer Zeitschrift einen Bericht über Sprach-KI. Dabei lernen Sie folgendes:

Seit 2017 sind sogenannte Transformer-Modelle wie etwa *Bert* und *GPT-3* in der Entwicklung. Diese Modelle werden auf sehr großen Datenmengen (Gigabytebereich) aus dem Internet wie z.B. Wikipedia, Zeitungsartikeln oder Webseiten trainiert und sind in der Lage, natürliche Sprache zu verarbeiten. Diese Transformer-Modelle zeigen beeindruckende Fähigkeiten. So sind sie beispielsweise in der Lage, Texte zusammenzufassen, Fragen zu beantworten, Emotionsanalyse durchzuführen, zu chatten und Programmcode zu erstellen. Die neuesten Modelle können sogar Witze erklären.

Der Vorstand Ihres Versicherungsunternehmens hat entschieden, dass solche Transformer-Modelle nun in Ihrem Unternehmen eingesetzt werden sollen. Konkret soll das Modell *GPT-3* des Anbieters *OpenAI* eingesetzt werden, das auf einem Server in den USA läuft und über eine API kostenpflichtig aufgerufen wird. Das Modell hat 175 Milliarden Parameter und wurde mit 570 GB Daten trainiert. Die Modellarchitektur ist bekannt, aber Sie haben keinen Zugriff auf die konkreten Modellparameter.

Hinweis: Bei den Transformer-Modellen handelt es sich um eine spezielle Klasse von Neuronalen Netzen. Für die Beantwortung werden aber keine Kenntnisse zu Neuronalen Netzen oder Transformer vorausgesetzt, da diese erst in den Modulen Immersion und Completion behandelt werden.

(a) [8 Punkte] Nennen Sie vier sinnvolle Einsatzfelder für das Modell in Ihrem Unternehmen und die jeweiligen Chancen.

(b) [8 Punkte] Nennen Sie vier Gründe, warum der Einsatz dieses Modells aufgrund der oben genannten Informationen problematisch ist.

Da der Vorstand Sie beauftragt hat, sich in das Thema weiter einzuarbeiten, lesen Sie nun wissenschaftliche Artikel zu Transformer-Modellen. Dabei erfahren Sie u.a. folgendes:

Transformer-Modelle zeigen einige Schwachstellen. Sie reproduzieren beispielsweise Stereotypen über Geschlechterrollen, Religionen und ethnische Gruppen. Außerdem können sie Beleidigungen und unangemessene Sprache benutzen. Die Ursache dafür sind in erster Linie die Trainingsdaten. Sie enthalten ungefilterten Text aus dem Internet, in denen Stereotypen vorkommen. In den Trainingsdaten sind weiterhin z.B. auch Kontaktinformationen von Personen enthalten, sodass

Transformer-Modelle auch datenschutzrechtlich relevante Daten reproduzieren können.

Die ökologischen und finanziellen Kosten von solchen Modellen sind auch sehr hoch. So verbraucht das Training eines aktuellen Transformer-Modells Energie in der Größenordnung eines Langstreckenfluges und kostet mehrere Millionen US-Dollar.

(c) [10 Punkte] Der Vorstand bittet Sie nun, fünf mögliche Risiken, die sich aus Nutzung von solchen Modellen ergeben, zu beurteilen.

(d) [10 Punkte] Machen Sie fünf Vorschläge, wie einige der von Ihnen in Aufgabe a) und c) analysierten Risiken minimiert werden könnten.

Lösungsvorschlag:

a)

(i) Chatbots: Chance ist Kundenbetreuung zu jeder Zeit und ohne Beschränkung durch nicht genügend Personal.

(ii) Automatisierte Schadenabwicklung: Chance ist eine schnellere Abwicklung von Standard-/Massenschäden.

(iii) Beschleunigung von Entwicklungsprozessen durch automatische Erstellung von Programmcode: Chancen sind die Einsparung von Entwicklertätigkeit und die Beschleunigung von Softwareentwicklung.

(iv) Zusammenfassung von komplexen Dokumenten wie beispielsweise Versicherungsbedingungen für den Versicherungsnehmer: Chance wäre eine bessere Information für den VN

(v) Weitere Nennungen möglich.

b)

(i) Die Übermittlung von Daten, die der DSGVO unterliegen, in die USA ist rechtlich problematisch.

(ii) Das Modell ist eine Blackbox. Das Modell wurde nicht selbst erstellt und man hat auch keinen direkten Zugriff auf das Modell.

(iii) Bei 175 Milliarden Parametern ist das Modell als solches schwierig zu überprüfen.

(iv) Der Versicherer wäre darauf angewiesen, dass der Service langfristig angeboten wird und der Anbieter seine Marktmacht z.B. durch Preiserhöhungen nicht ausnutzt.

(v) Weitere Nennungen möglich.

c)

(i) Datenschutz: Es dürfen keine persönlichen Daten ohne Einwilligung gespeichert werden. Wenn ein Modell solche Daten reproduziert, stellt es eine Datenschutzverletzung dar, da die Daten im Modell gespeichert sind und die Daten ohne Einwilligung verarbeitet werden. Es gibt auch nicht die Möglichkeit, einzelne Daten aus dem Modell zu löschen.

(ii) Diskriminierung: Es darf keine Benachteiligung beispielsweise aufgrund der ethnischen Herkunft oder des Geschlechts geben.

(iii) Reputationsrisiko: Der Einsatz von Modellen, die sehr viel Ressourcen verbrauchen und zu Diskriminierung und Datenschutzverletzung neigen, stellen auch ein hohes Reputationsrisiko dar. Außerdem kann der Einsatz solcher großen Modelle, die nur schlecht überprüfbar sind, einen Vertrauensverlust darstellen. Ein Modell, dass zu Beleidigungen neigt, kann nicht für den Kundenkontakt benutzt werden.

(iv) Weitere Nennungen möglich.

d)

(i) Einsatz von Anbietern innerhalb der EU (keine Daten in den USA).

(ii) Einsatz von Modellen mit Einblick in Architektur und Modellparameter (kein Einsatz von Black-Box-Modellen).

(iii) Bereinigung von Trainingsdaten (Entfernung persönlicher Daten und Beleidigungen).

(iv) Einsatz von angemessenen Trainingsdaten (ohne Stereotypen).

(v) Entwicklung eines eigenen Modells mit mehreren Versicherungsunternehmen (Einsparung von Energie und Erstellung eines angemessenen Modells für Versicherungsunternehmen).

(vi) Weitere Nennungen möglich.

Aufgabe 2 [6.3 Cloud Computing, 5.2 Datenschutz, 7.1 Regressions- und Clustermethoden 2] [28 Punkte]

Sie arbeiten in der Krankenversicherungssparte eines Versicherungsunternehmens. Gemeinsam mit Ihren Kolleg*innen überlegen Sie, ob es möglich ist, aus Ihren Bestandsdaten ein Vorhersagemodell für die Möglichkeit einer Corona-Infektion zu erstellen.

Hierfür haben Sie in einem ersten Schritt einen Datensatz mit folgenden Daten, sowie der Klassifikation in der Vergangenheit mit Corona infiziert Ja / Nein, für jeden ihrer Versicherten erhoben (n = 525.123):

Tabelle 1: Spaltenbezeichnungen kompletter Datensatz

Versicherungsnummer
Geburtsdatum
Summe der Leistungsausgaben der letzten 2 Jahre
Diagnosen der letzten 2 Jahre
Geschlecht
Vorname
Nachname
Adresse
Geburtsort
Raucher / Nichtraucher

Anzahl eingereichter Anträge in den letzten 2 Jahren
Anzahl Kundenkontakte in den letzten 2 Jahren
Summe gezahlter Beiträge der letzten 2 Jahre
Vorerkrankung der Lunge (J/N)
Allergien
Mit Corona infiziert (J / N)

Nachdem Sie den Datensatz noch weiter aufbereitet haben, erhalten Sie einen Datensatz mit weit über 10 GB Dateigröße. Sie stellen schnell fest, dass Sie mit diesem Datensatz auf Ihrem PC aus Performance-Gründen keine Modellierung durchführen können. Nach Rücksprache mit Ihrem Chef meint dieser, Sie sollen die Berechnungen in der „Cloud" durchführen. Leider kennt sich Ihr Chef mit der „Cloud" nicht aus.

a) [8 Punkte] Erläutern Sie Ihrem Chef tabellarisch die verschiedenen Bereitstellungsmodelle in der Cloud und ihre Unterschiede (Cloud-Services), sowie die verschiedenen Cloud-Typen. Sprechen Sie bezogen auf dieses Fallbeispiel eine konkrete Empfehlung für einen Cloud-Service, sowie einen Cloud-Typ aus.

Um einen ersten Eindruck der Daten zu erhalten, nehmen Sie sich eine kleine Teilmenge des Datensatzes heraus. Sie betrachten für 10 Datensätze lediglich das Alter x geteilt durch das durchschnittliche Alter $\bar{x}$ im Bestand, die Summe der Leistungsausgaben L geteilt durch die durchschnittliche Summe der Leistungsausgaben $\bar{L}$ im Bestand, sowie das Merkmal Corona-Infektion $y \in \{1, -1\}$. Sie erhalten folgenden Datensatz:

Tabelle 2: Teilmengen Datensatz

Id	1	2	3	4	5	6	7	8	9	10
$\frac{x}{\bar{x}}$	0,5	3	1	2,5	2	2	0,25	1	1,75	3
$\frac{L}{\bar{L}}$	0,3	2,25	1	1,5	1,75	3	2,5	2	3,5	4,5
y	-1	-1	-1	-1	-1	1	1	1	1	1

b) [10 Punkte] Zeichnen Sie eine geeignete Visualisierung des Datensatzes aus Tabelle 2. Aus der Visualisierung sollte schnell und einfach die Corona-Klassifikation, sowie der Zusammenhang zwischen skaliertem Alter und skalierter Leistungssumme ersichtlich sein. Achten Sie hierbei auch auf die grundlegenden Anforderungen an eine Visualisierung (bspw. Verwendung eines Lineals, Achsenbeschriftung, Titel, …).

Aufgrund der positiven Tests auf dem kleinen Datensatz modellieren Sie Ihr Modell in der Cloud und erhalten auf Ihren Testdaten eine Prognosequalität von 99,5%. Ihr Chef ist begeistert und präsentiert Ihre Ergebnisse auf der nächsten internationalen Konferenz. Einige Kolleg*innen aus den USA sind sehr angetan von den Ergebnissen und bitten um den Source-Code, sowie die Rohdaten, um die Modellierung nachvollziehen und auf den amerikanischen Bestandsdaten wiederholen zu können.

c) [10 Punkte] Erläutern Sie Ihrem Chef, warum er den amerikanischen Kolleg*innen die Rohdaten nach der aktuell gültigen Rechtsprechung nicht einfach überstellen darf. Nehmen Sie hierbei explizit Bezug auf aktuelle Gerichtsurteile, sowie nicht mehr gültige Abkommen zwischen der EU und den USA.

Lösungsvorschlag:

a) Cloud-Modelle

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

Bereitstellungsmodelle

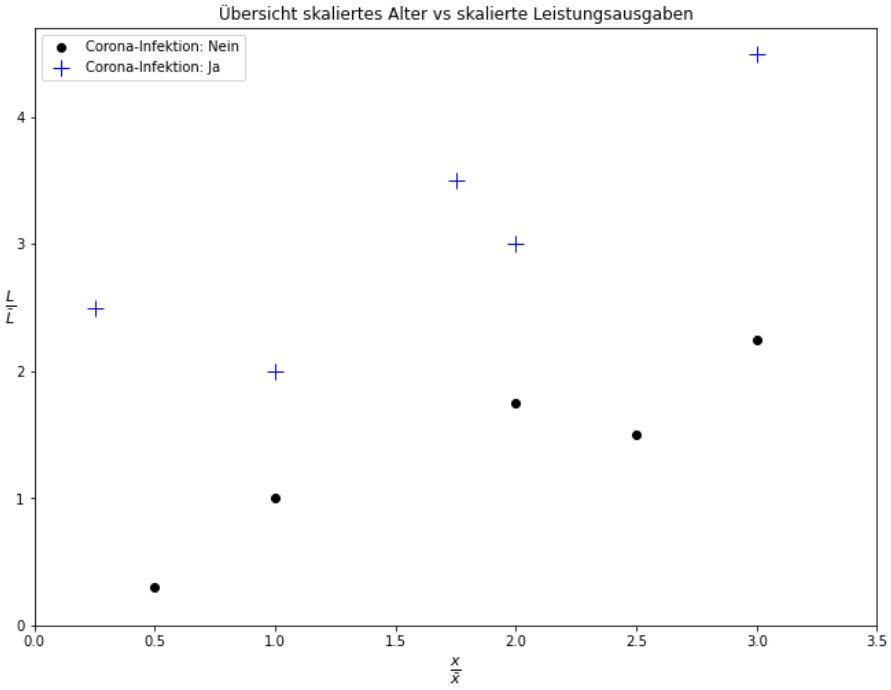
IaaS – Infracstructure as a Service	PaaS – Platform as a Service	Saas – Software as a Service
Infrastruktur Server, Netzwerk, Speicher	Infrastruktur Server, Netzwerk, Speicher	Infrastruktur Server, Netzwerk, Speicher
	Plattform OS, Entwickler-, Admin-Software	Plattform OS, Entwickler-, Admin-Software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

Im Prinzip ist jede Empfehlung richtig. Da es sich bei diesem Datensatz allerdings sehr deutlich um personenbezogene Daten inklusive Krankheitshistorie handelt, ist es unabdingbar sich Gedanken, um den Datenschutz zu machen. Ein Ablegen dieses Datensatzes in einer public cloud mit SaaS-Modell (bspw. bei kaggle.com) ist kritisch zu sehen. Selbst wenn dieser Datensatz dort als „privat" markiert wird. Diese Empfehlung wäre hier nicht angebracht. Liegt hingegen ein unternehmensindividueller Rahmenvertrag mit entsprechender Datenschutzvereinbarung vor, kann der Weg in eine public cloud (welche dann eher als private Wolke in der öffentlichen Wolke zu sehen ist), wie AWS, Azure etc., gewählt werden. Dann auch mit einem SaaS-Modell. Am sichersten wäre hier aber natürlich eine private cloud mit einem möglichst geringen Bereitstellungsmodell (allerdings muss dann auch sichergestellt sein, dass die firmeninterne IT über ausreichend Kenntnisse verfügt die eigenen Systeme abzusichern).

(7 Punkte, wenn alle Stichworte in den Tabellen genannt. Beispiele sind optional. 1 Punkt für eine Empfehlung.)

b) Die einfachste und wahrscheinlich auch beste Visualisierung ist ein ScatterPlot, wobei für die unterschiedliche Corona-Klassifikation unterschiedliche Farben, bzw. Symbole für die Punkte verwendet werden sollten.

Beispiel-Visualisierung:



(10 Punkte für die Visualisierung, Abzug für nicht erfüllte Basics.)

c) In der korrekten Antwort sollte Bezug genommen werden auf das **Safe-Harbor-Abkommen**, den **EU-US-Privacy-Shield**, die Schrems-Urteile (**Schrems I**, 6. Oktober 2015 und **Schrems II**, 16. Juni 2020), sowie den Ausweg von **Standardvertragsklauseln**, welcher auch nach Schrems II noch möglich ist. Eine korrekte Antwort könnte lauten:

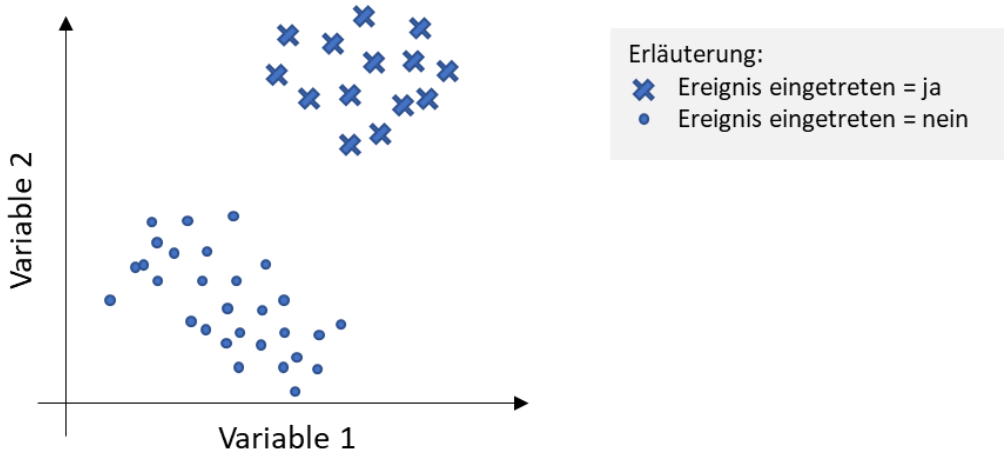
„Lieber Chef,

ich weiß, dass unsere Ergebnisse wirklich großartig sind, wir sollten die Rohdaten unseren amerikanischen Kollegen trotzdem nicht überstellen. Bis 2015 wäre das kein Problem gewesen, da das Unternehmen Ihres Kollegen ja dem Safe-Harbor-Abkommen beigetreten ist. Allerdings hat der Europäische Gerichtshof mit dem Schrems I Urteil vom 6. Oktober 2015 entschieden, dass das Safe Harbor Abkommen nicht mehr gültig ist. Es gab zwar im Anschluss das Folgeabkommen, den sogenannten EU-US-Privacy-Shield, über welchen wir wieder Daten mit unseren amerikanischen Kollegen hätten austauschen können, aber auch dieses wurde vom EuGh im Schrems II Urteil vom 16. Juni 2020 für ungültig erklärt. Damit können wir solche extrem schutzbedürftigen Daten, wie sie in unserem Datensatz enthalten sind, nicht in die USA übermitteln. Einziger Ausweg wäre ein gemeinsamer Vertrag, welcher sich an den EU-Standardvertragsklauseln (Durchführungsbeschluss (EU) 2021/914) orientiert. Dann müsste aber ihr amerikanischer Kollege für unseren Datensatz sicherstellen, dass die Daten dort ein gleichwertiges Schutzniveau wie in der EU genießen. Falls Sie so ein Papier haben, können wir die Daten über einen sicheren Datentransfer austauschen. Ansonsten würde ich vorschlagen, die Daten zu pseudonymisieren."

(10 Punkte. Pro erforderlichem Urteil/Abkommen/Stichpunkt 2 Punkte. 0,5 Bonuspunkte, aber trotzdem maximal 10 Punkte, falls der Vorschlag der Pseudonymisierung genannt wird)

Aufgabe 3 [6.1 Datenmanagement 2, 6.2 Datenverarbeitungstechnologien 2, 8.1 Data Mining 2, 8.2 Analytics 2, 7.1 Regressions- und Clustermethoden 2] [34 Punkte]

Im Rahmen eines Data Science Projektes soll ein Prognosemodell zur Vorhersage eines bestimmten Ereignisses erstellt werden. Zur Erstellung des Prognosemodells soll ein SVM (Support Vector Machine) Modell erstellt werden. In der nachfolgenden Grafik sind exemplarische Daten als Grundlage für die Modellierung dargestellt. Hierbei sind „Variable 1" und „Variable 2" metrisch skaliert.



a) [8 Punkte] Beschreiben Sie die grundlegende Methodik eines SVM Modells bei der Klassifikation von Daten. Erstellen Sie hierzu eine passende lineare Grenze in der dargestellten Grafik und erläutern Sie die Methodik zur Erstellung der Grenze.

Für die Erstellung des Support-Vector-Machine-Modells steht Ihnen – wie auch in Aufgabe 2 verwendet - der folgende Teildatensatz zur Verfügung

Id	1	2	3	4	5	6	7	8	9	10
$\frac{x}{\bar{x}}$	0,5	3	1	2,5	2	2	0,25	1	1,75	3
$\frac{L}{\bar{L}}$	0,3	2,25	1	1,5	1,75	3	2,5	2	3,5	4,5
y	-1	-1	-1	-1	-1	1	1	1	1	1

b) [9 Punkte] Berechnen Sie die lineare Grenze für den Datensatz aus der Tabelle. Gehen Sie davon aus, dass ein lineares SVM-Modell ($\Phi = Id$) mit dem Maximal-Margin-Ansatz verwendet wird.

Hinweis: Verwenden Sie die Datensätze mit der Id 3, 6 und 8 als Support-Vektoren.

Beim Fitting des SVM Modells kann ein Hyperparameter C zur Festlegung der Margin Violation festgelegt werden.

c) [5 Punkte] Erläutern Sie das Konzept und das Vorgehen zur Trennung der Daten beim Training und bei der Erstellung des SVM Prognosemodells unter Berücksichtigung des Parameters C.

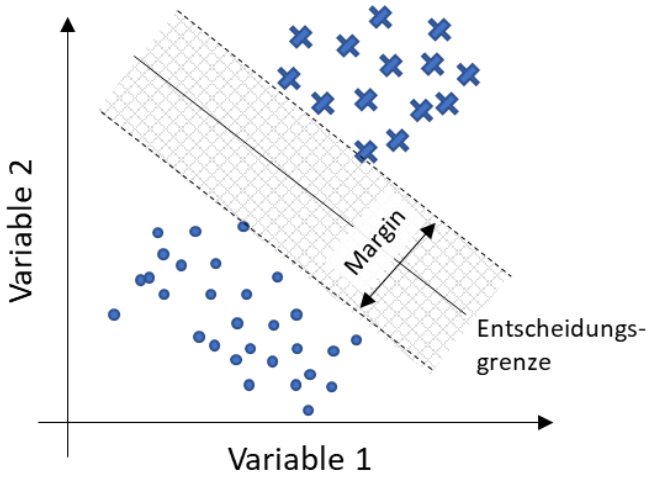
Für die Erstellung des Prognosemodells stehen Ihnen weitere Daten zur Verfügung. In der folgenden Tabelle sind für die erklärenden Variablen fünf Datensätze dargestellt:

ID	Variable 1	Variable 2	Variable 3
1	A	0,51	16
2	B	0,12	14
3	A	0,78	3
4	B	0,58	6
5	C	0,66	14

d) [12 Punkte] Benennen und beschreiben Sie zwei Datentransformationen, die vorab zur Modellierung eines SVM Modells durchgeführt werden sollten. Führen Sie die Datentransformationen für die exemplarischen Datensätze durch.

Lösungsvorschlag:

a) Zur Klassifikation der Daten wird bei einem SVM Modell eine Entscheidungsgrenze (decision boundary) zur Trennung der Daten erzeugt. Hierbei wird die Entscheidungsgrenze so erzeugt, dass der Abstand zwischen den beiden Klassen maximiert wird. Dies erfolgt über die Maximierung der sogenannten Margin, wie in der nachfolgenden Grafik dargestellt. In der folgenden Grafik ist die lineare Grenze für die gegebenen Datensätze dargestellt:



b) Wie aus der Visualisierung der Teilaufgabe b) aus Aufgabe 2. leicht ersichtlich ist, sind die Daten linear trennbar. Insofern existiert eine eindeutige Lösung des SVM-Optimierungsproblems:

$$Minimiere_{w,b} \quad \frac{\|w\|^2}{2}$$

$$so\ dass\ y_i(wx_i + b) \geq 1$$

Wobei $y_i$dem Corona-Klassifikator entspricht und die x_i aus den skalierten Daten zusammengesetzt sind, also $\tilde{x}_i = \begin{pmatrix} x_i/\bar{x} \\ L_i/\bar{L} \end{pmatrix}$. Es gilt $wx_i = w_1\, x_{i1} + w_2 x_{i2}$.

Lösungsmöglichkeit 1 *(Optimierungsproblem lösen)*:

Da wir wissen, dass es eine eindeutige Lösung gibt und die drei Support Vektoren kennen, muss für diese gelten: $wx_i + b = y_i$, mit $i = 3, 6, 8$. Damit ergibt sich folgendes lineares Gleichungssystem:

$$2w_1 + 3\, w_2 + b = 1$$

$$w_1 + 2w_2 + b = 1$$
$$w_1 + w_2 + b = -1$$

Die Lösung dieses Gleichungssystems lautet: $w_1 = -2, w_2 = 2, b = -1$ und damit ergibt sich als Gleichung für die trennende Hyperebene / lineare Grenze:

$$\binom{-2}{2} x - 1 = 0$$

Bzw. $-2x_1 + 2x_2 - 1 = 0$.

Lösungsmöglichkeit 2 *(Parallele Geraden bestimmen)*:

Da die Support Vektoren bekannt sind, ist aus der Skizze ersichtlich, dass es sich bei der gesuchten Hyperebene um eine parallele Gerade zur Gerade durch die beiden Punkte

$$P_6(2/3), P_8(1/2)$$

handeln muss, welche den Abstand zu Punkt $P_3(1/1)$ halbiert. Die Gerade H_0 durch P_6 und P_8kann über das Gleichungssystem

$$2m + c = 3$$
$$1m + c = 2$$

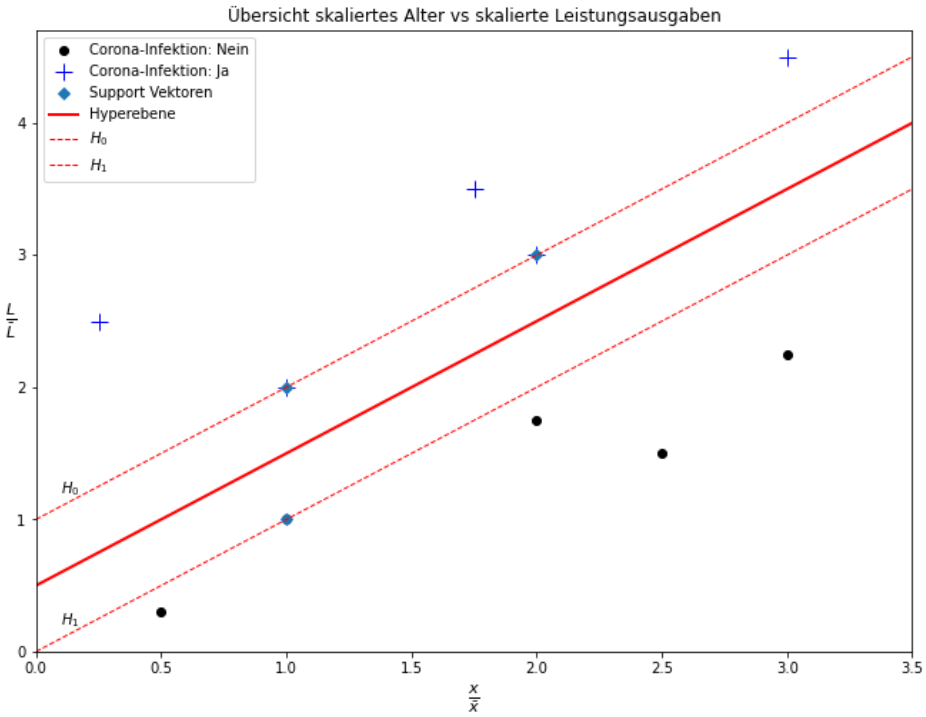
bestimmt werden, als $y = x + 1$. Die hierzu parallele Gerade H_1 durch den Punkt P_3hat damit die Gleichung $y = x$.

Nun kann die Gleichung für die Gerade, welche exakt zwischen diesen beiden Geraden liegt, bestimmt werden als: $y = x + 0{,}5$.

Lösungsmöglichkeit 3 *(Educated Guess)*:

Die Gleichung der Geraden kann auch durch scharfes Hinsehen oder durch Einzeichnen aus der Visualisierung von Teilaufgabe b) entnommen werden.

Visualisierung der Hyperebene:



(2 Punkte für die Angabe der korrekten Gleichung, 6 Punkte für den Rechenweg)

c) Die Trennung der Daten und Bestimmung des Parameter C kann wie folgt erfolgen:
Die Daten werden in Trainings-, Validation- und Testdaten aufgeteilt (z.B. im Verhältnis 60% / 20% / 20%). Die Daten werden anschließend zu folgenden Zwecken verwendet:

- Trainingsdaten: Die Trainingsdaten werden verwendet, um das Modell zu trainieren.
- Validationsdaten: Die Validationsdaten werden verwendet, um den Parameter C des Modells zu bestimmen. Hierzu wird das Modell mehrfach mit den Trainingsdaten trainiert und unter Verwendung der Validationsdaten wird der (optimalen) Parameter C für das Modell bestimmt.
- Testdaten: Die Testdaten werden zu einer unverzerrten Bewertung des finalen Modells verwendet.

d) Im Vorfeld zur Modellierung eines SVM Modells sollten folgende Transformationen durchgeführt werden:

- Feature Scaling: Durch ein Feature Scaling wird der Wertebereich der metrischen erklärenden Variablen auf eine einheitliche Skala gebracht. Hierzu gibt es verschiedene Methoden, wie z.B. die Standardisierung und Normalisierung (Min-Max-Methoden).

  Bei der Min-Max werden die Werte jeder numerischen Variablen auf dem Bericht [0, 1] transformiert. Hierbei wird folgende Formel auf jeden Datensatz angewandt:

$$x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

  Die Daten der Variablen 2 befindet sich im Wertebereich [0, 1] und daher ist keine Transformation notwendig, aber möglich. Für die Normalisierung der Variablen 3 gilt Min(x) = 3 und Max(x) = 16. Hieraus ergeben sich mit der Formel zur Normalisierung die folgenden Werte:

ID	**Variable 3**
1	1,00
2	0,85
3	0,00
4	0,23
5	0,85

- One-hot Encoding: Zur Verwendung der kategorialen Variablen 1 bei einer Modellierung eines SVM Modell sollte eine Transformation durchgeführt werden, die die Ausprägungen der Variablen in numerische Werte umwandelt. Bei dem One-hot Encoding wird für jede Ausprägung der Variablen 1 („A", „B" und „C") eine numerische (dummy) Variable mit dem Wertebereich 0 und 1 angelegt. Die Variable 1 wird mit folgenden drei Variablen ersetzt:

ID	**Variable 11**	**Variable 12**	**Variable 13**
1	1	0	0
2	0	1	0
3	1	0	0
4	0	1	0
5	0	0	1

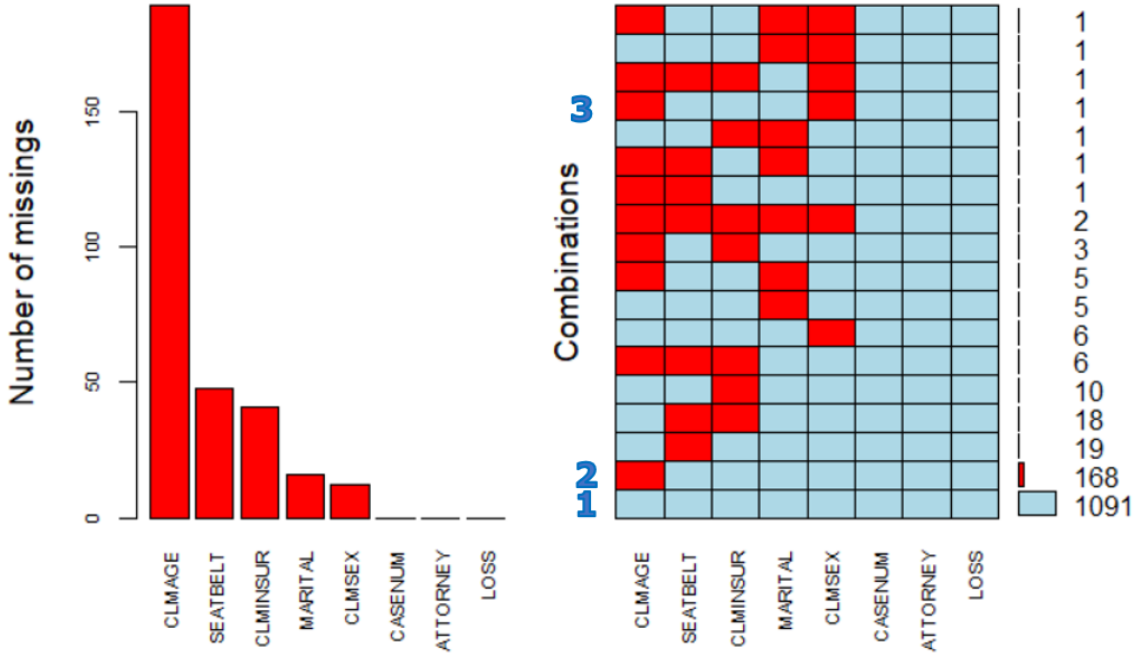
Aufgabe 4 [7.3 Datenaufbereitung zur Modellerstellung 2; 8.1 Data Mining 2] [40 Punkte]

a) [6 Punkte] Nennen und erläutern Sie die drei Typen von Missing Values nach der Typologie von Rubin (MAR etc.): jeweils Abkürzung, Langname, kurze Beschreibung in einem Satz.

b) [6 Punkte] Sie erheben Daten, indem Sie einigen Ihrer Kunden Fragebogen per Post zusenden. Nennen Sie für jeden der drei Typen von Missing Values eine mögliche Ursache, die bei Ihrer Fragebogenaktion auftreten kann, und erläutern Sie die von Ihnen gewählte Zuordnung.

c) [8 Punkte] Nennen Sie vier grundverschiedene Ansätze zum Umgang mit Missing Values.
Was versteht man unter Imputierung?
Erläutern Sie beispielhaft, inwiefern der Typ von Missing Values eine Rolle beim Einsatz von Imputierungsmethoden spielt. Ziehen Sie dazu eines Ihrer Beispiele aus b) heran.

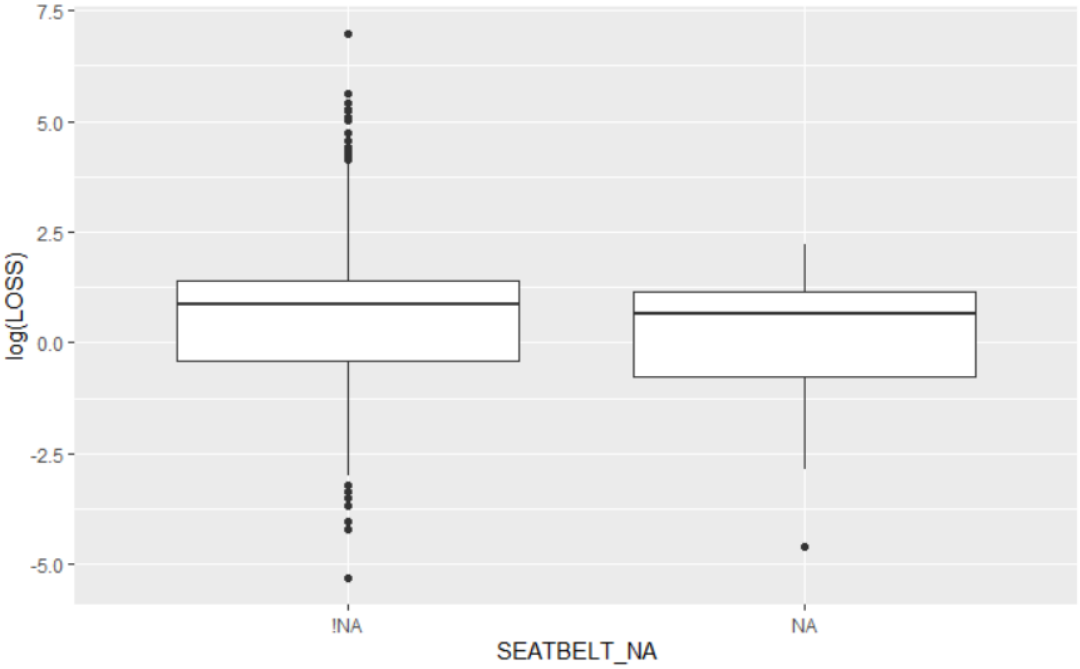
d) [10 Punkte] Sie haben die folgenden Visualisierungen von Missing Values erhalten (die drei fettgedruckten Zahlen in der Mitte der Grafik wurden lediglich für diese Aufgabenstellung eingefügt).



Erläutern Sie jeweils, wie diese Art von Grafik erzeugt wird (fachlich, nicht mit welchem Software-Paket), was man grundsätzlich ablesen kann und welche Schlussfolgerung(en) Sie im konkreten Fall aus der Grafik ziehen. Betrachten und erläutern Sie hierzu bei der rechten Grafik insbesondere

die für diese Aufgabe mit den drei fettgedruckten Zahlen von 1 bis 3 markierten „Zeilen".

e) [6 Punkte] Für den Zusammenhang zweier Merkmale haben Sie folgende Grafik erhalten:



HINWEIS: SEATBELT ist im Originaldatensatz ein binäres Merkmal, für das Einträge fehlen. Daher wurde ein binäres Merkmal SEATBELT_NA mit den Ausprägungen NA und !NA erzeugt, das anzeigt, ob SEATBELT fehlt oder nicht fehlt.
Erläutern Sie die Grafik. Welche Vermutung würden Sie hinsichtlich des Missing-Value-Typs aus der Grafik ableiten? Welche Zusatzinformationen oder welche alternative Visualisierung wären hilfreich, um die Vermutung (auch ohne statistische Tests) zu untermauern?

f) [4 Punkte] Erläutern Sie in einem Satz, was CRISP-DM ist. Wofür steht die Abkürzung? Wo in CRISP-DM würden Sie die oben gezeigten Visualisierungen zuordnen?

Lösungsvorschlag:

a) MCAR = „Missing Completely at Random": Es gibt keinen systematischen Zusammenhang zwischen dem Fehlen von Daten und den sonstigen Daten.

MAR = „Missing at Random": Das Fehlen von Daten ist unabhängig vom fehlenden Wert, ist aber abhängig von beobachteten Werten. M.a.W.: Die Wahrscheinlichkeit des Fehlens ist in bestimmten Gruppen von Beobachtungen konstant.

MNAR = „Missing Not at Random" (bzw. NMAR = „Not Missing at Random"): Keiner der beiden anderen Fälle liegt vor. Das Fehlen hängt vom fehlenden Wert ab.

b) *Die im Folgenden genannten Ursachen sind beispielhaft; andere Antworten sind möglich.*

MCAR: Durch einen Versandfehler haben einige Kunden die letzte Seite des Fragebogens nicht erhalten und somit die letzten drei Fragen nicht beantwortet. Es ist davon auszugehen, dass der Fehler beim Versand rein zufällig aufgetreten ist, d.h. der Fehler hat nicht bestimmte Kunden „ausgewählt", also MCAR.

MAR: Eine Frage war schwer verständlich formuliert und wurde deshalb vor allem von den Kunden nicht beantwortet, die nicht Deutsch als Muttersprache haben. Da die Muttersprache ein abgefragtes Merkmal war, ist das Fehlen dieser Antwort MAR.

MNAR: Kunden mit sehr hohem Einkommen haben die Frage nach dem Einkommen häufig nicht beantwortet. Hier besteht ein unmittelbarer Zusammenhang zwischen der fehlenden Antwort und ihrem Fehlen, also MNAR.

c) Ansätze zum Umgang mit fehlenden Daten *[nur vier grundverschiedene Ansätze waren gefragt]*:
- Entfernen aller Datensätze mit fehlenden Daten
- Entfernen von (einzelnen) Merkmalen, die fehlende Daten aufweisen
- Imputierung z.B. mit dem Median der beobachteten Werte des betreffenden Merkmals (sog. Einfache Imputierung)
- Imputierung z.B. mit der k-Nearest Neighbour Methode
- Multiple Imputierung (d.h. unter Nutzung von Simulationen)

Allgemein versteht man unter Imputierung den Ersatz fehlender Werte durch Werte, die in der Regel aus den vorhandenen Beobachtungen abgeleitet werden. Bei einfacher Imputierung beschränkt man sich bei dieser Ableitung auf Ausprägungen des fehlenden Merkmals. Bei komplexeren Imputierungsverfahren bezieht man dagegen auch die übrigen Merkmale mit ein.

Eine Imputierung z.B. mit dem Median ist nur dann vertretbar, wenn nur wenige Daten fehlen und vor allem, wenn MCAR vorliegt (was leider selten der Fall ist). Wenn das Fehlen der Daten nicht rein zufällig ist, verfälscht die Imputierung nicht nur die Streuung, sondern auch die Zentralitätsmaße. In dem MNAR-Beispiel aus b) wird durch Imputierung mit dem Median der vorhandenen Einkommen der Median der Einkommen der Gesamtstichprobe massiv verfälscht.

d) Die linke Grafik stellt ein Balkendiagramm der Anzahl der fehlenden Werte in den einzelnen Merkmalen dar. Hier werden also einfach die fehlenden Werte je Merkmal gezählt und durch entsprechende Balken visualisiert. Dabei sind die Merkmale nach der Anzahl fehlender Werte sortiert. Man erkennt unmittelbar, welche Merkmale besonders vom Fehlen von Daten betroffen sind.

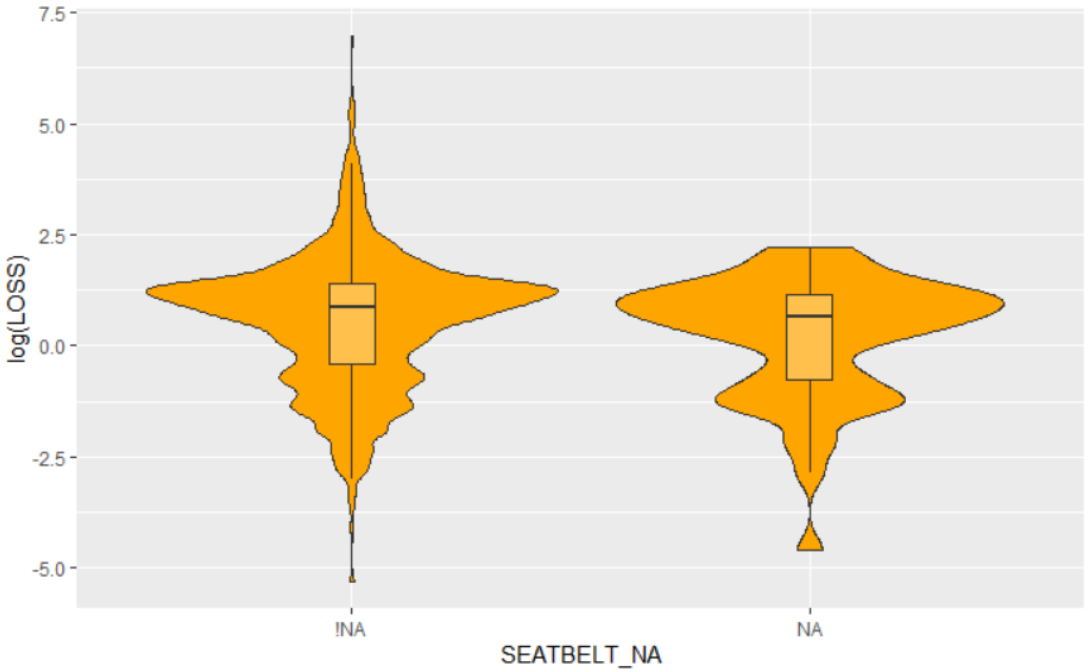
   - Das am häufigsten fehlende Merkmal ist CLMAGE (fast 200mal).
   - Die Merkmale CASENUM, ATTORNEY und LOSS fehlen in keinem Datensatz.

   Die rechte Grafik zeigt alle auftretenden Merkmalskombinationen von fehlenden Daten. Jede „Zeile" steht für eine solche Kombination. Die Zeilen sind von oben nach unten gemäß zunehmender Häufigkeit sortiert. Am rechten Rand der rechten Grafik signalisiert ein Histogramm, wie häufig diese Kombination fehlender Daten in der Gesamtmenge der Beobachtungen vorkommt. Die genaue Häufigkeit ist auch als Zahl angegeben. Rot signalisiert – entsprechend der linken Grafik – das Fehlen des jeweiligen Merkmals. Blau signalisiert, dass die Ausprägung des betreffenden Merkmals bekannt ist.

   - Die unterste Zeile (mit 1 markiert) steht für die 1091 Fälle/Beobachtungen ohne fehlende Werte.
   - Die Zeile darüber (mit 2 markiert) steht für die 168 Fälle/Beobachtungen, bei denen nur das Merkmal CLMAGE fehlt.
   - Die vierte Zeile von oben (mit 3 markiert) steht für den einen Fall, bei

dem das Merkmal CLMAGE und das Merkmal CLMSEX gleichzeitig fehlen.
- Die Merkmale CASENUM, ATTORNEY und LOSS fehlen in keinem Datensatz (das sieht man auch im Histogramm links).

e) Die Grafik zeigt zwei Boxplots für das Merkmal log(LOSS) – einmal für die Datensätze, bei denen SEATBELT fehlt (NA) und einmal für diejenigen, bei denen SEATBELT vorhanden ist (!NA). Wegen der Schiefe der Verteilung wurde die Grafik für log(LOSS) statt für LOSS erstellt. Man sieht, ob die Verteilung der Variablen LOSS abhängig vom Fehlen von SEATBELT ist. Bei den Datensätzen, bei denen das Merkmal SEATBELT fehlt, zeigt sich eine deutlich andere Verteilung im Merkmal LOSS als bei den Datensätzen, bei denen das Merkmal SEATBELT vorhanden ist. Daher ist das Fehlen von SEATBELT vermutlich nicht MCAR. Bei MCAR müsste die Verteilung des Merkmals LOSS unabhängig vom Fehlen oder Vorliegen von SEATBELT sein.
Zur Untermauerung der Vermutung „nicht MCAR" muss die Menge an Daten groß genug sein. Es geht also um die Frage, wie viele Datensätze beim Merkmal SEATBELT NA oder !NA sind. Diese Information bekommt man aus dem linken Teil der Grafik in Teilaufgabe d): knapp 50 Datensätze haben nicht das Merkmal SEATBELT. Demgegenüber sind – wie bereits festgestellt – 1091 Datensätze vollkommen ohne fehlende Werte bzw. knapp 1300 Datensätze haben das Merkmal SEATBELT. Somit ist die Schlussfolgerung „nicht MCAR" auch ohne statistische Tests hinreichend sicher.
Eine noch informativere Visualisierung bekommt man durch eine empirische Verteilungsfunktion oder noch besser durch einen Violinplot mit integriertem Boxplot. *Die folgende Grafik zeigt diesen Violinplot und ist nicht Bestandteil der geforderten Lösung, sondern dient nur zur Illustration der Musterlösung:*



f) CRISP-DM = Cross Industry Standard Process for Data Mining.
CRISP-DM ist ein Prozessmodell für Data-Mining-Projekte, das 1999/2000 von einem Konsortium von Firmen aus mehreren Branchen erstellt wurde und das auch heute noch aufgrund seiner Verbreitung als das Standardmodell angesehen wird.
Die Visualisierungen aus d) und e) gehören in die Phase „Datenverständnis", Aufgabe „Überprüfen der Datenqualität", können aber auch noch in der Phase „Datenvorbereitung", Aufgabe „Bereinigen von Daten" herangezogen werden.

Aufgabe 5 [7.4 Modellselektion und Regularisierung 2] (42 Punkte)

a) [5 Punkte] Beschreiben Sie im Kontext der linearen Modelle das Verfahren der Ridge Regression und das Lasso-Verfahren. Geben Sie die zugehörigen Optimierungsformulierungen (ausgehend von den Residuenquadratsummen (RSS), die bei der Methode der kleinsten Quadrate optimiert werden) an.

b) [15 Punkte] Was kann in der Ridge Regression unter den folgenden Annahmen über die Schätzer der Parameter für β_1 und β_2 mit dem Regressionsmodell $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ ausgesagt werden? $n = 2$ Beobachtungen, $p = 2$ Kovariablen (Features) und $x_{11} = x_{12}$ sowie $x_{21} = x_{22}$. (Hinweis: Zusätzlich wird angenommen, dass $y_1 + y_2 = 0$, $x_{11} + x_{21} = 0$ und $x_{12} + x_{22} = 0$ gilt. Dann gilt zudem $\hat{\beta}_0 = 0$, was Sie ohne Beweis verwenden können).

c) [5 Punkte] Beschreiben Sie folgende Verfahren zur Modellselektion:

1. Vollständige Modellselektion
2. Vorwärts-Selektion
3. Rückwärts-Selektion
4. Schrittweise Selektion

d) [17 Punkte] Der folgende R Source Code ist vorgegeben:

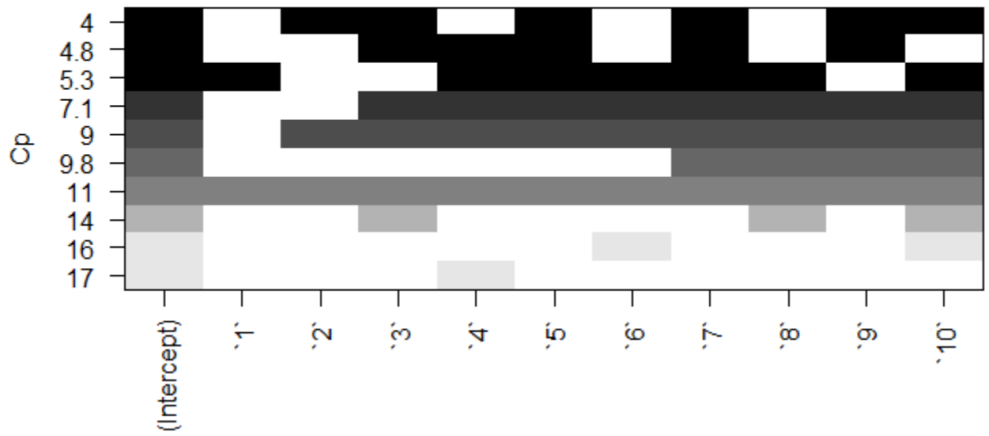
Codestück 1:

```
17    library(tidyverse)
18    library(leaps)
19    library(glmnet)
20
21    set.seed(42)
22    n <- 100
23    X <- rnorm(n,0,1)
24    err <- rnorm(n,0,1)
```

Codestück 2:

```
30    X_1 <- as_tibble(poly(X, degree=10, raw=TRUE))
31    Y <- 5 + 0.2 * pull(X_1, 4) + err
32    df <- X_1 %>% add_column(Y, .before = 1)
33
34    reg_subset <- regsubsets(Y ~ ., data=df, nvmax=10)
35    plot(reg_subset, scale="Cp")
```

Ausgabe zu Codestück 2:



Codestück 3:

```
38    lasso_cv <- cv.glmnet(as.matrix(df[-1]), df$Y, alpha = 1)
39    best_lambda <- lasso_cv$lambda.min
40    print(best_lambda)
41
42    lasso_model <- glmnet( as.matrix(df[-1]), df$Y, alpha=1, lambda=best_lambda)
43
44    coef(lasso_model)
```

Ausgabe zu Codestück 3:

```
[1] 0.4103011
11 x 1 sparse Matrix of class "dgCMatrix"
                    s0
(Intercept) 5.0424894
1           .
2           .
3           .
4           0.1642613
5           .
6           .
7           .
8           .
9           .
10          .
```

Beschreiben Sie den Ablauf, indem Sie auf die Codestücke 1 bis 3 eingehen. Interpretieren Sie im Anschluss das Ergebnis, indem Sie auf die Ausgaben zu Codestück 2 und 3 eingehen.

Hinweise zu den R-Codestücken:

- Der Befehl *poly* wird verwendet, um die Terme x^2, x^3 etc. zu erzeugen. So liefert der Befehl *poly(c(3,4), degree=3, raw=TRUE)* die Datenstruktur *num [1:2,1:3] 3 4 9 16 27 64.*
- Über *pull(X,i)* wird die i-te Spalte aus X extrahiert.
- Mit *cv.glmnet()* wird eine Kreuzvalidierung durchgeführt und damit ein Wert für λ ermittelt.
- Mit glmnet wird (mit alpha=1) das Lasso-Verfahren durchgeführt.

Lösungsvorschlag:

a) Sowohl die Ridge-Regression als auch das Lasso-Verfahren gehören zur Klasse der Schrumpfungsverfahren („Shrinkage Methods"). Mit diesen Methoden ist es möglich, die gesamte Anzahl an möglichen Kovariablen p für ein Modell zu verwenden, ohne bereits vorab Ausschlüsse vornehmen zu müssen. Die Koeffizienten werden bei diesen Verfahren regularisiert. Das kann die Varianz der Verfahren erheblich senken. Beide Verfahren sind ähnlich aufgebaut. Bezeichne mit RSS die zugehörige Residuenquadratsumme, d.h.

$$RSS = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2.$$

Bei der Ridge-Regression wird nicht die Residuenquadratsumme RSS minimiert, sondern der modifizierte Ausdruck

$$RSS + \lambda\sum_{j=1}^{p}\beta_j^2.$$

Der modifizierte Ausdruck beim Lasso dagegen lautet wie folgt:

$$RSS + \lambda\sum_{j=1}^{p}|\beta_j|$$

Beide Verfahren schrumpfen die Schätzungen der Koeffizienten gegen 0. Allerdings können beim Lasso-Verfahren die Schätzungen auch den Wert 0 annehmen (bei einem genügend großen Wert für λ). Damit kann über dieses Verfahren eine Variablenselektion vorgenommen werden.

b) Aus dem Ansatz zur Ridge Regression aus Teil a) ergibt sich

$$\min_{\beta_1,\beta_2}\sum_{i=1}^{2}(y_i - \beta_1 x_{i1} - \beta_2 x_{i2})^2 + \lambda(\beta_1^2 + \beta_2^2).$$

Setze $L(\beta_1,\beta_2) = \sum_{i=1}^{2}(y_i - \beta_1 x_{i1} - \beta_2 x_{i2})^2 + \lambda(\beta_1^2 + \beta_2^2)$.

Berechnung der partiellen Ableitungen, Nullsetzung und Vereinfachung liefert

$$\frac{1}{2}\frac{\partial L}{\partial \beta_1} = -x_{11}(y_1 - \beta_1 x_{11} - \beta_2 x_{12}) - x_{21}(y_2 - \beta_1 x_{21} - \beta_2 x_{22}) + \lambda\beta_1 = 0$$

sowie

$$\frac{1}{2}\frac{\partial L}{\partial \beta_2} = -x_{12}(y_1 - \beta_1 x_{11} - \beta_2 x_{12}) - x_{22}(y_2 - \beta_1 x_{21} - \beta_2 x_{22}) + \lambda\beta_2 = 0.$$

Unter den Voraussetzungen $x_{12} = x_{11}$ und $x_{21} = x_{22}$ liefern die letzten beiden Gleichungen

$$\lambda\beta_1 = \lambda\beta_2.$$

In diesem einfachen Setup, in dem die beiden Variablen abhängig sind, sind die Schätzwerte identisch. Prinzipiell bekommen bei der Ridge-Regression abhängige Variablen ähnliche Schätzwerte.

c) Vollständige Modellselektion: Betrachtung aller 2^p möglichen Modelle und Auswahl des Modells mit der größten Reduktion eines Modellwahlkriteriums (z.B. C_p, AIC, BIC etc.). Für $p < 40$ kann mit dem sogenannte „Leaps and bounds" Algorithmus (Furnival & Wilson (1974)) die Berechnung aller Modelle vermieden werden (siehe z.B. das Paket „leaps" in R).

Vorwärts-Selektion: Basierend auf einem Startmodell wird in jedem Schritt des Verfahrens eine weitere Kovariable aufgenommen, und zwar diejenige, welche die größte Reduktion eines Modellwahlkriteriums (z.B. C_p, AIC, BIC etc.) bewirkt. Falls keine Reduktion mehr möglich ist, endet der Algorithmus.

Rückwärts-Selektion: Start mit dem vollen Modell (alle Kovariablen sind enthalten). In jedem Schritt wird diejenige Kovariable aus dem Modell entfernt, welche die größte Reduktion eines Modellwahlkriteriums (z.B. C_p, AIC, BIC etc.) bewirkt. Falls keine Reduktion mehr möglich ist, endet der Algorithmus.

Schrittweise Selektion: Kombination aus Vorwärts- und Rückwärts-Selektion. In jeder Iteration kann sowohl eine Variable aufgenommen als auch eine Variable entfernt werden.

d) Im Codestück 1 werden zunächst vorbereitend die benötigten Bibliotheken geladen. Nach dem Setzen eines Seed-Wertes zur Reproduzierbarkeit wer-

den die Daten zur Simulation erzeugt: n, X und err werden nach einer Standard-Normalverteilung erzeugt. In Codestück 2 erhält X_1 die 100 Beobachtungen aus X

sowie die darauf aufbauenden Terme $x^2, x^3, \ldots, x^4$. Die Zielvariable Y wird aus dem Modell

$$Y = 5 + 0.2\, x^4 + err$$

gebildet und im Anschluss mit den Werten aus X_1 zu df zusammengefügt, wobei Y die erste Spalte darstellt (über .before=1). Über den Befehl regsubsets werden in diesem Fall die besten Modelle mit maximal 10 enthaltenen Variablen angezeigt. Die dargestellten Modelle sind, nach dem Modellwahlkriterium C_p, die besten Modelle mit einem bis zehn Parametern.

In Codestück 3 wird zunächst für das Lasso-Verfahren der Parameter λ über Kreuzvalidierung ermittelt. Im Anschluss wird das Lasso-Verfahren durchgeführt.

Aus der Ausgabe zu Codestück 2 geht hervor, dass die hier verwendete Modellselektion es nicht schafft, das ursprüngliche Modell (welches nur auf x^4 basiert) zu identifizieren. Anhand des Modellwahlkriteriums C_p ist das Modell ganz oben in der dargestellten Grafik (bestehend aus insgesamt 7 Variablen) das bevorzugte Modell.

Das Lasso Modell hingegen (welches wie in Teil a) bereits angesprochen auch zur Variablenselektion verwendet werden kann) ist in der Lage das ursprüngliche Modell zu ermitteln.