Lernziele im Vertiefungswissen *Actuarial Data Science*

Actuarial Data Science Immersion

Actuarial Data Science Completion

Autoren:

Die Mitglieder der DAV-Arbeitsgruppe „Vertiefungswissen Actuarial Data Science" sowie der AG Qualifizierung des Ausschusses ADS

Actuarial Data Science Immersion

1 ADS Grundlagen & Umfeld

1.1 Berufsständisches & Business Skills

Zur erfolgreichen Ausübung der Tätigkeit eines Actuarial Data Scientist sind neben dem technischen, fachlichen und mathematischen Wissen auch verschiedene Business Skills notwendig. Hierzu zählen unter anderem das Verständnis und das Wissen über die Wirtschaftlichkeit, sowie die Vorteile und die Grenzen von Data Science Aktivitäten. Die standesgemäße Ausübung der Tätigkeit ist ein weiteres Merkmal des Berufsbilds eines Actuarial Data Scientist.

Zielsetzung: Die Kandidaten können für Anwendungsfälle und Problemstellungen aus der Versicherungswirtschaft die Anwendung von Data Science Aktivitäten kritisch bewerten und prüfen, ob und inwieweit Konflikte mit den berufsständischen Grundsätzen bestehen.

1.1.1 Benennen und erläutern Sie die wichtigsten berufsständischen Grundsätze der DAV im Bereich Data Science. **(B2)**

1.1.2 Analysieren Sie für konkrete Anwendungen im Bereich Data Science die Konformität mit den berufsständischen Grundsätzen der DAV. **(B4)**

2 Informationstechnologie

2.1 Informationstheoretische Grundlagen

Die theoretische Informatik beschäftigt sich mit der abstrakten Beschreibung der Struktur, Verarbeitung, Übertragung und Wiedergabe von Informationen. Sie bildet somit die theoretische Grundlage unseres Verständnisses der Arbeitsweise von Computern.

Zielsetzung: Die Kandidaten kennen und verstehen die Begriffsbildung und grundlegende Resultate der theoretischen Informatik und sind sich der Relevanz für die alltägliche Arbeit bewusst.

2.1.1 Beschreiben Sie die Funktionsweise einer Turing Maschine und erklären Sie, wofür diese von Bedeutung ist. **(C2)**

2.1.2 Erklären Sie den Unterschied zwischen P-Schweren und NP-Schweren Fragestellungen. Gehen Sie dabei auch auf das P=NP-Problem ein und analysieren Sie dessen praktische Relevanz. **(C4)**

2.1.3 Formulieren Sie das Halteproblem und erläutern Sie, welche Konsequenzen sich daraus ergeben. **(C3)**

2.2 Systemarchitekturen

Anwendungen folgen in ihrem Aufbau, ihrer Schnittstellen und ihrer Komponenten einem Entwurf, der als Systemarchitektur bezeichnet wird. Moderne Systemarchitekturen unterstützen Entkopplung und Virtualisierung. Die Systemarchitektur hat stets auch Auswirkungen auf die von einer Anwendung verwalteten Daten.

Zielsetzung: Die Kandidaten haben einen Überblick über moderne Software-Systemarchitekturen und können die dafür notwendigen Technologien einordnen.

2.2.1 Erläutern Sie die wesentlichen Unterschiede zwischen einer Micro Service-Architektur und der klassischen, geschichteten Softwarearchitektur. Bewerten und beurteilen Sie, welche wesentlichen Vorteile Micro Service-Architekturen haben? **(C5)**

2.2.2 Wofür steht die Abkürzung REST bei Schnittstellen? Erklären Sie den Grundgedanken und die mit dieser Technologie verfolgten Ziele! **(B2)**

2.2.3 Erläutern Sie die Begriffe „Cloud ready" und „Cloud Native"! **(B2)**

3 Tools & Programme

3.1 Einführung und Überblick

Tools und Programme sind die zentralen Arbeitsmittel der Data Science.

Zielsetzung: Die Kandidaten erhalten einen Überblick über die wichtigsten Data Science Tools und Programme.

3.1.1 Nennen Sie die wichtigsten Programmiersprachen und Programmbibliotheken für Data Science. **(B1)**

3.1.2 Nennen Sie Kriterien für eine geeignete Data Science Programmiersprache und definieren Sie einen mindestens erforderlichen Funktionsumfang. **(B5)**

3.1.3 Charakterisieren Sie gängige Programmiersprachen im Kontext Data Science und geben Sie Empfehlungen basierend auf verschiedenen Einsatzgebieten ab. **(B5)**

3.2 Sprachübergreifende Data Science Tools

Einige leistungsfähige Tools können in verschiedene Programmiersprachen eingebettet werden oder codefrei über eine Benutzeroberfläche bedient werden.

Zielsetzung: Die Kandidaten lernen codefreie sowie sprachübergreifende Tools kennen.

3.2.1 Nennen Sie verschiedene sprachübergreifende Gradient-Boosting-Tools, beschreiben Sie die Unterschiede, nennen Sie die bevorzugte Wahl und begründen Sie diese. **(B3)**

3.2.2 Beschreiben Sie, wie sich Deep Learning Methoden mit State of the Art-Frameworks umsetzen lassen. **(B2)**

Bemerkung: Aktuelle Beispiele sind Keras und Tensorflow.

3.2.3 Nennen und beschreiben Sie Tools, die über eine Benutzeroberfläche bedient werden und inwieweit in diese Code (insbes. Python, R) eingebunden werden kann. **(B2)**

3.2.4 Nennen und beschreiben Sie Tools, die sowohl über eine Benutzeroberfläche als auch über R- und Python-API angewendet werden können. **(B2)**

Bemerkung: Aktuell wäre hier bspw. H2O zu nennen.

3.3 Data Science mit R

Die freie Programmiersprache R wurde von Statistikern für Analysen und graphische Darstellungen entwickelt, deckt die ganze methodische Bandbreite von klassischen Verfahren bis zum Deep Learning gut ab und ist eine der wichtigsten Data Science Sprachen.

Zielsetzung: Die Kandidaten kennen die wichtigsten Sprachelemente und Bibliotheken und sind in der Lage, mit R eigenständig Data Science Projekte durchzuführen.

3.3.1 Benennen und erläutern Sie die wichtigsten Sprachelemente von R. **(B2)**

3.3.2 Benennen und erläutern Sie die wichtigsten Bibliotheken zur Datenaufbereitung und Visualisierung. **(B2)**

3.3.3 Erstellen Sie Notebooks in R zur Analyse von Datenbeständen aus dem Versicherungsumfeld. **(C5)**

*Bemerkung: Beispielsweise unter Verwendung einer 2*2-Grid-Search zur Ermittlung geeigneter Hyperparameter am Beispiel eines Tree-Boosting-Verfahrens. (Grid Search kann bspw. mit dem Caret-Paket implementiert werden.)*

3.4 Data Science mit Python

Die freie Programmiersprache Python zeichnet sich durch ihre große Entwicklergemeinschaft und das breite Einsatzgebiet aus und ist mit ihren gut strukturierten und sehr leistungsfähigen Programmbibliotheken eine der wichtigsten Data Science Sprachen.

Zielsetzung: Die Kandidaten kennen die wichtigsten Sprachelemente und Bibliotheken und sind in der Lage, mit Python eigenständig Data Science Projekte durchzuführen.

3.4.1 Benennen und erläutern Sie die wichtigsten Sprachelemente von Python. **(B2)**

3.4.2 Benennen und erläutern Sie wichtige Methoden aus der Bibliothek scikit-learn. **(B2)**

*Bemerkung: Hinzu können Encoder, Normalizer, Regressions- und Klassifikationsmetriken, *Regressor- und *Classifier-Funktionen („Methoden"), Kreuzvalidierungsfunktionen und Pipelines gezählt werden.*

3.4.3 Entwerfen Sie für eine Fragestellung im Versicherungsumfeld einen Analyseablauf des überwachten Lernens mit scikit-learn. **(C5)**

3.4.4 Benennen und erläutern Sie die wichtigsten Bibliotheken zur Datenaufbereitung und Visualisierung. **(B2)**

3.4.5 Erstellen Sie Notebooks in Python zur Analyse von Datenbeständen aus dem Versicherungsumfeld. **(C5)**

4 Mathematik / Statistik

4.1 Überwachtes Lernen

Die zahlreichen Methoden des überwachten maschinellen Lernens weisen Gemeinsamkeiten auf und bilden gemeinsam noch stärkere Prädiktoren.

Zielsetzung: Die Kandidaten verstehen die grundlegenden Konzepte des überwachten maschinellen Lernens, haben einen Überblick über die wichtigsten Verfahren und können diese insbesondere für Klassifikations- und Regressionsfragestellungen sachgerecht anwenden.

Regularisierung

4.1.1 Nennen und beschreiben Sie lineare und nicht-lineare Verfahren, in denen Regularisierung angewendet wird. Erläutern Sie Gemeinsamkeiten und Unterschiede. **(B2)**

4.1.2 Erläutern Sie am Beispiel der linearen Regression die Unterschiede zwischen LASSO, Ridge Regression und Elastic Net und stellen Sie die jeweiligen Vorteile dar. Beurteilen Sie, für welches Anwendungsszenario welches Verfahren am besten geeignet ist. **(C5)**

Ensembles

4.1.3 Nennen Sie geeignete Startwerte und beschreiben Sie Verfahren für das Hyperparametertuning. **(C2)**

Bemerkung: Beispielhaft kann dies an GBM oder Random Forests demonstriert werden.

4.1.4 Erläutern Sie, wozu die „Variable Importance" dient und wie sie berechnet wird. **(B2)**

Blending und Stacking

4.1.5 Beschreiben Sie die Grundideen der Verfahren „Blending" und „Stacking" und nennen Sie die jeweiligen Vorteile. **(B2)**

4.1.6 Nennen Sie geeignete Blending-Berechnungsverfahren für Klassifikations- sowie für Regressionsfragestellungen. **(C2)**

Anwendung

4.1.7 Wenden Sie Verfahren des überwachten maschinellen Lernens auf konkrete Fragestellungen aus der Versicherung praktisch an, in dem Sie unter 4.1 genannte Verfahren verwenden. **(C5)**

4.2 Unüberwachtes Lernen

Falls keine klar definierbare Zielfunktion vorhanden ist, sind Methoden des unüberwachten Lernens von fundamentaler Bedeutung.

Zielsetzung: Die Kandidaten können Anwendungsfälle des unüberwachten Lernens abgrenzen, kennen die wichtigsten Methoden, können diese auf Beispieldaten anwenden und die Ergebnisse verstehen.

Methoden & Anwendungen

4.2.1 Nennen und beschreiben Sie erweiterte Methoden des unüberwachten Lernens und Clusterings, die über die in Basic genannten Methoden hinausgehen. (B2)

Clustering

4.2.2 Geben Sie einen Überblick über die Methoden k-means, k-modes und k-prototypes nach Art der zu analysierenden Variablen. **(B2)**

4.2.3 Diskutieren Sie die Algorithmen kritisch im Hinblick auf die zu wählenden Parameter, Interpretation der Ergebnisse und Komplexität. **(C5)**

4.2.4 Erklären Sie die Begriffe divisive und agglomerative Clusteranalyse im Kontext der Hierarchischen Clusterverfahren; verwenden Sie hierzu das Dendrogram. Vergleichen Sie die Hierarchischen Clusterverfahren mit k-means. **(B2)**

4.2.5 Erläuterten Sie das Prinzip und die Grundbegriffe des Density-Based Clustering. Skizzieren Sie den Algorithmus DBSCAN und diskutieren diesen kritisch, auch im Hinblick auf Zeit- und Speicherplatzaufwand. **(B5)**

Anwendung

4.2.6 Wenden Sie Verfahren des unüberwachten maschinellen Lernens auf konkrete Fragestellungen aus der Versicherung praktisch an. **(C5)**

4.3 Deep Learning 1: Grundlagen künstlicher neuronaler Netze

Tiefe neuronale Netze sind flexibel einsetzbare Modelle des maschinellen Lernens. Insbesondere in der Bildverarbeitung sind sie anderen Modellen deutlich überlegen.

Zielsetzung: Die Kandidaten verstehen die grundlegende Funktionsweise neuronaler Netze und sind in der Lage, neuronale Netze auf Klassifikations- und Regressionsprobleme mit strukturierten Daten sowie Bildern anzuwenden.

4.3.1 Beschreiben Sie die grundlegende Struktur eines vorwärts gerichteten neuronalen Netzes. **(B2)**

4.3.2 Stellen Sie mögliche Aktivierungsfunktionen für verdeckte Neuronen formal und graphisch dar. Analysieren und diskutieren Sie ihre Vor- und Nachteile. **(B4)**

4.3.3 Welche Funktion hat ein sogenanntes „Bias Neuron"? **(B2)**

4.3.4 Nennen und erläutern Sie eine geeignete Output-Aktivierungsfunktion für Klassifikationsprobleme mit mehr als zwei Klassen. **(B2)**

4.3.5 Beschreiben Sie die grundlegende Idee und Vorgehensweise des Backpropagation-Algorithmus und wenden diese in einem einfachen neuronalen Netz an. **(C3)**

4.3.6 Skizzieren Sie die Anwendung eines Konvolutions-Filters auf eine zweidimensionale Matrix. **(B2)**

4.3.7 Nennen und beschreiben Sie zwei Regularisierungsmethoden für tiefe neuronale Netze. **(B2)**

4.3.8 Beurteilen Sie, in welchen Anwendungsszenarien der Einsatz neuronaler Netze besonders vorteilhaft sein kann. **(C5)**

4.3.9 Gestalten Sie „Deep-Learning-Pipelines" für Problemstellungen im Versicherungsumfeld. Welche Netzarchitektur ist geeignet? Kann ein vortrainiertes Modell verwendet werden (transfer learning)? Wie werden die Hyperparameter bestimmt? Wie verläuft das „Lernen" der gegebenen Problemstellung? Wie lässt sich die Modellgüte beurteilen? **(C5)**

5 Insurance Analytics

5.1 Data Mining Vertiefung

Die richtige Vorgehensweise bei der Analyse von Daten und der gezielte Methodeneinsatz sind aufgrund der Fülle an Methoden komplex und bedürfen insbesondere bei der Arbeit mit großen Datenmengen fortgeschrittener Techniken.

Zielsetzung: Die Kandidaten sollen weitergehende analytische Kenntnisse zum Umgang, der Interpretierbarkeit und den Gefahren bei der Arbeit mit großen Datenmengen vermittelt bekommen. In diesem Zusammenhang werden methodische Kenntnisse zum Daten-Präprocessing und der Dimensionsreduktion vorgestellt. Über die in Mathematik und Statistik hinausgehenden Inhalte zum Clustering sollen komplexe und moderne Verfahren kennengelernt und deren Einsatzszenarien innerhalb der Versicherungswirtschaft verstanden werden.

Curse of Dimensionality, non-standard metrics

5.1.1 Erläutern Sie, welche Probleme sich ergeben, wenn Datensätze mit vielen Dimensionen/Merkmalen untersucht werden und wieso diese Probleme für niedrigdimensionale, wenngleich volumenreiche Datensätze nicht auftreten. **(A2)**

5.1.2 Geben Sie Beispiele für nicht-euklidische Metriken und deren Anwendungsszenarien im Data Mining an. **(B1)**

Allgemeines zur Mustererkennung

5.1.3 Definieren Sie die wesentlichen Eigenschaften von klassifizierbaren Mustern und erläutern Sie, welche formalen Voraussetzungen erfüllt sein müssen, um solche Muster zu erkennen. **(B2)**

5.1.4 Erläutern Sie die grundlegenden Herausforderungen nicht-überwachter Mustererkennung und skizzieren Sie geeignete Lösungsansätze. **(B2)**

Daten-Präprocessing / Fortgeschrittene Dimensionsreduktion

5.1.5 Erläutern Sie, wie und auf welche Weise die PCA auf tensorielle Daten erweitert werden kann und nennen Sie Vor- und Nachteile des Vorgehens. **(C2)**

5.1.6 Grenzen Sie die Independent Component Decomposition von anderen Dimensionsreduktionsmethoden ab und nennen Sie einen versicherungstechnischen Anwendungsfall. **(B4)**

5.1.7 Beschreiben Sie ein Beispiel für lösbares Reduktionsproblem und ordnen Sie die Vorteile der Methoden sowohl im Hinblick auf dafür geeignete Datenstrukturen als auch andere Methoden ein. **(C3)**

Bemerkung: Mögliche Verfahren sind schnelle Fourier-Transformation (FFT) oder Wavelet-Transformation.

Komplexes Clustering

5.1.8 Erläutern Sie, in welchen Fällen (abgesehen vom Dimensionsfluch) einfache Clusteringverfahren wie k-means nicht oder nur schlecht funktionieren und nennen Sie zumindest konzeptionelle Lösungsansätze. **(C2)**

5.1.9 Erläutern Sie Vor- und Nachteile von Ward's Hierarchischem Clustering und grenzen Sie es gegenüber bspw. auf Graphen basierenden Verfahren ab. **(C2)**

5.1.10 Beschreiben Sie das Konzept des MeanShift-Clusterings. **(B2)**

5.1.11 Geben Sie ein Beispiel dafür, welche Problemklasse sich mit Markov Chain Monte Carlo clustern lässt. **(B1)**

5.1.12 Beschreiben Sie die Algorithmen zu den oben genannten Verfahren anhand eines einfachen (händisch zu lösenden) Beispiels. Interpretieren Sie die Ergebnisse. Welche Beispiele aus der Versicherungspraxis könnten mit Hilfe von Clustering Algorithmen vereinfacht werden? **(B2)**

5.2 Visualisierung

Die Visualisierung von Daten ist eine wichtige Aktivität in den unterschiedenen Arbeitsschritten einer Data Science Anwendung.

Zielsetzung: Die Kandidaten sind in der Lage vertiefende Methode zur Visualisierung von Daten anzuwenden. Hierbei kennen sie verschiedene Tools und können diese im Scope voneinander abgrenzen. In der Darstellung von Daten und Ergebnissen können die Kandidaten weiterhin Darstellungsregeln zielgerichtet anwenden.

Darstellungsformen

5.2.1 Beschreiben Sie vertiefende Darstellungsmöglichkeiten und -formen sowie Konzepte zur Visualisierung von Daten in den verschiedenen Aktivitäten eines Data Scientists. **(B2)**

Bemerkung: Hierbei ist zu unterscheiden zwischen Visualisierung im Rahmen der Datenexploration (u.a. zur Identifikation von Auffälligkeiten in Daten), im Rahmen der Modellerstellung und -selektion (u.a. zur Bewertung der Modell- und Vorhersagequalität), sowie im Rahmen der Präsentation und Darstellung von Erkenntnissen und Ergebnissen.

5.2.2 Diskutieren und vergleichen Sie unterschiedliche Darstellungsmethoden und beschreiben Sie die Anwendungsmöglichkeiten und Vorteile der verschiedenen Methoden. **(B4)**

Bemerkung: Beispiele wären Dashboards.

Darstellungsregeln

5.2.3 Erläutern Sie Darstellungsregeln und -formen für die Visualisierung von Daten. **(B2)**

5.2.4 Benennen und beschreiben Sie Möglichkeiten, Visualisierungen barrierefrei zu gestalten. **(B2)**

Bemerkung: Zum Beispiel mittels Farbschemata für Farbenblinde oder „Gestalt Principles of Perception".

5.2.5 Optimieren Sie Datenvisualisierung bezüglich der Verständlichkeit und Lesbarkeit in der Darstellung von Daten. **(B5)**

5.2.6 Beurteilen Sie für Datenvisualisierungen, ob Darstellungsregeln eingehalten werden oder nicht. **(B5)**

Tools

5.2.7 Benennen Sie Tools zur Visualisierung von Daten und grenzen diese hinsichtlich Scope und Funktionsumfang voneinander ab. **(B4)**

6 Business Cases

Die konkrete Anwendung der erlernten Methoden und Verfahren auf konkrete Fragestellungen aus der Versicherung sind von zentraler Bedeutung für die Arbeit eines Actuarial Data Scientists (ADS).

Zielsetzung: Die Kandidaten sind in der Lage, einfache und umfängliche Data Science Analysen sowie Anwendungen des maschinellen Lernens selbstständig durchzuführen.

6.1.1 Basierend auf einer konkreten Fragestellung und einem gegebenen Datenbestand führen Sie eine Data Science Analyse selbstständig durch. Dabei durchlaufen Sie alle Phasen eines Data Mining Prozesses und erstellen ein Notebook in Python, R oder einer anderen geeigneten Sprache. Sie interpretieren und beurteilen die Ergebnisse und präsentieren diese Zielgruppengerecht. **(C5)**

6.1.2 Für eine konkrete Fragestellung wenden Sie Verfahren des maschinellen Lernens selbstständig an. Sie validieren die Güte der Ergebnisse und verbessern diese schrittweise durch Variation der Hyperparameter. Sie interpretieren und beurteilen die Ergebnisse und präsentieren diese Zielgruppengerecht. **(C5)**

Actuarial Data Science Completion

1 ADS Grundlagen & Umfeld

1.1 Digitalisierung 2

Die Digitalisierung ist eine der treibenden Kräfte hinter den globalen gesellschaftlichen Veränderungen, die wir seit ein paar Jahren bemerken und auf die sich die Versicherungsbranche einstellen muss.

Zielsetzung: Die Kandidaten kennen die wesentlichen Motivatoren für Digitalisierung und haben einen Überblick über die gebräuchlichen Techniken und ihre Einsatzgebiete.

1.1.1 Analysieren Sie, inwieweit der Einsatz von Digitalisierung es ermöglicht, das Kerngeschäft der Versicherung (Risikoübernahme und Entschädigung) um Services und Prävention zu erweitern. **(C4)**

1.1.2 Beschreiben Sie die Vorteile und Möglichkeiten von Cloud Computing für die Analyse und Verarbeitung von großen Datenmengen (Big Data). Welche Risiken stehen ihnen entgegen? **(B2)**

1.1.3 Beschreiben Sie Möglichkeiten, mit Hilfe von datenanalytischen Verfahren Informationen auf neue Weise miteinander zu verknüpfen, die bei der Erhebung und Speicherung der Daten nicht absehbar ist. Erläutern Sie einen konkreten Anwendungsfall für die Versicherungsbranche. **(C3)**

2 Informationstechnologie

2.1 Entwicklungsmethoden

Die Art und Weise, wie in der IT fachliche Anforderungen aufgenommen und umgesetzt werden, ist seit einigen Jahren im Wandel. Neben der immer noch verbreiteten klassischen Methode des geplanten Projektes haben sich sogenannte agile Methoden etabliert.

Zielsetzung: Die Kandidaten wissen, was sich hinter dem Begriff „Agilität" verbirgt, und kennen die wichtigsten agilen Vorgehensmodelle.

<u>Entwicklungsmethoden</u>

2.1.1 Benennen Sie verschiedene Organisationsmethoden in der IT-Entwicklung, analysieren Sie diese und grenzen sie gegeneinander ab. **(D4)**

Bemerkung: Methoden sind z. B. die klassische Wasserfall-Methode und Scrum in der Projektorganisation sowie der Design-Thinking-Prozess.

2.1.2 Benennen Sie die unterschiedlichen Ebenen des Testens von Anwendungssystemen. **(C1)**

<u>Verteilte Verarbeitungsmethoden</u>

2.1.3 Nennen Sie die verbreitetsten Ansätze zur verteilten Verarbeitung und Analyse großer Datenbestände und deren wesentliche Konzepte. **(C1)**

2.1.4 Skizzieren Sie die Funktionsweise des MapReduce-Algorithmus und seiner Vor- und Nachteile. **(C5)**

2.2 Kodierungstheorie

Kenntnisse der Kodierungstheorie bilden die Grundlage für den sicheren und effizienten Austausch von schutzwürdigen Daten über beliebige Medien und Kanäle.

Zielsetzung: Die Kandidaten verstehen die Grundzüge der Kodierungstheorie, haben Einblick in die wesentlichen Funktionsweisen sicheren Datenaustauschs sowie ihre Anwendbarkeit im Versicherungskontext.

2.2.1 Erläutern Sie jeweils ein Beispiel für reversible bzw. irreversible (sog. Hashing) Verschlüsselungsalgorithmen und ihren Nutzen im Versicherungskontext. **(C2)**

2.2.2 Analysieren Sie die rechtlichen und technischen Rahmenbedingungen, innerhalb derer Verschlüsselung zu Datensicherung und -austausch verwendet wird und welche, beispielsweise aus der Komplexitätstheorie erwachsenden, Abwägungen zwischen Sicherheit und Praktikabilität zu treffen sind. **(D4)**

2.2.3 Skizzieren und analysieren Sie ein Protokoll zur sicheren Übertragung einer binären Information zwischen zwei nicht vertrauenswürdigen Parteien (sog. Bit Commitment). **(C4)**

2.2.4 Nennen Sie Beispiele in der Versicherungsbranche, deren Geschäftsmodell ohne wirksame Verschlüsselung aus regulatorischen, formalen oder faktischen Gründen hinfällig wäre. **(C1)**

2.2.5 Geben Sie eine Übersicht über die konzeptuellen Grundlagen der Blockchain-Technologie, also Kettenlängen, Hashing, Verifizierbarkeit etc. **(C2)**

3 Tools & Programme

3.1 Big Data Analytics mit Apache Spark

Apache Spark ist in den letzten Jahren zu einem der wichtigsten Werkzeuge für verteiltes in-memory Rechnen geworden. Der Grund dafür liegt in dessen Open Source Charakter, den Schnittstellen zu mehreren Programmiersprachen und der relativ einfachen Bedienbarkeit. Spark wird von allen großen Cloud-Anbietern gehostet, sodass man die Clustergröße je nach Anwendung und Datenmenge konfigurieren und beliebig skalierbar halten kann.

Zielsetzung: Verständnis der Funktionsweise der Jobverteilung und der in-memory Berechnung von Spark. In der Lage sein, mit Spark SQL und einer der Programmierschnittstellen auf Dataframes zuzugreifen, diese zu verarbeiten und mit Spark MLib Machine Learning anzuwenden.

3.1.1 Warum spricht man bei Spark von einer DAG-Berechnungs-Engine? Erläutern Sie in diesem Zusammenhang die Begriffe „Lazy Evaluation", „Transformation" und „Action". **(B2)**

3.1.2 Benennen Sie Hauptbibliotheken von Spark und erläutern Sie deren Anwendungsgebiete jeweils kurz an einem Beispiel. **(B2)**

3.1.3 Analysieren und erläutern Sie die wichtigsten Repräsentationen für strukturierte tabellarische Daten in Spark (Data Collections). **(B4)**

4 Mathematik / Statistik

4.1 Korrelation & Kausale Inferenz

In der Praxis ist eine trennscharfe Abgrenzung zwischen Korrelation und kausalen Zusammenhängen unerlässlich.

Zielsetzung: Die Kandidaten lernen den Unterschied zwischen rein empirisch beobachteten und tatsächlich kausalen Zusammenhängen und können diese in Anwendungsbeispielen voneinander abgrenzen.

4.1.1 Nennen und beschreiben Sie verschiedene Techniken des probabilistischen Schließens. Geben Sie ein konkretes Anwendungsbeispiel einer der beiden Techniken in der Versicherungsmathematik. **(C2)**

4.1.2 Stellen Sie Probleme dar, die bei der Arbeit mit beobachteten Daten auftreten. **(B2)**

4.1.3 Erklären Sie die Begriffe „Survival Bias", „Outcome Bias", „Omitted-Variable Bias" und „Alternative Blindness" jeweils anhand eines konkreten Beispiels. **(C2)**

4.1.4 Grenzen Sie ausgehend von den philosophischen Grundlagen die Begriffe „Kausalität" und „Korrelation" voneinander ab. **(B2)**

4.1.5 Erläutern Sie die grundlegenden Prinzipien einer Kausalordnung und beschreiben Sie Grenzen der Kausalität. **(C2)**

4.1.6 Beschreiben Sie den modernen mathematischen Ansatz nach Pearl zur Beschreibung von Kausalität. **(C2)**

4.1.7 Erklären Sie den Begriff eines „Causal Experiments" und nennen Sie Techniken, die zum Design eines solchen Experiments verwendet werden können. **(C2)**

4.2 Deep Learning 2: Spezielle Anwendungsfälle tiefer neuronaler Netze

Neuronale Netze haben sich auch für Anwendungen bewährt, die über das klassische überwachte Lernen hinausgehen.

Zielsetzung: Die Kandidaten verstehen die Funktionsweisen und Anwendungsbereiche von rekurrenten neuronalen Netzen und Autoencodern. Sie sind in der Lage, diese speziellen neuronalen Netze auf Bilder, Texte und strukturierte Daten anzuwenden.

4.2.1 Erläutern Sie die grundlegende Idee eines Gated Recurrent Neural Network. (B2)

4.2.2 4.2.2 Skizzieren Sie den Aufbau einer rekurrenten Zelle. (B2)

4.2.3 Geben Sie einen vergleichenden Überblick über die Anwendungsmöglichkeiten eines Autoencoders (z.B. Dimensionsreduktion, Anomalie-Erkennung, Denoising). (B5)

4.2.4 Erläutern Sie das Sicherheitsrisiko einer Adversarial Attack. Wie kann man ihm begegnen? (B5)

5 Insurance Analytics

5.1 Textmining

Die Organisation von Dokumenten, unter anderem im Zusammenhang mit der Kundenkommunikation, gehört zu den originären Aufgaben von Versicherungsunternehmen. Die Möglichkeiten, die das Textmining als Anwendungsgebiet der Data Science bietet, sind dabei sehr vielseitig und können diverse Potentiale heben bzw. Prozesse verschlanken. Diverse Versicherer nutzen diese Möglichkeiten bereits und haben beispielsweise den Mehrwehrt der Sentimentanalyse für ihr Unternehmen bewiesen.

Zielsetzung: Die Kandidaten kennen analytische Modelle des Textmining und können die Vorgehensweise zur Analyse von Texten beschreiben. Es können Anwendungsfälle aus der Versicherungswirtschaft und Lösungsansätze beschrieben werden. Darüber hinaus sollen Textmining Tools bekannt und deren Umgang exemplarisch geübt worden sein.

Anwendungen

5.1.1 Beschreiben Sie typische Text-Mining-Anwendungen im Versicherungsunternehmen und zeigen Sie Möglichkeiten und Grenzen auf. **(B4)**

Techniken

5.1.2 Benennen Sie ein klassisches Relevanzmaß für Textdokumente bezüglich einer Suchanfrage. Erörtern Sie die praktische Relevanz klassischer Ranking-Metriken im Lichte der Fortschritte im NLP. **(C2)**

Bemerkung: Das Themenfeld besteht u.a. aus Volltextsuche sowie den klassischen Metriken BM25 / TFIDF.

5.1.3 Erläutern Sie die grundlegende Funktionsweise von Word-Vektoren. Nennen Sie (ein oder zwei) Einschränkungen und Ansätze, diese zu überwinden. **(C2)**

Bemerkung: Relevant sind die Repräsentation von Text-Word-Vektoren, Subword-Vektoren, Höhere Darstellungen mit Sprachmodellen, Bag-of-Words.

5.1.4 Erläutern Sie eine Text-Mining-Anwendung, der ein Unsupervised Learning-Verfahren zugrunde liegt. **(A2)**

5.1.5 Beschreiben Sie die Verarbeitungsschritte eines Textes, bevor er in einem RNN verarbeitet werden kann. **(C2)**

Bemerkung: Das Themenfeld besteht aus Supervised Learning-Anwendungen wie Sentiment Analysis / Opinion Mining, Intent Classification.

5.1.6 Ordnen Sie das Vorhersageproblem im Opinion Mining in die Anwendungen Neuronaler Netze ein. **(C3)**

Bemerkung: Auch hier sind Supervised Learning-Anwendungen wie Sentiment Analysis / Opinion Mining, Intent Classification gemeint.

5.1.7 Beschreiben Sie die grundlegenden Komponenten einer beispielhaften Chat-Bot-Architektur. **(B2)**

Tools

5.1.8 Sie sollen eine beispielhafte Sentiment-Analyse programmieren, welche Tools verwenden Sie? (C3)

Bemerkung: Aktuell würden sich Tools aus dem Themenfeld Natural Language Toolkit in Python, Word2Vec, NLTK, GenSim, Sentencepiece, Spacy, AllenNLP, Solr anbieten.

5.2 Anomaly Detection

Versicherungen aus allen Sparten haben bei der Prüfung ihrer Schäden mit dem Umgang und dem Erkennen von Betrugsfällen zu kämpfen, dies reicht von der Prüfung einzelner Rechnungen in der Krankenversicherung bis hin zur Bearbeitung von Leistungsfällen in der Lebens- und Unfallversicherung. Um Unregelmäßigkeiten in Daten zu finden, sind Expertenwissen und langjährige Erfahrung notwendig. Umso wichtiger ist es, moderne analytische Verfahren zur Erkennung von Anomalien einzusetzen. Die Anwendungsszenarien in Versicherungen beschränken sich nicht nur auf die Betrugserkennung und sind sehr vielschichtig.

Zielsetzung: Die verschiedenen Methoden zur Erkennung von Ausreißern können nach Art der vorliegenden Daten benannt und ihre Funktionsweise mathematisch beschrieben werden. Aktuelle Beispiele aus der Versicherungswirtschaft und Lösungsansätze für diese können skizziert werden.

5.2.1 Erläutern Sie, wie man durch Analyse von einfachen Ausreißern im Vorfeld der Modellierung bereits Aussagen zur Qualität der Analyse machen kann. **(B2)**

5.2.2 Grenzen Sie die Methoden des überwachten und unüberwachten Lernens im Kontext der Anomaly Detection ab. Nennen Sie Anwendungen aus der Versicherungswirtschaft. **(B2)**

5.2.3 Erklären Sie im Zusammenhang der Anomaly Detection den Begriff des Rauschens und unterscheiden Sie zwischen dem statistischen und systematischen Rauschen. Erklären Sie dabei den Einfluss auf die Analyse von Ausreißern. **(B2)**

5.2.4 Nennen und vergleichen Sie zwei Verfahren des unüberwachten Lernens im Kontext der Anomaly Detection. **(B4)**

5.2.5 Beschreiben Sie wie Hidden Markov Models, Support Vector Machines und Time Series Analytics zur Erkennung anormaler Zusammenhänge genutzt werden können. Stellen Sie die Modelle und deren Annahmen vor. **(C2)**

5.3 Interpretation (von Modellen und Ergebnissen)

Die Interpretation und Interpretationsfähigkeit von Modellen und Ergebnissen ist ein wichtiger Faktor für die Akzeptanz und die operative Anwendbarkeit von Data Science Verfahren.

Zielsetzung: Die Kandidaten sind in der Lage, Modelle und Ergebnisse unterschiedlicher Komplexität zu interpretieren. Hierbei können sie anwendungsfallbezogen bewerten und einschätzen, wie hoch die Komplexität und Interpretationsfähigkeit von Modellen sein soll.

5.3.1 Beschreiben Sie regulatorische und operative Anforderungen bezüglich der Interpretation und Nachvollziehbarkeit von Modellen innerhalb der Versicherungswirtschaft. **(B2)**

5.3.2 Erläutern Sie die unterschiedlichen Komplexitätsgrade in der Erstellung und der Interpretation von Modellen. **(B2)**

5.3.3 Erklären Sie die Bedeutung von Akzeptanz und dem Verständnis für Modelle in der (operativen) Anwendung von komplexen Modellen. **(B2)**

5.3.4 Erläutern Sie das Kosten-Nutzen-Verhältnis von steigender Komplexität in der Erstellung von Modellen und der Interpretation von Modellergebnissen. **(B2)**

5.3.5 Benennen und erläutern Sie Methoden (wie Surrogate Model, LIME, Maximum Activation Model, Variable Importance Measure, Shapley Value Explanation) zur Veranschaulichung und Erläuterung von Verfahren und Ergebnissen aus komplexen Modellen. **(B2)**

5.3.6 Beurteilen Sie für Anwendungsfälle aus der Versicherungswirtschaft, welche Modelle mit welchem Komplexitätsgrad anzuwenden sind, um ein optimales Verhältnis zwischen Genauigkeit und Interpretationsfähigkeit zu erhalten. **(B5)**

6 Business Cases

Die konkrete Anwendung der erlernten Methoden und Verfahren auf komplexe Fragestellungen aus der Versicherung sind von zentraler Bedeutung für die Arbeit eines Actuarial Data Scientists (ADS).

Zielsetzung: Die Kandidaten sind in der Lage, komplexe und umfängliche Data Science Analysen sowie anspruchsvolle Anwendungen des maschinellen Lernens selbstständig durchzuführen.

6.1.1 Basierend auf komplexen Fragestellungen und gegebenen Datenbeständen führen Sie Data Science Analysen selbstständig durch. Dabei durchlaufen Sie alle Phasen eines Data Mining Prozesses und erstellen eine Präsentation mit einem geeigneten Tool. Sie interpretieren und beurteilen die Ergebnisse und präsentieren diese Zielgruppengerecht **(C5)**

6.1.2 Für komplexe Fragestellungen wenden Sie Verfahren des maschinellen Lernens selbstständig an. Sie validieren die Güte der Ergebnisse und verbessern diese schrittweise durch Variation der Hyperparameter und vergleichen die Ergebnisse, die Sie mit verschiedenen anderen Verfahren ermittelt haben. Sie interpretieren und beurteilen die Ergebnisse und präsentieren diese Zielgruppengerecht. **(C5)**