Prüfungsordnung 5.0
Lernziele im Spezialwissen *Actuarial Data Science Immersion*

Inhalt

1 Informationstechnologie

1.1 Datenverarbeitungstechnologien 2

Zielsetzung: Die Kandidaten verstehen den Ablauf eines Map/Reduce-Jobs sowie in Grundzügen z.B. den technischen Ablauf auf einem Hadoop/Spark-Cluster.

1.1.1. Erklären Sie die Konzepte der Funktionalen Programmierung, insbesondere die Map/Filter/Reduce-Funktionale, und wenden Sie sie an konkreten Beispielen an. **(C3)**

1.1.2. Erläutern Sie den algorithmischen Teil des Grundablaufs eines Map/Reduce-Jobs. Diskutieren Sie die Einschränkungen und entwerfen Sie Beispiele, z. B. mit Hilfe von Hadoop. **(C3)**

1.1.3. Geben Sie einen Überblick über die Umsetzung der Job-Steuerung in verteilten Systemen und erläutern Sie, wie Skalierung und Ausfallsicherheit erreicht werden (z. B. anhand von Hadoop). **(B3)**

1.1.4. Nennen Sie wichtige Tools für verteilte Systeme und benennen Sie deren Kernaufgaben (beispielsweise HIVE aus dem Hadoop-Umfeld). **(A1)**

1.2 Informationstheoretische Grundlagen

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

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

1.2.2 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)**

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

1.3 Systemarchitekturen

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

1.3.1 Erläutern und diskutieren Sie die wesentlichen Unterschiede zwischen einer Micro Service-Architektur und der klassischen Softwarearchitektur. **(C5)**

1.3.2 Erklären Sie den Begriff und Grundgedanken der REST-Schnittstelle. **(B2)**

1.3.3 Erläutern Sie die Begriffe „Cloud ready“ und „Cloud Native“. **(B2)**

2 Insurance Analytics

2.1 Data Mining 3

Zielsetzung: Die Kandidaten sollen weitergehende analytische Kenntnisse zu Umgang, der Interpretierbarkeit und den Gefahren bei der Arbeit mit großen Datenmengen vermittelt bekommen. In diesem Zusammenhang werden methodische Kenntnisse zum Daten-Preprocessing und zur 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.

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

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

2.1.3 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. **(B2)**

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

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

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

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

2.2 Visualisierung 2

Zielsetzung: Die Kandidaten sind in der Lage, vertiefende Methoden 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.

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

2.2.2 Erläutern Sie die Konzepte zur Darstellung von Modellergebnissen und interpretieren Sie die Ergebnisse für unterschiedliche Modelle, wie Profit-Kurven, Lift-Kurven, ROC-Graphen und der Confusion-Matrix. **(B4)**

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

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

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

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

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

2.3 Innovative Produkte 2

Zielsetzung: Die Kandidaten sind mit den unterschiedlichen technischen und fachlichen Anforderungen in der Entwicklung von innovativen Produkten vertraut. Sie können die regulatorischen Rahmenbedingungen bei der Produktentwicklung beurteilen und kennen die Vorteile und Möglichkeiten von innovativen Produkten.

2.3.1 Erläutern Sie die rechtlichen Voraussetzungen und Einschränkungen zur Erhebung, Verarbeitung und Speicherung von risiko- und kundenbasierten Daten. **(B2)**

2.3.2 Erläutern Sie die Notwendigkeit zur Erhebung von risikobasierten Daten und die Einschränkungen aufgrund der Kundenakzeptanz. **(B2)**

2.3.3 Beurteilen Sie die technischen und wirtschaftlichen Restriktionen bei der Datenerhebung. **(B5)**

2.3.4 Beurteilen und prüfen Sie die Möglichkeiten und Einschränkungen der Erhebung, Speicherung und Verarbeitung von Daten für Anwendungsfälle aus der Versicherungswirtschaft. **(B5)**

2.3.5 Vergleichen Sie Methoden und Modelle des Data Minings zur Anwendung in der Produktentwicklung von Versicherungen und wählen Sie geeignete Methoden und Modelle aus. **(B4)**

2.3.6 Begründen Sie, warum Methoden und Modelle der „klassischen“ Produktentwicklung nur bedingt geeignet sind, die Anforderungen zur Entwicklung von innovativen Produkten zu erfüllen. **(B5)**

2.3.7 Benennen Sie regulatorische und aktuarielle Anforderungen zur Produktentwicklung in verschiedenen Sparten eines Versicherungsunternehmens. **(B1)**

2.3.8 Bewerten Sie Analysemethoden und Modelle exemplarisch kritisch hinsichtlich der folgenden Anforderungen: Wiederholbarkeit der Analyse, Nachvollziehbarkeit und Dokumentation der Analyse, Einhaltung von versicherungsrechtlichen Gleichbehandlungsgesetzen, Einhaltung von Datenschutzrichtlinien und gesetzlichen Vorgaben, Darlegung von statistisch signifikanten und validen Ergebnissen bzw. Risikogruppen und Möglichkeit zur Berechnung des Prognoserisikos. **(B5)**

2.3.9 Diskutieren Sie die Einhaltung von regulatorischen und versicherungstechnischen Standards in der Entwicklung von innovativen Produkten für exemplarische Anwendungsfälle. **(B4)**

2.3.10 Benennen Sie Anwendungsfälle von innovativen Produkten und Kundenmanagementprojekten in der Versicherungswirtschaft. Erläutern Sie die Funktionsweise und Eigenschaften der Produkte und Projekte anhand von folgenden Eigenschaften: Art und Umfang der Erhebung und Verarbeitung von Daten, Anwendung von Methoden und Modellen und Umsetzung von regulatorischen und rechtlichen Standards. **(B2)**

2.3.11 Vergleichen Sie „klassische“ mit „innovativen“ Produkten der Versicherungswirtschaft in verschiedenen Versicherungssparten. Beurteilen und begründen Sie hierbei die möglichen Vor- und Nachteile von innovativen Produkten. **(B5)**

3 Mathematik / Statistik

3.1 Unüberwachtes Lernen 2

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

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

3.1.2 Diskutieren Sie die Algorithmen des k-means, k-modes und k-prototypes kritisch im Hinblick auf die zu wählenden Parameter, Interpretation der Ergebnisse und Komplexität. **(C5)**

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

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

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

3.1.6 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)**

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

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

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

3.1.10 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)**

3.2 Deep Learning 3

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.

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

3.2.2 Skizzieren Sie den Aufbau einer rekurrenten Zelle. Erläutern Sie die Vorteile am Beispiel eines LSTM-Netzes (Long-Short-Term-Memory). **(B2)**

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

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

3.3 Anonymisierung / Pseudonymisierung 1

Zielsetzung: Die Kandidaten kennen die Begriffe der Anonymisierung und Pseudonymisierung und verstehen deren Notwendigkeit im Kontext der Gesetzgebung. Erste Methoden können definiert, beispielhaft angewendet und bewertet werden.

3.3.1 Erklären Sie die Begriffe der Anonymisierung und Pseudonymisierung aus Sicht der Datenschutzgrundverordnung und grenzen Sie diese voneinander ab. Erklären Sie, welche Anforderungen es an Verfahren der Anonymisierung gibt. **(B2)**

3.3.2 Zeigen Sie anhand eines Praxisbeispiels, warum die Anonymisierung und Pseudonymisierung in der Versicherungswirtschaft relevant sind. **(B3)**

3.3.3 Klassifizieren und erklären Sie grundlegende Verfahren der Anonymisierung und Pseudonymisierung von strukturierten Daten. **(B2)**

3.3.4 Wenden Sie grundlegende Verfahren der Anonymisierung und Pseudonymisierung jeweils auf Beispiele mit Bezug aus der Versicherungswirtschaft an. **(C3)**

3.3.5 Bewerten Sie die einzelnen Verfahren im Hinblick auf die Eignung zur Anonymisierung bzw. Pseudonymisierung von Daten und vergleichen Sie diese. **(B5)**

3.4 Modellselektion & Regularisierung

Zielsetzung: Die Kandidaten verstehen die Notwendigkeit der Dimensionsreduktion in der Modellbildung und haben einen Überblick über die wichtigsten Verfahren zur Parameter- bzw. Modellselektion wie Shrinkage, Early Stopping, Drop-Out etc. Sie sind in der Lage, eine strukturkonforme Modellauswahl zu treffen und wissen um die Grenzen der spezifischen Gütemaße.

3.4.1 Erläutern Sie verschiedene Konzepte zur (semi-) automatisierten Modellbildung z. B. über Stepwise Regression oder über Kennzahlen zur Modellgüte. **(C2)**

3.4.2 Identifizieren Sie laufzeitintensive Problemstellungen, die parallelisierbar sind und schätzen Sie die Auswirkungen und Vorteile einer Parallelisierung der Rechenschritte ab. **(C2)**

3.4.3 Nennen Sie die Konzepte zur Auswahl von signifikanten Merkmalen (Signifikanztests, AIC, BIC, etc. und weitere Kennzahlen zur Prädiktionsgüte von Merkmalen) und setzen Sie sie in Anwendungsfällen um. **(C3)**

3.4.4 Nennen Sie Shrinkage-Methoden (z. B. Ridge und Lasso) und interpretieren Sie die Ergebnisse. **(B3)**

3.4.5 Nennen und diskutieren Sie Methoden zur Reduktion der Dimensionen (Hauptkomponentenanalyse, partielle kleinste Quadrate etc.) und interpretieren Sie die Ergebnisse. **(B3)**

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

3.4.7 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)**

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

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

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

4 Tools & Programme

4.1 Big Data Analytics

Zielsetzung: Die Kandidaten haben ein Verständnis über die Funktionsweise der Jobverteilung und der in-memory Berechnung von verteilten Systemen. Sie sind in der Lage, mit einem verteilten System und einer der Programmierschnittstellen auf verteilte Daten zuzugreifen, diese zu verarbeiten und mit Machine Learning-Methoden zu bearbeiten.

4.1.1 Erläutern Sie die Begriffe „Ausführungsgraph“, „Lazy Evaluation“, „Transformation“ und „Action“ an einem konkreten Beispiel wie etwa mit Spark mit der DAG-Berechnungs-Engine. **(B2)**

4.1.2 Benennen Sie für ein konkretes System die Hauptbibliotheken und erläutern Sie deren Anwendungsgebiete jeweils kurz an einem Beispiel. **(B2)**

4.1.3 Analysieren und erläutern Sie die wichtigsten Repräsentationen für strukturierte tabellarische Daten für ein konkretes Beispiel wie etwa Data Collections in Spark. **(B4)**

5 Use Cases

5.1 Use Case

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

5.1.1 Basierend auf einer anspruchsvollen Fragestellung und einem gegebenen Datenbestand führen Sie eine Data Science-Analyse selbstständig durch. Dabei durchlaufen Sie alle Phasen eines Data Mining-Prozesses und dokumentieren den Prozess und das Ergebnis in geeigneter Form, z. B. in einem Notebook unter Beachtung üblicher Programmierstandards (Wartung, Zuverlässigkeit, Effizienz, Benutzerfreundlichkeit). Sie wenden Modelle an, interpretieren und beurteilen die Ergebnisse und präsentieren diese zielgruppengerecht. **(C5)**