european actuarial academy

EAA Certificate in Actuarial Data Science

Learning objectives

September 2023

european actuarial academy

Contents

Note: Behind each learning objective you will see the assignment according to Bloom's taxonomy

european actuarial academy

Subject: Actuarial Data Science Basic

1 ADS basics & environment

1.1 Basics

Objective: Candidates are familiar with the basic concepts in the field of Actuarial Data Science and can categorise the underlying data appropriately.

1.1.1. Explain the key terms in the context of data science, such as data engineering, data mining, digitalisation, big data, machine learning and artificial intelligence, taking into account their historical development, and differentiate between them. **(B2)**

1.1.2. Explain the idea of artificial intelligence. Discuss general possible applications in the insurance industry. **(B2)**

1.1.3. Explain the idea of machine learning and the fundamental differences to classical programming. **(C3)**

1.1.4. Explain the most important categories of machine learning and differentiate between them. **(C3)**

1.1.5. Describe the impact of the development of actuarial data science on the insurance industry as a whole, on an insurance company and on the actuarial profession in particular. **(C3)**

1.1.6. Explain the differences between structured, unstructured and semi-structured data and the significance of this distinction in the context of data science. **(B2)**

1.2 Digitisation

Objective: Candidates are familiar with the key topics of digitalisation and can assess and evaluate the impact on the insurance industry.

1.2.1. Name and explain the key terms, topics and technologies in the context of digitalisation in the insurance environment. **(B2)**

1.2.2. Name and explain possible applications of data science and machine learning in insurance. Give examples for each line of business and explain the procedure. **(B2)**

1.2.3. Explain the effects and requirements of digitalisation on the application landscape and data processing in an insurance company. **(B2)**

1.2.4. Using specific examples, analyse and discuss the effects of digitalisation on the business model and business processes of the insurance industry and derive possible effects on the tasks of the actuary. **(D5)**

1.2.5. Explain the importance and tasks of associations (e.g. GDV, DAV) in the context of digitalisation and the use of artificial intelligence. **(B2)**

1.3 Social environment & ethics 1

Objective: Candidates are familiar with the most important arguments and reservations against data science and are sensitised to ethical issues in the context of data science.

1.3.1. Using suitable examples, explain how the application of data science can have a positive or negative impact on marketing. **(B2)**

1.3.2. Using a suitable example, explain the extent to which the application of data science can harbour a reputational risk. **(B2)**

1.3.3. Explain the regulatory system in connection with the use of artificial intelligence in the insurance industry. Name the relevant legal sources. **(B2)**

1.3.4. Explain the ethical problems that can arise from the use of artificial intelligence procedures in the insurance industry. In addition, explain the possible effects on society. **(B2)**

1.3.5. Decide on a case-by-case basis whether an actuarial data science application harbours ethical problems and assess it with regard to regulatory requirements. **(B4)**

1.4 Data protection 1

Objective: Candidates are familiar with the legal system and the most important data protection regulations in Germany and the European Union and can apply these to simple case studies or assess when further legal advice should be sought.

1.4.1. Explain the legal system in Germany in the context of data protection. Name and explain the relevant legal sources in Germany and the European Union. **(B2)**

1.4.2. Name and explain the principles of the EU General Data Protection Regulation and relate them to the other sources of law. **(B3)**

1.4.3. Name and explain the central principles of the Code of Conduct - Data Protection for the Insurance Industry. Apply these principles to the processing of data in insurance companies. **(B3)**

1.4.4. Apply the German and European data protection regulations in concrete case

studies. **(B3)**

1.4.5. Use specific business cases to decide whether they are prohibited, permitted or a borderline case. **(B5)**

2 Information Technology

2.1 Data management 1

Objective: The candidates know the most important concepts used to store and transfer data in file systems. They can evaluate the advantages and disadvantages of various options on a case-by-case basis.

2.1.1. Give an overview of the most important aspects and technologies of data storage and explain their advantages and disadvantages. Discuss the applicability of the technologies in different scenarios. **(B2)**

2.1.2. Name and define the most common (open) data formats for storing data in files and discuss their specific applications. Apply data transformations in specific example cases. **(C3)**

2.2 Data processing technologies 1

Objective: Candidates will be familiar with important standards, tools and practices that contribute to reliable and efficient code development in the field of analytics.

2.2.1. Explain the quality assurance and documentation tasks of software tests and explain the paradigm of "Test Driven Development". Describe how the execution of tests is supported by software platforms (e.g. JUnit, Test-Runner) and develop criteria that support the effective testability of codes. **(C2)**

2.2.2. Name the core functionalities that data science platforms should provide. **(A1)**

2.2.3. Explain the most important processing steps in data preparation and demonstrate them using examples. **(B3)**

2.2.4. Define the terms horizontal and vertical scalability. Explain, classify and delimit different concepts for the implementation of parallel data processing in data science applications such as multiprocessing /-threading, parallel collections, message-based systems and GPU computing. **(C2)**

2.2.5. Explain the concept of "reproducible research" and explain the implementation using the example of so-called "notebook documents". **(B2)**

2.3 Information processing in insurance companies

Objective: Candidates know the core processes of an insurance company and the components of the system landscape in which they are implemented. They know the key criteria for assessing data quality and can apply them.

2.3.1. Name the most important business processes in an insurance company. Discuss for which business processes data science applications are useful. Give examples of data science applications and assign them to a business process. **(C2)**

2.3.2. Name the most important application systems in an insurance company and explain how processes are mapped to the application systems. Explain the functional components of an overall architecture (business processes, architecture, software architecture, software components). **(C2)**

2.3.3. Explain typical existing software architectures in insurance companies and the typical requirements such as auditability, maintainability, scalability, flexibility and affordability. **(A2)**

2.3.4. Explain the importance and structure of portfolio management and its integration into the system landscape of an insurance company. **(B2)**

2.3.5. Explain the importance of the policy administration system as a data source for data science applications. Name other systems in an insurance company that can be important as a data source for data science applications. **(C2)**

2.3.6. Define the term ”data quality”. Name and explain criteria for checking data quality and use them to analyse data quality in specific data science applications. **(B4)**

2.3.7. Name measures to improve data quality and apply them in specific examples. **(B4)**

3 Insurance Analytics

3.1 Data Mining 1

Objective: Candidates are familiar with the basic concepts and fundamental methods of data mining and are able to prepare and analyse data as the basis of the data mining process.

3.1.1. Explain the term data mining in general and in the context of the insurance industry. **(B2)**

3.1.2. Describe the characteristics and possible applications of descriptive, predictive

and prescriptive analyses in data mining. **(C2)**

3.1.3. Apply descriptive, predictive and prescriptive analyses to examples of problems in the insurance environment. **(C3)**

3.1.4. Explain the characteristics and differences between supervised and unsupervised learning. Name exemplary applications in insurance companies. **(C2)**

3.1.5. Name and explain the basic application classes of data mining. **(B2)**

3.1.6. Name and classify the characteristics of data and categorise data according to type. **(B4)**

3.1.7. Explain the characteristics and differences of dependent and independent variables (target and covariates) and categorise data according to the types of variables. **(B4)**

3.1.8. Outline the selection and integration of different data sources to provide a database for data mining. Describe the possible problems that arise and explain possible solutions. **(C2)**

3.1.9. Explain the causes and the need for data cleansing, data reduction and data transformation as the basis for data mining. **(B2)**

3.1.10. Describe methods for cleansing data, reducing data and transforming data. **(C2)**

3.2 Analytics 1

Objective: Candidates are familiar with the concepts and (basic) methods of modelling and can select and apply methods and procedures for generating models.

3.2.1. Explain the functionality and possible applications of (basic) machine learning and statistical learning methods. **(B2)**

3.2.2. Explain the properties of modelling to predict decisions, rankings or to determine estimated values. **(B2)**

3.2.3. Describe a procedure with (analytical) models for tasks from the insurance industry and determine the objectives of modelling and forecasting. **(B2)**

3.2.4. Explain the necessity and benefits of selecting informative attributes for modelling. **(B2)**

3.2.5. Explain how redundant and irrelevant attributes are identified and excluded for the creation of a predictive model. **(B2)**

3.2.6. Determine parameters for the prediction quality of the attributes for exemplary use cases and select informative attributes on this basis. **(B4)**

3.2.7. Explain the model fitting process. **(C2)**

3.2.8. Describe various methods from machine and statistical learning for determining model parameters and model structures. **(B2)**

3.2.9. Apply procedures to create predictive models. **(B3)**

3.2.10. Explain the concept (techniques and objectives) for separating data into training, test and validation data. **(B2)**

3.2.11. Explain the options and concepts for assessing the goodness of fit of models, model performance and checking model assumptions. **(B2)**

4 Mathematics / Statistics

4.1 Supervised learning 1

Objectives: Candidates understand the basic concepts of supervised machine learning, have an overview of the most important methods and can apply them appropriately, especially for classification and regression problems.

4.1.1. Describe the algorithmic functioning of decision trees (classification and regression trees) and different split criteria. **(B2)**

4.1.2. Explain the methodology for extended tree methods such as random forest and bagging. **(B2)**

4.1.3. Explain the principle of (stochastic) gradient boosting. **(B2)**

4.2 Unsupervised learning 1

Objective: Building on the basic knowledge of applied stochastics, candidates acquire initial knowledge of unsupervised learning methods. In particular, they will be able to interpret the results of common data mining software packages and critically examine the modelling.

4.2.1. Describe basic methods of cluster analysis. **(B2)**

4.2.2. Give an overview of the methods used in unsupervised learning and describe the basic solution approach of the respective algorithm. **(C2)**

4.3 Deep Learning 1

Objective: Candidates understand the basic functioning of neural networks and are able to apply neural networks to classification and regression problems with structured data and images.

4.3.1. Define the term "artificial neural network" mathematically. Describe the basic structure of a feedforward neural network. **(B2)**

4.3.2. Present possible activation functions for neurons formally and graphically. Analyse and discuss their advantages and disadvantages. **(B4)**

4.3.3. Explain the function and intuition behind a so-called bias neuron. **(B2)**

4.3.4. Describe the basic idea and procedure of the backpropagation algorithm and apply it to a simple neural network. **(C3)**

4.3.5. Explain how a feedforward artificial neural network is trained or how it "learns". **(B2)**

5 Tools & Programmes

5.1 Introduction and overview

Objective: Candidates receive an overview of the most important data science tools and programmes.

5.1.1. Name the most important programming languages and program libraries for data science. **(A1)**

5.1.2. Explain how notebooks work. Discuss the advantages of using them in data science projects. **(B5)**

5.1.3. Name common tools and software to support data science applications. **(A4)**

5.2 Cross-language data science tools

Objective: Candidates learn about code-free and cross-language tools.

5.2.1. Name different cross-language gradient boosting tools, describe the differences, name your favourite choice and give reasons for it. **(B3)**

5.2.2. Describe how deep learning methods can be implemented with state of the art frameworks. **(B2)**

5.2.3. Name and describe tools that are operated via a user interface and the extent

to which code (especially Python, R) can be integrated into them. **(B2)**

5.2.4. Name and describe tools that can be used via a user interface as well as via R and Python Application Programming Interface. **(B2)**

5.3 Programming languages for data science

Objective: Candidates are able to carry out data science projects independently using common languages and are familiar with their most important language elements and libraries.

5.3.1. Name and explain the most important language elements of R. **(B2)**

5.3.2. Name and explain the most important R libraries for data preparation and visualisation. **(B2)**

5.3.3. Create notebooks in R to analyse data sets from the insurance environment in compliance with common programming standards (maintenance, reliability, efficiency, user-friendliness). **(C5)**

6 Use Cases

6.1 Use Case

Objective: Candidates are able to carry out data science analyses and machine learning applications independently.

6.1.1. Based on a simple question and a given dataset, you will carry out a data science analysis independently. You will go through all phases of a data mining process and document the process and the result in a suitable form, e.g. in a notebook, taking into account common programming standards (maintenance, reliability, efficiency, user-friendliness). You apply models, interpret and evaluate the results and present them in a manner appropriate to the target group. **(C5)**

Subject: Actuarial Data Science Advanced

1 ADS basics & environment

1.1 Social environment & ethics 2

Objective: Candidates are familiar with the most important reputational risks associated with the use of data science and are familiar with the risks to the insurance company's business model. They know the professional principles of the DAV in the area of data science and are familiar with the regulatory requirements for the use of AI.

1.1.1. Explain the concept of disruption using examples from other industries. **(B2)**

1.1.2. Analyse and discuss specific risks of disruption for the insurance industry. **(C4)**

1.1.3. Name and explain the most important professional principles of the DAV in the field of data science. **(B2)**

1.1.4. Analyse the conformity with the professional principles of the DAV for specific applications in the field of data science. **(B4)**

1.1.5. Explain the regulatory requirements arising from the relevant European regulations, such as the Artificial Intelligence Act, for the insurance industry and apply them to specific use cases. **(C4)**

1.2 Data protection 2

Objective: Candidates are familiar with the basic principles of the central legal sources for data protection in Germany and the European Union and can apply these on a case-by-case basis. The inhomogeneity of data protection regulations worldwide is known and the effects on internationally active companies can be assessed.

1.2.1. Explain the position and implementation of data protection in the insurance industry in a global comparison. Give examples of different data protection regulations. **(B2)**

1.2.2. Assess the possible effects for internationally active insurance companies. **(C5)**

1.2.3. Name and explain the basic principles of the central legal sources for data protection in insurance companies in Germany and the European Union. **(B2)**

1.2.4. Use specific business cases to decide whether they are permitted or prohibited

and what measures need to be taken to ensure that they can be implemented in compliance with data protection regulations. **(B5)**

2 Information Technology

2.1 Data management 2

Objective: Candidates are familiar with the concept of relational databases and the basics of Structured Query Language (SQL). They know important types/representatives from the area of NoSQL databases. They understand the necessity of dispositive data storage for analysis purposes and are familiar with the concepts of data warehouse and data lake.

2.1.1. Explain the most common concepts from the field of relational database systems and apply them in example cases. **(B3)**

2.1.2. Interpret an entity-relationship diagram. **(B2)**

2.1.3. Distinguish between the different types of SQL statements, explain basic operations and apply them to example cases. **(B3)**

2.1.4. Explain the concept of the transaction and the basic principles of ACID compliance. **(C2)**

2.1.5. Define the terms operational and dispositive data management and explain why both are needed in the insurance environment. **(A2)**

2.1.6. Explain the organisation and structure of a data warehouse. Discuss the advantages and disadvantages compared to the traditional standard form in relational database management systems. **(B2)**

2.1.7. Explain the principles of the CAP theorem. **(C2)**

2.1.8. Explain other important types of databases and discuss specific applications. **(B2)**

2.1.9. Explain the concept of the data lake. Differentiate it from traditional relational database management systems (schema on read vs. schema on write) and draw connections to NoSQL technologies. **(B3)**

2.1.10. Explain the concept of distributed file systems, e.g. using the Hadoop Distributed File System. **(B2)**

3 Insurance Analytics

3.1 Data Mining 2

Objective: Candidates are familiar with the data mining process and the results. They can interpret and apply the results of data mining and visualise data within the data mining process.

3.1.1. Name different process models of data mining and outline the individual steps of the process models. **(B1)**

3.1.2. Explain the steps and processes of the data mining process models. Transfer the steps of the data mining process to the process flows of typical applications in the insurance environment. **(B3)**

3.1.3. Describe how the results of data mining can be interpreted and evaluated and explain this using specific applications in insurance. **(C2)**

3.1.4. Give an example of how model results can be integrated into the operational application. **(C2)**

3.2 Analytics 2

Objective: Candidates are familiar with (in-depth) methods for modelling and can assess the applicability of the methods and models. They can apply procedures for recognising and avoiding overfitting.

3.2.1. Explain the functionality, possible applications and practical application of (advanced) methods of machine learning and statistical learning. **(B2)**

3.2.2. Apply (advanced) methods of machine learning and statistical learning to data from the insurance environment. **(C3)**

3.2.3. Compare (advanced) methods of machine learning and statistical learning and explain the advantages and disadvantages of the individual methods with regard to applicability, prerequisites, results and computational effort. **(B4)**

3.2.4. Assess the applicability of modelling methods to actuarial problems. **(B5)**

3.2.5. Explain the terms underfitting, overfitting, generalisation and model complexity. **(B2)**

3.2.6. Describe how overfitting can be recognised during model creation. **(C2)**

3.2.7. Explain concepts and practical application of different methods (such as cross-validation, prediction-actual comparison on validation data, bootstrap sampling, parallel development, backtesting) to recognise and avoid overfitting. **(B2)**

3.2.8. Interpret the results of "Fitting Graphs" and "Learning Graphs" when creating and selecting models. **(B4)**

3.3 Visualisation 1

Objective: Building on the basic knowledge of applied stochastics, candidates can use data visualisation tools to prepare and analyse information for better understanding. Candidates can use data visualisation methods in the run-up to modelling to assess data quality and prepare the data. Candidates can use graphical methods to check the model quality and separation properties of models.

3.3.1. Apply graphical methods to recognise relationships between covariates and target variables (e.g. mosaic plots, effect plots, etc.) and explain their significance for modelling (variable selection, continuous vs. categorical characteristics, classification/grouping of covariates). **(C3)**

3.3.2. Assess the prerequisites and the goodness of fit of models with regard to estimation quality and separation properties by applying quantitative (interpretation of parameters such as coefficients of determination) and graphical methods (e.g. residual plots, lift curves or Lorenz curves) to training and validation data. **(C3)**

3.3.3. Explain basic exploratory procedures for graphically analysing covariates and target variables that allow an assessment of extreme values, outliers and spurious values. **(B2)**

3.4 Innovative products 1

Objective: Candidates are familiar with the requirements of data collection and processing for innovative products and know the advantages and possibilities of innovative products.

3.4.1. Explain the application of data science in the development of "classic" insurance products in various insurance lines. **(C2)**

3.4.2. Describe the importance of data science for the development of "innovative" products that go beyond the application of processes and methods of "classic" product development. **(B2)**

3.4.3. Name possible advantages of innovative insurance products. **(B1)**

3.4.4. Explain the advantages and possibilities of using data science in product development using various examples such as risk differentiation. **(B2)**

3.4.5. Compare the possibility of risk code differentiation for different data bases and,

based on this, assess the possibilities for improving the selection of risks. **(B5)**

3.4.6. Explain the possible influence of customer view models on the design and marketing of innovative products. **(B2)**

3.4.7. Name different data sources as the basis for product development and describe the characteristics of the data. **(B2)**

3.4.8. Explain the technical possibilities and requirements for collecting and processing (individual) risk data. Assess the effects and requirements of processing large, heterogeneous data volumes and heterogeneous data on the product development process. **(B4)**

3.4.9. Describe the requirements for integrating and linking external data with (company) internal data. Link external data with internal data as a basis for the product development of an insurance company for exemplary use cases. **(B3)**

3.4.10. Explain the necessary data basis and the basic procedure for creating a cross-divisional customer view and for managing the customer relationship. **(B2)**

4 Mathematics / Statistics

4.1 Supervised learning 2

Objective: Building on the ADS Basic modules Supervised Learning 1 and Unsupervised Learning 1, candidates deepen their knowledge of the methods of linear and generalised linear models, which can also be additive or mixed.

4.1.1. Describe the basic concepts of multiple linear regression. **(B2)**

4.1.2. Explain the extension of the regression methods to the generalised additive models with spline estimators. **(B2)**

4.1.3. Name extensions of regression methods such as generalised estimating equations, mixed models or non-parametric regression methods and their possible applications. **(B1)**

4.1.4. Explain the estimation and test theory of special generalised linear models used in risk modelling and in the analysis of dichotomous target variables. **(B2)**

4.2 Deep Learning 2

Objective: Candidates understand the advanced concepts of neural networks and are able to apply neural networks to classification and regression problems with structured data as well as images.

4.2.1. Sketch the application of a convolutive filter to a two-dimensional matrix. **(B2)**

4.2.2. Name and describe two regularisation methods for deep neural networks. **(B2)**

4.2.3. Assess in which application scenarios the use of neural networks can be particularly advantageous. **(C5)**

4.2.4. Design ”deep learning pipelines” for problems in the insurance environment. Which network architecture is suitable? Can a pre-trained model be used (transfer learning)? How are the hyperparameters determined? How does the ”learning” of the given problem proceed? How can the quality of the model be assessed? **(C5)**

4.3 Correlation & Causal Inference

Objectives: Candidates learn the difference between purely empirically observed and actual causal relationships and can differentiate between them in application examples.

4.3.1. Name and describe different techniques of probabilistic inference. Give a concrete example of an application of one of the two techniques in actuarial mathematics. **(C2)**

4.3.2. Describe problems that occur when working with observed data. **(B2)**

4.3.3. Explain the terms ”survival bias”, ”outcome bias”, ”omitted-variable bias” and ”alternative blindness” using a specific example. **(C2)**

4.3.4. Based on the philosophical foundations, differentiate between the terms ”causality” and ”correlation”. **(B2)**

4.3.5. Explain the basic principles of a causal order and describe the limits of causality. **(C2)**

4.3.6. Describe Pearl’s modern mathematical approach to describing causality. **(C2)**

4.3.7. Explain the concept of a ”causal experiment” and name techniques that can be used to design such an experiment. **(C2)**

4.3.8. Interpret and analyse the modelling results of the classical methods. **(D2)**

4.4 Data preparation for modelling

Objective: Candidates are familiar with the main data preparation techniques used in modelling. They can interpret and assess the necessity and functionality of the techniques.

4.4.1. Explain methods for replacing missing or incorrect values and procedures for

modelling with missing covariate values. **(B2)**

4.4.2. Describe the concept of dummy variables and implement it in use cases. **(C3)**

4.4.3. Interpret the results of data visualisations and assess whether data derivations (such as class formation, interaction, missing category) make sense. **(C4)**

4.4.4. Describe the functionality and practical application of the selection of informative attributes, including an exemplary explanation of various concepts (such as AIC and BIC). **(B2)**

4.4.5. Explain the common data transformations (polynomial, log, Box-Cox, etc.) and, if a data transformation is necessary for modelling, identify which one is suitable. **(C4)**

4.4.6. Name the basic properties and functionality of resampling methods (e.g. bootstrap) and describe their use in regression and model validation procedures (e.g. cross-validation). **(B4)**

5 Tools & Programmes

5.1 Programming languages for data science

Objective: Candidates are able to carry out data science projects independently using common languages and are familiar with their most important language elements and libraries.

5.1.1. Name and explain the most important language elements of Python. **(B2)**

5.1.2. Name and explain important methods from the scikit-learn library. **(B2)**

5.1.3. Design an analysis process of supervised learning with scikit-learn for a question in the insurance environment. **(C5)**

5.1.4. Name and explain the most important Python libraries for data preparation and visualisation. **(B2)**

5.1.5. Create notebooks in Python to analyse data sets from the insurance environment in compliance with common programming standards (maintenance, reliability, efficiency, user-friendliness). **(C5)**

6 Use Cases

6.1 Use Case

Objective: Candidates are able to carry out data science analyses and machine learning applications independently.

6.1.1. Based on a realistic question and a given set of data, you will carry out a data science analysis independently. You will go through all phases of a data mining process and document the process and the result in a suitable form, e.g. in a notebook, taking into account common programming standards (maintenance, reliability, efficiency, user-friendliness). You apply models, interpret and evaluate the results and present them in a way that is appropriate for the target group. **(C5)**

european actuarial academy

Subject: Actuarial Data Science Immersion

1 Information Technology

1.1 Data processing technologies 2

Objective: Candidates understand the process of a map/reduce job and the basics of the technical process on a Hadoop/Spark cluster, for example.

1.1.1. Explain the concepts of functional programming, in particular the map/filter/reduce functions, and apply them to specific examples. **(C3)**

1.1.2. Explain the algorithmic part of the basic process of a map/reduce job. Discuss the limitations and design examples, e.g. with the help of Hadoop. **(C3)**

1.1.3. Give an overview of the implementation of job control in distributed systems and explain how scaling and fail-safety are achieved (e.g. using Hadoop). **(B3)**

1.1.4. Name important tools for distributed systems and name their core tasks (e.g. HIVE from the Hadoop environment). **(A1)**

1.2 Fundamentals of information theory

Objective: Candidates know and understand the conceptualisation and basic results of theoretical computer science and are aware of their relevance for everyday work.

1.2.1. Describe how a Turing machine works and explain why it is important. **(C2)**

1.2.2. Explain the difference between P-hard and NP-hard problems. Also discuss the P=NP problem and analyse its practical relevance. **(C4)**

1.2.3. Formulate the holding problem and explain the resulting consequences. **(C3)**

1.3 System architectures

Objective: Candidates have an overview of modern software system architectures and can categorise the technologies required for this.

1.3.1. Explain and discuss the main differences between a micro service architecture and the classic software architecture. **(C5)**

1.3.2. Explain the term and basic idea of the REST interface. **(B2)**

1.3.3. Explain the terms ”cloud ready” and ”cloud native”. **(B2)**

2 Insurance Analytics

2.1 Data Mining 3

Objective: Candidates should be taught advanced analytical knowledge of handling, interpretability and the dangers of working with large amounts of data. In this context, methodological knowledge of data preprocessing and dimension reduction will be introduced. In addition to the content on clustering in mathematics and statistics, complex and modern methods should be familiarised with and their application scenarios within the insurance industry should be understood.

2.1.1. Define the essential characteristics of classifiable patterns and explain which formal requirements must be met in order to recognise such patterns. **(B2)**

2.1.2. Explain the basic challenges of unsupervised pattern recognition and outline suitable solutions. **(B2)**

2.1.3. Explain the problems that arise when analysing data sets with many dimensions/features and why these problems do not occur for low dimensional, albeit high volume, data sets. **(B2)**

2.1.4. Give examples of non-Euclidean metrics and their application scenarios in data mining. **(B1)**

2.1.5. Explain how and in what way principal component analysis can be extended to tensorial data and name the advantages and disadvantages of this approach. **(C2)**

2.1.6. Differentiate independent component decomposition from other dimension reduction methods and name an actuarial use case. **(B4)**

2.1.7. Describe an example of a solvable reduction problem and categorise the advantages of the methods with regard to both suitable data structures and other methods. **(C3)**

2.2 Visualisation 2

Objective: Candidates are able to apply in-depth methods for visualising data. They are familiar with various tools and can differentiate between them in scope. When visualising data and results, candidates can continue to apply visualisation rules in a targeted manner.

2.2.1. Describe in-depth presentation options and forms as well as concepts for visualising data in the various activities of a data scientist. A distinction must be made between visualisation in the context of data exploration (e.g. to identify anomalies in data), in the context of model creation and selection (e.g. to evaluate model and prediction quality), and in the context of the presentation and display of findings and results. **(B2)**

2.2.2. Explain the concepts for visualising model results and interpret the results for different models, such as profit curves, lift curves, ROC graphs and the confusion matrix. **(B4)**

2.2.3. Discuss and compare different visualisation methods and describe the possible applications and advantages of the different methods. **(B4)**

2.2.4. Explain display rules and forms for the visualisation of data. **(B2)**

2.2.5. Name and describe ways of making visualisations accessible. **(B2)**

2.2.6. Optimise data visualisations with regard to the comprehensibility and readability of the presentation of data. **(B5)**

2.2.7. For data visualisations, assess whether presentation rules are observed or not. **(B5)**

2.3 Innovative products 2

Objective: Candidates are familiar with the various technical and professional requirements in the development of innovative products. They can assess the regulatory framework for product development and know the advantages and possibilities of innovative products.

2.3.1. Explain the legal requirements and restrictions for the collection, processing and storage of risk- and customer-based data. **(B2)**

2.3.2. Explain the need to collect risk-based data and the limitations due to customer acceptance. **(B2)**

2.3.3. Assess the technical and economic restrictions on data collection. **(B5)**

2.3.4. Assess and check the possibilities and limitations of collecting, storing and

processing data for use cases from the insurance industry. **(B5)**

2.3.5. Compare methods and models of data mining for use in the product development of insurance companies and select suitable methods and models. **(B4)**

2.3.6. Explain why methods and models of "classic" product development are only partially suitable for fulfilling the requirements for the development of innovative products. **(B5)**

2.3.7. Name regulatory and actuarial requirements for product development in different lines of an insurance company. **(B1)**

2.3.8. Critically evaluate analysis methods and models as examples with regard to the following requirements: Repeatability of traceability and documentation of the analysis, compliance with equal treatment laws under insurance legal requirements, compliance with data protection guidelines and law, the analysis, presentation of statistically significant and valid results or risk groups and the possibility of calculating the forecast risk. **(B5)**

2.3.9. Discuss compliance with regulatory and actuarial standards in the development of innovative products for exemplary use cases. **(B4)**

2.3.10. Name use cases of innovative products and customer management projects in the insurance industry. Explain the functionality and characteristics of the products and projects based on the following properties: Type and scope of data collection and processing, application of methods and models and implementation of regulatory and legal standards. **(B2)**

2.3.11. Compare "classic" and "innovative" insurance products in various insurance sectors. Evaluate and justify the possible advantages and disadvantages of innovative products. **(B5)**

3 Mathematics / Statistics

3.1 Unsupervised learning 2

Objectives: Candidates will be able to delineate use cases of unsupervised learning, know the most important methods, apply them to example data and understand the results.

3.1.1. Give an overview of the methods k-means, k-modes and k-prototypes according to the type of variable to be analysed. **(B2)**

3.1.2. Critically discuss the algorithms of k-means, k-modes and k-prototypes with regard to the parameters to be selected, interpretation of the results and complexity. **(C5)**

3.1.3. Explain the terms divisive and agglomerative cluster analysis in the context of hierarchical cluster methods; use the dendrogram for this. Compare the hierarchical clustering methods with k-means. **(B2)**

3.1.4. Explain the principle and basic concepts of density-based clustering. Outline the DBSCAN algorithm and discuss it critically, also with regard to time and memory requirements. **(B5)**

3.1.5. Apply unsupervised machine learning methods to specific questions from the insurance industry. **(C5)**

3.1.6. Explain in which cases (apart from the curse of dimensionality) simple clustering methods such as k-means do not work or only work poorly and name at least conceptual solutions. **(C2)**

3.1.7. Explain the advantages and disadvantages of Ward's Hierarchical Clustering and differentiate it from methods based on graphs, for example. **(C2)**

3.1.8. Describe the concept of MeanShift clustering. **(B2)**

3.1.9. Give an example of which problem class can be clustered with Markov Chain Monte Carlo. **(B1)**

3.1.10. Describe the algorithms for the above methods using a simple example (to be solved manually). Interpret the results. Which examples from insurance practice could be simplified with the help of clustering algorithms? **(B2)**

3.2 Deep Learning 3

Objective: Candidates understand the functionalities and application areas of recurrent neural networks and autoencoders. They are able to apply these specialised neural networks to images, texts and structured data.

3.2.1. Explain the basic idea of a gated recurrent neural network. **(B2)**

3.2.2. Sketch the structure of a recurrent cell. Explain the advantages using the example of an LSTM network (Long Short Term Memory). **(B2)**

3.2.3. Give a comparative overview of the possible applications of an autoencoder (e.g. dimension reduction, anomaly detection, denoising). **(B5)**

3.2.4. Explain the security risk of an adversarial attack. How can it be countered? **(B5)**

3.3 Anonymisation / pseudonymisation 1

Objective: Candidates know the terms anonymisation and pseudonymisation and understand their necessity in the context of legislation. Initial methods can be defined, applied as examples and evaluated.

3.3.1. Explain the terms anonymisation and pseudonymisation from the perspective of the General Data Protection Regulation and differentiate between them. Explain the requirements for anonymisation procedures. **(B2)**

3.3.2. Use a practical example to show why anonymisation and pseudonymisation are relevant in the insurance industry. **(B3)**

3.3.3. Classify and explain basic methods of anonymisation and pseudonymisation of structured data. **(B2)**

3.3.4. Apply basic anonymisation and pseudonymisation procedures to examples from the insurance industry. **(C3)**

3.3.5. Evaluate the individual methods with regard to their suitability for anonymising or pseudonymising data and compare them. **(B5)**

3.4 Model selection & regularisation

Objectives: Candidates understand the necessity of dimension reduction in modelling and have an overview of the most important methods for parameter and model selection such as shrinkage, early stopping, drop-out etc. They are able to make a structurally compliant model selection and are aware of the limits of specific quality measures.

3.4.1. Explain various concepts for (semi-) automated modelling e.g. via stepwise regression or via key figures on model quality. **(C2)**

3.4.2. Identify runtime-intensive problems that can be parallelised and estimate the effects and advantages of parallelising the calculation steps. **(C2)**

3.4.3. Name the concepts for the selection of significant characteristics (significance tests, AIC, BIC, etc. and other key figures for the predictive quality of characteristics) and implement them in use cases.. **(C3)**

3.4.4. Name shrinkage methods (e.g. ridge and lasso) and interpret the results. **(B3)**

3.4.5. Name and discuss methods for reducing the dimensions (principal component analysis, partial least squares, etc.) and interpret the results. **(B3)**

3.4.6. Name and describe linear and non-linear methods in which regularisation is used. Explain similarities and differences. **(B2)**

3.4.7. Using linear regression as an example, explain the differences between LASSO, Ridge Regression and Elastic Net and outline the advantages of each. Make a judgement as to which method is best suited to which application scenario. **(C5)**

3.4.8. Name suitable starting values and describe procedures for hyperparameter tuning. **(C2)**

3.4.9. Describe the basic ideas of the ”blending” and ”stacking” processes and name the advantages of each. **(B2)**

3.4.10. Name suitable blending calculation methods for classification and regression questions. **(C2)**

4 Tools & Programmes

4.1 Big Data Analytics

Objective: Candidates have an understanding of how job distribution and in-memory computation work in distributed systems. They are able to use a distributed system and one of the programming interfaces to access distributed data, process it and work on it using machine learning methods.

4.1.1. Explain the terms ”execution graph”, ”lazy evaluation”, ”transformation” and ”action” using a concrete example such as Spark with the DAG calculation engine. **(B2)**

4.1.2. Name the main libraries for a specific system and briefly explain their areas of application using an example. **(B2)**

4.1.3. Analyse and explain the most important representations for structured tabular data for a concrete example such as data collections in Spark. **(B4)**

5 Use Cases

5.1 Use Case

Objective: Candidates are able to carry out simple and comprehensive data science analyses and machine learning applications independently.

5.1.1. Based on a challenging question and a given database, you will carry out a data science analysis independently. You will go through all phases of a data mining process and document the process and the result in a suitable form, e.g. in a notebook, taking into account common programming standards (maintenance,

reliability, efficiency, user-friendliness). You apply models, interpret and evaluate the results and present them in a manner appropriate to the target group. **(C5)**

european actuarial academy

Subject: Actuarial Data Science Completion

1 Information technology

1.1 Coding theory

Objective: Candidates understand the basics of coding theory, have an insight into the essential functions of secure data exchange and their applicability in an insurance context.

1.1.1. Explain one example each of reversible and irreversible (so-called hashing) encryption algorithms and their use in the insurance context. **(C2)**

1.1.2. Analyse the legal and technical framework within which encryption is used for data protection and exchange and which trade-offs, for example those arising from complexity theory, must be made between security and practicability. **(D4)**

1.1.3. Outline and analyse a protocol for the secure transmission of binary information between two untrusted parties (so-called bit commitment). **(C4)**

1.1.4. Outline and analyse a protocol for the quantum-safe transmission of information (so-called quantum cryptography). **(C4)**

1.1.5. Give examples in the insurance industry whose business model would be invalid without effective encryption for regulatory, formal or factual reasons. **(C1)**

1.1.6. Provide an overview of the conceptual foundations of blockchain technology, i.e. chain lengths, hashing, verifiability, etc. **(C2)**

1.2 Cloud computing

Objective: Candidates are familiar with the common service and deployment models of the various providers and know the criteria that should be considered when implementing an application in the cloud.

1.2.1. Explain different service models such as Iaas, Paas and SaaS. **(B2)**

1.2.2. Name deployment models such as private cloud, hybrid cloud, community cloud and public cloud and differentiate between them. **(B2)**

1.2.3. Discuss the advantages and disadvantages of implementing processes and calculations in the cloud. **(C4)**

1.2.4. Explain the possibilities and limitations of using cloud computing for actuarial use cases. **(B2)**

1.2.5. Name the most common approaches to distributed processing and analysis of large data sets and their main concepts. **(C1)**

1.2.6. Outline how the Map/Reduce algorithm works and its advantages and disadvantages. **(C5)**

1.3 Development methods

Objective: Candidates know what the term ”agility” means and are familiar with the most important agile process models.

1.3.1. Name different organisational methods in IT development, analyse them and differentiate between them. Methods include, for example, the classic waterfall method and Scrum in project organisation as well as the design thinking process. **(D4)**

1.3.2. Name the different levels of application system testing. **(C1)**

2 Insurance Analytics

2.1 Anomaly Detection

Objective: The various methods for recognising outliers can be named according to the type of data available and their mode of operation can be described mathematically. Current examples from the insurance industry and possible solutions for these can be outlined.

2.1.1. Explain the terms anomaly and outlier and differentiate between them. **(B2)**

2.1.2. Explain how an analysis of simple outliers in the run-up to modelling can be used to make statements about the quality of the data and the impact on the analyses. **(B2)**

2.1.3. Differentiate between the methods of supervised and unsupervised learning in the context of anomaly detection. Name applications from the insurance industry. **(B2)**

2.1.4. Explain the concept of noise in the context of anomaly detection and differentiate between statistical and systematic noise. Explain the influence of noise on the analysis of outliers. **(B2)**

2.1.5. Name and compare two methods of unsupervised learning in the context of anomaly detection. **(B4)**

2.1.6. Name supervised methods (e.g. Hidden Markov Models, Support Vector Ma-

chines and Time Series Analytics) for recognising anomalous correlations and describe how these can be used. Present the models and their assumptions. **(C2)**

2.2 Interpretation (of models and results)

Objective: Candidates are able to interpret models and results of varying complexity. They can evaluate and assess the level of complexity and interpretability of models on a case-by-case basis.

2.2.1. Describe regulatory and operational requirements regarding the interpretation and traceability of models within the insurance industry. **(B2)**

2.2.2. Explain the different degrees of complexity in the creation and interpretation of models. **(B2)**

2.2.3. Explain the importance of acceptance and understanding of models in the (operational) application of complex models. **(B2)**

2.2.4. Explain the cost-benefit ratio of increasing complexity in the creation of models and the interpretation of model results. **(B2)**

2.2.5. Name and explain methods (e.g. Surrogate Model, LIME, Maximum Activation Model, Variable Importance Measure, Shapley Value Explanation, localGLMnet, ICEnet, MACQ) for illustrating and explaining procedures and results from complex models. **(B2)**

2.2.6. For use cases from the insurance industry, assess which models should be used with which degree of complexity in order to achieve an optimum balance between accuracy and interpretability. **(B5)**

3 Mathematics / Statistics

3.1 Deep Learning 4

Objective: Candidates are familiar with analytical models of text mining and can describe the procedure for analysing texts. Use cases from the insurance industry and solution approaches can be described. In addition, text mining tools should be known and their use should have been practised using examples.

3.1.1. Describe typical text mining applications in the insurance company and show the possibilities and limitations. **(B4)**

3.1.2. Name a classic relevance measure for text documents with regard to a search

query. Compare classic ranking metrics with the results of other NLP (natural language processing) methods. **(C2)**

3.1.3. Explain the basic functionality of word vectors. Name limitations and approaches to overcome them. **(C2)**

3.1.4. Explain a text mining process that is based on an unsupervised learning process. **(A2)**

3.1.5. Describe the usual processing steps of a text before it can be processed using NLP. **(C2)**

3.1.6. Explain the differences between feedback (RNN), convolution (CNN) and attention with regard to the data section under consideration and discuss the advantages and disadvantages of attention compared to other methods. **(C2)**

3.1.7. Outline the structure of the transformer architecture and explain which parts are required for speech coding and speech generation. Name use cases within and outside of language processing. Categorise the advantages and disadvantages against the background of keywords such as ecological footprint, hallucinations and data leakage. **(C3)**

3.2 Anonymisation / pseudonymisation 2

Objective: Candidates are familiar with ways of measuring and increasing the anonymity of data and methods. These can be determined using examples and weaknesses identified.

3.2.1. Explain and motivate methods for measuring the anonymity of data (such as k-anonymity or l-diversity). **(B2)**

3.2.2. Apply methods for measuring anonymity of data to examples related to insurance. **(B3)**

3.2.3. Name the weaknesses of the methods and explain them using examples. **(B5)**

3.2.4. Explain differential privacy and discuss various distribution assumptions. **(B2)**

3.2.5. Demonstrate the benefits of differential privacy using an example. Explain the choice of parameters. **(B5)**

3.2.6. Critically discuss the concept of anonymity. Name known examples from practice where data breaches have occurred. **(B5)**

3.3 Quantum computing

Objectives: Candidates are familiar with the special features and advantages of quantum mechanical data processing in relation to machine learning and are able to apply these in a targeted manner to optimise model structures and predictions.

3.3.1. Describe the qualitative and quantitative differences between classical data processing and quantum-based data processing, especially with regard to entangled states of the qubit information units. **(B4)**

3.3.2. Describe the correspondences in the quantum gate system compared to classical gate theory. In particular, give examples of the non-classical operation of the Hardamard gate and the phase-shift gate. **(C3)**

3.3.3. Explain how the Shor algorithm and the Grover algorithm work, how they are categorised in the BQP complexity class and how they can be used in principle for machine learning. **(D3)**

3.3.4. Explain how Quantum-Enhanced Support Vector Machines work. **(C3)**

3.3.5. Explain how Quantum-Enhanced Neural Nets work. **(C3)**

3.3.6. Use an example to demonstrate the practical implementation of Quantum-assisted machine learning. **(D5)**

4 Use Cases

4.1 Use Case

Objective: Candidates are able to carry out complex and comprehensive data science analyses and sophisticated machine learning applications independently.

4.1.1. Based on a complex question and a given dataset, you will carry out a data science analysis independently. You will go through all phases of a data mining process and create a notebook in Python, R or another suitable language or use a framework, observing the usual programming standards (maintenance, reliability, efficiency , user-friendliness). You apply models, interpret and evaluate the results, explain how the model works and present these findings in a way that is appropriate for the target group. **(C5)**