AAE | ACTUARIAL ASSOCIATION OF EUROPE
ADVISING · ACHIEVING · ENGAGING

ACTUARIAL ASSOCIATION OF EUROPE
ASSOCIATION ACTUARIELLE EUROPÉENNE

SILVERSQUARE NORTH
BOULEVARD ROI ALBERT II 4
B-1000 BRUSSELS, BELGIUM
(+32) 2 274 06 61
INFO@ACTUARY.EU
WWW.ACTUARY.EU

CPD in Data Science

INTRODUCTION

The following document has been produced to support actuarial associations and individual actuaries in identifying topics in the field of Data Science (DS) to enhance corresponding knowledge and skills. Specifically, the document may be applied by associations in their development of activities related to continuing professional development. This paper has been produced by the AAE Education Committee in its scope of work.

We would also like to mention the paper "What should an actuary know about Artificial Intelligence" (January 2024), This discussion paper offers, among other things, additional perspectives on continuing professional development within the application of AI and DS methods by actuaries.

Due to the fast evolution of data science related techniques, this document should be understood as a "living" one, in the sense that it may require adaptations to future breakthroughs in the field. This paper specifically addresses member associations to provide appropriate guidance to their members and to organize adequate CPD activities for them. Actuaries benefiting from these actions can be qualified actuaries with a limited background in Data Science but also actuaries with more experience in the field that want to deepen their knowledge of the topic. A scale of technical complexity of the topics has been introduced to help actuaries and associations to better shape CPD programs in accordance with the initial background and the expected output.

We list Data Science techniques in six different areas, ranging from methods already familiar to most actuaries, to artificial intelligence solutions at the boundary of the unexplored, for potential usage in actuarial applications.

Modern Data Science methods can be among others applied in the following areas[1]:

- Product development and pricing,
- Marketing, distribution, and CRM,
- Underwriting,
- Claims and benefits management,
- Operations and IT,
- Finance and actuarial,
- ALM and investment.

Furthermore, we briefly touch on the specific environment actuaries regularly operate in, especially managing potentially conflicting interests of companies, regulation, and customer protection. Through this, this paper also addresses Data Science specialists that would like to apply their skills beneficial to solving actuarial tasks.

---

[1] *What should an actuary know about Artificial Intelligence (https://actuary.eu/wp-content/uploads/2024/01/What-should-an-actuary-know-about-Artificial-Intelligence.pdf)*

The starting point for all presented Data Science techniques is the AAE Core Syllabus and Guidelines for Actuarial Training in Europe (see table 1 below). Within its Learning Areas, the Core Syllabus already includes many Topics and Techniques that can be seen as Data Science related.

To further structure potential training programs and individual activities, the related topics of the AAE Core Syllabus can be groups into four categories:

- Descriptive statistics and data preparation,
- Probability theory concepts, econometrics, and statistical inference methodologies,
- Machine learning skills and simulation,
- Ethics and professional standards.

Descriptive statistics	Probability, Statistical inference and Econometrics	Machine Learning and simulation	Ethics and professional standards
Data as a resource for problem solving (6.1)	**Random variables** (1.1)	**Simulation** (1.7)	**Professional and risk management issues** (6.4)
Data analysis (6.2)	**Statistical Inference** (1.2)	**Statistical learning** (6.3)	**Professional standards** (9.3)
Visualizing data and reporting (6.5)	**Graduation and Statistical tests** (1.3)		**Professionalism in practice** (9.4)
	Regression (1.4)		**International and institutional awareness of professional standards** (9.5)
	Bayesian statistics and credibility theory (1.5)		

Table 1: List of main topics related to data science that are present in the AAE Core Syllabus

A) DATA SCIENCE TECHNIQUES AND DATA ETHICS

The list of mentioned topics cannot be seen as complete as the field of Data Science is evolving rapidly. The levels of complexity are given with letters from A to C (A being the easiest), CS means that the subject is already in the Core Syllabus.

1 - Clustering

General description: Clustering can be understood as summarizing the information contained in a dataset by gathering its observations within sufficiently homogeneous classes. This provides a general toolbox for actuaries aiming, for example, to filter the heterogeneity of a population of policyholders by identifying some groups sharing similarities in their characteristics, to determine categories of similar financial products, to perform a synthesis of the information contained in a database, etc.

The subtopics in this section start with standard data analysis tools, like Principal Component Analysis and Hierarchical Clustering, that are prerequisite already part of the Core Syllabus, but which should be deepened, to K-means and Support Vector Machines.

Principal Components Analysis (CS)

- Apply Principal component analysis to summarize information contained in data.

Hierarchical Clustering (CS)

- Understand the meaning of classical dissimilarity measures.
- Apply hierarchical agglomerative clustering to data.
- Apply Divisive clustering (DIANA algorithm) to data.

K-means (A)

- Understand the concept of quantization and the heuristic behind K-means algorithm.
- Apply K-means algorithm to data.
- Select the appropriate number of clusters to describe data.

Support Vector Machines (A or B)

- Understand the role of kernels in SVM, the « kernel trick », and how to choose the appropriate one for a given problem.
- Formulate appropriate optimization problems related to SVM and use algorithms to solve them.
- Apply Bayesian SVM to data.

2 - Regression - Prediction

General description: Regression techniques are related to measuring the impact of covariates (or observable risk factors) on a response variable. They are widely used in the context of a posteriori pricing, and of reserving. The historical regression approach is related to the explanation of a phenomenon. It coexists with a more “prediction-oriented” approach in machine learning, where the understanding of the model is less considered.

Both approaches are key for actuaries: prediction is intimately related to the fact that a premium or an amount of reserve are anticipations of future claims. But the interpretation is also key: for communication (towards policyholders, regulators, non-actuaries within the insurance or finance ecosystem), but also for risk management and prevention. The question of interpretability (see item 3) will insist on this particular point.

Since regression is a cornerstone of actuaries’ statistical methodologies, the first subtopics of this section are implicitly part of the Core Syllabus: model selection via information criteria is usually linked with Linear Model or Generalized Linear Models (1.4.2, 1.4.3, 6.2.7). Regularization approaches are not explicitly in the Core Syllabus, but became very popular since the beginning of 2000’s. They can be considered as a compromise between modern machine learning and intelligible parametric techniques, since they are used to deal with parametric regression models in high dimension (that is when the number of covariates is high, potentially higher than the number of observations).

Bagging and boosting methods are key ingredients that are part of many machine learning algorithms. Bagging (for Bootstrap Aggregating) show the importance of bootstrap techniques (item 1.7.5 of the Core Syllabus), while boosting presents the general algorithmic methods used to iteratively improve the quality of a predictor.

The section finishes with classical categories of machine learning techniques: tree-based methods and neural networks. These techniques are already mentioned in the Core Syllabus (item 6.3.4), we here give precision on the directions that should be either refreshed or deepened.

Regression modelling and fitting (CS)

- Define a proper regression model and determine the appropriate loss function to be optimized.
- Determine proper scoring rules to measure the quality of a model.
- Use a test sample or a cross-validation approach to select a model.
- Implement auto-calibrated predictors.

Model selection (AIC, BIC) (CS)

- Use penalized model selection techniques like Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) in the context of regression or time series analysis.
- Apply model selection heuristic like backward / forward / stepwise selection.

Sparsity and Regularization (A)

- Explain the concept of sparse models.
- Apply regularization methods (LASSO, Elastic-net, SCAD…) to produce sparse models.

Bagging (A)

- Use the bootstrap method to estimate properties (e.g. the mean squared error) of an estimator. (item 1.7.5 of the Core Syllabus).
- Understand the principle of aggregation of estimators / predictors.
- Use bootstrap aggregating techniques (bagging) to improve the quality of the predictor.

Boosting methods (B)

- Understand the general principle of boosting algorithms.
- Apply classical boosting algorithms (AdaBoost, XGBoost…).

Tree-based methods (regression trees (A), random forest (B))

- Fit a tree to data using CART algorithm.
- Use pruning algorithm to select a subtree from the « maximal » regression tree.
- Understand the problems caused by the potential instabilities of regression trees.

Neural networks: multilayer perceptron (A)

- Understand the structure and concepts of neural networks (weights, activation function).
- Understand back-propagation algorithm to optimize the weights of a network.
- Fit a multilayer perceptron to data using a grid search to optimize the hyper parameters.

3 - Interpretability

General description: As mentioned in section 2, the specific missions of actuaries require to go beyond the “black-box” structure of many machine learning techniques. This is of course a question of communication, but also of responsibility and ethics. Interpretable machine learning is a fast-evolving field, and the present section restricts itself to sufficiently well diffused techniques. It contains will require updates in the near future, according to the evolution of the area.

This section starts with data visualization techniques, that are a first way to explore the information contained in a dataset. These tools can be seen as extensions of exploratory statistical methods that are already in the toolbox of actuaries (see item 6.2.1, 6.2.2, 6.5.1), but the rise of more complex machine learning techniques lead to the development of more advanced techniques in this field. The concept of “Data storytelling”, although already related to the actuarial approach to data, is more complex in the field of machine learning, and potentially requires to be deepened in the context of a data science training program.

Variable importance are relatively simple methods (that are for example used in tree-based techniques and available in several packages) that can provide simple ways to measure the impact of a covariate on a response variable, while Shapley values are concept from game theory applied to interpreting the impact of variables that are getting more and more popular (in the data science and actuarial literature).

Data visualization (A)

- Select appropriate software and chart types to model and visualize data for research projects based on the intended audiences.
- Produce visual reports and interactive charts and dashboards.
- Data storytelling

Variable importance (B)

- Understand the concept of “out-of-bag” error.
- Understand and compute variable importance metrics.

Shapley values (B)

- Understand and use the concept of Shapley values to measure the impact of features on the prediction made by the model.
- Use Partial Dependence Plots and Local Interpretable Model-agnostic Explanation.

4 – Natural Language Processing

General description: Actuaries have to deal very often with unstructured data like text. Text is considered as “unstructured data” because it is not a quantitative variable, nor a qualitative variable which would have a limited number of modalities. Machine learning produces tools that considerably facilitate the incorporation of text data into quantitative actuarial methods.

The first step to deal with text is to transform this data into numerical vectors in a proper space where words (or group of words) are close to each other when they have a similar meaning (or are concepts that are related to each other). This approach is called embedding. Developing these types of representations from scratch can be hard, but many publicly available embedding techniques can be used to easily perform this task, as long as one understands the whereabouts of the methodologies and their limits (one of them being the fact that embedding techniques are trained on a specific corpus of texts, which do not rely on the same vocabulary as actuaries, limiting their performance in the actuarial field).

In addition to the use of word embedding, deep learning techniques like Convolutional Neural Networks (initially used in image analysis) or recurrent neural networks like LSTM or Transformers can be used to automatically analyse some large corpus of text but require more advanced skills.

Embedding: use of pre-trained models (A),

- Understand the general methodology of embedding and simple algorithms in this field (Cbow, Skipgram).
- Use pre-trained embedding (like FastText, word2vec) to transform text into numerical variables.

Embedding: design of embedding matrices (C)

- Train embedding model on a corpus of text.

Convolutional neural networks (B for the definition, C for the practical use)

- Understand the structure and hyper parameters of convolutional neural networks.
- Apply and calibrate CNN for text or image analysis.

LSTM (C)

- Apply and calibrate recurrent networks like LSTM to perform text analysis.

Transformers (C)

- Understand the principles of transformers for text processing.
- Use BERT type models.

5 - Further advanced topics

General description: Data science methods produce applications and algorithms for many fields. In this section, we list some topics whose practical applications in insurance and finance are not completely stabilized (although already used for specific application in the actuarial scientific literature and practice), and which are direct extensions of methodologies widely used by actuaries. Essentially, this section lists introductory fields related to generative AI, and to the concept of Reinforcement learning.

Generative Adversarial Networks (C)

- Understand the practical behaviour of Generative Adversarial Network.
- Train a Generative Adversarial Network.

Variational Auto Encoders (C)

- Understand the meaning of the ELBO distance and its role in VAE.
- Train VAE.

Reinforcement learning (C)

- Understand the principles of multi-armed bandits (contextual bandits, stochastic bandits…).
- Q-learning.
- Use reinforcement learning frameworks like open aigym.
- Understand and implement actor-critical algorithm.

6 – Data ethics and fairness

General description: Fairness has always been a critical concern in insurance, given that insurance addresses inequalities. Fairness in insurance is essential for social justice and equal access to services. Actuaries must continue to address biases to maintain trust in the industry and to contribute to a more equitable future. The current technological and scientific advances related to data science, machine learning, and artificial intelligence methods pose new data ethical challenges and reignite past discussions. Furthermore, new notions of fairness based on advanced probabilistic and statistical concepts are appearing at an alarming rate – matched perhaps only by an increasing awareness by regulators and supervisory authorities alike. It is critical for actuaries to understand the relations between different notions of fairness, not only from a mathematical-statistical point of view, but also from an economical as well as a juridical frame of reference. Data ethics and fairness are to some extent already addressed in the Core Syllabus. This includes Topic 6.4 on *professional and risk management issues* within *data and systems*, Topics 9.3-9.5 related to professionalism, and the section on Advanced Skills. However, current developments call for additional interdisciplinary training, not least in big data ethics.

Data ethical principles (A)

- Understand and discuss proposed fundamental principles of data ethics, including privacy, and how they relate to human rights.
- Be able to make judgments based on at least a selection these ethical principles.
- Reflect on in which way one can transgress norms to create more progressive and socially responsible norms in the application of data science.

Regulation, governance, and compliance (B)

- Be familiar with the relevant legal framework applicable for actuaries as it pertains to fundamental rights, personal data protection, and the use of artificial intelligence. For actuaries operating in the jurisdiction of the European Union, this includes the General Data Protection Regulation and the upcoming AI Act
- Be able to distinguish between principles and regulatory measures related to data protection, such as the General Data Protection Regulation of the European Union, and the use of artificial intelligence, such as the upcoming AI Act of the European Union, respectively.
- Understand the role of regulators, supervisory authorities, and insurance providers in ensuring fairness and equity in insurance.
- Demonstrate readiness to comply with existing professional standards related to data and data ethics.
- Apply a risk-based approach to assess the negative impact specific uses of data has on people, the environment, and society.

Fairness and discrimination (B)

- Be familiar with various notions of mathematical-statistical fairness, both group fairness criteria and individual fairness criteria.
- Employ test procedures and explainability tools to assess, for selected criteria, whether a premium or a decision is fair.
- Be able to compare different fairness criteria, including their relative advantages and limitations, with a focus on insurance applications.
- Discuss the role of causality, and causal methodology, in establishing fairness.

Data availability and privacy (B)

- Understand the limitations of *fairness by unawareness*, which is essentially just the omission of protected characteristics in the statistical analysis, and its impact on data collection and the processing of protected attributes such as gender.
- Discuss the impact of increased data collection and application on insurability.

Products and stakeholders (A)

- Characterize how the role of an insurance product as a social good, an economic commodity, or something in between relates to data ethics and affects fairness.
- Discuss the impact fairness and accountability measures, both external and internal, can have on stakeholder welfare.

B) PROFESSIONAL FOUNDATION

The following considerations in this section are primarily aimed at the group of data scientists mentioned in the introduction, who wish to apply their professional skills to actuarial issues and problems in the insurance industry.

In order to be able to deal with specific issues in an actuarial environment with the means of data analysis, it is necessary to understand the context. This requires sufficient knowledge of the business environment.

In the following topics are listed whose knowledge significantly contributes to a good understanding of actuarial issues:

Basic knowledge of insurance

- Definition and classification of risks,
- Special risks, such as adverse selection or moral hazard, which play a significant role in the context of insurance,
- Definition of an insurance (as plan-based coverage of a requirement for funds which is uncertain in detail but can be estimated on the whole on the basis of risk compensation collectively and over time).
- Basic (economic) functioning of insurance,
    - risk transfer,
    - risk transformation, i.e. risk compensation in the insured collective (based on the laws of large numbers) and over time,
    - Financing of insurance, in particular
        - essential financial systems,
        - solidarity communities and risk carriers,
        - insurance premiums, reserves and risk capital.
- Main lines of insurance.
- Reinsurance.

Basic knowledge of the legal and regulatory framework.

Basic knowledge of qualitative risk management,

- risk management process,
- the concept and key elements of Enterprise Risk Management (ERM),
- risk culture (governance),
- professional standards.

In order to deal with an actuarial issue, it is often necessary to be able to

- assess, compare and select different approaches and models in the context of the problem with regard to their suitability,
- understanding, interpreting and verifying results (such as analyses or forecasts) (based on reality),
- to discuss the methods used and the results achieved (also with non-specialists), as well as to document and communicate them in a comprehensible manner.

It is as well important to have a solid knowledge and be able to apply key actuarial concepts and models.

For a comprehensive list of relevant skills, see AAE's "Core Syllabus for Actuarial Training in Europe".

Approved by the AAE Board on 16 September 2024.

APPENDIX

1. **Mission**

   Actuarial education is an ever-developing field. Central topics have been relevant for many years, other fields have been added to the AAE Core Syllabus for Actuarial Education as the basis for actuarial education in Europe. Over the last years methods of Artificial Intelligence, Data Science and Data Analytics have become more and more important for the insurance industry and therefore also for the actuarial profession. The AAE Education Committee has been tasked by the AAE Board to draft a proposal for CPD in Data Science for actuaries as a support for members associations to offer CPD opportunities for their members.

2. **Objectives**

   - To draft a proposal for CPD in Data Science as a support for member associations.
   - To liaise with member associations, universities, and other educational bodies in the actuarial and data science field to get an understanding which programs already exist.
   - To draft a structure on how CPD in Data Science could be presented to potential participants, especially as a CPD offering.

3. **Reporting authority**
   The task force will report to the AAE Education Committee.

4. **Communication**
   The task force will present a proposal for CPD in Data Science to the AAE Education Committee.

5. **Members**
   - Jean-Claude de Pooter
   - Colm Fitzgerald; left Task Force in 10/2023
   - Christian Furrer
   - Olivier Lopez
   - Phillip Miehe; left Task Force in 06/2023
   - Rafael Moreno Ruiz; joined Task Force in 10/2023
   - Frank Weber
   - Henning Wergen; joined Task Force in 10/2023