Commission de Surveillance
du Secteur Financier

ARTIFICIAL INTELLIGENCE

OPPORTUNITIES, RISKS AND RECOMMENDATIONS FOR THE FINANCIAL SECTOR

December 2018

Table of Contents

“*If the Human Brain Were So Simple That We Could Understand It,*
We Would Be So Simple That We Couldn’t”

— Emerson M. Pugh

1 PREFACE

Today, Artificial Intelligence (“AI”) is one of the most promising technologies, and different kinds of practical applications, especially in the financial sector, are emerging.

This topic attracts a lot of attention, but at the same time, there is still a sense of ambiguity about what kind of technology is hidden behind this term.

The potential benefits that AI can bring are enormous, but these can only be achieved if the fundamentals of this technology and its underlying risks are well understood and an adequate control framework is put in place.

In this context, the CSSF has performed a research study in order to better understand what Artificial Intelligence is and the related risks. The result is a document which intends to provide some basic knowledge about Artificial Intelligence, describe the different types of AI and some practical use cases for the financial sector. Furthermore, this study covers the analysis of the main risks associated with AI technology and provides some key recommendations to take into account when implementing AI inside a business process.

Given the increasing adoption of AI in the financial sector and the relative lack of practical guidance from a risk perspective, the CSSF has decided to share the results of this study with the public, for the benefit of the financial sector.

This document is published in the form of a “white paper” and has no binding value vis-à-vis the supervised institutions. Nevertheless, it provides the foundations for a constructive dialogue with all the stakeholders of the financial sector for a deeper understanding of the practical implementations of AI technology and its implications.

2 INTRODUCTION

2.1 What is AI?

What is Artificial Intelligence?

Several definitions of Artificial Intelligence exist, and it is difficult to find a standard one. This is due to the fact that it is still not clear what human *intelligence* is.

In one of its papers, the Financial Stability Board has defined Artificial intelligence as “The theory and development of computer systems able to perform tasks that traditionally have required human intelligence.” This definition, although quite generic, summarizes the main concept behind artificial intelligence, which is the execution of *intelligent* tasks.

Such *intelligent* tasks, normally executed by humans, are somehow related to the formation of knowledge and then using that knowledge as a basis to take decisions or actions to achieve a specific goal.

Artificial Intelligence is therefore that branch of Computer Science focusing on making machines execute tasks like:

- Reasoning
- Problem solving
- Perception
- Learning
- Planning
- Ability to understand language and speech
- Ability to manipulate and move objects

The activities listed above are just examples of intelligent tasks: AI systems can focus on only one task or combine multiple capabilities, resulting into a varying degree of autonomy.

Many AI applications nowadays, and particularly in the financial sector, are “**augmented intelligence**” solutions, i.e. solutions focusing on a limited number of intelligent tasks and used to support humans in the decision-making process.

AI systems showing multiple intelligent capabilities and able to take decisions on their own in order to achieve specific goals are called “**autonomous systems**”.

2.2 Current context

Nowadays, AI is everywhere. From the applications recommending us what songs to listen to or what products to buy online, to highly accurate language translators, autonomous cars or military drones. AI is helping doctors to quickly analyze radiographies to detect tumors, and police officers to identify criminals using face recognition programs.

In the financial sector, AI is introducing some revolutions like algorithmic trading (to which regulation like MiFID II has brought much attention), robo-advisors, automatic facial recognition in the video KYC processes, or even AML tools detecting frauds on blockchain transactions.

The investments done globally in AI research and development are notable and especially in the United States and China, with the EU making a call to increase investments[1].

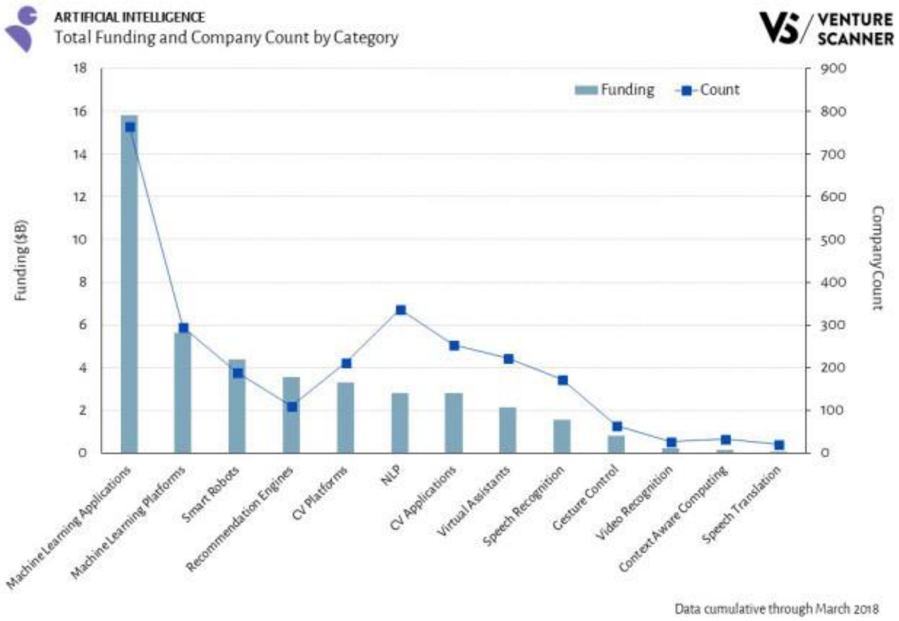


Figure 1: Total funding and count of AI companies by category[2]

[1] "Europe is behind in private investments in AI which totaled around EUR 2.4-3.2 billion in 2016, compared with EUR 6.5-9.7 billion in Asia and EUR 12.1-18.6 billion in North America. The EU Commission will work with Member States on a coordinated plan to help align and step up investments, building on the declaration of cooperation signed on 10 April 2018" (source: European Commission, "Artificial Intelligence for Europe", April 2018).

[2] Source: https://www.venturescanner.com/artificial-intelligence

Digital innovation hubs have flourished around Europe, as part of the digital EU strategy, acting like incubators of AI startups and as accelerators of synergies between the private sector and innovating companies for a quicker adoption of AI and other new technologies.

Furthermore, given that new specific skills will be more and more required, governments also focus on initiatives aimed at training young students in the AI disciplines, in order to become the AI experts of the future.

Luxembourg is following the same trend, and strategic initiatives like the LHOFT (Luxembourg House of Financial Technology), the implementation of the European HPC (High Performing Computing) in Luxembourg and the partnership with Nvidia (the leader provider of GPUs[3], at the basis of AI architectures) to develop training programs for students confirm that the strategy is to build a favorable environment for this promising technology.

2.3 AI is not new

We may think that Artificial Intelligence is something new, a new trend of technology that has appeared only lately and that is revolutionizing our world. However, AI has existed for many years, the first promises (or hopes at that time) being around already in the 1950s after WWII.

The first AI conference was held in 1956 at Dartmouth College, where the term “Artificial Intelligence” was first coined. The AI experts at that time had a very optimistic view, thinking that Artificial intelligence could be achieved in no more than a generation, and received a lot of funding for their research projects. Some years later, the AI pioneer Marvin Minsky predicted “In from three to eight years we will have a machine with the general intelligence of an average human being”.

Needless to say, the time required was largely underestimated and progress did not come as expected, mainly due to insufficient computing resources. AI suffered from a few so-called “AI winters”, i.e. periods where funding decreased and progress slowed down.

[3] Graphics Processing Units

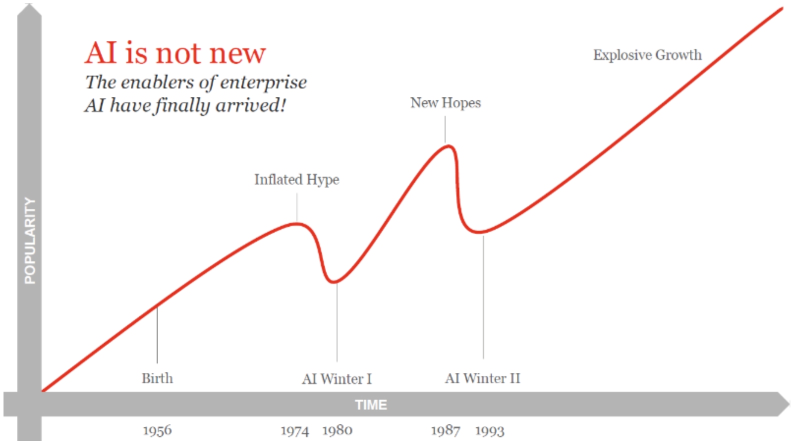


Figure 2: AI winters (source: PWC)

Luckily, in the last few years, AI has shown a clear resurgence, and today global investments in AI are soaring. As a result, major progresses and advanced commercial solutions in a wide spectrum of sectors have appeared.

This is due to several factors:

- The cost of storage has dramatically decreased during the past few years.
- The processors are increasingly dense, fast and powerful.
- The availability of high-performance parallel computing, as found in GPUs (Graphic Processing Units, designed for parallel operations) for example, has further increased the overall computational speed, allowing to analyze big, complex data sets in shorter times.
- The widespread availability of cloud computing has simplified the access to cheaper, scalable, flexible and easily interconnected computing environments.
- With the advent of big data, there has been an incredible increase of data available, including large data sets available for learning (i.e. to "train" AI algorithms).
- The democratization of the access to AI and especially Machine Learning[4] ("ML") libraries providing easy access to AI/ML algorithms (e.g. ML libraries available in the cloud from Amazon, Google, Microsoft, etc.).

[4] Machine learning will be introduced in more detail in section 3.3 Machine learning.

3 TYPES OF AI

Several different categorizations of AI exist: it is difficult to draft a list that is at the same time mutually exclusive and collectively exhaustive as there is some overlap between different AI categories.

3.1 Narrow and broad AI

A first basic classification of AI technologies consists in grouping them into two main categories: narrow AI and broad (general) AI.

The first category refers to AI systems intending to solve specific problems, while the second refers to a form of Artificial Intelligence that can meet or even exceed human intelligence by demonstrating an intelligent behavior across a variety of tasks and being capable of solving different types of problems.

Among the most relevant attempts to build general intelligent systems are Google's DeepMind[5] or IBM Watson[6].

DeepMind's objective is to develop an AI program that can learn to solve any complex problem without needing to be taught how. By applying their research in the field of games, they were able to create a single program that taught itself how to play and win at 49 completely different games. Based on that, in 2015 their AlphaGo[7] program beat the world's best player at Go, one of the most complex games ever created.

IBM Watson is a high performing "question answering" machine capable of answering questions posed in natural language. The machine can learn from different data sets pertaining to different fields and therefore provide answers related to a variety of topics.

In the field of robotics, the impressive videos of the robots from Boston dynamics[8] are a demonstration of what AI applied to physical robots can do[9] and how humanoid robots could soon become an everyday reality.

These attempts to build multi-purpose systems are, however, far from being a broad type of AI, which at the moment does not really exist. Although today it is just hypothetical, broad/general AI represents what common people think about AI as well as their worst fears, including AI being responsible for job losses, or even taking over humans and dominating the entire world.

Artificial Intelligence today, and in particular in the financial sector, is instead represented by various forms of *narrow* AI, highly specialized models addressing specific problems.

[5] https://deepmind.com/about/
[6] https://www.ibm.com/watson/
[7] https://deepmind.com/research/alphago/
[8] https://www.bostondynamics.com/
[9] https://www.vision.ee.ethz.ch/en/publications/papers/proceedings/eth_biwi_01223.pdf

3.2 AI subfields

Artificial Intelligence has a wide range of subfields. The picture below illustrates some of the main branches which are currently very well advanced.

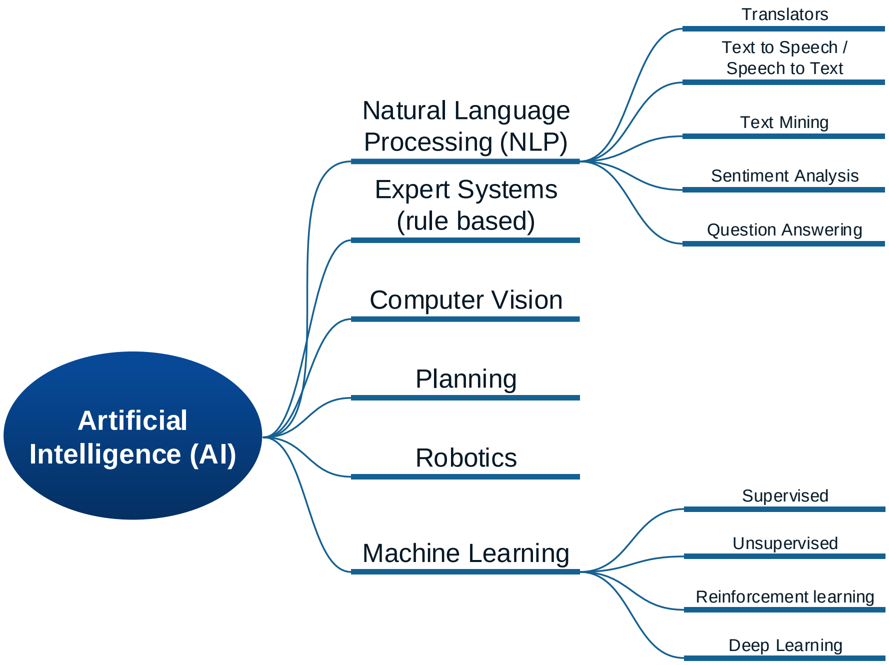


Figure 3: Main AI subfields

3.2.1 Natural Language Processing (NLP)

Natural Language Processing is the branch of AI enabling computers to analyze, understand and generate human language, in both written and spoken form.

Examples of applications of NLP technology are:

- Text to speech applications
- Text translators
- Real time speech translation
- Sentiment analysis
- Text mining

Regarding **speech translators**, market solutions are so advanced that currently they are even able to detect the accent of the speaker and transform it into the accent of the listener[10]. These solutions may be helpful for example in a call center.

Sentiment analysis techniques aim at identifying and categorizing sentiments or opinions expressed in written texts or by speech, in order to determine the attitude of the person toward a particular topic (e.g. positive, neutral, or negative). For example, such techniques are already applied by wealth advisors to build a cognitive profile of their clients and use this information to propose more tailored investments.

Text mining (or Text analytics) is a type of analytics that identifies the key contents in large quantities of documents and transforms them into actionable insights. Text mining extracts patterns and information from large sources of unstructured text and transforms them into structured data.

For example, this technology enables the identification of concepts (e.g. company names, addresses, dates, etc.) in raw text, or summarizing key information contained in long documents. It can also allow the identification of inconsistent information in several documents (e.g. between the prospectus and other marketing documents of an investment fund).

3.2.2 Expert systems (rule-based systems)

Expert systems, also called rule-based systems, are systems that store and manipulate knowledge in the form of rules, and derive new knowledge (new rules) by applying an inference engine to the existing knowledge base. The term “rule-based system” is normally used to identify systems where the set of rules are pre-defined by humans, as opposed to machine learning systems where the “rules” are automatically learnt by the system.

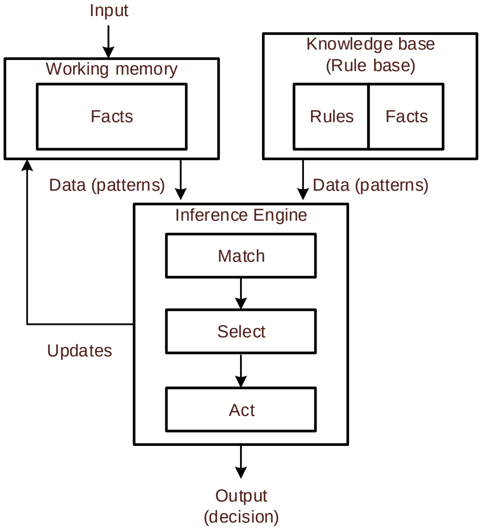


Figure 4: Rule based (expert) system

[10] https://techcrunch.com/2018/08/02/amazon-patents-a-real-time-accent-translator/?guccounter=1

3.2.3 Computer vision

Computer vision includes methods for acquiring, analyzing and understanding images and videos in digital format. A classic example of computer vision task is the image recognition and classification.

3.2.4 Planning

Planning is the ability of an intelligent agent to act autonomously by constructing a series of actions to reach a final goal. Automated planning is a complex task since it requires a system to constantly adapt based on the surrounding environment.

These systems create a representation of the surrounding environment and make predictions about how their actions will change it, and then make choices that maximize the return among a set of available choices.

Examples of applications can be found in autonomous systems or robots.

3.2.5 Robotics

Robotics is the field that connects perception to action, building robots that are equipped with sensors to interact with the physical world. Robotics has very strong connections with the other AI subfields in order to build the intelligent “brain” that move the robots.

3.3 Machine learning

AI is a broad field, of which machine learning is a subcategory. Machine learning is certainly one of the most prominent AI technologies nowadays. Most of AI innovative solutions, especially in the financial sector, rely on machine learning, so that today the two terms are often used interchangeably. For this reason, a large part of this document will focus on this technology.

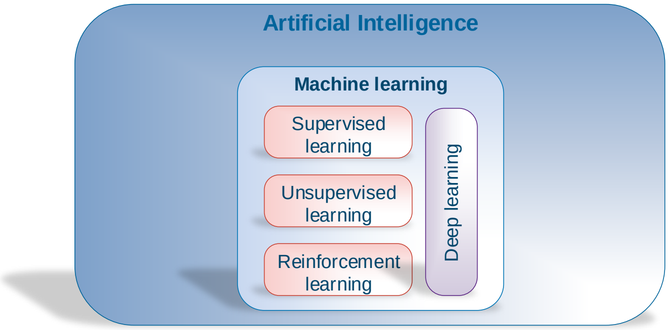


Figure 5: AI and machine learning

The standard ISO/IEC 38505-1:2017 on IT Governance defines **Machine Learning** as a “process using algorithms rather than procedural coding that enables learning from existing data in order to predict future outcomes”.

Indeed, machine learning is a class of learning **algorithms** that can learn from examples by identifying patterns in available data and can then apply such knowledge to new data to make **predictions**. The learning is done by means of suitable algorithms, which are used to create predictive **models**, representing what the algorithm has learnt from the data in order to solve the particular problem. Their performance improves as more data is available to learn (to *train* the model).

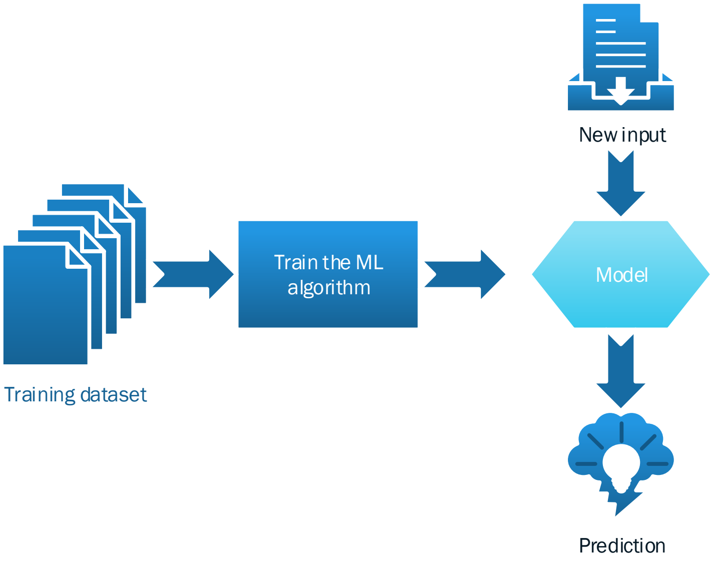


Figure 6: A model is a representation of what the algorithm has learnt from the training data and is used to make prediction on new input data

Machine learning has therefore enabled the transition from descriptive analytics, purely observing the facts, to **predictive analytics**, using past data to predict what will happen in the future. Predictive analytics represents one of the major applications of machine learning technology nowadays.

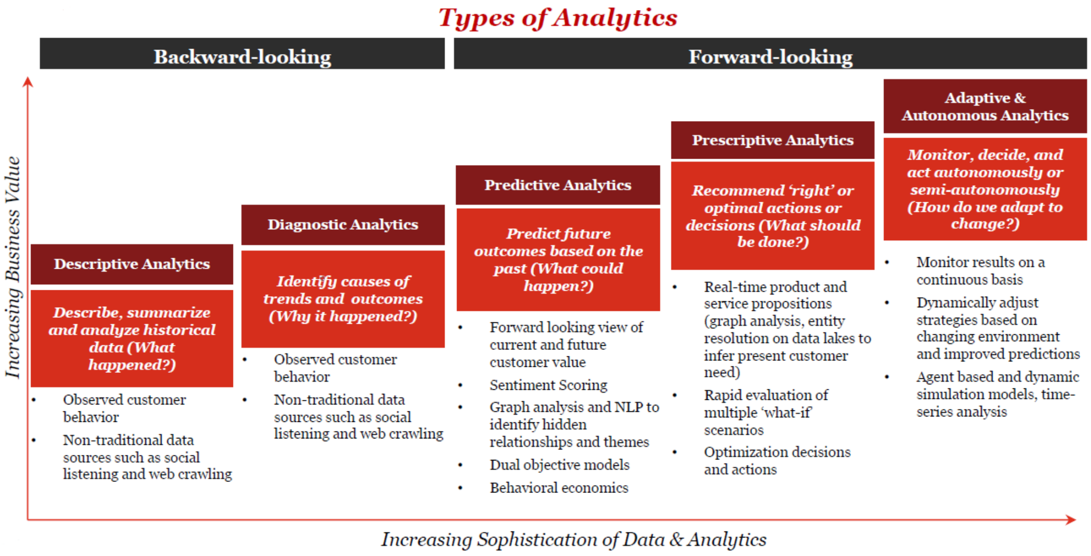

Types of Analytics
Backward-looking
Forward-looking
Increasing Business Value
Descriptive Analytics
Describe, summarize and analyze historical data (What happened?)
• Observed customer behavior
• Non-traditional data sources such as social listening and web crawling
Diagnostic Analytics
Identify causes of trends and outcomes (Why it happened?)
• Observed customer behavior
• Non-traditional data sources such as social listening and web crawling
Predictive Analytics
Predict future outcomes based on the past (What could happen?)
• Forward looking view of current and future customer value
• Sentiment Scoring
• Graph analysis and NLP to identify hidden relationships and themes
• Dual objective models
• Behavioral economics
Prescriptive Analytics
Recommend 'right' or optimal actions or decisions (What should be done?)
• Real-time product and service propositions (graph analysis, entity resolution on data lakes to infer present customer need)
• Rapid evaluation of multiple 'what-if' scenarios
• Optimization decisions and actions
Adaptive & Autonomous Analytics
Monitor, decide, and act autonomously or semi-autonomously (How do we adapt to change?)
• Monitor results on a continuous basis
• Dynamically adjust strategies based on changing environment and improved predictions
• Agent based and dynamic simulation models, time-series analysis
Increasing Sophistication of Data & Analytics

Figure 7: Different types of analytics (source: PWC)

This type of forward-looking analytics can further evolve by suggesting the next actions to perform (prescriptive analytics) and even by reaching an increased level of autonomy in the decision-making (autonomous analytics).

3.3.1 Supervised, unsupervised, reinforcement learning

Machine learning algorithms can be classified into three main categories:

- Supervised learning
- Unsupervised learning
- Reinforcement learning

These three categories differ according to the level of human intervention required in labelling the data:

- In **supervised learning**, the algorithm learns from a set of 'training' data (observations) that have labels. For example, a data set composed of transactions may contain labels identifying the fraudulent transactions and those that are not. The algorithm will 'learn' a general rule for classification (model) that it will use to predict the labels for the remaining observations in the data set.

- **Unsupervised learning** refers to algorithms that will learn from a data set that does not have any label. In this case, the algorithm will try to detect patterns in the data by identifying clusters of similar observations that have some common characteristics. For instance, an unsupervised algorithm for fraud detection may try to segment customers based on their standard (transactional) behavior and will detect as anomalies those transactions that lead to a sudden change of behavior, for which it suspects that the transaction was not executed by the same client. Another illustrative example of unsupervised learning is an algorithm that would try to price an illiquid security. The algorithm would group (clustering) the securities based on similar characteristics, and when it finds an appropriate cluster for the illiquid security, the price of another security in the cluster would be used to estimate the price of the illiquid security.
- **Reinforcement learning** is different from the supervised and unsupervised learning since rather than learning from a sample data set, it learns by interacting with the environment. In this case, the algorithm chooses an action starting from each data point (in most cases the data points are collected via sensors analyzing the environment), and receives feedback indicating whether the action was good or bad. The algorithm is therefore trained by receiving positive and negative rewards, and adapts its strategy in order to maximize the rewards. These types of algorithms are often used in robotics (e.g. robots able to run and jump over obstacles[11]: in this case the data points are collected via video cameras and GPS sensors gathering information about the surrounding environment), game playing[12] or autonomous cars.

3.3.2 Deep learning

In addition to the three main categories of machine learning described above, there is currently one type of machine learning, called Deep Learning, that, given its importance, is often described as a fourth category.

Deep learning stands for deep *neural networks*.

Neural networks are a particular type of machine learning algorithms that generate models inspired from the structure of the brain. The model is composed of several layers (at least one of which is hidden), with each layer being composed of units (called **neurons**) interconnected between each other.

Each connection, like the synapses in a biological brain, can transmit a signal from one neuron to another. A neuron that receives a signal can process it and then reacts by signaling (“activating”) additional neurons connected to it. This is done by means of an activation function which takes as input the weighted sum of the input values (i.e. the connections from the previous layer) and calculates the output values as the weights associated to each connection to the next layer.

The learning process is based on the **backpropagation** algorithm, which basically consists of iteratively measuring the predicted output against the correct answer, and inputting back the error values into the network in order to tune the weights at each layer. After several runs, the prediction converges to the correct answer and the errors are reduced.

[11] https://www.bostondynamics.com/atlas

[12] The AlphaGo machine described earlier in section 3.1 is a combination of reinforcement learning and deep learning.

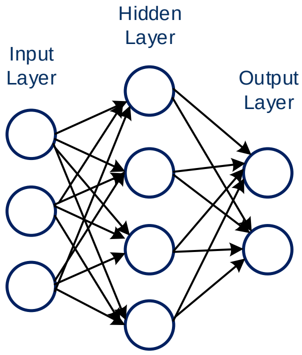


Figure 8: example of a basic neural net. Each node represents a neuron and an arrow represents a connection from the output of one neuron to the input of another.

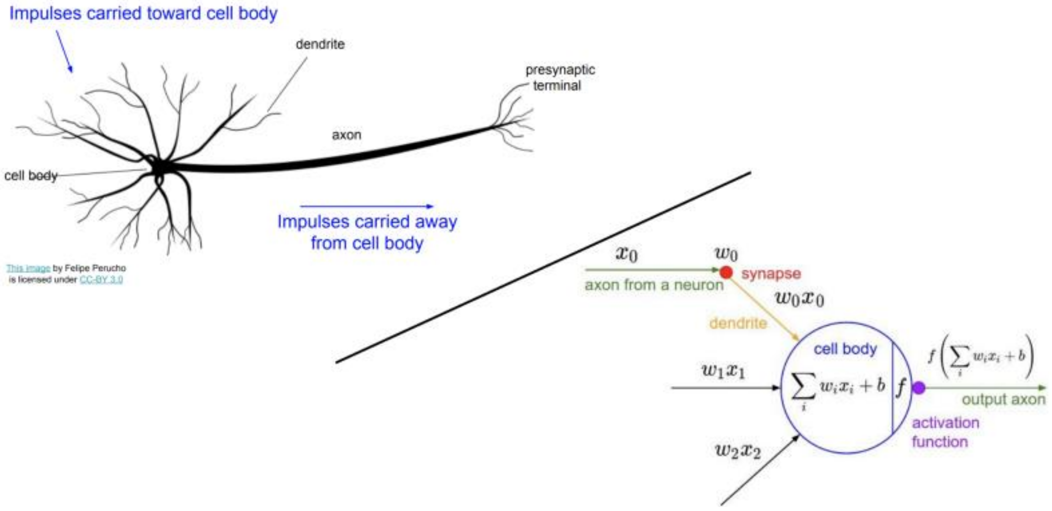


Figure 9: Side-by-side illustrations of biological and artificial neurons[13]

Deep learning algorithms are neural networks that have many hidden layers (the number of layers can vary from tens to thousands), which can make their structure very complicated, so much so that their functioning can easily become a "black box"[14]. Deep learning algorithms are especially efficient at image recognition (included in **computer vision**): the input values are the pixels

[13] Source: Stanford's CS231n
[14] This risk is discussed in more detail in chapter 6.

composing an image and the output is the category that the image represents (in the figure below for example, the algorithm classifies the image as a “giraffe”).

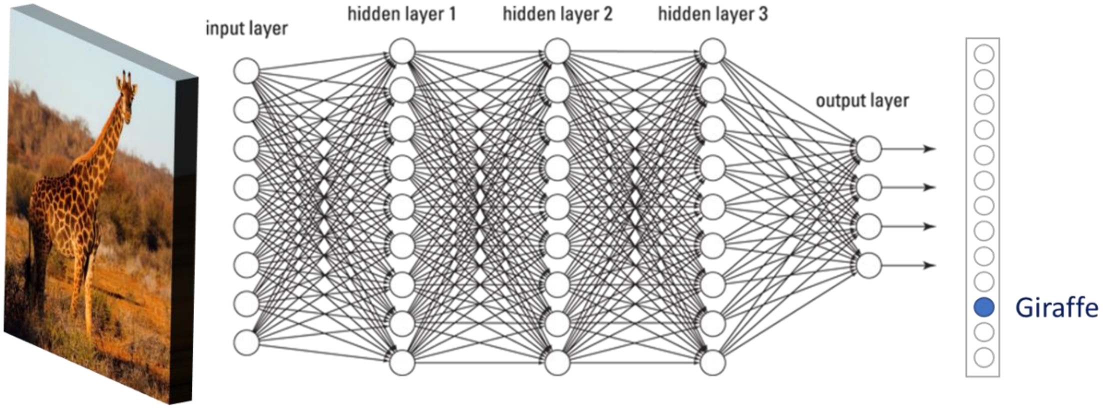


Figure 10: Example of neural network applied at image recognition (or “computer vision”)

In the middle layers, the algorithm will try to recognize distinctive elements composing the giraffe such as the long neck, legs, horn-like ossicones, coat patterns, tail, etc.

Deep learning algorithms (and neural networks in general) can be used for supervised, unsupervised, or reinforcement learning.

3.3.3 The learning algorithms

For many years, humans focused on programming, i.e. coding a set of rules to solve a problem. However, this approach has proved inefficient at solving complex problems, for which it is very difficult to find all the underlying rules.

Nowadays, the approach has changed and instead of coding a program that, given some input X, will calculate the output Y, the learning algorithm will build (“learn”) by itself the “program” (i.e. the function mapping X into Y) by analyzing a big amount of data. The program or mapping function learned constitutes the **model**. Such model will then be used, given a new input, to predict the output.

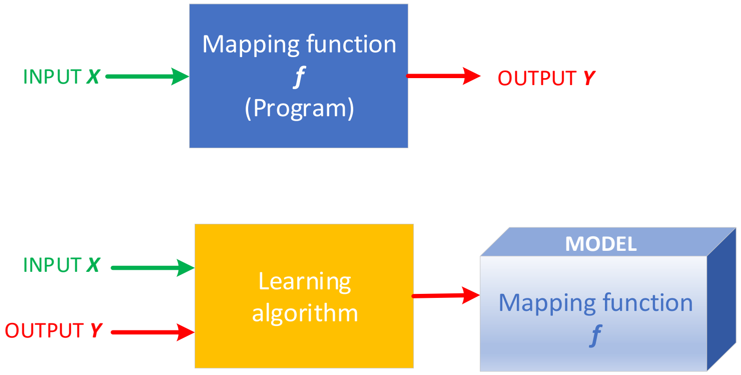


Figure 11: Change of paradigm: in machine learning, the "program" (i.e. the mapping function f) is "learnt" by the algorithm

The common principle underlying all supervised machine learning algorithms can therefore be summarized as finding (learning) a function ***f*** that best maps some input variables **X** (where X can represent more than one variable) to an output variable **Y**:

$$\boldsymbol{Y = f(X)}$$

The function ***f*** will then be used to make predictions of ***Y*** for a new set of ***X***. This is called predictive modeling or **predictive analytics**.

In unsupervised learning, the objective is similar and aims at identifying the target function ***f(X)*** representing the relationship among the input variables, with the only difference being that the output variable (which corresponds to the labels in the training data) is unknown.

Machine learning has solid roots in mathematics and statistics: the functions are represented by their mathematical formulas, and the performance of the model (i.e. the function learnt by the algorithm) is measured using statistical metrics.

3.3.4 The types of problems

The main objective of machine learning algorithms is to solve some common types of problems. Each type of problem is best addressed by one of the learning categories (supervised, unsupervised, reinforcement learning) described in section 3.3.1.

The most common types of problems addressed by machine learning and some of the algorithms typically used to solve those problems are summarized in the table below (the list is not exhaustive). It should be noted that some algorithms can solve different types of problems, depending on how they are parameterized.

	Type of problem	Examples of algorithms
Supervised learning	Classification	Logistic Regression[15] Decision tree Naive Bayes Classifier K-nearest neighbors (KNN) Support Vector Machine (SVM) Neural network Deep Learning Markov decision processes
	Regression	Linear regression Non-linear regression
Unsupervised learning	Clustering	K-Means Dbscan Principal Component Analysis (PCA) Hidden Markov Model
	Anomaly detection	K-nearest neighbors (KNN) Bayesian Belief Network (BBN) Decision tree Support Vector Machines (SVM)
	Association	Bayesian Belief Network (BBN) Decision tree Neural networks

Table 1: Common type of problems solved via ML and examples of algorithms

3.3.4.1 Supervised learning

3.3.4.1.1 Classification

Supervised learning is often used to classify data into categories or *classes*. Starting from a set of labeled training data, **classification** algorithms will classify the data based on the labels and it will then use what it has learnt to predict the labels (representing the category/class) for new, unknown data. In classification problems, the target to predict is a discrete variable, i.e. having only a finite number of possible values.

[15] Despite its name, Logistic Regression is a method for binary classification.

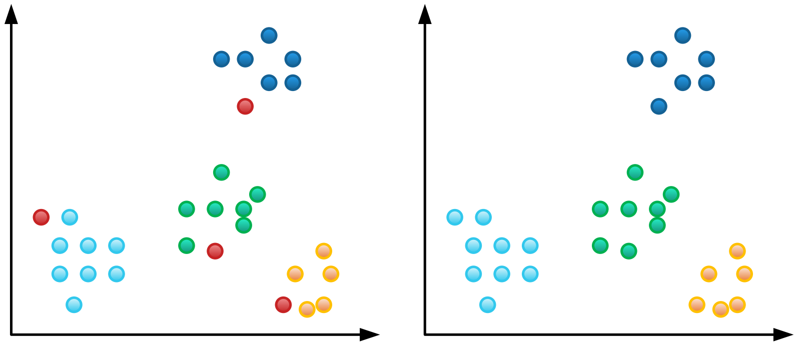


Figure 12: Example of classification of new data points (red circles); each class has a different color.

3.3.4.1.2 Regression

Regression problems are similar to classification in that they both use labeled past data to predict the value of new data, with the exception that regression methods will predict a variable that is a real number, meaning that it can have continuous possible values (as opposed to only a discrete set of values such as in the classification methods).

Regression methods try to model the relationship existing between variables by establishing a function that mimics that relationship and can be used to predict new values, which will sit along that function.

Regression methods can be linear (**linear regression**) when the model can be represented by a linear function[16] or non-linear (**non-linear regression**) if the model is a non-linear function.

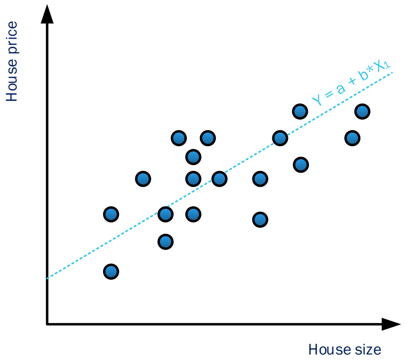


Figure 13: Example of linear regression

[16] Linear equations are equations with *linear* parameters, i.e. of type Y = constant + $\text{parameter}_1 * X_1 + \ldots + \text{parameter}_n * X_n$, where Y is the outcome variable and X_i are the input variables (or "independent variables"). In its simpler version (Y = constant + parameter * X), such equation can be represented by a straight line, however the formula can lead to a curve if the variables are raised by an exponent (e.g. X_i^2).

3.3.4.2 Unsupervised learning

3.3.4.2.1 Clustering

Clustering is a form of unsupervised learning that is used to identify patterns in unlabeled data and group data into clusters (that have a common pattern). A popular algorithm for clustering is k-means, where the data set is divided into k clusters, each of which with a centroid that minimizes the distance from all other data points in the same cluster.

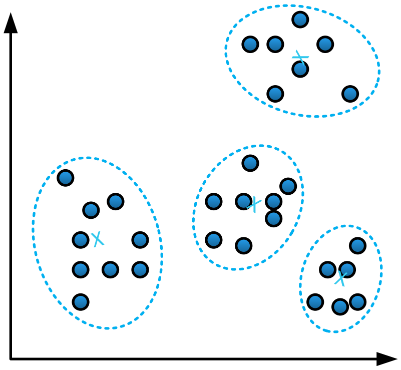


Figure 14: example of k-means with k=4

3.3.4.2.2 Anomaly detection

Anomaly detection is done by first detecting the structure of most of the data, for example by clustering, and then looking for the data points that do not follow any cluster, i.e. the "**outliers**". This technique is particularly useful when there is a need to identify unusual activity, like for examples transactions linked to Terrorism Financing.

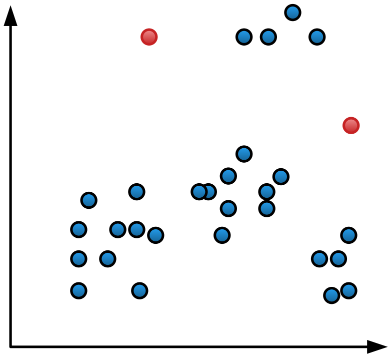


Figure 15: Example of anomaly detection: the red dots are the "outliers" (not belonging to any cluster)

3.3.4.2.3 Association (recommender systems)

Association is a particular type of clustering for which the common pattern is a rule (e.g. if customer purchased item_1, then he/she purchased also item_2). This technique is especially used in recommender systems (e.g. Amazon) to recommend to customers additional items that other customers already bought.

3.3.5 Which algorithm for which use case

The previous paragraphs are just a brief introduction of what kind of problems can be solved by using machine learning. Indeed, there are many different algorithms that can solve similar problems and is not easy to find the right one for a specific use case. Although comparative studies exist that evaluate different algorithms given specific use cases, the best approach consists in trying and comparing a set of shortlisted algorithms in order to select the one that best models the underlying problem.

4 THE DATA SCIENCE PROCESS (ML DEVELOPMENT LIFECYCLE)

4.1 The process flow

A classic project to develop a *supervised* machine learning ("ML") model follows an iterative process as the one described in figure 16 below. Compared to a standard software development lifecycle, we can notice that the phases are different and focus mainly on the data (extraction and preparation) and on the model (training, validation, testing).

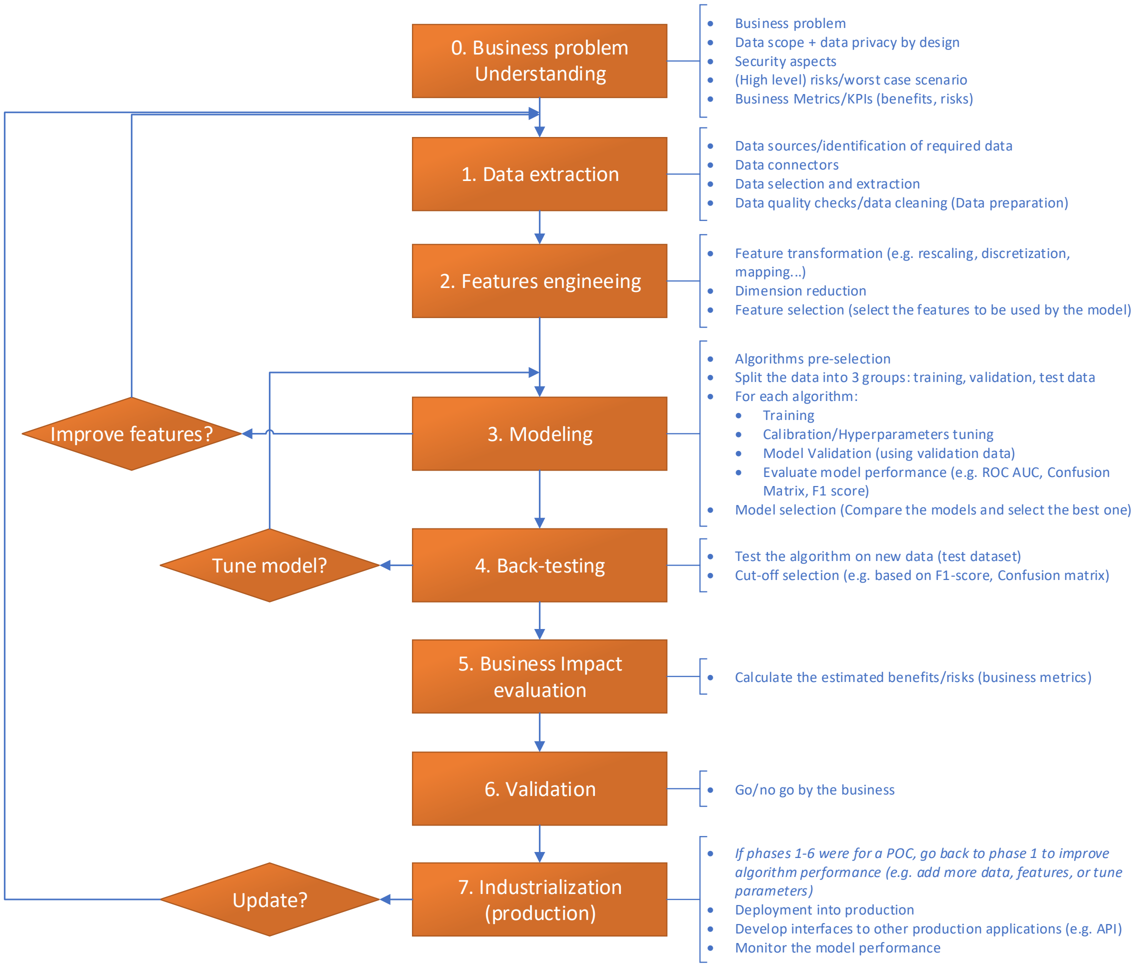

0. Business problem Understanding
Business problem
Data scope + data privacy by design
Security aspects
(High level) risks/worst case scenario
Business Metrics/KPIs (benefits, risks)
1. Data extraction
Data sources/identification of required data
Data connectors
Data selection and extraction
Data quality checks/data cleaning (Data preparation)
2. Features engineeing
Feature transformation (e.g. rescaling, discretization, mapping...)
Dimension reduction
Feature selection (select the features to be used by the model)
3. Modeling
Improve features?
Algorithms pre-selection
Split the data into 3 groups: training, validation, test data
For each algorithm:
Training
Calibration/Hyperparameters tuning
Model Validation (using validation data)
Evaluate model performance (e.g. ROC AUC, Confusion Matrix, F1 score)
Model selection (Compare the models and select the best one)
4. Back-testing
Tune model?
Test the algorithm on new data (test dataset)
Cut-off selection (e.g. based on F1-score, Confusion matrix)
5. Business Impact evaluation
Calculate the estimated benefits/risks (business metrics)
6. Validation
Go/no go by the business
7. Industrialization (production)
Update?
If phases 1-6 were for a POC, go back to phase 1 to improve algorithm performance (e.g. add more data, features, or tune parameters)
Deployment into production
Develop interfaces to other production applications (e.g. API)
Monitor the model performance

Figure 16: Typical ML development process

The process flow for an *unsupervised* learning process is very similar with the exception that the performance of the model will not be measured in terms of percentage of accurate results (e.g. false positive rate) but more in terms of statistical metrics. Indeed, given that the data is unlabeled, it is not possible to verify what the model predicts against the value observed in historical data. However, the steps performed to build the ML model are more or less the same.

4.2 Data science vs. machine learning

The process described in figure 16 is often called “Data science process” since machine learning applied to predictive analytics (i.e. to extract predictions from data) falls under the much broader term of “Data Science”. Compared to a pure machine learning development process focusing just on the creation of the ML model, the data science process covers the entire lifecycle including the important (and most time consuming) phases required to prepare the data to be used by the machine learning algorithm.

Data science is an interdisciplinary field that aims at extracting information and insights from data available in both structured and unstructured form, similar to data mining. However, as opposed to data mining, data science includes all steps associated with the cleaning, preparation and analysis of the data. Data science combines a large set of methods and techniques encompassing programming, mathematics, statistics, data mining and machine learning.

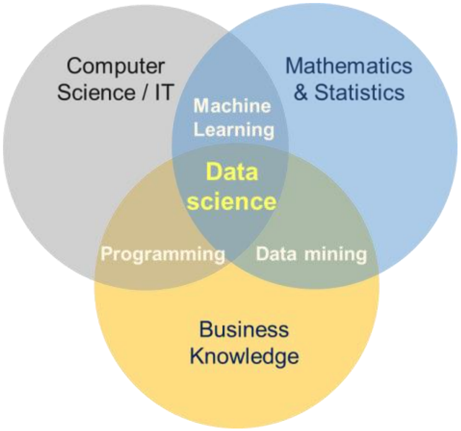


Figure 17: Data science is an interdisciplinary field covering programming, mathematics, statistics and machine learning

A **data scientist** is therefore a profile requiring several different skills ranging from mathematics and statistics to programming and machine learning.

4.3 BUSINESS UNDERSTANDING

The first important step to start an ML project is to have a clear understanding of the business problem that requires a solution.

The problem needs to be described in terms of current issues and expected solution, as well as in terms of benefits, including operational efficiency gains (the more the problem is related to a current “pain point”, the higher the expected return will be), data scope and risks.

The data scope, i.e. the data that will be used by the algorithm for the training and for the subsequent prediction, needs to be defined by listing the type of data to consider, the data sets to be excluded from the scope (for example transaction records with a particular code need to be excluded because of data quality reasons) and therefore the related filters to apply to the data.

During this initial step, particular attention should be paid to the aspects of data confidentiality and data privacy. For example, the following points, easy to be overseen, need to be considered:

- Does personal data need to be in scope? If so, what are the business and legal[17] justifications?
- Would the inclusion of personal data into the scope generate potential bias? How is this controlled?
- How would fairness (absence of discrimination) be ensured? How are deviations detected?

Regarding in particular the data bias, the analysis should identify already at an early stage whether the data to be included in the scope is, by its very nature, including a **human bias**. For example, if the problem to be solved is to predict whether an account overdraft can be approved, the inclusion in the training data set of the transactions that were approved in the past by the account manager may include human bias, since the account manager took the decision in the past based on his/her own criteria, and the final decision he/she took might have been wrong. The risk would be that the algorithm will learn what the account manager used as decision criteria instead of being neutral.

The **security** aspects need to be evaluated at the beginning of the project, to ensure that security by design is applied consistently and that, for example, the Need to Know principle is applied. For example, data scientists should have access only to the data strictly required to develop the models, and safeguards should be implemented throughout the ML development process. Among the security safeguards, production data has to be accessible in read-only mode, and any exception needs to be justified and adequately approved.

Furthermore, the environments used to develop the ML model should be appropriately segregated from the production environment and potentially contained in a sandbox.

The overall **risk** also needs to be analyzed. In particular, simulations of the worst-case scenario can help identify what could be the maximum level of tolerable risk in case the ML solution would be adopted. For example, in the case of the overdraft approval mentioned above, if the transaction would be approved only if the customer has sufficient balance considering the sum of all his/her accounts, then the overall risk would be limited.

[17] According to the GDPR, a data processing is permitted only if it has a lawful basis (ref art. 6 GDPR), for example if it is necessary for compliance with a legal obligation, such as the AML obligation.

The risk related to the specific ML implementation, together with the benefits, need to be measurable via specific KPI, which will be used to measure the performance of the model that will be developed and to decide whether the model can be deployed into production.

The results of the analysis need to be documented to ensure the traceability and auditability of the decisions made.

As a consequence, the involvement of the Compliance, DPO (Data Protection Officer), Risk Management and Information Security functions during this phase is recommended.

4.4 Data extraction and data preparation

In the data extraction phase, the data sources required to solve the problem are identified and connectors to the data are implemented, in order to be able to extract the required data.

Starting from the business description of the data scope from the previous phase, the corresponding databases, tables and fields need to be identified. In this phase, having a well-documented data dictionary and the collaboration of the business departments (data owners) represents a real added value in order to find the information required.

After finding the required data, a set of tasks, often called "**data preparation**" tasks, are performed in order to clean the data and put it into the right format to be exploited by the model.

An important aspect verified in this phase is the quality of the data. Data is carefully inspected in order to evaluate whether there are **data quality** issues, like for example missing data, duplications, inconsistent data or data in the wrong format. In case of issues, the concerned records with insufficient quality are either excluded from the data scope (this could be an easy and quick solution especially if the model to be developed is just a Proof of Concept which will be further improved later on anyway) or included only after being fixed. In both cases, it is important to note that data quality issues identified need to be escalated to the business departments in order to be fixed at the source (production data).

Depending on the particular use case, the data required may be from **external data sources**, for example because the institution does not have enough historical data or the internal data present data quality issues or is not in an easy to exploit format. In this case, the institution should perform a due diligence analysis before using the data, to verify the trustworthiness of the data vendor, the quality of the data and also the pertinence of that data for the particular use case. For example, historical credit data specific to the French market may not be appropriate in a Luxembourg context.

The tasks performed in this phase, which at first sight may seem trivial, are very time-consuming due to the difficulties in finding the right data inside the multitude of databases and tables, setting up the connections with legacy systems and finally cleansing and organizing the data. Indeed, this phase together with next phase of Features engineering often represents more than 80% of the entire project efforts.

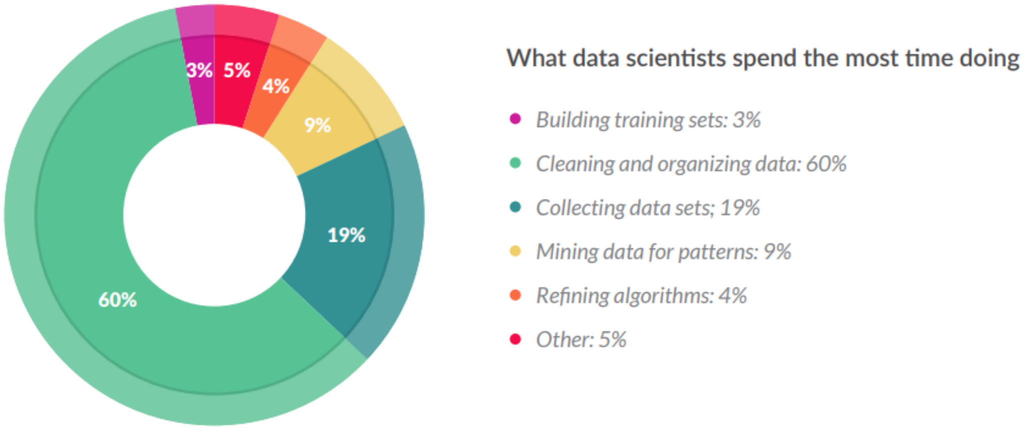


Figure 18: Time spent on the different phases of a data science project[18]

4.5 Features engineering

In machine learning, a **feature** is an input variable that will be used by the model to make predictions.

Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data[19].

In other words, feature engineering is about creating new input features from the existing data, in order to extract more useful information that describes more precisely the problem being predicted and that will ultimately improve pattern detection and the overall performance of the model.

A citation from Prof. Pedro Domingos, author of the famous book “The Master Algorithm”, helps to understand the importance of this phase:

“At the end of the day, some machine learning projects succeed and some fail. What makes the difference? Easily the most important factor is the features used.”

— Prof. Pedro Domingos

Features are created via a sequence of data transformation steps usually involving some programming (examples of languages often used in this phase are SQL and Python). Data transformation operations may include rescaling, discretization, normalization, data mapping, aggregations, ratios etc... Furthermore, dimensionality reduction[20] techniques can be applied to reduce the number of input variables.

[18] Source: CrowdFlower Data Science Report, 2016.
[19] Source: Jason Brownlee, Ph.D., Machine Learning Mastery.
[20] An example of dimensionality reduction technique is the Principal component analysis (PCA).

Given that the features can be the result of several data transformation steps, the link with the original raw data can be difficult to reconstruct. For this reason, it is important to well document the steps applied to generate the features.

Currently, most common data science development platforms[21] allow the creation of graphical flows representing all steps of data extraction/preparation and feature engineering in order to track all the steps performed, document them (for auditability) and be able to re-execute them if needed. Nevertheless, the reasoning behind each step of the feature engineering needs to be thoroughly documented.

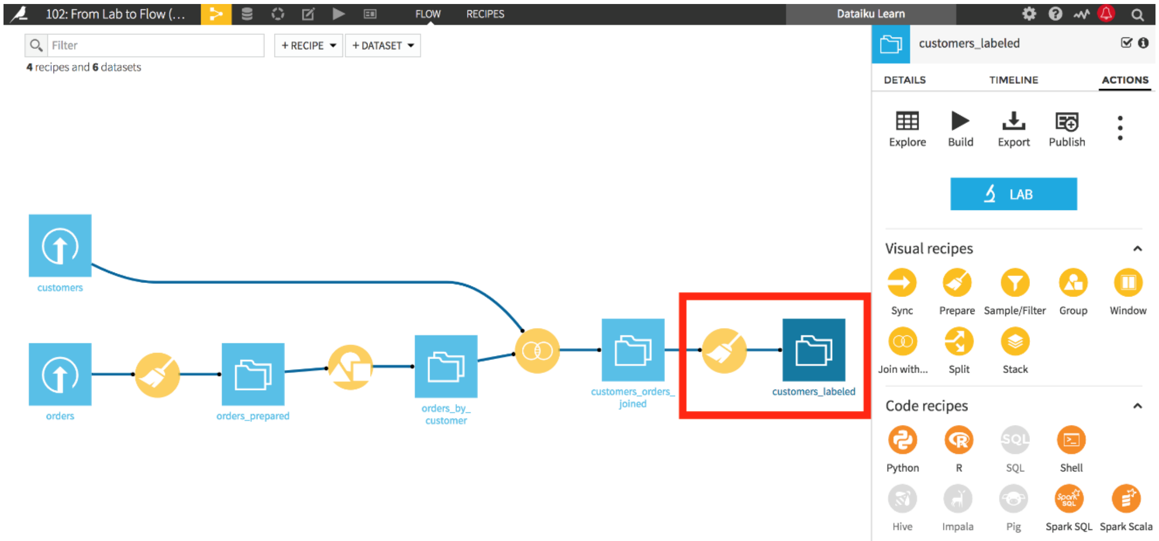


Figure 19: Example of visual representation of a data flow for data preparation and feature engineering in DataIKU (source: DataIKU)

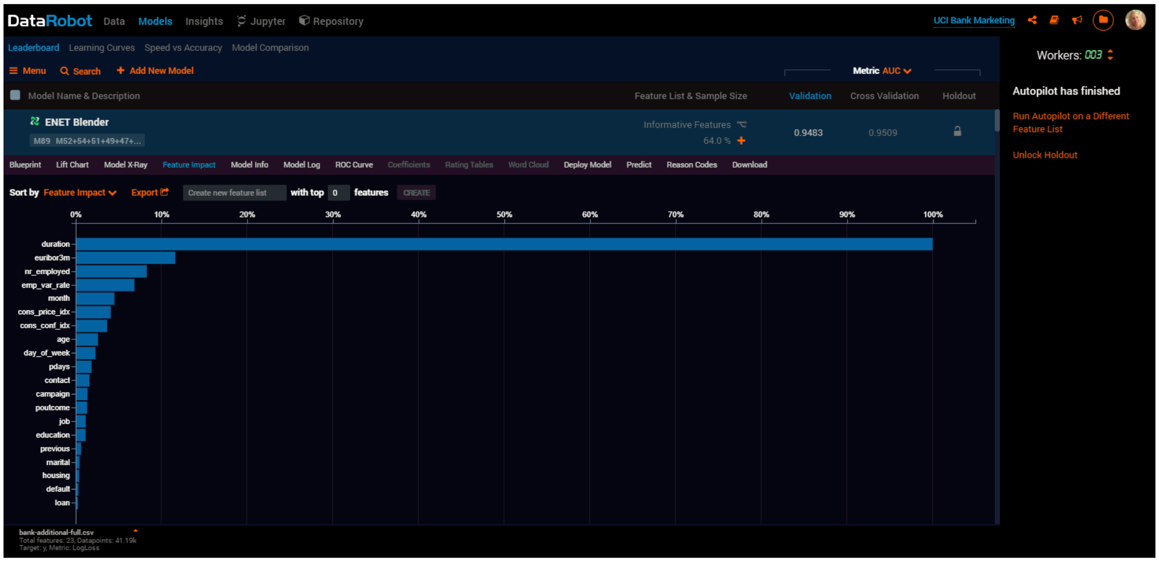
Features can very heavily impact the performance of the model, in a good or in a bad way. For example, a feature that contributes alone in a predominant way to the prediction of the model is not a good sign, since it means that the final prediction will strongly depend on the value of only that variable instead of being linked proportionally to all features. This could lead to inaccurate results and even to discrimination[22], for example when the variable contributing too much to the result is something like "gender".

The importance of each feature as contributor to the prediction of the model can be measured via a feature ranking report (also called Feature importance diagram or Feature impact diagram).

[21] The data science development platforms are integrated platforms covering all the steps of a data science/ML development project. Examples are DataRobot, DataIKU, Microsoft Azure Machine Learning Studio, Google Cloud Machine Learning, Amazon AWS SageMaker, etc.

[22] More details are provided in section 8.2.3.

In the extreme case, an issue that could occur is the **leakage**, which is the accidental inclusion of outcome information into the input features (outcome information should not be used as input data). This can be easily spotted in a feature ranking report since that feature will have an exceptionally high impact (example in the diagram below).



Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
ENET Blender M89 M52+54+51+49+47+...	Informative Features 64.0 %	0.9483	0.9509	

bank-additional-full.csv
Total features: 23, Datapoints: 41.19k
Target: y, Metric: LogLoss

Figure 20: Feature impact diagram – leakage (source: DataRobot)

4.6 Modeling

In the modeling phase, selected algorithms are trained in order to generate the models.

A **model** is the representation of what a machine learning algorithm has learnt from the training data.

In order to create the model, some steps need to be followed.

First of all, a few algorithms need to be selected (shortlisted) to be trained and evaluated. Indeed, as said above, several algorithms exist that can solve a certain class of problems. It is generally a best practice to try a few in order to pick the best one for the specific use case[23].

[23] Note: there is no perfect algorithm for a specific use case, but certainly some algorithms are more appropriate than others. The only way to find the right model is by trying different algorithms. Work on this topic has been initiated by the SnT (University of Luxembourg).

4.6.1 Training data, validation data, testing data

Then, the training data needs to be identified. A common technique is to split the data available into 3 groups: training data, validation data, test data[24].

The **training dataset**, as the term suggests, is the portion of data that will be used to train the model and ensure it is “fit”.

The **validation dataset** will be used in a following step to validate the predicting capacity of a trained model and tune the model.

Finally, the **test dataset** will be used during the back-testing phase in order to provide an “unbiased” evaluation of the final (trained, fit and tuned) model.

4.6.2 Model training

The algorithms shortlisted at this stage need to be trained on the learning data in order to build the models.

The first thing to check after having built a model is whether it is fit. A model is **fit** when it provides a good representation of the underlying problem without being overfit or underfit. Overfitting and underfitting are common issues that can be faced during the modeling phase.

Overfit models are models that during the learning phase have also captured the noise in the data, leading to an overly complex, unreliable predictive model; the model has learnt too many details and is so tightly fit to the underlying data set (including its noise or inherent error in the dataset), that it performs poorly at making predictions when new data comes in. This problem often happens when too many features have been selected as input for the model.

On the other hand, **underfitting** occurs when the model has not captured the underlying patterns in the data and is therefore too generic for good predictions. This often happens when the model does not have enough relevant features.

Therefore, in order to avoid becoming too specific (with too many features) or too vague (with not enough features), it is important to select the right features with the right amount of predictive information.

[24] A typical split can be 50% for training, and 25% each for validation and testing. When the validation step is not performed, the percentage can vary till up 80% for training and 20% for testing (this comes from Pareto Law, for which 80% of the data is representative of the entire population).

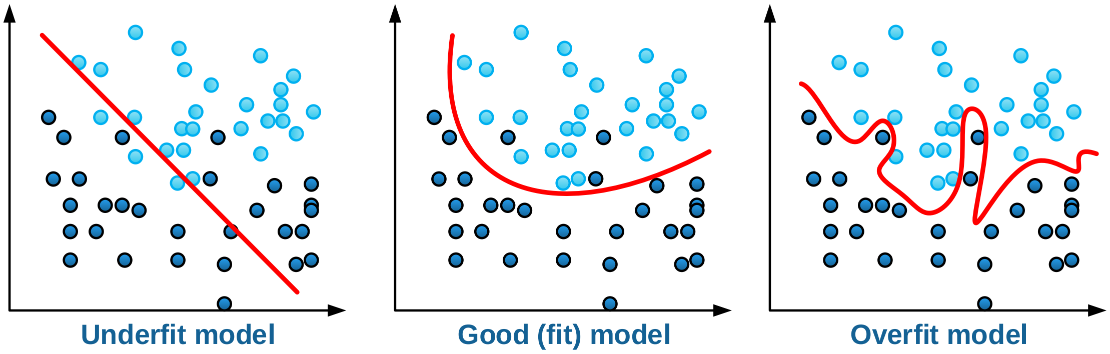


Figure 21: Common Data Preparation Issues

4.6.3 Model validation

After having trained the models, the models are tested against the validation dataset (which, as said above, is not part of the training data, and therefore a good base for testing) in order to check the quality of the prediction. This phase is called "**model validation**".

During this phase, models are calibrated/**tuned** by adjusting their **hyperparameters**[25]. Examples of hyperparameters are the depth of the trees in the Decision tree algorithm, the number of trees in a Random Forest algorithm, or the number of clusters k in the k-means algorithm.

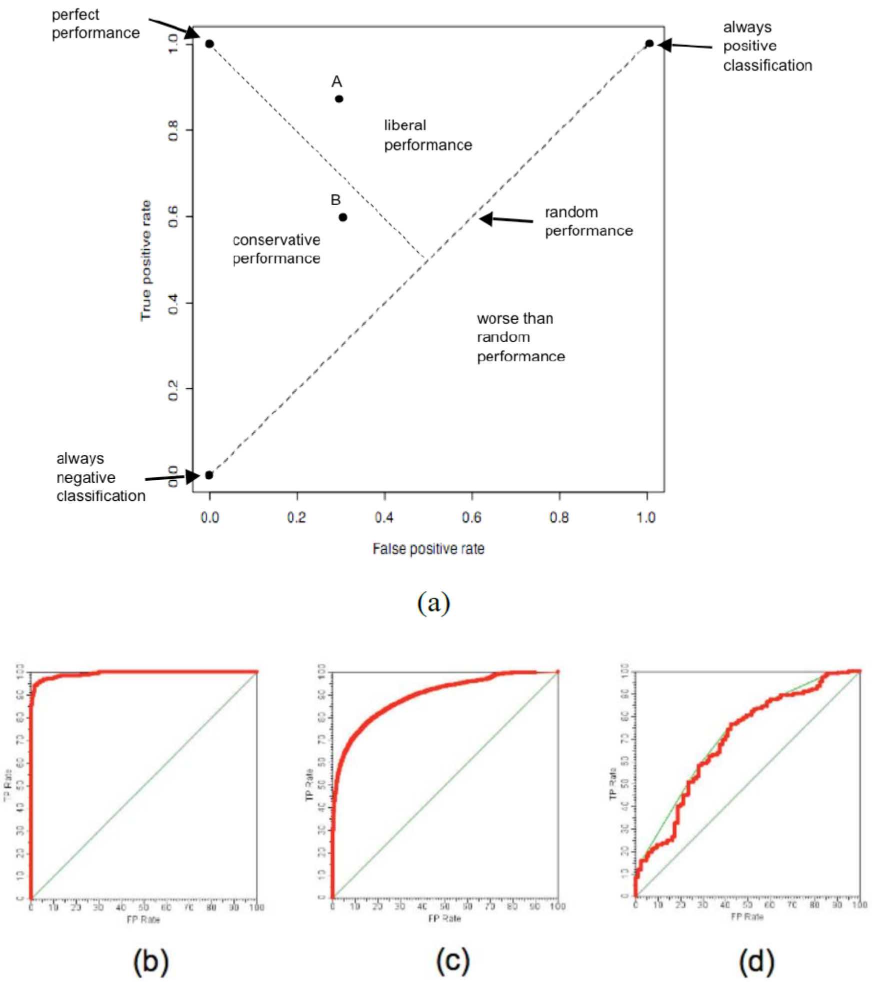
In addition to the simplest model validation technique that uses only one validation dataset, there is an interesting technique called "K-fold cross validation" which allows to build more robust models. This technique consists of dividing the data into k subsets, each one of which is used as the validation set while the other k-1 subsets are combined to form the training set. The results of the k validation tests are compared to identify the most performant model and to check its robustness (in this particular context, the term "robust" means its sensitivity to the noise in the training data). Of course, this technique is also more time and resource consuming.

4.6.4 Model evaluation and selection

After passing the validation step, the performances of the different models are measured and compared in order to select the best one.

Examples of statistical metrics that can be used to evaluate the model performance are the ROC AUC (Area Under Curve), the Confusion Matrix (which compares the predicted values with the actual values from the test dataset), or the F-1 score (which is calculated based on the confusion matrix and represents the ideal cut-off between Precision and Recall).

[25] **Hyperparameters** are settings of the algorithm that are fixed and can be adjusted manually in order to tune the performance, as opposed to the model parameters whose value is set by the algorithm via training/learning.



(a)

(b)

(c)

(d)

Figure 22: ROC curves: (a) regions of a ROC graph (a) an almost perfect classifier (b) a reasonable classifier (c) a poor classifier[26]

[26] Figure credits: (a) Lutz Hamel, "Model Assessment with ROC Curves"; (b), (c), and (d) Peter Flach, "ICML'04 tutorial on ROC analysis" - International Conference on Machine Learning, 2004.

		ACTUAL LABEL	
		Positive	Negative
PREDICTED LABEL	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Figure 23: Confusion matrix

METRIC	DEFINITION	FORMULA
ACCURACY	Percentage of predictions that are correct	(TP +TN) / (TP+TN+FP+FN)
PRECISION	Percentage of positive predictions that are correct	TP / (TP+ FP)
SENSITIVITY (RECALL)	Percentage of positive cases that were predicted as positive	TP / (TP+ FN)
SPECIFICITY	Percentage of negative cases that were predicted as negative	TN / (TN + FP)

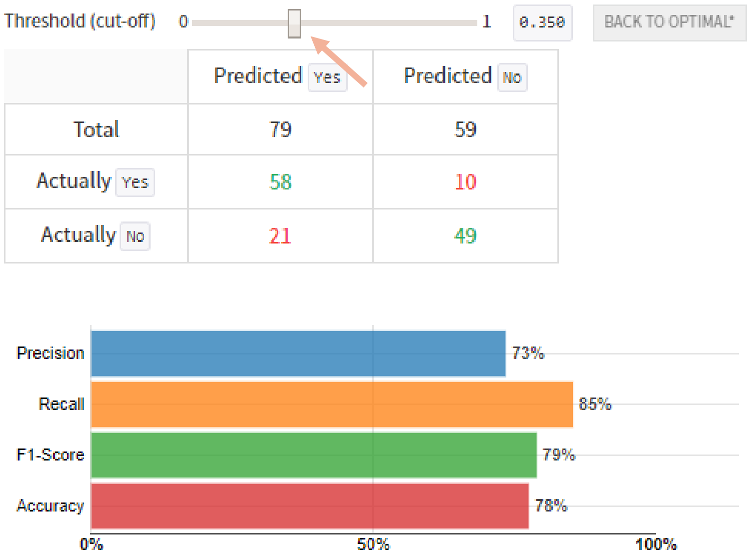
Figure 24: Statistics evaluation metrics based on confusion matrix (for supervised learning)

4.7 Back-testing

During the back-testing phase, the model that has been selected after the modeling phase is once again tested with fresh new data (the testing dataset). The **cut-off** is selected together with the business users (and the model is tuned accordingly) in order to define the optimum, acceptable balance (from a business perspective) between false positives and false negatives.

For example, if the model is used to predict a fraud, too many false positives would lead to a high number of cases being investigated (and therefore high operational cost), but, on the other hand, too many false negatives would mean that the model is not accurate and several frauds will be missed, leading to high financial losses. It is important that the cut-off is selected and approved by the business users.

The F1 score can be used in order to find the right compromise between accuracy, precision and recall.



Threshold (cut-off) 0 1 0.350

	Predicted Yes	Predicted No
Total	79	59
Actually Yes	58	10
Actually No	21	49

Figure 25: Cut-off and F-1 score. The cut-off can be adjusted in order to find the right balance of accuracy, precision and recall (source: DataIKU)

It is important to note that although the flow has been described as sequential, the steps can be reiterated multiple times (e.g. tune model, define new features, etc.) in order to progressively improve the model performance.

4.8 Business Impact evaluation

After the model has reached a satisfactory level of performance, measured based on purely statistical measures, it is important to measure the overall business benefits and risks that the implementation of the model into production would bring, i.e. the business metrics.

As an example, if the model is used to detect frauds, the number of true positives can be used to calculate the benefit (in terms of missed losses), while the number of false negatives and false positives can be used to estimate the risk in terms of financial losses and operational impact (Full Time Equivalent for the time spent/lost on the investigation).

The KPIs (business metrics) used in this phase are those that were defined initially in the phase Business Understanding.

4.9 Validation

Based on the KPIs/business metrics evaluated in previous phase, the choice to go into production has to be based on the evaluation of the overall net gain brought by the introduction of the model. If the estimated net gain is not sufficient (also considering the accuracy of the model, based on the cut-off choice), the model will be abandoned, otherwise it is validated. The choice of validating the model must be made by the business departments and be based on a business evaluation.

4.10 Industrialization (production)

Finally, after the model has been validated, it can be deployed into production (according to the standard change management process).

As part of the industrialization process, the interfaces with the other applications (e.g. via standard APIs[27]) are built in order to integrate the machine learning model inside the business workflow and related applications.

It is important, during the integration, to think about the overall business process and the level of automation required, taking into consideration the underlying risks. In particular, a fully automatic flow (with no human in the loop) means that the choice is completely left to the ML model. This can be acceptable in certain contexts/use cases, but not in others. The choice has to be justified and documented considering the overall risks.

Another important aspect of the industrialization phase is that once in production, the performance of the model needs to be continuously monitored, to promptly detect if the model accuracy would suddenly worsen. In that case, the model might need to be updated by retraining it on new data.

In the example of fraud detection, the model will most probably need to be periodically retrained once new patterns of frauds are discovered. In other use case implementations, the frequency of the model update/re-learning can even be in real time (this is called “online learning” as opposed to “offline learning”), like in the example of the recommender system[28]. This can be made possible only if the underlying architecture allows that. It should be noted that a learning phase, depending on the quantity of the data used for learning, can take a reasonable amount of time (even a couple of weeks), therefore limiting the possibility for online learning.

[27] Application Program Interfaces

[28] Recommender systems are used to recommend commercial products to customers based on previous customer consumption patterns. Such patterns evolve so rapidly and the market volume is so high that even a small performance improvement can bring a huge gain, therefore justifying the investment for an online training.

4.11 A paradigm shift

As described in previous paragraphs, the development process of an AI/ML project presents some unique characteristics which differ from a standard development lifecycle, although the overarching change management process remains more or less the same.

For example, as part of a standard **change management** process, it is important that a preliminary risk analysis phase (including the analysis of data privacy, security and legal and compliance risks) is performed during the early stages of the project. The development should then be carried out on different development/test environments segregated from the production environment, like it is the case for any other project. Similarly, at the end of the development cycle the AI/ML model needs to undergo the integration test phase to ensure that there is no connectivity issue with the other systems before going live. Finally, before deploying the AI/ML project, there should be an approval step involving a change management committee (e.g. Change Management Board) including business, IT, risk and information security stakeholders which is the same for all other projects.

Therefore, an AI/ML project needs to follow the standard change management process of the company in order to be efficient (e.g. from a prioritization and resource/budget allocation point of view) and centrally monitored, also from a risk perspective.

Nevertheless, a lot of things differ in an AI/ML project compared to a classic IT development project, as described in the next paragraphs.

4.11.1 The importance of data

The main difference between an AI/ML project and a classic IT development project is certainly that data is playing a central role, so much so that even the notion of “test data” has changed. Fake data purely created for test purposes or anonymized data would in most cases not be sufficient to train the algorithm correctly. As a consequence, the data in the development/test environments is often a copy of production data (refreshed with a certain frequency, e.g. daily). It is therefore important to carefully select the data required and limit the access to sensitive data by ensuring that the “Need to know” principle is respected also among the data scientists working on the AI/ML project. Some recent progress on privacy preserving technologies, like homomorphic encryption techniques for example, can help to ensure that confidential data is accessed only by the algorithm for the training needs while keeping the user access to such data limited to the strict minimum (the data is accessed only by the algorithm). Furthermore, as more extensively explained in section 4.5 and 8.2.3, the data to be used as input for the AI/ML project should be carefully selected to limit the risk of unfairness.

4.11.2 A new development lifecycle

In AI/ML projects, the development lifecycle does not follow “classic” development approaches like waterfall or even agile methodologies, but a specific data science process. There is no design phase to define the functionalities of the end product (e.g. via the definition of the functional and technical specifications), since the model (i.e. the equivalent of the end product) is not being developed by the developers but built, together with its parameters, by the algorithm. The focus of the entire approach is on the data (being fed to the algorithm and the model) rather than on building a program which is a predefined set of rules. Here, the rules are defined by the algorithm itself by learning from the data.

Nevertheless, the documentation of the chain of steps applied to manipulate the raw data until the feature creation (i.e. the data ultimately fed to the model) is very important, and without an integrated platform[29] this task can be very challenging and might raise some risks in terms of traceability and explainability of the model outputs.

4.11.3 Simple software code, complicated mathematical concepts

Generally, the portion of software code developed as part of an AI/ML project is relatively small and fairly simple, compared to the amount of code composing a classic software application. On the other hand, such small portions of code in an AI/ML project hide complex mathematical concepts and formulas which are much more complicated than the code itself, and require solid mathematical skills in order to be understood. This is the reason why a standard IT developer profile would not be able to fully understand the code and also the reason why a separate team composed of "data scientists" is required to work on AI/ML projects.

4.11.4 The data scientist role

Due to the mathematical (and statistical) concepts embedded in AI/ML algorithms, **data scientists** working on an AI/ML project should have a strong mathematical background, most of the time stronger than the IT programming skills. Nevertheless, strong IT programming skills are required as well when the model needs to be deployed into production and performances and production systems stability are at stake, to ensure that the code is performing well and fast, without causing any unnecessary slowness or impact on other common resources in integrated production environments (e.g. network, CPU, database, etc.). Finally, data scientists also require some good business understanding, without which the entire AI/ML project would not succeed. Given that it is difficult to find all these characteristics in one person, often the data scientist role is represented by a team of specialists with complementary profiles, a larger part with a stronger mathematical background and with basic programming knowledge, few others with more advanced programming and database skills, guided by a team leader with a more holistic view and a deeper understanding of the business requirements.

4.11.5 Deployment into production

Due to the same complex mathematical concepts embedded in the model, in most cases the IT department is not able to estimate the real impacts of the code being deployed into the production environment. Hence, it is important that technical tests are performed before the go-live to validate the interfaces with the other systems and that safeguards are taken in order to closely monitor in

[29] Ref. data science development platforms described at section 4.5

real-time the resource consumption of the model. For example, the execution time of the model at each run should not exceed certain thresholds and in case it exceeds them, technical safeguards should be in place in order to automatically bypass the model and avoid any impact in cascade on the rest of the environment (this risk being particularly relevant in integrated environments). Similarly, a good practice consists of executing parallel runs for a certain time period, which would constitute the "user acceptance testing" for the end users and at the same time validate that the charges (e.g. resource consumption) on the production environment are acceptable.

5 SELECTED USE CASES (SPECIFIC TO THE FINANCIAL SECTOR)

The following paragraphs describe some of the most relevant use cases applicable to the financial sector. The goal is to describe the business purpose and at the same time highlight some of the main risks associated with such kind of implementations. The risks so identified are presented in more detail in the section 8.2.

5.1 RPA & Intelligent Process Automation

RPA, standing for Robotic Process Automation, are systems allowing to automate highly repetitive tasks which normally represent low value-added tasks for humans. In the financial sector, these systems are usually employed for automating back office tasks, where they can bring high returns on investment. Furthermore, they are particularly useful in use cases involving highly heterogeneous systems that are not very integrated with each other (e.g. legacy systems or proprietary systems), where they can be used to remove the need for human intervention at each step of the workflow and replace it with an automatic agent, which is just simulating exactly what the human does but in an automated way. The result is a more efficient workflow requiring less human interventions, leading to lower risks of manual errors and reduced operational costs.

Examples of tasks that can be automated via RPA are:

- Create temporary MS Excel/Word documents
- Export data from one office application to another (from example from MS Word into MS Excel)
- Formatting data
- Generate tables and diagrams
- Launch MS Excel macros/VBA code
- Save a file in a specific folder
- Open emails and export the attachments
- Send emails
- Login into web application by inputting the credentials via the web interface
- Click on specific buttons on web applications
- Export files
- …

The automatization of the tasks often consists in simply simulating via a software agent (bot) the user interaction to launch the process while passing on as input parameters the results of the previous task. As such, the pure automatization via RPA is mostly a rule-based system not requiring advanced AI.

However, nowadays RPA has moved into Intelligent Process Automation (IPA), combining AI and machine learning functionalities, such as **NLP** and **text mining**. For example, the NLP/text mining engine can analyze a scanned document and automatically classify it according to its category (e.g. ID document, invoice, payment receipt, ….), making it possible to automatize entire parts of middle and back office processes like account opening, loan granting, invoicing and payment reconciliation, etc...

It is important to note that RPA can make some manual steps automatic but it should not be used to replace the critical steps requiring human validation with a machine.

Furthermore, standard security controls should be applied to restrict and secure the access to the credentials used by the RPA for launching the various steps of the workflow, and ensure proper tracking and auditability of all actions. Actions performed by the bots should be clearly flagged as such in the audit log.

Additional aspects to be considered during RPA implementation projects are the monitoring of the execution of RPA automated processes and related error management, which should be integrated into the standard IT operations processes and procedures of the institution.

Usually RPA comes in the form of predefined offers from vendors providing the tool and the expertise for implementation. However, the dependency on the external providers for the maintenance of the RPA should not be underestimated and the institution should as well evaluate the need to train internal staff for the ongoing maintenance and evolution of the RPA process.

Finally, it is important to note that although RPA/IPA can strongly improve an operational process by automating it and therefore making it more rapid and less prone to manual errors, these technologies do not improve the process itself. Sometimes, more efficiencies could be gained by reengineering the entire process, i.e. reviewing all the phases and interactions composing the process and implementing an STP (Straight Through Processing).

5.2 Chatbots

Chatbots are virtual assistants that are implemented by institutions in order to help their customers with some common frequent questions.

The technology used by chatbots usually relies on an NLP (Natural Language Processing) engine and on machine learning to learn new knowledge and improve over time. The NLP engine is used to interact with the customer, capture the main concepts included in the question raised by the customer and the related context.

In their simpler versions, the machine learning algorithm used is the decision tree. The decision tree is configured in order to start from generic questions and then guide the user through a set of more specific questions in order to reach a conclusion. In this case, the concept of explainability is quite well represented, since the decision tree allows to easily follow the branches of the tree till the answers on the leaves. Nevertheless, more complicated algorithms may be used (e.g. neural networks), leading to the risk of “black boxes”.

Currently, chatbots are frequently used to help customers with solving simple questions and requests (e.g. “tell me what my account balance is”), or providing a first level of help desk in order to better redirect the customer to the appropriate customer service. In both cases, the operational risk is reduced due to the limited scope of the chatbot, and because there are no automatic decisions generated by the chatbot. The level of integration of the chatbot within a business process remains limited, and there is a human control after the customer has interacted with the chatbot.

Nevertheless, the reputational risk can quickly become devastating in case the bot can automatically learn new knowledge from the user interactions without control. An extreme example is Microsoft chatbot Tay, who had to be removed from Twitter after a few hours since it learnt racist behavior from other users.

Therefore, it is important that the input data used by the machine learning algorithm to learn is subject to scrutiny, and that performances of the bot are constantly monitored to identify potential deviations.

Another aspect to consider when implementing a chatbot is that the user should be informed at the beginning that he/she is interacting with a chatbot, for reasons of transparency towards the user as well as to avoid any sort of misunderstanding.

Finally, it is also important that the scope of the application of the chatbot remains limited and that humans are taking over when more critical topics are being discussed. The responsibility and accountability for the information provided and actions done by the chatbot, even when of purely informational nature, remain with the institution.

5.3 Robo-advisors

More advanced forms of intelligent assistants and types of usage are emerging, like the robo-advisors, which provide advice to clients, especially regarding proposed investments. This technology enables access to investment advice for a larger customer base and therefore favors inclusion, by reaching retail customers that normally cannot afford the cost of a specialized investment manager.

Today, a large part of the robo-advisors existing on the market have limited capabilities consisting in simply allocating to clients one among a set of predefined portfolios (most of the time calibrated by humans) based on the data obtained via a questionnaire. Nevertheless, enhanced versions of robo-advisors based on artificial intelligence are emerging, where the automated advice relies on ML models (which may leverage on big data[30]) and include dynamic portfolio optimization features.

Particular attention should be paid to the algorithm used by the robo-advisor to ensure that it does not favor investment funds with higher commissions. Moreover, the financial institutions offering robo-advisors should regularly monitor the effectiveness and appropriateness (in line with MiFID II requirements) of the advice provided to avoid mis-selling. Precautionary mechanisms should be in place to be able to suspend the provision of advice should errors or bias be detected.

Finally, poor investment advice generated by robo-advisors could quickly be amplified on large scale (even leading to financial stability risks) especially when different robo-advisors use the same algorithms (e.g. using packages from external providers) with insufficient customization.

[30] An example of big data used by advanced robo-advisors is the alternative data gathered from news feeds, social media, blogs, online communities, satellite images, online reviews, etc.

5.4 Fraud detection/money laundering investigations

Recently, AI has been applied with satisfactory results to assist with the detection of frauds and of suspicious activities that may be linked with money laundering more generally.

In particular, historical data about past transactions and confirmed frauds are used to train a supervised machine learning algorithm. The model so generated is able to identify the patterns of past frauds and use them to detect new ones more efficiently, with a higher accuracy rate (and therefore a lower number of false positives) compared to the standard methods, simply based on a limited number of rules and predefined criteria identified by human specialists. In order for the accuracy to remain high, the model needs to be periodically re-trained when new patterns of fraud are identified. Similarly, the accuracy needs to be constantly monitored to promptly identify deviations.

The suspected frauds detected by the ML model are then manually verified by the compliance experts who can now concentrate on less cases to investigate. This brings a higher detection rate, high efficiency, reduced operational costs, and also lower reputational risk (due to less frauds as well as less investigations on cases of suspected fraud that end up being a false positive, which still leaves a negative impact on the client relationship).

New ideas are appearing regarding the possibility to share such advanced fraud detection models among different financial entities of the same group. This can create synergies (smaller entities of the group can benefit from using advanced models that they would not be able to develop by themselves) and further improve the detection rate when the set of fraudulent transactions inputted into the model is fed by multiple entities, contributing with different fraud schemes. At the same time, this approach raises concerns regarding the anonymization of the data before being shared.

5.5 Terrorism financing

As regards investigations related to the terrorism financing in particular, AI is used more in the form of unsupervised machine learning.

With unsupervised learning, data is grouped into clusters according to different criteria. The goal is to identify the "outliers", i.e. those individuals or transactions that do no share common patterns with the rest of the population and represent anomalous behavior. For example, individuals buying several airplane tickets or renting several cars abroad without ever leaving the country may look suspicious. This is in contrast with the method of supervised learning described above for fraud detection, which is used to detect patterns already seen in historical data.

For terrorism financing, supervised learning is difficult to apply due to the lack of sufficient historical data and confirmed events (transactions related to terrorism financing are only rarely identified).The unsupervised learning and the anomaly investigation can instead help to efficiently identify unusual activities, although the investigations are more laborious and time consuming.

It should be noted that the grouping (clustering) of customers and transactions according to common criteria described above corresponds to a profiling activity. Although the use of personal data may be allowed if required to comply with a legal obligation (such as AML obligations), the challenge for the bank/institution collecting the profiling information is not to use the information for other commercial purposes (which may not be allowed under data protection[31] regulation). Furthermore, adequate protection measures should be implemented to limit as far as possible the access to personal data to data scientists during the various phases of model development[32] (like for example data encryption or pseudonymization).

5.6 Credit scoring (automated loan decisions)

Automated credit scoring is not something new, and probably one of the first financial problems to which statistical modeling was applied[33].

Today, credit scoring methods are often based on machine learning using mainly regression, decision trees, and statistical analysis. Credit scoring is now used for automating loan decisions (or at least large part of the process), for both corporate and retail customers.

Starting from transactional data, credit history and account balance information, the machine learning models allow to get a better understanding of the cash flows and give a more accurate estimation of the credit risk.

Results are more reliable when a larger set of transactional data is available. In this sense, PSD2 (and its incorporation into national Law) makes this possible by forcing the banks, subject to prior consent from the customer, to share the customer transactional data with other financial sector players.

As regards retail customers, new types of data have been exploited recently, especially by FinTechs, including, for instance, social media activity and mobile phone usage[34]. Big data analytics is also applied to analyze big quantities of data in a short time, like for example the analysis of the timely payment of telephone, electricity and other utility bills. This enables access to new variables such as the “consumption behavior” and “ability to pay” in order to calculate a credit score even for individuals that do not have enough credit history to be “scorable” with traditional models.

While this new trend allows greater access to credits and more accurate scoring, concerns have been raised regarding the protection of personal data, which should be used only in the context of the lending process and subject to clear and informed consent from the customer.

Automated (or partially automated) loan processes based on these models will have reduced operational costs and will therefore allow a faster credit decision for the customer. In addition,

[31] Ref. to General Data Protection Regulation (GDPR): data collected for a specific processing cannot be used for a new processing unless its purpose is compatible with the original (authorized) processing or a new legal basis is applicable.
[32] See section 4 for more details.
[33] Already in 1960 the FICO credit score was created, mainly based on statistical analysis
[34] Source: FSB, “Artificial intelligence and machine learning in financial services” pages 12-13, November 2017.

customers will have easier and quicker access to offers from competitors, enhancing market transparency.

Nevertheless, attention should be paid to avoid that the model includes bias. For example, populations not sufficiently represented in the input dataset used to train the model may lead to discrimination. As an example, until the recent past (1972), married women in Luxembourg were not allowed to hold a bank account (and even less a loan) under their name, without the husband's signature which was required. Taking historical data as representative of the credit capacity of women would probably lead to wrong conclusions.

Furthermore, complex models may result in a lack of transparency for customers, to which it would be difficult to explain why a credit request was rejected. It should be noted that according to data protection regulation, the customer (data subject) has the right "not to be subject to a decision based solely on automated processing" (GDPR, art. 22).

Finally, if high volumes of credit decisions are automated, errors eventually included in the model would be amplified. Similarly, if a few models developed by external providers gain large adoption, design flaws or wrong assumptions contained in the models may have systemic effects.

5.7 Other use cases

The use cases of AI and machine learning implementation listed above are only a limited part of those that can be seen today, but they are representative of some of the main issues and risks that this technology brings.

Some other AI applications worth mentioning are the following:

- NLP and Text mining to analyze big quantities of legal documents
- Reinforcement learning applied to algorithmic trading
- NLP for customer sentiment analysis used at customer support centers and as well to assist wealth managers in proposing more tailored offers to clients
- Use of computer vision technology for facial recognition in video KYC processes
- Deep learning applied to track cryptocurrencies and perform money laundering investigations on blockchain transactions[35]
- Forensic solutions to perform in-depth fraud investigations based on analysis of emails and other log files
- Machine learning and deep learning for advanced IT security solutions (e.g. SIEM, antivirus, email filters, etc.) used for detecting cyberattacks or suspicious activity via behavioral monitoring, or to filter undesired emails such as emails containing malware or SPAM
- AI and machine learning tools to assist with risk, compliance and auditing tasks, as well as to assist regulators in their supervision tasks (RegTech and SupTech). For example, the BdI, MAS, and SEC are exploring the use of AI to investigate frauds and money laundering suspicious activities.

[35] Example of a company active in this sector: https://www.neutrino.nu/

6 EXPLAINABLE AI

Depending on the criticality of the use case to which AI is applied and the level of automatic decision made, there is a need for explainability. Indeed, while non-critical systems for which the risk associated with a single wrong prediction is very low (e.g. recommender systems suggesting articles to buy to potential customers), for more critical systems or systems that take decisions autonomously, explainability is fundamental in order to understand and validate the internal behavior of the system.

Machine learning algorithms have a varying degree of intrinsic explainability (ref. figure 26 below) and in most cases they contribute to create "black box" models. Furthermore, more accurate algorithms (e.g. deep learning) are those that usually lead to less explainable models. To help understand the model despite the opacity of the underlying algorithms, techniques exist that generate explanations for the models.

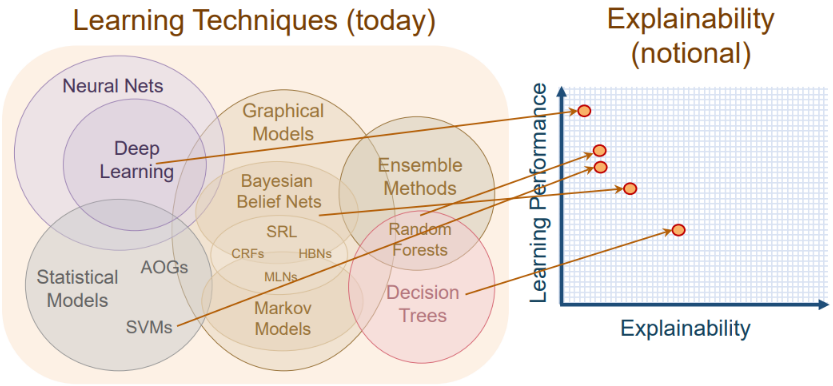


Figure 26: algorithms explainability[36]

Explainability techniques can be either **model agnostic** or **specific to the learning algorithm**. The former are essentially techniques that can be applied to any "black box" model without knowing how it works internally, whilst the latter are techniques that directly probe the internal functioning of the model. Some examples of both types are illustrated below.

6.1 Model agnostic explanations

6.1.1 Sensitivity analysis (feature importance)

Sensitivity analysis is a simple technique already used in other fields (e.g. to understand complex electrical circuits) that consists of slightly perturbing a single input feature and measuring the

[36] Source : DARPA, "Explainable Artificial Intelligence (XAI)", 2017

change in the output. The process is repeated iteratively for many input values and for each of the different input features in order to come up with a representation of the model behavior. This method is often used to create diagrams illustrating the importance of features.

This approach is fairly simple and can be applied to any type of model, including deep learning where it can be used to extract the importance of the single pixels in image classification problems.

Nevertheless, this technique is so simple that it does not capture the correlation among different features.

6.1.2 Local Interpretable Model-agnostic Explanations (LIME)

Unlike Sensitivity Analysis, the LIME (Local Interpretable Model-agnostic Explanations) technique tries to capture the interaction amongst features. It does so by generating a simplified model which approximates the underlying one, learned from perturbations (applied to multiple input features) around a particular prediction. The simpler model so generated represents the *explanation* of the model.

The key idea behind this method is that it is much easier to provide a *local* explanation of the model (i.e. related to the specific prediction we want to explain), instead of providing a *general* explanation (i.e. valid for all predictions of the model).

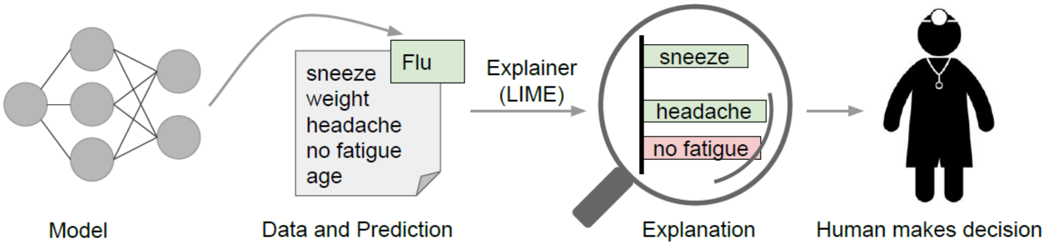


Figure 27: Explaining individual predictions by using a LIME model[37].

In the example depicted in figure 27, the model predicts that the patient has a flu, and the LIME model highlights the symptoms (input features) that have contributed to that prediction (in this case: sneeze and headache) and those that represent evidence against that conclusion (in this case: "no fatigue"). With these elements, the doctor can finally take an informed decision on whether or not to trust the model conclusions. It should be noted that the symptoms highlighted, as explained above, represent a *local* explanation valid for the particular patient, but may not represent how the model works for *all* patients.

[37] Source: Marco Tulio Ribeiro, "Why Should I Trust You? - Explaining the Predictions of Any Classifier", 2016.

This method may be helpful also to identify instances of data leakage. Data leakage[38] is the inadvertent leakage of output information (or information strongly correlated with the output) into the input data (training and validation datasets). An example[39] is one where the patient ID, included in the training and validation dataset, was found to be heavily correlated with the target class (output). This issue could have been easily spotted by using techniques such as LIME, since the patient ID would be listed in the explanation[40].

6.2 Explanations specific to the learning algorithm

More detailed explanations can be achieved by using techniques, such as interpreters, which are specific to the learning algorithm to be explained.

6.2.1 Tree interpreter

Decision trees are probably the type of algorithm that by their very nature are more easily interpretable. Furthermore, their internal behavior can be further explained by using a tree interpreter. Indeed, even in cases where the decision tree is very deep (contains many levels), an interpreter may be able to analyze it and draw the main decision steps within that tree that predominantly contribute to the final decision/prediction.

A **random forest** algorithm is a variation of the decision tree consisting in training many similar versions of decision trees and making the final decision via a majority vote of the individual trees. This makes the random forest algorithm much more accurate than the decision tree. Also in this case, a tree interpreter can be used to analyze the single trees and then aggregate the results to draw the decision path. The explanations provided are valid both *locally* and *globally*.

6.2.2 Neural network interpreter

As opposed to decision trees, neural networks and deep neural networks ("DNN") are probably the most complex and difficult to explain type of algorithm. Nevertheless, methods exist to provide some explanation as to how they work. Recent research[41] proposed a method for interpreting DNNs based on the **Relevance Propagation** technique. Basically, this technique is the opposite of sensitivity analysis, in the sense that it starts with the output and then goes back through the different layers of the DNN, analyzing at each layer the *relevance* of the input from the preceding layer, until it reaches the input layer. The result is a sort of heatmap of the input features that mainly contributed to the output.

[38] See section 4.5 for more details.

[39] Source: S. Kaufman, S. Rosset, and C. Perlich., "Leakage in data mining: Formulation, detection, and avoidance", 2011.

[40] Note: this example of data leakage could have been also spotted by using the feature sensitivity analysis described at section 4.5 since it is due to a single feature, but (as opposed to sensitivity analysis) LIME could also help with other more complicated examples where the data leakage is represented by a combination of multiple features.

[41] Montavon et al., "Methods for Interpreting and Understanding Deep Neural Networks", 2017.

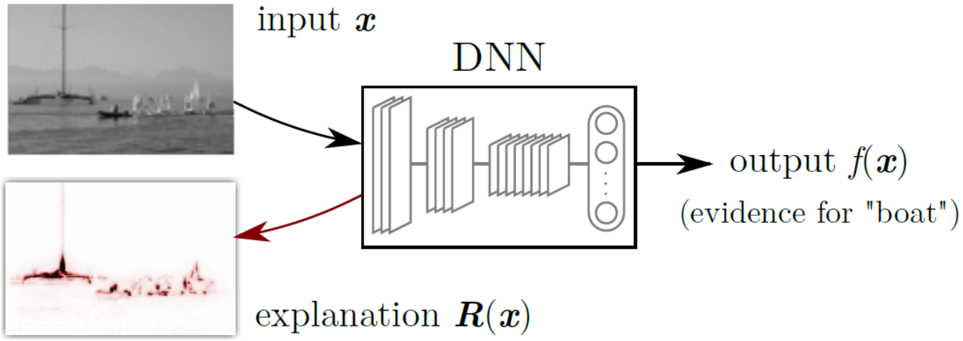

input $\boldsymbol{x}$
DNN
output $f(\boldsymbol{x})$
(evidence for "boat")
explanation $\boldsymbol{R}(\boldsymbol{x})$

Figure 28: Explanation of the DNN prediction "boat"[42]

In the example depicted in figure 28 above, an image x is classified by the DNN as a “boat”. This output is then mapped back to the input, and the result is a heatmap where the pixels with high relevance are colored in red.

6.3 The need for explanations

The techniques listed above are just examples of possible ways to generate explanations of machine learning models. Even when fairly basic, explanations can help to get a functional understanding of the behavior of machine learning systems and check whether bias is incorporated, which may lead to substantial errors or discrimination. For instance, they allow to spot whether data that may lead to discrimination (e.g. gender, age, address, etc.) have too much relevance in the final prediction.

These methods should therefore be used in conjunction with more quantitative methods to measure accuracy (e.g. confusion matrix, ROC curves[43], etc.) to validate a model before deploying it into production.

In models that are retrained and therefore updated in continuous mode, automatic controls can be defined to check for example that a single or a group of a few features is not a major contributor for the final decision. For example, an image classifier that has a single pixel as major contributor to the decision may be subject to adversarial attacks, and therefore an automatic rule such as “if a single pixel is the major contributor of a decision then raise an alert” may help to prevent this issue.

Furthermore, these techniques may allow off-the-shelf models included in commercial packages to be more understandable, and allow humans to understand the decisions proposed and make informed decisions accordingly.

[42] Source: Montavon et al., “Methods for Interpreting and Understanding Deep Neural Networks”, 2017.
[43] See section 4.6.4 for more details.

Finally, it should be noted that in case the output of the machine learning model is a decision affecting physical persons, according to GDPR[44] the person which is the subject of that decision has the right to be informed on how that decision was reached. Having explainability embedded into the model from the start can help satisfy this requirement.

[44] GDPR art. 22 and recital 71.

7 SECURITY AND ROBUSTNESS

Whenever there is a new technology trend, there are also new attack techniques exploiting security vulnerabilities, and AI is also victim of this universal rule.

Some of the main attacks affecting in particular machine learning are the following:

- Data poisoning
- Adversarial attacks
- Model stealing

The degree of resistance of an AI/ML solution to such security attacks is usually called **robustness**.

Furthermore, advanced technology developed using AI and machine learning can be misused for malicious purposes. An example is presented at the end of this chapter regarding video forgery.

7.1 Data poisoning

In poisoning attacks, attackers deliberately influence the training data to manipulate the results of a predictive model[45]. This type of attack is especially valid in case the model is exposed to the Internet in ***online*** mode, i.e. the model is continuously updated by learning from new data.

A fairly simple defense technique to implement against this attack is to properly filter the training data to detect and filter out anomalous data.

7.2 Adversarial attacks

Most machine learning models used for image classification (usually built with Deep Neural Networks or “DNN”) are highly vulnerable to adversarial attacks. An adversarial attack consists in providing a sample of input data which has been slightly perturbed in order to cause the model to misclassify it. In most cases, these perturbations (basically representing a noise) can be so subtle that a human eye does not even notice it.

In the example below, a noise (“nematode”) has been added to the input image of a panda in order to induce the classifier into a recognizing it as a gibbon with high confidence.

[45] See Jagielski et al., “Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning”, April 2018.

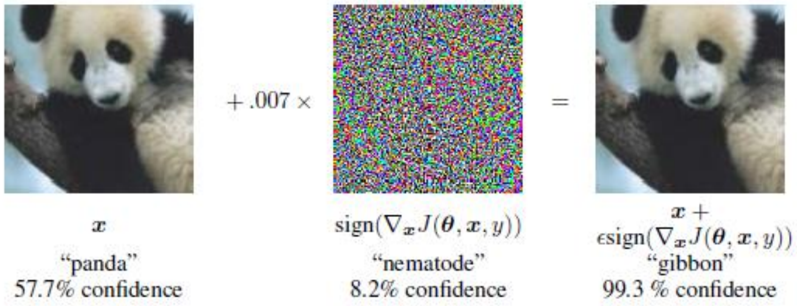

+ .007 ×
=
$\boldsymbol{x}$
"panda"
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Figure 29: Example of adversarial attack[46]

To realize how dangerous such attacks may be, it is worth mentioning that adversarial attacks may even be run on physical computer vision sensors in order to induce them into not seeing an obstacle or a traffic stop sign. Similarly, attackers could evade face recognition systems by wearing specially designed glasses.

Researchers have recently developed some new defense techniques to defeat these attacks[47], for example by adding a de-noiser before the input.

IBM Research[48] has recently released an open source library called Adversarial Robustness Toolbox ("ART") which includes state of the art adversarial (and poisoning) attack and defense techniques. The library, written in Python (the most common programming language for machine learning) can be used by developers to check the **robustness** of their models (in particular Deep Neural Networks – "DNN") against such attacks. The approach is three-fold:

- **Measuring model robustness.** Firstly, the robustness of a given DNN can be assessed either by recording the loss of accuracy on adversarially altered inputs, or by measuring how much the internal representations and the output of a DNN vary when small changes are applied to its inputs.
- **Model hardening.** Secondly, a given DNN can be "hardened" to make it more robust against adversarial inputs. This can be achieved by preprocessing the inputs, augmenting the training data with adversarial examples, or by changing the DNN architecture to prevent adversarial signals from propagating through the internal representation layers.
- **Runtime detection.** Finally, runtime detection methods can be applied to flag any inputs that an adversary might have altered. Those methods typically try to detect abnormal activations in the internal representation layers of a DNN caused by the adversarial inputs.

[46] Source: Goodfellow et al., "Explaining and Harnessing Adversarial Examples", 2015.
[47] Refer to the Google brain competition "Adversarial Attacks and Defenses Competition", paper released in 2018.
[48] Source: https://www.ibm.com/blogs/research/2018/04/ai-adversarial-robustness-toolbox/

7.3 Model stealing (model extraction attack)

Model stealing attacks are used to "steal" models by replicating their internal functioning. This is done by simply probing the targeted model with a high number of prediction queries and using the response received (the prediction) to train another model. The cloned model can reach a high level of accuracy even above 99.9%.

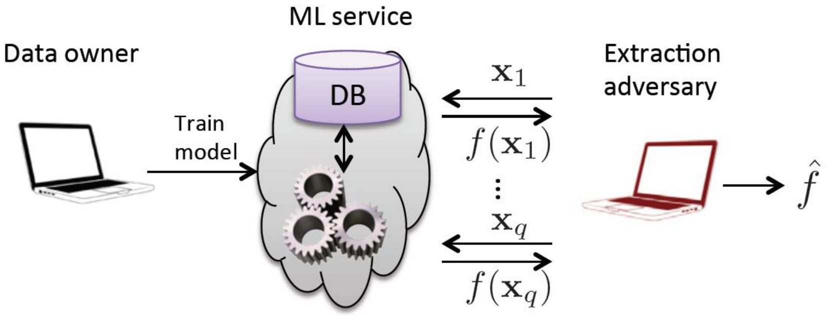


Figure 30: ML model stealing attacks.[49] A data owner has a model f; the attacker uses q prediction queries to extract the prediction function f ˆvery similar to f

The attack can be used in theory to reproduce for example models used for stock market prediction or high frequency trading; however, the attacker needs to have access to the API[50] of the model first.

The attack would especially work on those cloud platforms (e.g. Amazon, Google, Microsoft) that allow users to share their models by asking other users to pay for each query sent to the model, in order to clone the model and avoid paying.

In other cases, attackers may use this technique to reverse engineer security systems (e.g. email SPAM filters or malware prevention systems) and use the knowledge acquired to defeat them.

Among the defense techniques, a recent research[51] proposes to add a layer to the model that introduces a small perturbation (noise) to the output, which would slow down the stealing attack while still preserving the accuracy.

7.4 Video forgery

Recent research has produced an advanced technique to forge portrait videos[52]. Such technique, using deep neural networks, enables photo-realistic re-animation of portrait videos using only an

[49] Source: Tramèr et al., "Stealing Machine Learning Models via Prediction APIs", 2016.
[50] Application Program Interface.
[51] Source: Lee et al., "Defending Against Model Stealing Attacks Using Deceptive Perturbations", May 2018.
[52] Kim et al., "Deep Video Portraits", May 2018.

input video. In particular, starting from a video of a source actor, its facial expressions, head position, eye gaze and eye blinking can be transferred to the video of a target person. The output video has a very high quality and the forgery is difficult to be spotted by a human eye.

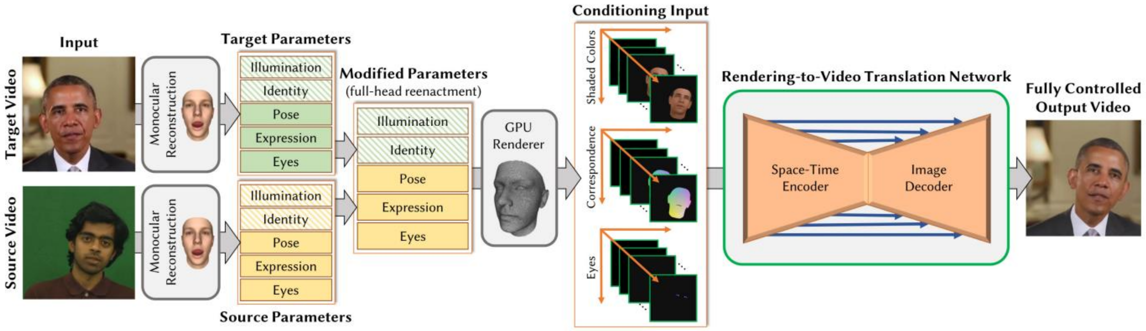


Figure 31: The output video has been modified (forged) to apply the facial expressions of the actor in the source video[53]

Despite the many positive use cases where such technology can be applied (e.g. movie video editing), attackers may use it to modify videos with malicious intent.

Although the practical applications of this technology and related malicious use to the financial sector are difficult to foresee, a point of attention shall be raised concerning the video-KYC solutions.

On the positive side, it is worth noting that even when the forged videos are indistinguishable from the real ones to a human eye, algorithms will still be able to spot that. Indeed, the detection of forged videos is an easier problem than the video generation, and it is possible to train a machine learning detector to be able to spot the forgery with high accuracy.

[53] Source: Kim et al., “Deep Video Portraits”, May 2018.

8 OPPORTUNITIES, RISKS AND RECOMMENDATIONS

8.1 Opportunities

AI and especially deep learning and machine learning are currently at the top of the hype curve[54]. This is due to the easier access to these technologies, which has brought a trend of "**AI democratization**". For example, cloud services like Amazon, Google and Microsoft offer entire AI libraries which enable developers to easily apply machine learning algorithms to their problem (including for example advanced libraries for NLP or computer vision or deep neural networks) and create models by using integrated development platforms assisting them throughout all phases of the development process[55]. This is called "Machine Learning as a Service".

Cloud services also enable easy access to rapid and scalable computing power (including GPUs) and easier sharing of the models created.

Platforms like Kaggle[56] allow data scientists to compete to create the best of breed of machine learning model, and being top ranked at Kaggle competitions is a real asset on the curriculum of the developer.

Whenever companies prefer not to make use of cloud services for confidentiality reasons, integrated platforms exist[57] that can be installed on-premises to assist with the entire data science/machine learning development lifecycle.

On the data side, big data has enabled the access to large quantities of data that are needed to train machine learning models.

Furthermore, recent EU initiatives to promote data sharing[58] in particular fields (like the healthcare field), still in compliance with data protection regulation, are just part of a series of initiatives to boost research and promote a digital single market.

There are a lot of opportunities for AI, but there are also risks associated with the implementation of this technology. Such risks are presented more in detail in the next section.

[54] Reference to Gartner technology hype cycle.
[55] See section 4 for more details.
[56] https://www.kaggle.com/
[57] Examples: DataIKU, DataRobot, etc.
[58] http://europa.eu/rapid/press-release_IP-18-3364_en.htm

8.2 Key risks and recommendations

8.2.1 Data related risks

8.2.1.1 Data quality and data governance

The power of AI and especially machine learning ("AI/ML") relies primarily on the data that is used to train the systems. Data quality is therefore paramount to ensure the success of an AI/ML project.

In particular, as seen in section 4, data can present several data quality issues such as wrong format, missing data or inconsistent data. Furthermore, finding the right data to feed to the ML model can be challenging.

Without the right data, the entire project can go wrong.

8.2.1.1.1 Key Recommendations

Institutions willing to start an AI project need to have solid data governance in place. This includes setting clear roles and responsibilities for data ownership and defining the process to identify data quality issues and fix them at the source (in the production database).

Data quality issues identified during the data preparation phase of the machine learning project should be escalated to the appropriate business owner to be fixed at the source.

Furthermore, data dictionaries should be used to document the data ownership, business meaning, confidentiality level, format, location, validation rules and relationships with other data.

8.2.1.2 External data sources

If an institution does not have the data required for the AI/ML project internally, or if the data it has is not sufficient or of insufficient quality, it may decide to use external data sources. If the data from the external source is not reliable or not of good quality, the impacts can be devastating.

Similarly, if the data is not appropriate for the specific problem being treated, the entire project could lead to wrong results. For instance, data used for credit risk valid in the US may not be representative for the Luxembourg market.

8.2.1.2.1 Key Recommendations

Institutions willing to use external data sources should perform a due diligence of the data provider and verify that data is reliable and of good quality. Furthermore, institutions should carefully evaluate the pertinence and applicability of the data to the target context.

8.2.2 Governance

8.2.2.1 No human in the loop

Although full automation can bring operational efficiency and reduced costs, lack of human oversight can lead to huge risks depending on the criticality of the task or process being automated.

8.2.2.1.1 Key Recommendations

Institutions should carefully review the integration of the AI/ML into the business process and establish controls performed by humans whenever there is a critical decision step. Humans should be able to receive appropriate information in order to make informed decisions[59]. This analysis should also consider the need for accountability[60], which cannot be delegated to a machine.

8.2.2.2 Skills

Having a team with the right skills can determine the success or not of an AI/ML project.

Institutions may have difficulties in finding the right people to develop or supervise the solution (in case it is implemented with the help of external consultants), or may rely on only a few key people. Indeed, an AI/ML project requires new types of profiles (e.g. data scientist[61]) competent in statistics, mathematics and programming which may be difficult to recruit[62].

Furthermore, lack of adequate internal skills to maintain the AI solution may lead to difficulties in updating the model or over-reliance on external parties.

Finally, lack of AI knowledge and involvement of internal auditors may lead to difficulties in auditing the final solution.

8.2.2.2.1 Key Recommendations

Institutions should ensure a sufficient level of internal AI skills to understand, develop or supervise the solution, including via specific training and knowledge transfer from external consultants when required.

Institutions willing to implement machine learning projects internally should set up multidisciplinary teams involving data scientists, IT infrastructure and database administration staff, business representatives (for the understanding of the business problem to be addressed), risk and compliance teams (for the early analysis of risks and compliance aspects, e.g. data privacy).

[59] See section 6 for more details.
[60] See section 8.2.3.3 for more details on accountability risk.
[61] See section 4.2 and 4.11.4.
[62] IBM predicts that the demand for data scientists will increase by 28% by 2020.

Institutions willing to implement off-the-shelf packages integrating AI/ML technology (e.g. RPA/IPA[63]) should carefully evaluate the skills required internally to maintain the solution once in production.

Furthermore, early involvement of internal auditors in the project can help to facilitate knowledge transfer about this new technology, efficiently embed control points into the project, and improve the overall auditability of the final solution. Alternatively, the institution may be supported by external experts to audit the AI/ML solution.

8.2.2.3 Cultural change

The adoption of AI technology may be disruptive. People may fear the change or even be afraid of losing their job.

Business users may not understand, or refuse to understand, the results produced with AI[64], incorrectly interpreting the results from the AI and leading therefore to poor final outcomes.

Finally, AI experts may remain isolated, not integrated with the other IT teams and the business teams, impacting therefore the business adoption of the AI product.

8.2.2.3.1 Key Recommendations

Institutions should prepare their employees for the cultural changes brought by AI, in particular when AI will strongly impact the business processes.

Business users should be involved in AI projects from the beginning to ensure early adoption and also to make sure that the final product correctly answers the business needs and is well integrated inside the existing business processes.

8.2.3 Ethical & societal concerns

8.2.3.1 Bias and Discrimination

If bias is incorporated into the model, the output of the model will be unfair and discriminative for certain populations. The most common source of bias is the one included, unconsciously or consciously, in the training and validation data sets (data bias). For example, populations that are not well represented in the dataset may have less chances to have a favorable outcome in a credit scoring system, simply because the algorithm has learnt that in the past this category was not granted many loans.

[63] Robotic Process Automation/Intelligent Process Automation.
[64] See also section 6: difficulty to extract explanations on the results produced with AI.

An extreme example, useful to understand the importance of bias and how this can bring discrimination, is the one related to the COMPAS[65] system. This system, based on machine learning, was used by US judges to estimate the probability of recidivism of people who were convicted to jail. It was discovered that the system was attributing a better score (i.e. less probability of recidivism) to white people than to black people, simply because in the data that was used for the training there were more black criminals than white.

Furthermore, bias can also be incorporated by choosing the wrong model (algorithmic bias) or the wrong input features (not correctly representing the population) or by coding some discriminatory rules into the model.

Finally, bias can also be simply due to humans (human bias). For example, in case historical data about past loan requests is used to train a credit score system, this data can be biased because the loans were authorized in the past based on the personal evaluation of the bank's account officers.

8.2.3.1.1 Key Recommendations

Institutions should aim at building systems respecting the principle of fairness and non-discrimination. This target can be achieved through several actions, e.g.:

- document and publish an ethical code of conduct policy to promote non-discrimination principles[66];
- seek for diversity in the input data, especially with regards to representation of different populations (**active inclusion[67]**);
- carefully review training and validation data during the data preparation phase to identify and remove data bias (for example, via the feature importance analysis[68], or measuring skewness[69] of data and normalizing it when required, etc.);
- craft specific validation datasets to check for bias/discrimination[70];
- evaluating different algorithms and comparing the results;
- embed non-discriminatory rules into the algorithm when required (e.g. if predictions are always negative for a certain class of population then raise an alert);
- constantly monitor the performance of the model to identify deviations, and investigate the reason.

[65] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[66] A notable example is the AI principles recently published by Google: https://www.blog.google/technology/ai/ai-principles/

[67] The concepts of active inclusion and as well fairness, right to understanding and access to remedy are the main recommendations included in the paper of the World Economic Forum "How to prevent Discriminatory Outcomes in Machine Learning" (March 2018).

[68] See section 4.5.

[69] Skewness is a measure of the asymmetry of the probability distribution of a variable around its mean. For example, if data is skewed left or right, it means that is not following a normal distribution (since it is not equally distributed around its mean). To ensure a fair representation of a certain variable/feature (e.g. gender), often it is recommended to apply some data transformations to normalize it (ensuring it has a normal distribution).

[70] For example, for image recognition models there are specific datasets including all skin colors variations.

8.2.3.2 Data privacy

The data used by the AI/ML model can include personal data, i.e. data that allows to identify, directly or indirectly, physical persons[71]. Apart from the most obvious personal data like name and address, plenty of data (especially when multiple fields are combined) may allow to indirectly identify a person [72], such as identity card number, bank account number, age, birthplace, gender, etc.

Furthermore, nowadays valuable information is extracted by profiling the users, i.e. by collecting data related to their behavior (e.g. consumption behavior, payment timeliness behavior, etc.). These alternative sources of data may be used by ML algorithms to make automatic decisions[73] (e.g. accept or refuse a loan request) which could positively or negatively impact the user and condition his/her access to the (financial) service proposed.

Finally, companies may tend to think that they can do whatever they want with the personal data collected (including behavioral data) and forget that instead each personal data processing needs to have a lawful basis, most commonly based on the user’s explicit consent or based on the fact that the processing is necessary for a contract or a legal obligation.

It should also be noted that whenever consent has been received for a specific data processing, the data collected cannot be used for a new data processing which has not been explicitly authorized.

Insufficient protection of personal data or improper use (data processing) of personal data may lead to privacy risks for the persons impacted and reputation and compliance risks for institutions, including high fines for not complying with principles included in GDPR regulation.

8.2.3.2.1 Key Recommendations

Institutions should limit the use of personal data in their AI/ML projects and apply protection measures, including:

- challenge the need for personal data as input to the AI/ML model;
- restrict access to personal data on a **Need to Know and Least Privilege** basis (restrictions should also apply to data scientists/ML developers working at the project);
- involve compliance and data protection representatives (e.g. Data Protection Officer) from the early stages of the AI/ML project to perform a data protection impact analysis of the project and verify compliance aspects such as:
  - check that the users are properly **informed** on the data processing performed;
  - for each data processing, check that proper **consent** has been received from the user or there is another valid legal basis[74] (e.g. required to perform AML obligations);
  - in case automatic decisions are made via the AI/ML model, ensure that the user can receive valid, human understandable explanations[75] on how the decision was reached (**right to understanding**)[76] ;

[71] Refer to art. 4(1) GDPR.

[72] A study reported that in the United States, in the year 2000, 87% of the whole population was uniquely identifiable by using only the combination of zip code, birthdate and gender. This rate may even increase if applied to small countries like Luxembourg.

[73] See for example the use case on credit scoring (automated loan decisions) described at section 5.6.

[74] Art. 6 GDPR.

[75] See also section 6.

[76] Art. 22 and Recital 63 GDPR.

  - ensure that persons have the possibility to rectify or erase (when not contradicting other legal obligations) data related to them that reveals to be wrong (**access to remedy/redress**);
- apply **data protection by design** principles[77], such as data minimization, data pseudonymization[78], data anonymization, data encryption, and the related technical security measures;
- carefully design the **infrastructure** used for hosting the AI/ML development and production environments and the connections to production data according to the principles listed above. (e.g. use sandboxed environments).

8.2.3.3 Accountability

Although AI/ML can be used to automatize processes and business decisions, this cannot result into delegating the responsibility of the action taken to a machine.

Unclear definition of responsibilities for AI/ML automated decisions, for example when the institution is embedding off-the-shelf packages, may result in disputes related to liability when wrong decisions are made.

Finally, unclear roles and responsibilities throughout the entire AI lifecycle may result in the absence of continuous engagement and oversight, ultimately causing project failure.

8.2.3.3.1 Key Recommendations

Institutions should assume clear responsibility and accountability for the actions and decisions taken by automated AI//ML systems and processes. Ultimate responsibility should rely on senior management of the institution which integrates the AI/ML logic into its business processes. Whenever off-the shelf packages are acquired, clear liability provisions should be defined at contractual level.

Furthermore, clear roles and responsibilities should be defined along all the AI lifecycle, including the development and operations activities, to ensure continuous engagement and accountability.

8.2.3.4 Explainability

AI/ML systems are complex systems that can quickly become black boxes: the system can generate predictions with a high level of accuracy but how the system arrives to a certain conclusion is unknown and very difficult to understand. For example, algorithms such as deep neural networks are inherently complex due to their internal structure based on hundreds or even thousands of different layers, almost impossible to be comprehensible for a human being.

[77] Art. 25 and Recital 78 GDPR.

[78] Art. 4 GDPR: "'pseudonymization' means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information…".

Black box behavior can represent an issue depending in particular on the criticality of the use case in scope. The more the latter is critical, the more there is a need for explainability because:

- in order to trust the result of the system and validate its correct functioning, humans need to understand how that conclusion was reached;
- business users will tend to refute the results produced by the AI/ML system if they cannot understand the logic behind it, potentially hindering the business adoption of the new system and the entire success of the AI/ML project;
- people that should be accountable for the results will refuse to take responsibility if they do not trust the system;
- black box systems are difficult to maintain and have portability issues.

Furthermore, as seen in section 8.2.3.2 on Data Privacy risks, GDPR imposes the requirement to provide explanations to persons subject to automatic decisions.

8.2.3.4.1 Key Recommendations

Institutions should implement measures to ensure explainability of their AI/ML systems from the design phase. Even when full transparency cannot be achieved due to the intrinsic nature of the algorithm employed (e.g. deep neural networks), steps can be taken to identify and isolate in a human understandable format the main factors contributing to the final decision. Among the measures recommended there are:

- Thoroughly document the development process leading to the construction of the AI/ML model since the design phase, together with the assumptions and choices made at each step. For example:
  - Document the blueprint of the data preparation flow used to build the input features, including the transformations applied on the raw data (e.g. normalization, dimensionality reduction, exclusion of correlated features[79], aggregation, etc.)[80], and the analysis performed to check the features importance. Indeed, the information about which are the main input features influencing the model is already providing a first basic explanation of the model behavior.
  - Document the algorithms assessed and the comparative results justifying the selection made, including the analysis done on the model to ensure it is fit[81], the validation methodology employed and the results obtained. Indeed, different algorithms have different degrees of inherent complexity and the choice of the algorithm directly influences the model explainability.
  - Document the accuracy metrics used to monitor the model performance and promptly identify deviations.
- When appropriate, embed interpreters into the model design to ensure traceability of the main internal steps leading to the final prediction.
- According to the criticality of the underlying use case and the need for transparency, evaluate the possibility to implement explainability techniques such as those described in chapter 6.
- Finally, ensure that the solution implemented in production is auditable (see point below on auditability), since audit logs can help to understand how data is processed.

[79] When different input variables/features are highly correlated, the correlated ones are often dropped during the feature selection process to improve the model performance by focusing only on the features with more predictive power.

[80] See section 4.4 and 4.5 on Data extraction/preparation.

[81] See section 4.6.2 on Model Training.

8.2.3.5 Auditability

As AI/ML is integrated into business processes, there is a need to track the main actions performed and gather evidence allowing investigations to be performed in case of incidents, or simply to better understand and validate the process during the development phases.

8.2.3.5.1 Key Recommendations

Institutions should ensure that the solution implemented in production is equipped with audit logs that are sufficiently detailed to follow the data flow through the AI/ML process (e.g. from the raw input data to the calculation of the input features to the AI/ML output) and be able to re-simulate the input data if required.

Audit log retention should allow for sufficient history to perform investigations on past events when required and should be in line with applicable regulatory requirements.

8.2.3.6 Safety

Although difficult to apply to AI solutions currently available in the financial sector, the concept of human safety is fundamental whenever autonomous systems are designed or important automatic decisions are taken.

The most representative example of the safety risk is the one related to autonomous cars (where they cause mortal accidents), although other examples could be related to simpler decisions that could still have a physical or psychological impact on humans (e.g. the impacts of an automated portfolio management system that would lose all the money of the investor...).

8.2.3.6.1 Key Recommendations

Institutions should perform impact assessments to determine the possible (physical or psychological) impacts that the solution may have on humans and consider the level of autonomy of the AI/ML solution. Whenever appropriate, institutions should embed adequate safety protections into the design of the AI/ML product.

8.2.4 Technology

8.2.4.1 Change Management

The entire AI project may fail if the underlying change management process is not solid.

For example, lack of security and data protection analysis from the design phase may lead to security and compliance risks, and unclear testing procedures will generate inaccurate results.

8.2.4.1.1 Key Recommendations

First of all, institutions should challenge the real need for AI/ML and justify the pros and cons in a Business Case, considering the maintenance aspects and the skills required internally for such kind of projects. For example, RPA/IPA projects could often be avoided or the perimeter could be reduced by reengineering and optimizing the process first, before automating tasks of a workflow that are organized in an inefficient way anyway.

Secondly, the **infrastructure** that will be hosting the AI/ML systems needs to be carefully designed including the security aspects. For example, the connectors to production data should allow a granular access in read only mode, and AI/ML development environment should be properly segregated from the production one.

For a more efficient project, AI/ML teams should involve IT infrastructure and database administration from the early stages, together with the business users that will ultimately benefit from the solution.

Security and data privacy by design concepts should be adopted from the early stages, and access to production data for data scientists and ML developers should be attributed on a Need to Know basis.

All the steps of the change management process (AI/ML development lifecycle) should be duly **documented**, and the choices made at each step justified. This includes, among others, the data transformations applied to raw input data and feature selection, the algorithms evaluated and the criteria for selection, the choice of hyperparameters, the testing methodology, the accuracy and KPI criteria used to monitor performances.

Whenever software **programming** is involved[82], best practices on secure and quality software development should be followed. For example, before using a third party library, appropriate due diligence on the origin and quality of the code should be applied. Portions of code developed internally should be subject to peer reviews and quality checks.

Versioning should be applied, whenever possible by means of integrated AI/ML development platforms.

Metrics to measure the model accuracy and business KPIs should be defined in the design phase and then used to evaluate the model. Performance of the model deployed in production should be continuously monitored to identify deviations. The frequency of the **model update** (re-train the model based on new training data) should be assessed in the design phase and adjusted if needed (for example, if accuracy decreases then the model may need to be updated).

Finally, whenever possible, a “parallel” run should be adopted in order to compare the results of the old model with the ones of the proposed new model, in order to validate the latter for deployment into production.

[82] For example, Python and R are programming languages often used in ML projects.

8.2.4.2 Model update

The predictive power of ML is limited to what can be learnt from past observations and cannot predict entirely new things that do not correspond to any pattern seen in the past (something never seen before). For example, a completely new fraud scheme in a fraud detection system or a new risk scenario in a risk model simulator (similar to what happened during the financial crisis) cannot be easily predicted.

8.2.4.2.1 Key Recommendations

Institutions should continuously monitor ML model performance and update the model (re-training) when new (disruptive) events occur.

Furthermore, institutions should, as far as possible, “think out of the box” to imagine extreme scenarios that could not be predicted by the ML model and prepare for it.

8.2.4.3 IT operations

Once deployed into production, aspects related to the daily IT operations may be underestimated. For example, an RPA/IPA system without proper error management may render the process automation completely useless.

8.2.4.3.1 Key Recommendations

Institutions should consider the IT operations aspects related to the implementation of an AI/ML solution in production, including for example:

- integration with legacy systems;
- error management (especially in case of robotic process automation);
- continuous monitoring of the performance of the model and raising alerts when deviations are detected.

8.2.4.4 Robustness and Security

AI/ML systems involve access to high quantities of data, including production data, personal data, confidential data. Without adequate security measures in place, sensitive data can be exposed to unauthorized access.

Furthermore, AI/ML systems can be vulnerable to specific security attacks, such as those described in chapter 7. Finally, it should be noted that attackers can also use artificial intelligence for their malicious activities, and new types of attacks may be discovered.

8.2.4.4.1 Key Recommendations

Institutions willing to implement or adopt AI/ML technologies should apply security measures throughout the development lifecycle, starting from the design phase. The Need to Know principle should be applied to the AI/ML development team as well to restrict access to sensitive and confidential data.

Depending on the sensitivity of the AI/ML model and its exposure to online attacks, appropriate defense techniques, such as those described in chapter 7, should be evaluated and potentially implemented. Similarly, independent reviews should be performed to check the level of robustness of the AI/ML system.

Finally, institutions should monitor the progress of security attack and defense techniques to ensure the adequacy of their security protection and robustness against attacks.

8.2.5 External providers

8.2.5.1 External AI service providers and outsourcing

Not all companies possess the adequate skills and resources internally to be able to start an AI/ML project on their own. An easier option is to purchase off-the-shelf packages, or ask external consultancy companies to assist them with the implementation. This approach may create dependencies with the external service provider when maintenance activities are concerned.

8.2.5.1.1 Key Recommendations

Institutions should evaluate the risks related to the maintenance of off-the-shelf packages and to the outsourcing of the AI/ML development and maintenance activities. Adequate controls should be implemented in line with best practices and other regulatory requirements applicable to outsourcing (e.g. Circular CSSF 12/552, sub-chapter 7.4 on outsourcing).

8.2.5.2 Systemic risks

In the near future it is probable that advanced AI/ML solutions developed by some specialized companies or FinTechs may be very successful and may be adopted by many financial institutions. This could lead to a concentration amongst a few service providers and thus a high dependency on their services.

In addition, depending on the particular use case, the use of a common algorithm by many institutions may be dangerous and engender systemic risks (e.g. institutions acting identically to other market participants). For instance, if the algorithm contains errors, these will be amplified by the large number of implementations of the algorithm. A particular case is the one of automated portfolio management systems or algorithmic trading: if the same algorithm would be used by several banks, theoretically there is a risk that market movements can be generated with potential amplification effects.

8.2.5.2.1 Key Recommendations

When purchasing an AI/ML product, institutions should consider the risks associated with using AI/ML solutions already in use by a large number of other financial institutions and ensure that the product is sufficiently customizable.

Regulators should monitor potential systemic effects and issue warnings when appropriate.

A summary of the key risks and recommendations is available in Annex 3.

8.3 Key success factors

The success of an artificial intelligence and in particular a machine learning project depends on a few key factors.

Firstly, the **business case**: the AI/ML project should tackle a pain point for the business users, for which the AI solution should bring a real added value. AI should not be used just because it is trendy, but because it can bring strong improvements compared to more traditional solutions. The business benefits and risks should be measurable via clear KPIs that should be used to validate the effectiveness of the AI solution before it is deployed into production. The employment of Proof of Concepts (“POC”) can help to confirm the validity of the business case more rapidly. Business users should be involved from the start and throughout the entire project to ensure strong engagement. They should perceive the AI/ML solution as *their* solution, thereby also ensuring clear accountability.

Secondly, the importance of **data**. The power of AI and especially ML solutions relies primarily on data. Therefore, aspects like data quality and data governance are a fundamental prerequisite for the success of the project. The connectivity to the data sources is equally important, and should be verified at the beginning of the project.

Even when performing a Proof of Concept, the model should be based on real data and not on data specifically crafted for the POC, since this will only delay the discovery of possible issues with the data sources (e.g. data quality, insufficient historical data, difficult integration of legacy systems, etc.) that could make the entire project fail if discovered too late.

Thirdly, the **team** working on the AI/ML project is also important, as it needs to efficiently address all the challenges encountered during the different phases of the project. It should be an interdisciplinary team composed not only of data scientists but also of IT infrastructure and database administrators and strongly involve business users and risk, compliance and security representatives. This will ensure that principles like privacy by design and security by design are addressed from the early stages of the project.

Finally, the performance of the model should be continuously monitored and the **model** should be updated whenever new disrupting events (not corresponding to any existing pattern in the past observations) are encountered.

9 CONCLUSION

Artificial intelligence is a broad field which today is benefiting from favorable conditions (like the availability of high amounts of data and of the computing resources required to process them efficiently) that have contributed to its impressive development.

Among the AI technologies most commonly applied to the financial sector is machine learning. Machine learning has demonstrated its validity in several financial use case applications where traditional approaches are struggling with performances, like fraud detection.

Machine learning has revolutionized the way to solve certain problems and has enabled more efficient predictive analytics, where models are trained based on past observations and are then used to predict new outcomes. Nevertheless, it should not be forgotten that although very innovative, this technology is not a magic crystal ball and its predictive power is limited to what it has learnt from the past. Disruptive events, like those that happened during the last financial crisis, cannot be predicted and the models need to be updated when such events occur.

AI and machine learning are subject to various types of risks, some of which, like ethical concerns, are generally not found in more traditional approaches and are instead very common in AI/ML projects. Also, the vulnerabilities and related security attacks affecting these solutions are quite new and constantly evolving. Similarly, accountability aspects are very important for an efficient integration into the business processes.

Institutions willing to implement an AI project therefore need to address these new risks from the start and continuously monitor their evolution, to ensure a reliable implementation and business integration of the AI solution while maintaining a sound control environment.

ANNEX 1: LIST OF ACRONYMS

Acronym	Meaning
AI	Artificial Intelligence
GPU	Graphics Processing Unit
ML	Machine Learning
NLP	Natural Language Processing
GDPR	General Data Protection Regulation
AML	Anti-Money Laundering
KYC	Know Your Customer
KYT	Know Your Transactions
RPA	Robotic Process Automation
IPA	Intelligent Process Automation
POC	Proof of Concept
PWC	Price Waterhouse Coopers
FSB	Financial Stability Board
BdI	Banca d'Italia
MAS	Monetary Authority of Singapore
SEC	U.S. Securities and Exchange Commission

ANNEX 2: GLOSSARY

Term	Definition	Source
Artificial intelligence	The theory and development of computer systems able to perform tasks that traditionally have required human intelligence.	FSB
Augmented intelligence	Augmentation of human capabilities with technology, for instance by providing a human user with additional information or analysis for decision-making.	FSB
Big data	A generic term that designates the massive volume of data that is generated by the increasing use of digital tools and information systems.	FSB
Big data analytics	Analysis of large and complicated datasets ("big data").	FSB
Chatbots	Virtual assistance programs that interact with users in natural language.	FSB
Cluster analysis	A statistical technique whereby data or objects are classified into groups (clusters) that are similar to one another but different from data or objects in other clusters.	FSB
Computer vision	The process of pulling relevant information from an image or sets of images for advanced classification and analysis.	PWC

Data Science	Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured, similar to data mining. Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data. It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and computer science.	Wikipedia
Deep learning	A subset of machine learning, this refers to a method that uses algorithms inspired by the structure and function of the brain, called artificial neural networks.	FSB
Explainability	Explainability describes the ability to determine the main factors influencing a specific individual decision that has been reached by a system.	PWC
Feature	An input variable used in making predictions.	developers.google.com
Feature engineering	Process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.	J. Brownlee, Machine Learning Mastery
Hyperparameter	The "knobs" that you tweak during successive runs of training a model. For example, learning rate is a hyperparameter. Contrast with parameter.	developers.google.com
Machine learning	A method of designing a sequence of actions to solve a problem that optimize automatically through experience and with limited or no human intervention.	FSB

ML-as-a-Service (MLaaS)	Machine learning as a service (MLaaS) is an array of services that provide machine learning tools as part of cloud computing services. This can include tools for data visualization, facial recognition, natural language processing, image recognition, predictive analytics, and deep learning. Some of the top ML-as-a-service providers are: • Microsoft Azure Machine Learning Studio • AWS Machine Learning • IBM Watson Machine Learning • Google Cloud Machine Learning Engine • BigML	www.analyticsvidhya.com
Model	The representation of what a machine learning system has learned from the training data.	developers.google.com
Model Accuracy	The fraction of predictions that a classification model got right.	developers.google.com
Model overfitting	Situations in which a model is so tightly fit to its underlying dataset, as well as the noise or random error inherent in that dataset, that the model performs poorly as a predictor for new data points.	www.dummies.com
Model Overgeneralization	Overgeneralization is the opposite of overfitting: It happens when a data scientist tries to avoid misclassification due to overfitting by making a model extremely general. Models that are too general end up assigning every category a low degree of confidence.	www.dummies.com
Natural Language Processing (NLP)	Algorithms that process human language input and convert it into understandable representations.	PWC
Neural network	A model that, taking inspiration from the brain, is composed of layers (at least one of which is hidden) consisting of simple connected units or neurons followed by nonlinearities.	developers.google.com

Neuron	A node in a neural network, typically taking in multiple input values and generating one output value. The neuron calculates the output value by applying an activation function (nonlinear transformation) to a weighted sum of input values.	developers.google.com
Outlier	Outlier is an observation that appears far away and diverges from an overall pattern in a sample.	www.analyticsvidhya.com
Parameter	A variable of a model that the ML system trains on its own. For example, weights are parameters whose values the ML system gradually learns through successive training iterations. Contrast with hyperparameter.	developers.google.com
Text mining	Text mining are techniques to retrieve information from text documents by extracting key phrases, concepts, etc. and prepare the text processed for further analyses with data mining techniques.	PWC
Weight	A coefficient for a feature in a linear model, or an edge in a deep network. The goal of training a linear model is to determine the ideal weight for each feature. If a weight is 0, then its corresponding feature does not contribute to the model.	developers.google.com

ANNEX 3: SUMMARY OF AI/ML KEY RISKS AND RECOMMENDATIONS

Category	Risk	Risk description	Recommendation
Data	Data quality and data governance	• Difficulties in finding the right data • Data quality issues: wrong format, missing data, inconsistent data • Difficulties in connecting data sources in legacy systems	• Implement a solid data governance framework (clear roles and responsibilities for data ownership, data dictionaries, etc.) • Involve business data owners to find the right data • Identify data quality issues and escalate them to business owners to fix them at the source (production data)
Data	External data sources	• External data is not appropriate to the local (Luxembourg) context • Data of insufficient quality/not reliable	• Perform due diligence reviews on the data source provider • Verify the adequacy of the data to the target context
Governance	No human in the loop	• Lack of oversight can lead to uncontrolled automated actions directly impacting the business process	• Ensure that all AI tasks are under control • Depending on the criticality of the task being automated, implement a human oversight/dual control
Governance	Skills	• Lack of AI specific skills (e.g. data scientist) to develop or supervise the solution • Teams developing AI solutions need as well business understanding, risk and compliance • AI teams lack the required business, risk, compliance knowledge to ensure a suitable and compliant solution • Lack of skilled team to maintain the solution • Over-reliance on a few skilled staff • Lack of auditors with AI competencies to audit the solution	• Ensure a sufficient level of internal AI skills to understand, develop or supervise the solution, including via specific trainings and knowledge transfer from external consultants when required • Set a multidisciplinary team to start an AI project (involve business, risk, compliance) • Ensure auditors have the right skills to audit the AI solution (involve them from the start of the project), or use external auditors (AI experts)

Governance	Cultural change	• Adoption of AI technology may be disruptive (people afraid of losing jobs) • Fear of change, concerns on changing job profile • Business users may not understand or may refuse to understand the results produced with AI, incorrectly interpreting the AI results and leading to poor final outcomes • Lack of integration of AI experts with the other teams	• Prepare the company for cultural change of its employees, especially when the adoption of AI technology will strongly impact the business processes • Involve business users (that will need to use the solution) from the start and throughout the entire project (key success factor)
Ethics	Bias and discrimination	• Data bias (incorporated in training and validation data sets) • Algorithmic bias (wrong model, wrong selection of hyperparameters; wrong selection of features, bias incorporated in the model) • Human bias (data includes past decisions made by humans, which were subjective and not objective) • Discriminative results (populations not well represented in the training data may be discriminated)	• Implement an AI code of conduct • Active inclusion: seek for diversity in the training and validation data • Identify and remove data bias during the data preparation phase (e.g. data skewness analysis; feature importance analysis; etc.) • Create specific validation datasets to test against discrimination • Evaluate several algorithms and compare results • Code specific non-discriminatory rules if required (e.g. too many women or black people refused ==> alert) • Continuously monitor model performance
Ethics	Data protection	• Personal data collected and processed without proper consent • User not adequately informed • Unauthorized processing of behavioral data (including to make automated decisions) • Non-compliance with GDPR	• Challenge the effective need for using personal data in the AI/ML solution • Restrict access (e.g. for data scientists) on Need to Know basis • Involve compliance and data protection teams in order to effectively implement GDPR recommendations including: o Obtain consent when required o Information to the user o Right to understanding (provide clear explanations re. automated decisions) o Access to remedy/redress (to update or delete wrong personal data)

Ethics	Accountability	• Lack of accountability over the actions performed by AI • Unclear roles and responsibilities throughout the whole AI lifecycle may lead to the absence of continuous engagement and oversight and the entire project may fail	• Accountability cannot be delegated to a machine: the ultimate responsibility for AI/ML solutions relies with the senior management of the supervised institution • Define clear roles and responsibilities along the entire AI lifecycle
Ethics	Explainability	• Black box models • Lack of trust since humans do not understand internal functioning • Non-compliance with data protection regulation (right to understanding)	• Document the data preparation workflow and model blueprint • Document the choice of the algorithm; choose more interpretable algorithms depending on the criticality of the system • Measure the model accuracy • Use explainable AI techniques (e.g. interpreter, etc.) when required
Ethics	Auditability	• Lack of audit trails to track the flow of data through the AI system and the automated decision taken by AI	• Implement detailed audit logs to track all phases from raw data to feature creation and till the AI/ML outcome; • Implement technical means to be able to re-simulate the input data into the AI to perform investigations in case of need • Ensure adequate audit log retention in line with business and legal requirements
Ethics	Human safety	• Automated decisions taken by AI may physically or psychologically harm human people	• Perform impact assessments and implement safety controls when required

Technology	Change management	• Lack of involvement of business users leading to poor outcomes • Insufficient data preparation checks leading to errors or data leakage (output data in input data) • Security vulnerabilities • AI/ML model not accurate	• Document the entire development and testing methodology • Document the choices made at each step of the development process and the related justifications (e.g. feature selection, choice of algorithm, etc.) • Apply security by design (e.g. read only production connectors, need to know/granular access, development environment segregation, etc.) • Apply versioning (prefer using integrated platforms including versioning and simplifying the release deployment) • Apply best practices for those parts requiring software coding (e.g. check quality of open source libraries; peer reviews, software quality checks, etc.) • Measure model performance (via accuracy metrics and business KPIs, etc.) • Whenever possible, perform parallel runs to compare the old with the new AI/ML model
Technology	Model update	• Predictive power of ML is limited to what can be learnt from past observations: cannot predict something never seen before!	• Continuously monitor ML model performance and update the model (re-training) when new (disruptive) events occur • Think “out of the box” to imagine extreme scenarios that could not be predicted by the ML model and prepare for it
Technology	IT operations	• Insufficient error and incident management • Technical operational issues (e.g. interfaces with legacy systems)	• Properly prepare the design phase to ensure integration with legacy systems • Plan for error management and incident management (e.g. processes automated via RPA/IPA could generate frequent operational errors) • Integrate technical monitoring of the model within the IT monitoring system
Technology	Robustness/Security	• Security vulnerabilities • Example of attacks: ○ data poisoning ○ adversarial attack ○ model stealing ○ video forgery	• Apply security by design • Apply specific security defense techniques (especially if the model does online learning or it is exposed on the Internet) • Technological watch: monitor improvements in the attack and defense techniques (remember that attackers are also using AI to improve their attacks!)

			• Perform independent security reviews according to the criticality and exposure of the system
External providers	External providers & outsourcing risks	• Dependency on few providers • Outsourcing risks	• Evaluate the maintenance requirements of the AI/ML product once in production and prepare for it (e.g. ensure to have the right resources/ skills internally or a service level agreement with an external service provider) • Apply best practice and regulatory recommendations on IT outsourcing (e.g. circular CSSF 12/552)
External providers	Systemic risks	• If the same model is used by many institutions, market movements and errors may be amplified	• Customize the AI/ML product to own needs • Monitor systemic effects

Author :

Anna Curridori
Domaine surveillance des systèmes d'informations
Service surveillance des systèmes d'informations et des PSF de support

Commission de Surveillance du Secteur Financier

Commission de Surveillance du Secteur Financier

283, route d'Arlon

L-2991 LUXEMBOURG

Tél. : (+352) 26 251-1

Fax : (+352) 26 251-2601

E-mail : direction@cssf.lu

Internet : http://www.cssf.lu