

Machine Learning

Evaluation guidelines for financial services business cases

Company	Finalix AG
Author	Matthias Laube
File	Machine Learning Evaluation Guidelines
Version	V01.01
Date	19.02.2023
Status	final
Classification	public
Pages	43

Management Summary

Introduction	<p>The goal of these evaluation guidelines is to provide an expert-driven and practice-oriented framework to critically evaluate the business case behind machine learning projects in the financial services industry.</p> <p>For this, I have interviewed machine learning domain experts, of which most have implementation experience in financial services.</p>
Centerpiece	<p>The centerpiece of this paper is the Business case assessment forms, which consist of four tables with a total of 22 distinct criteria. The first table is to determine the machine learning category. The second table deals with potential showstoppers & promoters – the big topics which need to be dealt with before diving deeper into the business case assessment. The third table considers costs & benefits, and the fourth table completes the assessment by looking at additional success factors. All of this is condensed on 2 pages.</p> <p>These four tables are then followed by in-depth background for each criterion with all the critical thoughts to be considered and enriched with some real-world examples. To round everything off, the last chapter walks you through an exemplary case.</p>
Framework can be applied by beginners...	<p>While I have initially intended to make these guidelines something a person without machine learning background can apply, this endeavor was only partially successful. This is because some evaluation criteria do require a basic understanding of machine learning in order to enable critical thinking. However, this basic understanding is something a consultant should be able to acquire within 1-2 days.</p> <p>Armed with these guidelines, a machine learning beginner should at least be able to critically think through a machine learning business case and, if necessary, challenge it.</p>
...but serious evaluation requires an expert	<p>However, due to the complex nature of the field, hundreds of already existing algorithms and the fast progress being made on developing new solutions and improving old ones, some of the criterions will require a data scientist to assess them in a serious manner.</p> <p>Such a person should of course be available if a machine learning project is seriously considered.</p>
...and business inputs	<p>Last but not least, it is also of key importance to include your business stakeholders early on. Understand the desired business outcome and, if necessary, show them cost-benefit tradeoffs to manage expectations and negotiate towards a positive return on investment.</p>

Table of Content

- 1 Introduction..... 5
 - 1.1 Motivation 5
 - 1.2 Goal..... 6
 - 1.3 Scope 6
 - 1.4 Relevance 7
- 2 Approach 8
 - 2.1 Method 8
 - 2.2 Contributing experts & practitioners 8
- 3 Quick introduction to machine learning..... 10
 - 3.1 Definition 10
 - 3.2 Distinction from other fields 10
 - 3.3 The three main categories of machine learning..... 11
 - 3.3.1 Determine the machine learning category 11
 - 3.3.2 Supervised learning 11
 - 3.3.3 Unsupervised learning 12
 - 3.3.4 Reinforcement learning..... 12
 - 3.4 Important terms 13
- 4 Evaluation guidelines..... 14
 - 4.1 Business case assessment forms 14
 - 4.1.1 Table 1: Machine learning category 14
 - 4.1.2 Table 2: Potential showstoppers and promoters 14
 - 4.1.3 Table 3: Cost and benefit drivers 15
 - 4.1.4 Table 4: Additional success factors 15
 - 4.2 How to use the assessment forms 16
 - 4.3 Machine learning category 17
 - 4.4 Potential showstoppers & promoters 17
 - 4.4.1 The Gartner analytic continuum 17
 - 4.4.2 Insufficient data volume 18
 - 4.4.3 Insufficient data quality or availability..... 20
 - 4.4.4 Unstable environment..... 21
 - 4.4.5 Regulatory barriers 22
 - 4.4.6 Lack of explanatory factors 22
 - 4.4.7 Strategic investment case..... 23
 - 4.5 Cost and benefit drivers..... 24
 - 4.5.1 Data 24
 - 4.5.2 Technical complexity 25
 - 4.5.3 Project complexity..... 25
 - 4.5.4 Acceptable error margins..... 26
 - 4.5.5 Maturity of employed capabilities..... 27

- 4.5.6 Data sourcing & response time 27
- 4.5.7 Scalability 28
- 4.5.8 Proof of reliability 28
- 4.5.9 Explainability and interpretability..... 28
- 4.5.10 Expected benefits 29
- 4.6 Additional success factors 30
 - 4.6.1 Time criticality 30
 - 4.6.2 Business acceptance & change management..... 30
 - 4.6.3 Focus on smart data over big data 31
 - 4.6.4 Design actionable output 32
 - 4.6.5 Difficulty level of output prediction 32
 - 4.6.6 Identify and evaluate alternative options..... 32
- 5 Applying the evaluation guidelines to an exemplary case 34
 - 5.1 Case introduction 34
 - 5.2 Table 1: Determine machine learning category 34
 - 5.3 Table 2: Potential showstoppers and promoters 35
 - 5.3.1 Assessing the evaluation criteria 35
 - 5.3.2 Assessing the table results 36
 - 5.4 Table 3: Cost and benefit drivers..... 37
 - 5.4.1 Assessing the evaluation criteria 37
 - 5.4.2 Assessing the table results 39
 - 5.5 Table 4: Additional success factors 39
 - 5.5.1 Assessing the evaluation criteria 39
 - 5.5.2 Assessing the table results 40
 - 5.6 Overall assessment 40
- 6 Appendix..... 41
 - 6.1 Common financial services use cases..... 41
 - 6.2 Sources 42
 - 6.3 Abbreviations..... 42

1 Introduction

1.1 Motivation

Focus on business side

As consultants in the financial services industry, finalix employees from time-to-time face projects with machine learning (ML) deliverables.

finalix is a business consultancy and not an IT consultancy. Hence, our focus area is not about knowing the latest state-of-the-art technical solution. Our expertise is on the business side: Understand business needs, ensure that proposed solution(s) do address those needs, and that the solution generates a positive return on investment.

In the context of machine learning, this means that we do not focus on knowing the best methods of cleaning & preparing data, or to know all the machine learning models and understand which one to choose for which use case. However, we should be able to critically question if a machine learning solution makes sense for a given project.

Personal experience

I have already experienced projects which considered machine learning. In one of them, the return on investment (ROI) was questionable. In two other cases, ROI was clearly negative.

- A requirements engineer for a risk engine kept talking about “creating the required business rules with machine learning”. When challenged on his approach, he kept insisting others are thinking of regression rather than machine learning, but evaded to specify what he was actually doing. In the end, he did not deliver and the business rules had to be created manually by business experts under time pressure. This could have been avoided, if the product owner would have had some base knowledge in machine learning.
- For a regulatory project, legal forms had to be analyzed for certain phrases. The legal team was able to provide certain key words as a filter criterion, to reduce the number of forms that had to be reviewed, i.e. only forms where such a key word was found had to be reviewed by a legal expert. Since the legal forms were available in ‘machine readable’ PDFs, the business analyst thought it would be a good use case for an ‘AI Scanning Tool’ instead of manual review by a paralegal. An RfP was initiated with three legal firms, two consulting firms and one internal solution provider to learn about their capabilities in this area. A provider was chosen, and the search executed. However, this was not worth the money. First, only a simple word search algorithm was required (like Ctrl+F). Second, rather generic terms like ‘information’ caused a lot of false positive results. This meant that in the end, almost every form had to be reviewed by the legal team anyway.
- A business transition required to find four specific words in a large pool of client contracts. The solution team came up with a rather high budget. 11% of which was for OCR (making the documents machine readable) and 45% for a complex natural language processing (NLP) model. The NLP model was questioned, but ultimately accepted with the promise that only the stand-alone instances of these words would be identified (e.g. only ‘REP’ and not ‘repetition’ or ‘republic’). The promise, however, was not delivered and many false positives were highlighted. It turned out, that a simple document search function using spaces around the three-letter word would have performed better.

1.2 Goal

Practice-oriented framework As business consultants, we should be able to critically evaluate the business case behind a project. For example, if the promised benefits are realistic in the given business context and that the efforts of a certain endeavor are worth the benefit it generates.

However, this is not enough. The suboptimal outcomes in the above examples were foreseeable and in two cases even challenged in advance. Nevertheless, these endeavors were executed. Of course, internal politics in the company played a key role. A key reason they were able to push their request through approval was an authority imbalance due to expert knowledge on one side and 'only common sense' on the other side.

Therefore, the goal of this guideline is **to provide an expert-driven and practice-oriented framework to critically evaluate the business case behind machine learning projects in financial services.**

Please note that by critical evaluation I do not mean to have a purely negative view on a business case and only look for reasons to discard it. It also means to notice a potentially hidden gem, to remove obstacles, save costs, clarify fuzzy benefits and/or realize benefits faster. Sometimes there are low hanging fruits but at other times this may require questioning the scope or even the goal of your project.

1.3 Scope

Focus on business case For a project to be successful and generate value once it's completed, the business case is only one puzzle piece – although an important one. It is the key focus of this paper, with the guidelines provided in section "4 Evaluation guideline" centered around it.

In addition, these guidelines also touch on benefits management and change management topics.

Not in focus Not in focus of this paper, but also important for a successful project are other project management activities and themes, for example the method; planning; progress tracking; handling of risks, assumptions, issues, dependencies & changes, or quality control.

Also, this will not be an introduction to the topic machine learning. Section "3 Quick introduction to machine learning" provides a brief overview of the key topics. The interested reader can find much more insightful resources online. Two that I have enjoyed are Andrew Ng's Stanford lesson series on [YouTube](#) and [Coursera](#), and Steve Brunton's intro to data science on [YouTube](#).

1.4 Relevance

A critical business view When looking through the machine learning literature, I found a lot of material on research (academic and corporate); implementation, software & code; and education.

When it comes to the business side, most of the published material I have found is marketing and sales driven. This inevitably leads to a rather one-sided view on machine learning, where benefits are vastly overpromised. For example, its ‘tremendous potential’ or ‘required to remain competitive’ or enable ‘unprecedented rapid innovation’ or ‘driving success’.

While one can find detailed pros and cons of a selected machine learning algorithm, I was unable to find something like a ‘value for money’ view for the business side. This paper aims to close this gap.

When it can be used When writing these guidelines, I had the following use cases in mind where they could be applied:

- Projects, where machine learning itself is a key deliverable (e.g. anti-money laundering, signal generation for trading, robo-advice, client onboarding, credit scoring, brand sentiment, damage assessment for insurance companies)
- Projects, where machine learning is a supportive deliverable (e.g. searching documents for words / phrases, refine financial forecasting models)
- Comparing vendor offers for a machine learning request for proposal (RfP)

2 Approach

2.1 Method

Expert interviews Most of the content in this paper is based on expert interviews. The word expert is referencing

- i. machine learning implementation experience, and
- ii. domain knowledge in financial services.

The insights gained from the interviews were then complemented with the author's own experience and some literature research.

2.2 Contributing experts & practitioners

Background of interviewed experts

I am very grateful to all my interview partners, whose inputs greatly contributed to this paper. In total, 10 experts provided valuable inputs to this paper. Below, I am providing a short description of their relevant background (i.e. data analytics and financial services). Kindly note, that some preferred to stay anonymous (e.g. in order to avoid negative 'publicity' for their company when having contributed a negative example or due to being an employee of a competing firm).

- Daniel Perruchoud: Professor for Data Science at FHNW (4 years), Team Head Client Analytics for Sales and Marketing at UBS (11 years), Data Scientist at UBS (7 years), Consultant Data Warehousing & Data Mining at NCR Teradata (1 year).
- Markus Grob: Consulting Lead Schweiz at b.telligent (2 years), Chief Analytics Officer (Member of Executive Board) at Argus Data Insights Schweiz AG (2 years), Head of Business Intelligence & Analytics at Helsana (3 years), Head Business Engineering Products at Helsana (2 years), Head Corporate Development at Sanitas (1 year), Senior Manager with focus on business intelligence, analytics & digital marketing at Accenture Interactive (8 years), Manager in Financial Services at Accenture Interactive (5 years).
- Robert Schumacher: Director & Member of the Executive Board at gateB (7 years), Customer Intelligence Solutions Manager at SAS (6 years), CRM Manager at Bank One (2 years), Associate Director Marketing at UBS (4 years).
- Senior data scientist at a large consulting firm. Over 5 years of experience in data science, with project experience in financial services.
- Partner at a small data consulting firm and lecturer on machine learning. Roughly 5 years of experience in financial services.
- Data advisory team lead at a major bank. Over 10 years of machine learning experience in financial services.
- Senior project manager at a major bank. Over 20 years of experience in financial services and roughly 2 years of project experience with artificial intelligence.

Machine Learning - Evaluation guidelines for financial services business cases



-
- Partner at a medium-sized consulting company. Over 10 years of experience in financial services and some project experience with machine learning.
 - Project lead at a major bank. Over 5 years of experience in financial services and some project experience with machine learning.
 - Managing Partner at a medium-sized consulting company. Over 20 years of experience in financial services and some project experience with machine learning.

3 Quick introduction to machine learning

3.1 Definition

The probably first and apparently most popular definition of machine learning reads as follows:

“Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed” – Arthur L. Samuel, AI Pioneer

While other definitions for machine learning can be found, most of them are close to the one cited above. Variations may for example include that machine learning is an application of artificial intelligence, generalize from computers to methods / models or expand the learning to include improvement from experience.

3.2 Distinction from other fields

There are some highly related terms, which popular sources and news sometimes even use interchangeably to machine learning. This can cause confusion, hence it makes sense to provide their definition & delineation from machine learning.

Artificial Intelligence

This is the most difficult one to nail down. There are several deviating definitions and on top of that, marketing has appropriated this term for many applications and promises. Due to this, there is a saying that “If it’s written in Python, then it’s Machine Learning – if it’s on PowerPoint, then it’s AI”.

The Oxford English Dictionary (and many tech firms such as Microsoft and Google) define AI as:

“[T]he theory and development of computer systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.”

Often, AI is divided into ‘narrow AI’ designed to solve a specific task, and ‘general AI’, which would be multi-purpose and mimic human intelligence across tasks.

An example of AI that is not machine learning would be a set of logical rules, such as nested if-else statements, rule engines, expert systems, knowledge graphs, sensoric mappings, remote sensing, or robot control.

Usually, Machine Learning is seen as a subset of AI. However, there are different views, which see AI and ML as overlapping sets since it could be argued that not every ML technique fits the above definition of AI.

Deep Learning

Deep learning is quite well defined and refers to a subset of machine learning methods based on artificial neural networks with representation learning.

Many (although not all) classical machine learning algorithms require features, which are usually defined in a separate step called ‘feature extraction’. Deep learning algorithms on the other hand conduct this step by themselves. For example, to recognize cat pictures, they do not require pre-defined features like ‘cat ear’, ‘legs’, ‘body’, ‘tail’ etc. but rather generate internal representations with which they are able to recognize cats.

Big Data Big data was originally associated with three key concepts: high volume, greater variety, and more velocity. Often it also means data sets that are too large or complex to be dealt with by traditional data-processing applications.

There is a close relationship between machine learning and big data. While there are machine learning methods which work with small data sets, bigger volumes increase robustness and validity – and often bigger volumes are required to properly generalize a problem. On the other hand, big data volumes often require machine learning techniques to analyze them. This is also why deep learning algorithms usually require more data than ‘classical’ machine learning algorithms.

3.3 The three main categories of machine learning

3.3.1 Determine the machine learning category

To evaluate machine learning use cases, it is helpful to understand the three main categories of machine learning.

The flowchart below outlines in a ‘quick-and-dirty’ way their key differences, followed by a more detailed sub-chapter for each category.

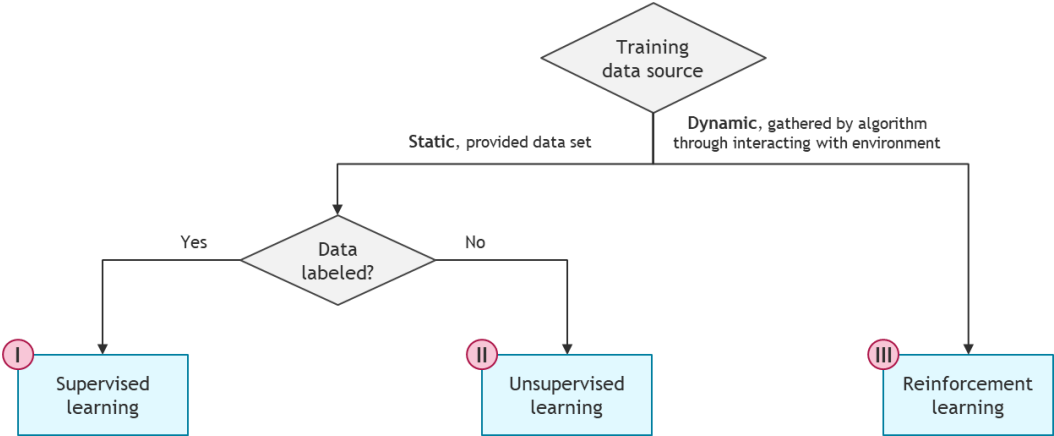


Figure 1 – Machine learning category decision tree

3.3.2 Supervised learning

Short description Supervised learning refers to algorithms which use labeled input data to train. For example, the input might be a picture of a triangle {matrix of pixel values}, or a vector of property valuation attributes {square meter (living area), square meter (property), zip code, year of construction, year of last renovation, closeness to a waterfront}. The corresponding labels could be the category {triangle} and a number for {property price}.

During the training phase, the supervised algorithm infers a function (‘model’) from the training examples. Roughly speaking, this works by comparing predictions for a given input to its label, and then correcting the model parameters to improve those predictions.

The goal is that the final model resulting from the training phase can correctly predict the 'labels' for new, unlabeled input data.

Precondition You know how to classify (i.e. label) the input data as well as the type of behavior you want to predict.

Properties

- Input Data: Labeled, static (provided set of data).
- Feedback for learning algorithm: Direct.
- Goal: Predict an outcome or even recommend an action for new data points.

Semi-supervised For the purpose of this paper, semi-supervised algorithms are subsumed in the 'supervised' category. First, because they are usually relevant in a supervised learning setting, and second it can be argued that semi-supervised is a special case of 'weak supervision'.

3.3.3 Unsupervised learning

Short description Unsupervised learning refers to algorithms which use unlabeled input data to train. It explores the input data without being given an explicit output variable. The algorithm captures patterns, clusters, or other (hidden) relationships in the data through self-organization.

Preconditions You may have to specify e.g. number of dimensions of the output space.

Properties

- Input Data: No labels, static (provided set of data).
- Feedback for learning algorithm: Usually none, although there are algorithms that can provide a feedback loop (e.g. mimicry-based approaches).
- Goal: Find the inherent structure of the input data without using explicitly provided labels.

3.3.4 Reinforcement learning

Short description Reinforcement learning refers to algorithms which learn through interacting with an environment to maximize a (cumulative) reward. Hence, it does not explore big data sets as supervised and unsupervised algorithms usually do, but rather the input data is generated through the series of actions the algorithm takes.

It is useful when learning can only happen from interacting with the environment, or when an expert cannot clearly define an ideal end state. A key strength is the explorative capabilities.

Preconditions Reinforcement learning requires a well-defined reward function, a closed world and 'many attempts' when exploring that world. The 'closed world' condition often requires a simulation or digital twin.

Properties

- Input Data: Dynamic, usually generated by interacting with environment (through decision process of the algorithm).
- Feedback for learning algorithm: Through the reward function, which is determined by the series of actions the algorithm took.
- Goal: Learn a (optimal) series of actions to achieve an optimal outcome (maximized reward).

3.4 Important terms

Algorithm	In machine learning, an algorithm is the mechanism which learns. This means, a procedure that runs computations on the training data and calculates a model (or at least most of the model parameters).
Model	<p>In science, a model represents a simplified / abstracted description of reality. They usually have a descriptive and an explanatory aspect.</p> <p>In machine learning, a model is usually the output of the training (learning) phase of the algorithm. The model is a mathematical construct which can be used with new inputs to e.g. classify (label) those inputs, predict a target variable or action, or whatever else the model is intended for.</p>

4 Evaluation guidelines

4.1 Business case assessment forms

4.1.1 Table 1: Machine learning category

Based on the machine learning category of the chosen algorithm, some criteria are evaluated differently. Therefore, it is helpful to first think about which algorithms you plan to assess and potentially implement in your project – and in which main category they belong to.

Supervised learning	Unsupervised learning	Reinforcement learning
Yes / No	Yes / No	Yes / No

4.1.2 Table 2: Potential showstoppers and promoters

Before going into the detailed analysis of your business case, this table helps you to gauge if there might be any major roadblocks – or vice versa, an imperative to proceed.

Evaluation criterion	Weight	Machine learning business case viability						
		-3	-2	-1	0	+1	+2	+3
The Gartner analytic continuum								
Insufficient data volume								
Insufficient data quality or availability								
Unstable environment								
Regulatory barriers								
Lack of explanatory factors								
Strategic investment case								
Total (weighted sum)	n/a							

4.1.3 Table 3: Cost and benefit drivers

This table dives into the core business case drivers – costs and benefits. Assessing those will take some additional time and effort. Ideally, this will make your business case more robust by helping to address some potential issues early.

Evaluation criterion	Weight	Impact on costs					Expected benefits				
		-3	-2	-1	0	Estimate	0	+1	+2	+3	Estimate
Data											
Technical complexity											
Project complexity											
Acceptable error margins											
Maturity of employed capabilities											
Data sourcing & response time											
Scalability											
Proof of reliability											
Explainability and interpretability											
Expected benefits											
Total (weighted sum)	n/a										

4.1.4 Table 4: Additional success factors

The last business case assessment table aims to estimate the impact of business case relevant points other than direct costs and benefits.

Evaluation criterion	Weight	Impact on business case						
		-3	-2	-1	0	+1	+2	+3
Time criticality								
Business acceptance & change								
Focus on smart data over big data								
Design actionable output								
Difficulty level of output prediction								
Identify and evaluate alternative options								
Total (weighted sum)	n/a							

4.2 How to use the assessment forms

Background When interviewing the experts and consulting some of the literature on machine learning, it became clear that it is not easy to define objective, case-independent evaluation criteria to assess a machine learning business case.

Instead, it was possible to define evaluation criteria along with qualitative answer options, as seen in the assessment forms.

Customization required The actual answers, as well as the weight of each evaluation criterion, can drastically vary between different projects. Hence, whoever uses this questionnaire must tailor it to his/her project and ultimately weigh and tally up the answers by him/herself in a manner deemed reasonable in the project context.

Scales Table 1 does not use any scales or weights. It just serves to identify the machine learning category. The scales for table 2 which intend to measure the business case viability, and table 4 which intends to measure business case impact, are going from negative three to positive three, with a neutral mid-point.

Table 3 has two scales, one for costs which goes from negative three to zero, while the scale for benefits goes from zero to positive three.

The intention to design the scales in this way was, for example, when all business case criteria are added up, the sign of the sum will already give an important indication (i.e. positive meaning a sustainable business case). With regards to costs and benefits, the same applies.

The scale representation table below provides a rough indication for what the scores are meant to indicate.

Scales	-3	-2	-1	0	+1	+2	+3
Table 2: Business case viability	almost unviable	strong reservations	some reservations	neutral	some endorsement	strong endorsement	full endorsement
Table 3: Impact on costs / benefits	high cost impact	medium cost impact	low cost impact	no impact	small benefits	medium benefits	high benefits
Table 4: Impact on business case	highly negative	negative	slightly negative	no impact	slightly positive	positive	very positive

Weights Different criteria are more or less important depending on project context or business case goals. Hence, the weights allow you to consider your specific context by giving more or less importance to different criteria. For criteria that are not applicable, a weight of zero can be applied.

Estimates (optional) If money estimates are available for the costs and benefits, a more intuitive evaluation approach is to sum up the estimates and calculate their grand total.

However, it is not uncommon for projects in their early stage to have estimates that are not yet available or have a high uncertainty attached to them. Hence, the scale & weight approach is the baseline and using estimates is optional. One possibility to decide how a cost is rated would be to think of the cost impact in relative terms of your overall budget. For example, if it's less than 5% of overall budget it's -1; between 5% to 10% it's -2; and higher than 10% it's -3.

Implication	<p>Of course, choosing your own weights allows you to ultimately control the outcome of the questionnaire. Hence, the weights should be agreed with key stakeholders to avoid the impression of either pushing or discouraging a certain project.</p> <p>The main intention of these guidelines is to allow you to critically think through your business case and optimize it. To maximize your return on investment and minimize unnecessary efforts and pain.</p>
Evaluation help	<p>The following sections are going to provide some helpful information and insights into the thoughts that went into each evaluation criterion. The intention is to facilitate applying your project context onto the assessment forms and provide a framework, or at least a feeling, on how to answer them.</p> <p>Each criterion contains a paragraph called “evaluation implications” that conveys a rough idea on how to score the criterion on the assessment form.</p> <p>To make this more tangible, this will be wrapped up with an exemplary ‘walk-through’ case.</p>

4.3 Machine learning category

Determine which machine learning category applies to your project as this will be helpful for some evaluation criteria in the following section.

Section 3.3 “The three main categories of machine learning” provides a short description and some rough guidance on their categorization.

4.4 Potential showstoppers & promoters

This section contains the criteria which can ‘make or break’ a machine learning business case. This ‘business case viability assessment’ includes two dimensions:

First, the use case viability, i.e. if it is even possible or reasonable to apply a machine learning algorithm to the assessed problem.

Second, how the expected return on investment is impacted.

In most of the cases, these will coincide anyway. However, you can imagine the scenario where a project might promise a tremendous ROI, but it might not be allowed by regulations or not be possible due to a lack of data.

4.4.1 The Gartner analytic continuum

In 2010, Gartner plotted the ‘Analytic Continuum’ in terms of value and difficulty. The four key categories are quickly introduced below. Note that the examples do not strictly fall into one category, e.g. regression analysis can be diagnostic or predictive.

Hindsight	<p>Descriptive Analytics</p> <p>Describing what happened. This is applied heavily across all industries and can be done with standard business intelligence tools. Some examples are: Automated reports, queries and ad-hoc reports, alerts, data discovery, OLAP, drill down / across.</p>
-----------	--

Insight	<p>Diagnostic Analytics</p> <p>Investigating why something happened. Again, this is already applied across most industries and is covered by most business intelligence tools. Some examples are: Sentiment analysis, data / text mining, visualization, regression analysis.</p>
Foresight (passive)	<p>Predictive Analytics</p> <p>Describing what will happen – or more generally, prediction tasks (like classifying new / unknown data). Typical business intelligence tools might cover the lower end of this field. But for most use cases, advanced statistical suites, or outright machine learning (or deep learning) tools are required. Some examples are: Multivariate statistics, forecasting, pattern matching, predictive modeling, optimization, complex anomaly detection.</p>
Foresight (active)	<p>Prescriptive Analytics</p> <p>Generating advice, recommendations, or actions on how to make a desired outcome happen. In most instances, these use cases require sophisticated machine learning or deep learning tools, although rule engines or knowledge graphs might be applied as well. Some examples are: Simulation, network analysis, complex event processing, automated incident remediation.</p>

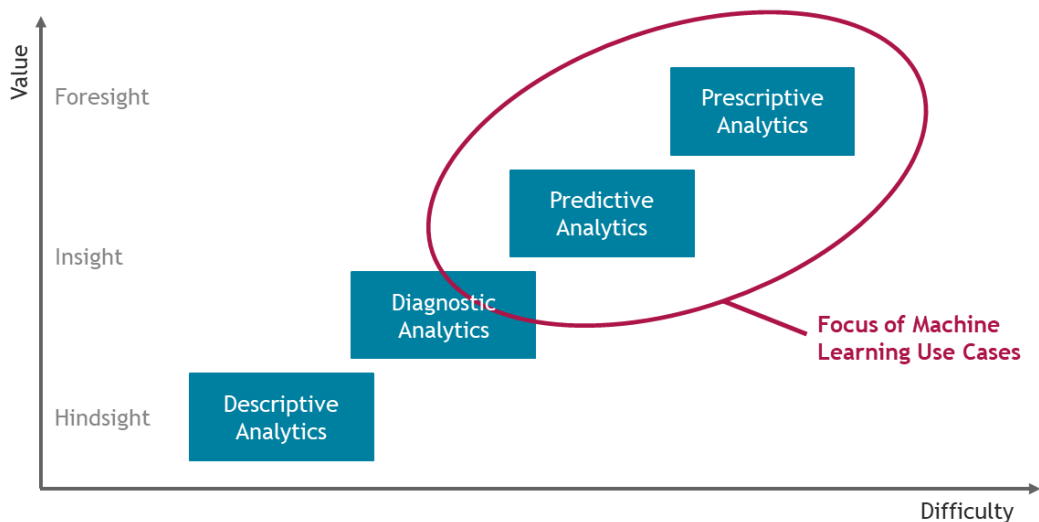


Figure 2 – The Gartner analytic continuum

Evaluation implications

If the use case in question is in the **descriptive or diagnostic domain**, there are most likely **already established tools** in the company which can be leveraged (with lower cost and effort) instead of machine learning. Hence, business case viability can be rated in the negative area, with the exact score depending on the already available tools.

If it's in the predictive or prescriptive area, on the other hand, it can initially be assumed that there is a potential machine learning use case. Therefore, business case viability can be rated in the positive area, where prescriptive analytics should score higher than predictive analytics.

4.4.2 Insufficient data volume

Background

Machine learning requires 'enough' 'relevant' data of 'good' quality. For supervised learning, these requirements also extend to the training data being labelled.

If there is not enough data, or only in insufficient quality, the resulting model will perform poorly, and the project will not realize the intended benefits. In the worst case, it might even produce negative benefits if business blindly follows low quality or biased predictions.

Unfortunately, it is not so easy to provide clear numbers or hardcoded instructions of what constitutes 'enough' data or 'good' quality. However, there is a set of factors which impacts the amount of data required.

- Considerations for supervised and unsupervised learning
- For supervised and unsupervised learning, the amount of data required depends on several factors:
- The selected machine learning algorithm. More complex algorithms require a larger amount of data. Standard algorithms that use structured learning (pre-defined features) are not going to need the same amount of data as deep learning algorithms, which have to figure out all the parameters by themselves. In case the algorithm is already known at this stage, computational learning theory provides insights into its data requirements.
 - The complexity of the model. Simplified, the number of parameters that the algorithm should learn.
 - The complexity of the output. This is related with the model complexity, as a more complex output usually requires a more complex model. Factors to consider regarding the expected outputs are its variability, number of properties (e.g. animal pictures are expected to have more properties than simple shapes) and number of classes (e.g. binary classification 'either car or dog' vs multiclass 'which type of animal').
 - Variability of the environment. The more uncontrolled an environment is, the more diverse input data it will generate. This means a model needs to achieve a higher degree of generalization and therefore requires more data. For example, a virtual assistant can receive a variety of questions, and similar questions can be asked in different styles, with different choice of words, with more or less spelling and grammar mistakes.
 - A more intuitive hypothesis usually requires less data.
 - Dealing with rare events requires more data to be collected.
 - Non-linear relationships usually require more training data.
 - The structure of the input data, like its number of properties and its distribution also influence the amount of data required.
 - Narrower acceptable error margins, i.e. more accurate model predictions, usually require more training data.

Considerations for reinforcement learning

Reinforcement learning is based on actions an algorithm takes in an environment and is (sometimes) rewarded or punished for. This algorithm does not require (many) training data as input, but rather collects this data actively through own actions and observing the results. This means, for reinforcement learning, the data requirement is that 'many attempts' are possible, in order for the algorithm to test the outcome of different action sequences. If only few attempts are possible, a reinforcement algorithm will most likely produce a poor performance.

Risks

Overfitting is a well-known issue, which occurs when a model has been trained on the training data set for so long, that it starts to 'memorize' the correct answers instead of building a general model of the data. This can be mitigated through a separate 'validation data' set which is not used for training.

A rather newly discovered issue is 'underspecification'. This refers to the gap between the requirements the engineer had in mind and the requirements actually enforced by the final model. This might often be overlooked because the model will perform well on the validation data. If searched for, it can be observed through the fact that using the same training data with the same algorithm and the same training process, the resulting models can make vastly different predictions with real-world data. This can have two root causes: Either a highly complex model, with billions of

parameters, where initial seeds and hyperparameters can strongly influence the final model – or insufficient validation data to represent all real-world scenarios.

For example, when trying to predict the reproduction number R_0 during the COVID pandemic, many models suffered from this. While it was easy to build a model explaining the currently available data, the models of different researchers made vastly different predictions.

Rule of thumb To get a rough idea on data requirements in an early project stage, there are some rules of thumb. These should be used with caution and critical thinking.

A commonly applied rule of thumb is the '10 times rule'. It means that there should be 10 times more training examples than parameters (or degrees of freedom) of the model. There are more complex rules to determine data requirements, however, they would go beyond the scope of this discussion.

For supervised learning, on top of the required training data, validation and test data is usually required. Most sources recommend 25% to 30% of the overall available data set to be used for validation and testing.

Evaluation implications In case of doubt, it is likely that you would need to consult with a data scientist to determine the data requirements of your project.

In case the amount of available data is deemed to be insufficient, the follow up question then is how costly it would be to obtain (e.g. collect, generate, make available or buy) additional data. The answer can then be factored into your business case.

If there is already enough data available, a +3 rating can be applied. If additional data is required, but can be gathered with low cost & effort, this can still constitute a positive rating. Additional data requirements which are costly, but also produce additional benefits outside the project may get a neutral rating, but if the cost-benefit ratio is clearly negative, this would constitute a negative rating. Finally, if additional data cannot be acquired or collecting it is prohibitively expensive, a -3 rating can be applied.

4.4.3 Insufficient data quality or availability

Considerations for supervised and unsupervised learning Data quality and data availability are not exactly the same topic but can have a common link and are often used interchangeably. For example, data is 'per-se' available, but not in the desired format.

Data quality is very important because machine learning algorithms do not study the data but rather the relationships and patterns behind the data. Bad data will therefore result in bad models (aka the 'garbage in, garbage out' principle).

Data quality Regarding data quality in the classical sense, key aspects to pay attention to are:

- Data should be relevant. For example, if there was a regime change in a time series, data before the regime change might be less relevant. Or if the use case is about exchange traded options, then OTC option data should be removed from the data set.
- Data should be representative for the different output classes (i.e. free of bias). If a model should classify financial transactions into 'potential fraud' and 'no fraud' classes but has 99'995 'no fraud' and 5 'fraud' cases to learn from, chances are high that the model prediction is distorted towards 'no fraud' because this would be the correct prediction most of the times.
- Data should be (mostly) complete and (mostly) free of errors and outliers. If the algorithm is fed too many errors or outliers, it will learn to model incorrect predictions.
- Be wary of poorly documented data and make sure it is what you expect it to be.

Data preparation	<p>Best practices in data cleaning and preparation should be applied. For example, making the data machine readable, transforming it into the correct data types, normalizing it where necessary, and dealing with outliers and missing data points. This should include some sanity checks (e.g. expected data distribution, labels matching the examples etc.).</p> <p>This might be abbreviated for algorithms which do not require much data preparation, transformation, normalization and/or are robust to outliers.</p>
Data availability	<p>The required data might be available, but not in the desired format. For example, it's on physical paper, documents stored as pictures or unstructured data fields.</p> <p>One example of unstructured data and its challenges are in extracting information from text. A structured date field would be, for example, "31.03.2024", which an algorithm can extract quite well. However, a date can also be defined like this: "the maturity date will be the second business day following the day on which the final share price is determined". Values could also be conditional, for example: "threshold amount means with respect to Party A, the higher amount between USD 12'000'000 or 3% of the stockholders' equity of Party A".</p> <p>What can make matters even more challenging is, that documents can exist in different versions and/or with amendments which might have to be considered as well in order to extract the correct values.</p>
Considerations for reinforcement learning	<p>Since the data is collected through interacting with an environment, the data quality itself is less in the focus. More important however, is a well-defined reward function.</p>
Evaluation implications	<p>If you are dealing with poor data quality or low availability, the question is at what cost (if at all) such data can be improved. The interviewed experts warned, that structuring unstructured data is often very expensive. Hence, this is a major determinant for the business case. A -3 rating would be applied if data quality/availability is that bad, that fixing it would make the project ROI clearly negative. A less negative or neutral rating can be applied if data quality is bad, but fixing it is not too expensive and/or might potentially create positive side-effects for other users of that data.</p> <p>On the other hand, if data is already structured and available in tabular form, it might be overkill to use complex deep learning algorithms. In such cases, simpler machine learning algorithms can perform as well, or even better, than their more complex and expensive cousins. Such a constellation can be rated in the positive area of the scale.</p>

4.4.4 Unstable environment

Considerations for supervised and unsupervised learning	<p>Often, machine learning algorithms are only trained once or twice. This is fine, if the resulting model is only used once, or if it's used multiple times in the same context.</p> <p>In general, however, the predictions of un-/supervised learning algorithms will deteriorate (and therefore require re-training) if the environment or underlying conditions change (for example, the market regime). Similar if a covariate shift occurs, i.e. if the data fed into the trained model to make predictions structurally differs from the data used to train the model. Such re-training is rather costly, because it usually has to be done from scratch, using the new, increased data set.</p> <p>Incrementally adding new batches of non-stationary data (i.e. data for which properties like mean and variance change over time, which is often the case with real-world data) to train an existing neural network (or other machine learning algorithm) can lead to 'catastrophic interference', i.e. lower model performance. There are new approaches to address this, however these are currently still in the research and development phase.</p>
Considerations for reinforcement learning	<p>Reinforcement learning requires a closed world. This means an environment without interfering factors from outside, so that the same series of actions produces predictable results (not</p>

necessarily the same, but within a probabilistic stable range). This means, it can adapt to change within a certain range.

Evaluation implications

It is worthwhile to assess how the relevant environment (and by extension the data) has changed in the past and/or consider how much change can be expected in the future.

If you are dealing with a fast-changing environment, machine learning may not be the right answer – or at least, it would require a sound business case. Hence, the faster changing your industry or environment is, a more negative rating should be applied.

If you are dealing with slow and gradual changes, it would make sense to add model re-training to the maintenance plan. And in a very stable environment, it is well possible that the initial model can stick around with good performance for a long while. These constellations warrant a neutral to positive rating.

4.4.5 Regulatory barriers

Background and examples

Especially in Switzerland, but also in other countries, laws on banking secrecy, data protection and related areas can impede or even block potential machine learning use cases.

One example which was mentioned was of a bank running an ad campaign with a third party vendor (e.g. Google). Due to the bank client secrecy, it must not give feedback on specific conversion cases, i.e. the third party algorithm cannot improve through feedback.

Another example would be health insurance where client data is often classified as 'sensitive'. Sensitive data may not be processed or exchanged (even company-internal) without a strong justification.

But even general data protection laws can block certain use cases. For example, in a country with low data protection standards, a bank or insurance could buy personal data from prospective client segments and use this to ease the onboarding process with 'auto-fill' options – even for prospects which are not yet bank clients.

Other regulatory and compliance topics which might impact machine learning use cases are cross-border data sharing, consumer protection, SOX controls or competition laws.

Evaluation implications

Check with your compliance and/or legal team, if there are any barriers or obstacles due to regulations or internal policies.

If there are such obstacles, the additional effort to fulfill extra requirements, approval processes etc. should be factored into your business case.

A -3 rating would imply, that regulations forbid your business case, while other negative ratings would mean a cost impact directly for the project or on the 'business as usual' (BAU) organization (e.g. requiring extra audit controls). The neutral point would be if regulations do not meaningfully impact the project or business. A positive rating could be applied in cases, where the machine learning project would actually solve a regulatory compliance problem or audit issue.

4.4.6 Lack of explanatory factors

Applicable for supervised learning only

Given the training data, if a knowledgeable human would be able to categorize it, make predictions or take decisions, then it obviously has informational value.

If this is not the case, then the prize question is, if there could be patterns in the data which humans have not yet recognized. If this is a realistic possibility, the data likely possesses informational value.

Example	In a gastronomy use case, the goal was to predict revenues per menu item to optimize the food purchases. They had very good training data available and used state-of-the-art algorithms to learn from it. However, it turned out, that there simply weren't any explanatory factors to predict menu item revenues better than the gut feeling of the kitchen staff.
Evaluation implications	<p>For unsupervised and reinforcement cases, this criterion can be omitted (weighted zero or rated neutral).</p> <p>For supervised cases, the rating would be in the neutral area, if there is at least potential informational value in the data. The rating should be more positive, the clearer i) the informational value is and ii) the business value of that data.</p> <p>On the other hand, if there is not even the potential for informational value in the data, supervised learning cannot generate value and therefore this would generate a negative rating.</p>

4.4.7 Strategic investment case

Learning experience	<p>Besides the business case of the individual project, it could be, that the company made a strategic decision to build machine learning capabilities. This of course requires a certain learning journey, where experience has to be gained. So, within that strategic goal, a business case which on its own is not going to return a positive ROI, might still be viable. This doesn't mean to push any far 'out of the money' business case with this argument, but it should also be considered that gaining experience has value in itself.</p> <p>Also, one interviewed expert pointed out that it is rare for a machine learning project to have no effect. As an example, he mentioned a marketing project with only sparse data in which they were still able to achieve a lift (i.e. higher response rate) of three. Of course, there might be projects which do have an impact, but still were not 'worth their money'.</p>
Explorative projects	Another consideration is, that some projects can have an R&D (research and development) character. In other words, the benefit is unknown or speculative at best. Again, this should not be used as an excuse to push any machine learning project, however it is worth considering that true innovation is often attached to uncertainty.
Evaluation implications	<p>If your company does already have standing machine learning capabilities, then it makes sense to only undertake projects which promise a positive ROI. Hence, the rating on this criterion should most likely be neutral.</p> <p>However, if such capabilities are currently built as part of a strategic effort, then an investment case can make sense and a positive rating is warranted.</p> <p>An explorative project can depend on market opportunities or strategic goals. If these conditions are fulfilled, a positive rating can be applied.</p> <p>On a further note, considering the nature of explorative projects in the context of the Stacy matrix, an agile approach is recommended (since agile methodologies have a proven track record of dealing with projects in the 'complex' domain). And it is worth pointing out that methodologies like Scrum work with a capped budget. In other words, when the benefits are unclear, you should limit the amount of money spent to chase them.</p>

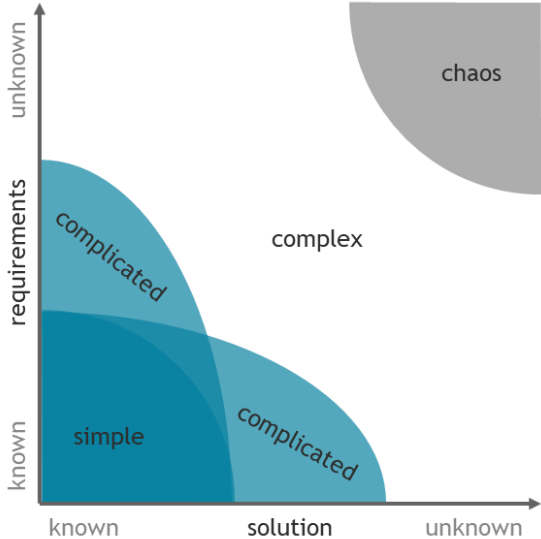


Figure 3 – Adapted Stacy matrix (own depiction)

4.5 Cost and benefit drivers

4.5.1 Data

Volume, quality & availability Data and its impact on the business case has already been discussed in Sections 4.4.2 “Insufficient data volume” and 4.4.3 “Insufficient data quality or availability”.

However, every interview partner has pointed out the importance of data. Hence, I thought it’s worthwhile to mention it again in the cost / benefit context.

While there is often a cost associated with obtaining or improving data, it can also generate some benefit. Your company now possesses additional data, or has cleaned up some databases, or made some data that was previously useless, available. This might be useful for future projects or daily operations.

Evaluation implications If data is not a showstopper, it may be worthwhile to spend some extra thoughts on the cost and benefit impact it is expected to generate.

However, this criterion could also be weighted down (or even omitted) in order to avoid giving data too much weight in your project evaluation.

In many projects, especially if their nature is explorative, there will be an effort required to obtain the data and / or ensure that the data quality & availability reaches a useful level. This might include things like labeling the data, pulling documents out of an (electronic) archive, or aligning the format of data pulled from different sources. A rough cost impact can usually be determined based on days of effort to obtain / clean the data, or if using the rating scale, then the rating could be applied in relation to the project budget.

Regarding the benefits, this might be rated neutral if the obtained and / or cleaned data is only useful for the project itself. If the benefits would apply to the wider BAU organization, for example paper documents are digitized and uploaded into a commonly accessible electronic archive, or the

data quality improvement facilitates the financial accounting process, then a positive rating could be applied relative to the project benefits.

4.5.2 Technical complexity

Implementation	The technical complexity refers to the type of algorithm applied, as well as the complexity of the resulting model. Two key cost drivers in that regard are the technical infrastructure and skilled talent. More complex algorithms / models require more computational resources. In addition, they also require higher skilled talent, than a simple regression model.
Operation and maintenance	Besides the implementation cost, once a model is deployed into the productive environment, it will also create operational expenses (OpEx) and maintenance costs. As a rule of thumb, more complex models will generate higher OpEx and maintenance costs. For example, they might require a certain supervision / sanity checks, trigger business questions or are simply more prone to potential bugs and issues. Furthermore, depending on the environment, periodic re-training might be required to keep the model outputs within the desired error margins.
Hidden costs	It is often the case that for a machine learning project, data from different sources is fed into the algorithm. While this unlocks potential by combining different kinds of data, it can create dependencies between those data sources (e.g. if certain attributes are required to link data from different sources together). At the very least, it creates a dependency from the deployed model to its data sources.
Evaluation implications	To estimate potential costs, you should have a rough idea of the potential algorithm(s) that might be utilized, as well as the expected complexity of the final model. As outlined above, higher technical complexity will almost certainly drive up the cost (i.e. more negative rating). With regards to the benefits rating, it depends on the use case. While there are cases where (almost) the same benefits can be achieved with a simpler algorithm / model, it can certainly be the case, that a more complex algorithm / model can deliver higher benefits – or in fact is required to attain any benefit at all. In such cases, the benefits rating would be neutral to positive.

4.5.3 Project complexity

Some reference points	Besides the technical complexity, there are other factors which can increase the complexity of a project and therefore its cost and timeline: <ul style="list-style-type: none">▪ Business & organizational: As more products, business areas and / or functions are involved, the more complex a project becomes.▪ Environmental: The environment can drive complexity. In financial services this is often regulatory driven but can also arise from market changes or societal trends.▪ Uncertainty: For example; unclear scope, ill-defined assumptions, vague or even opposing requirements, pending decisions, and missing vision.
Implication	While some complexity drivers can and should be dealt with (such as vision, scope etc.), other like existing regulations must be accepted and the resulting cost implications considered in the business case. Higher complexity implies more parts of the organization are involved and / or higher risks, which would warrant a higher change in budget. Therefore, the more complex your project, the more negative the cost rating. The benefits rating in my opinion should be neutral, as I adhere to the

governance principle that a project should only be as complex as necessary. From that baseline, there would be no benefit from additional complexity. Also, fewer complexity would result in dis-benefits because it would mean that certain success factors are omitted.

4.5.4 Acceptable error margins

Desired model traits

When talking about machine learning, and in particular artificial intelligence, business users often expect perfection (because “machines don’t make mistakes” right?).

In fact, the model of a well-trained algorithm often performs better than a standard human (lower error margin), although rarely perfect. A human can perform his / her tasks perfectly – but often not consistently. It might depend on who in the team is performing the task and / or the mental fitness of the task performer on that workday.

Some key traits a model should have are:

- ‘Useful’ predictions, i.e. performing well with real-world data (no model overfitting)
- Robust predictions, i.e. reliable, producing similar outputs for similar inputs, robust against errors or outliers in the data
- ‘Correct’ predictions, i.e. precise & accurate

Error types

The two statistical error types are:

- Type I error: False positive, i.e. miss / exclude a true sample / fact / element
- Type II error: False negative, i.e. detect / include a wrong sample / fact / element

Which error type should be minimized, and what the acceptable error margins are, depends on the use case. In fraud detection, a bank might want to be overcautious and minimize the type I error at the cost of accepting more type II errors (‘overdetect’) while in churn rate reduction it might be more reasonable to not ring false alarms and focus on reducing the type II error at the cost of accepting more type I errors (‘underdetect’).

Often used measures in machine learning are recall and precision, while in statistics the errors are often measured through significance level α and the ‘power of the test’ $1-\beta$. Some metrics are not direct measurements of an error type. So, when deciding on which error type to focus on, it helps to be aware of what your metrics are actually measuring.

Evaluation implications

A very important point is to clearly manage the expectations of the business side. First, that the machine won’t be perfect and second, that getting close to perfection will increase the cost of the project. Many interview partners pointed out, that many use cases are perfectly viable with an error margin of 20% - 30%. There are two counterbalancing factors to consider:

- The cost of an error. This might be rather low when generating cross-selling opportunities or churn warnings for a relationship manager who usually does not have much time to spare. Generating 100 or 500 recommendations might not make a big difference in the bottom line. On the other hand, missing a fraud case can be expensive.
- The cost of reducing the error margin. More accurate predictions require more training data which is often costly to obtain and prepare. In addition, it often also requires more complex machine learning algorithms to be applied. Training a complex convoluted or recurrent neural network can quickly cost several hundred kCHF.

The lower the acceptable error margin, the more negative the cost rating.

On the other hand, when the acceptable error margin is medium or even high, this can, but doesn’t have to, yield a positive benefits rating. Keep in mind the cost of an error in the productive environment should be considered as well.

4.5.5 Maturity of employed capabilities

Background	<p>The Fraunhofer Institute (2018) published a paper on machine learning. In one of the chapters, they looked at the maturity levels of different capabilities, fields of application and related algorithms.</p> <p>To keep this overview simple & business relevant, I decided to stick to the capabilities.</p>
Level 1: Well established	<ul style="list-style-type: none">▪ Create groups of similar data▪ Classify objects▪ Estimate and predict values▪ Recognize images▪ Recognize speech / language▪ Extract pre-specified information from a text and execute simple instructions▪ Learn with very large amounts of data
Level 2: Demonstrators exist	<ul style="list-style-type: none">▪ Choose promising actions for a 'bot'▪ Understand pictures and videos in context▪ Combine multi-modal contents (e.g. text, audio, images, sensor data etc.)▪ Transparent, explainable, and robust models
Level 3: Early R&D phase	<ul style="list-style-type: none">▪ Understand language and text, and communicate▪ Generate new content▪ Learn with additional knowledge (e.g. expert input, logical rule sets, knowledge graphs)▪ Learn with sparse data▪ Adapt to changing environment▪ Automated learning (partial automation of model development)
Out of reach	<p>Something that is still far out of reach is a 'general AI', i.e. a human-like artificial intelligence. Also, innovative thinking which can result in generalized rules are still out of reach for machines. For example, an algorithm can be trained to predict the trajectory of arrows in an archery contest. However, it will not be able to derive Newton's Law from this.</p>
Evaluation implications	<p>While this categorization is already four years old, it can at least provide a good indication with respect to your own use case. The more standardized and well established an algorithm, the higher the chances for success, and the less likely they are to generate cost overruns. Hence, they could be rated cost neutral.</p> <p>Less established algorithms will not only be harder to implement but are also more likely to cause cost overruns due to unforeseen problems. Hence, negative cost ratings are appropriate.</p> <p>The benefits rating should, in the base case, be rated neutral and might be rated positive if your business case can be covered with only minimal code customization and / or feature engineering.</p>

4.5.6 Data sourcing & response time

Factors to consider	<p>Another cost driver can be the data sources and required response time of the productive system.</p> <ul style="list-style-type: none">▪ Are there multiple data streams?▪ What is the daily data volume?
---------------------	---

- Will the productive system consume historical data, or process real-time data? The latter often requires additional interfaces and technical capabilities.
- What is the required response time? Is it measured in milliseconds, or can requests be processed in (overnight) batches?

Evaluation implications The higher the demands, the higher the cost impact (i.e. more negative rating). On the benefits side, it depends on the requirements. If you have a service level agreement which prescribes an 80ms response time, it will certainly generate a benefit to fulfill this. However, there are none or not much extra benefits in achieving a 70ms response time.

4.5.7 Scalability

Factors to consider Will your final model have to be scalable, is it maybe a pilot or 'proof-of-concept' that will be extended, if it fulfills expectations? If yes, it might influence decisions regarding system architecture, but also regarding the applied algorithm. Scalability might be required in different forms:

- Number of requests to the productive systems
- Amount of training data for the algorithm
- Numbers of businesses / products / functions within an organization

Evaluation implications As with the previous criteria, higher demands will lead to a higher cost impact while the benefit impact depends on the requirements.

4.5.8 Proof of reliability

Factors to consider Sometimes there is a heightened emphasis on the reliability of employed model or systems. One example is the health industry, which is quite heavily regulated when it comes to treating patients. But there are also fields in the financial services industry with heightened regulatory or audit scrutiny. In such cases, more extensive testing regimes might be required or even something akin to a certification. This, of course, will increase implementation costs, but also maintenance & future updates.

Evaluation implications As described, achieving some kind of 'proof' of reliability will increase the cost. However, it can also generate benefits. This could be regulatory compliance, but it might also increase the trust from users and / or clients.

4.5.9 Explainability and interpretability

Terminology In the context of machine learning, explainability and interpretability are often used interchangeably. However, they do have different meanings. Explainability refers to the understanding of how models are created by the algorithm, while interpretability refers to the understanding of the model output. With the right effort, interpretability can be ensured. Explainability on the other hand has a dependency on the chosen algorithm, although progress is being made to improve this (see 4.5.5 "Maturity of employed capabilities").

Enhancing trust For example, a Gartner article on AI Ops Platforms states that many IT operation staff previously made experience with misbehaving automation tools that have misfired or otherwise wreaked havoc. Hence, to earn back trust, explainability and transparency are important.

And given the 'Responsible AI' movement, it might be important for certain use cases to support end user comprehensibility of the system to enable monitoring and human intervention. A classic example would be to avoid bias in credit scoring against certain demographics.

Enhancing acceptance In most cases, however, explainability is required from the business side. Especially when model outputs are used in marketing or customer contact.

It is often also important to have edge cases explained. Even if it might not be worth the effort, from a pure 'value for money' perspective, it will help with business acceptance. Because a client advisor may also be the one who has to explain this edge case to the client...or at least fears the client may inquire about it.

Evaluation implications It may be worthwhile to check who is using the model outputs and if they do require explainability. While this does cause extra effort on the project side, it is in most cases absolutely worth it from a change management perspective.

The cost impact is usually more apparent, with the negative score depending on the additional effort caused. However, the benefits this generates should not be forgotten and might in fact be rated higher than the costs.

4.5.10 Expected benefits

Identify benefits A business case is justified through benefits, which can include things like innovation, higher client or employee satisfaction, cost savings, the avoidance of costs / dis-benefits or achieving regulatory compliance.

These benefits should of course be properly identified as part of the business case.

Realistic expectations In the case of 'machine learning' or 'artificial intelligence', there are many glossy marketing brochures which tend to overpromise on what can be achieved or how valuable those achievements really are. Therefore, it is important to have a grounded view and, if necessary, manage expectations towards your senior management.

Define metrics With the benefits identified, the next step is to define the desired business and corresponding KPIs to measure them.

For example: Increase amount of detected anti-money laundering (AML) cases by x%, increase amount of detected credit card fraud cases by x%, increase alpha of a fund by x basis points, reduce churn rate by x%, increase client conversion rate by x% or increase product cross-selling by x%.

Evaluation implications Obviously, the core benefits expected to be generated by the project should be considered in the business case. Their impact should be calculated and considered here. Since the rating scale is capped at +3, the weight of this criterion should be adjusted in accordance with the number and magnitude of expected benefits your project intends to realize.

4.6 Additional success factors

4.6.1 Time criticality

Factors to consider

Depending on the timeline for a machine learning project, it may restrict the types of algorithms under consideration. A short exploratory project or a 'proof of concept' will not have the time to set-up a deep learning neural network. Instead, a tree-based algorithm could be the way to go. On the other hand, if a project has a specialized long-term use case (e.g. cancer recognition), it is much more reasonable to invest time and resources into a sophisticated algorithm which is likely to perform better, once it's properly parametrized and trained.

Evaluation implications

A time critical project is more likely to generate cost overruns or delays due to unforeseen circumstances. Hence, the business case impact would be rated negative, while a generous timeline warrants a positive rating. As outlined above, the rating can be mitigated depending on the chosen algorithm. A short timeline may be perfectly feasible with a simple algorithm and could therefore be rated with a neutral business case impact.

4.6.2 Business acceptance & change management

Reasons for change management

It is one thing to technically implement a machine learning algorithm and having it perform as per specifications. However, the best model estimates & predictions are useless if they are not used by the target stakeholders. For example, if a client advisor does not trust the investment recommendations of a machine learning algorithm, or feels he / she will not be able to explain them to the client, chances are high that these recommendations are ignored. Hence, a project should not declare victory after the successful technical implementation, but rather ensure through proper change management, that the target stakeholders will use the output produced by the model. This often requires a certain cultural change, to accept and trust output produced by a 'machine'.

Measures to improve acceptance

Stakeholder acceptance can often be increased through explainability of the model outputs. Another measure to increase business acceptance is if the system produces choice. For example, one project intended to output discussion topics for the client advisor with his / her clients. Instead of 'the one next best' topic, it suggests three options. This was helpful because not all client advisors feel confident in every topic, or they might already know that the client will not be willing to go ahead with certain suggestions. With reference to section "4.5.4 Acceptable error margins", it also builds trust to inform the business users beforehand how reliable the model is (error margins) and which error types were deliberately not minimized, including the reasons. Sometimes it can also take more emotional things like 'beautification' steps to ensure business adoptions, such as a nice summary or factsheet with good looking graphics. And finally, it can never hurt to make positive impact perceptible to the end user. For example, if a model improves the conversion rate in a call center from 2.2% to 2.7%, it might very well be that a

call center agent does not notice this directly in his / her daily business. But if the conversion rate is linked to the compensation, it will certainly be noticed when the next salary is paid out.

Evaluation implications With regards to business case impact, it could be argued that in order to ensure the benefits identified in section 4.5.10 “Expected benefits”, some change management budget is required. This means somewhat counterintuitively a negative rating for this criterion (depending on the cost impact) because all the benefits gained from it are already addressed in another criterion.

4.6.3 Focus on smart data over big data

Domain knowledge Something that has also been pointed out when conducting the expert interviews, was that domain experts should work together with the technical experts.

In theory, ‘big data’ has the advantage of its sheer volume and therefore a huge number of possibilities to find patterns in the data. However, in practice it turns out that bringing in domain knowledge to pre-define even a vague goal on what to find can help in two ways: It often makes the technical implementation easier, as certain constellations can be excluded or enforced. But it also often helps with business acceptance, because the result is more relevant and more explainable.

Define a smart question In many use cases, the problem statement (i.e. the question to answer through data analysis), can influence the result. For example, an open question like “Which product is the most profitable?” depends on (not exhaustive):

- The granularity, i.e. which product group level are we comparing? Derivatives with loans and advisory mandates or FX options with equity reverse barrier convertibles and UK retail mortgages?
- The time horizon (e.g. one month, two years, ten years)
- The KPI to measure profitability (e.g. gross margin, net profit margin, ROA, ROE)
- ... Or maybe we wanted to know “Which product generates the most profit?” because the top 5 most profitable products are all small niche products?

Comparing this to the question “Which product generated a higher ROA in 2018, plain vanilla FX swaps or plain vanilla FX options?” shows that the kind of question that is asked will influence the precision of the answer.

Know the data It also helps to have a rough understanding of the data from a business view. In the above example, when looking for the most profitable product: Are all products even reflected in this data set? And are they grouped in similar comparable hierarchies? A technical expert might not pick up on the fact that the data included loans and structured products, but advisory mandates are missing.

Or when conducting a sentiment analysis of your brand through analyzing the news, it helps to be aware that you’re getting insight into the published opinion, not necessarily the public opinion.

Considerations for unsupervised learning Certain types of unsupervised learning algorithms require as input the number of desired factors or clusters. Choosing 10, 15 or 20 explaining factors can quite substantially alter the result through the way the algorithm has to aggregate the data.

While there are mathematical tools to help, it can also be worthwhile to experiment if time and resources allow. And it certainly helps to have a domain expert involved at this step.

Evaluation implications While ‘smart data’ considerations do constitute an extra work step, it is often a rather small effort compared to its potential benefits. Therefore, if your project is taking this step, a small positive business impact can be expected.

If your project will not specifically consider domain knowledge but business is included through requirements engineering, a neutral rating can be applied.
For a purely IT driven project, a neutral to negative business case impact may be the consequence.

4.6.4 Design actionable output

Fulfilling KPIs vs generating benefits Once an algorithm has been trained and the developed model is now productive, it will start to produce outputs. Often, this is the long-awaited moment of truth to validate the business case. However, even if a KPI is fulfilled, it does not mean value is automatically generated.

Examples One example brought up was a customer churn prediction model established at a major bank. The model was successfully implemented and started to do its job, i.e. predicting potential customer attrition. However, even while those predictions were within the range of the pre-defined metrics, it was not quite clear what to do with the clients where ‘probability of attrition’ was high. How to approach them and with what message, was not easy to figure out. There have been some successful cases of avoiding attrition, but the overall picture was patchy.
Another example where benefits realization aligned with the metrics, is of a model which was supposed to increase net new money (NNM) per year. It generated proposals which clients should be invited to a meeting. While it suggested less client meetings than the client advisors did in earlier years, those meetings were much more successful and more than doubled NNM.

Evaluation implications Therefore, when defining the metrics and benefits, make sure they are aligned (i.e. predicting churn does not necessarily prevent churn). Furthermore, it should be clear how the benefits will be realized – or at the very least having a project stream dedicated to identifying ways for realizing the benefits.
With regards to rating on the assessment form, it could be argued that thinking about this can save cost later on and therefore yield a positive rating. It can also be argued that benefits outlined in section 4.5.10 “Expected benefits” are the baseline and ensuring their realization is a neutral outcome with respect to this criterion.

4.6.5 Difficulty level of output prediction

Output classes The type of output class influences the algorithms which can be used but can also limit the possible accuracy. For example, machine learning algorithms are good at classifying pictures or producing cross-selling recommendations. Predicting a quantitative variable (e.g. sales prediction) is the hardest task, with the most uncertainty attached to it.

Evaluation implications It helps to be aware of the difficulty level for your output class and the impact this will have on the error margins (or confidence intervals). It can be worth discussing alternative output classes with your business stakeholders, in conjunction with section “4.5.4 Acceptable error margins”.
Depending on how easy or hard the output prediction is, a slightly positive to slightly negative business case impact can be assumed.

4.6.6 Identify and evaluate alternative options

Business case definitions There is no singular definition of the term business case. While most of what consists of a business case is ‘common sense’, I have seen definitions differ in a key element.

According to PRINCE2, a business case is *“the business justification for undertaking a project, based on the estimated costs (of development, implementation and incremental ongoing operations and maintenance costs) against the anticipated benefits to be gained and offset by any anticipated dis-benefits and associated risks.”*

According to APM, a business case *“provides justification for undertaking a project, programme, or portfolio. It evaluates the benefit, cost and risk of alternative options and provides a rationale for the preferred solution.”*

Identify alternative options The definition by APM stresses to evaluate alternative options. This is also something I can identify with. With a vision or desired business outcome in mind, we should be open to different ways of implementation. In the machine learning context, two typical alternative considerations come to mind:

- Execute it manually / with human brainpower
- Use a simpler / faster / less fancy model (e.g. tree-based model instead of a deep neural network)

Evaluation implications If there are other viable options, they should be considered in your business case. A true machine learning use case will prevail as the best option. However, the example mentioned in the motivation section, that common words like ‘information’ will trigger a manual review, should have been replaced by the alternative options to directly review all documents manually.

With regards to business case impact, the options should be rank ordered according to their ROI and potentially some other metrics like quality, timeline, or risk. When comparing for example three options, the best option would get a positive rating for this criterion, the second best a neutral rating and the worst a negative rating.

5 Applying the evaluation guidelines to an exemplary case

5.1 Case introduction

Fraud and error prevention As part of an audit issue, a major bank was overhauling their fraud detection processes. This included major updates in their application landscape, usability improvements and enhancing automation. They also planned to update the current ruleset which was used to alert potentially fraudulent transactions (payments, trades, credit card) or entry mistakes ('fat-finger errors', e.g. the decimal point slipped to the right (or left)).

One goal was to decrease false alerts because it was planned to start automatically blocking potentially fraudulent transactions. De-blocking would only happen manually once the alert was clarified. Furthermore, the second goal was to improve identification patterns for potentially fraudulent transactions (so far it was only based on the transaction amount).

The project was well funded, and the bank had their own machine learning specialists in a 'center of excellence' (CoE) set-up.

Apply evaluation guidelines I am now walking through the business case assessment forms, filling them in for this exemplary case in order to give the reader an idea of how it could be applied to your own project.

5.2 Table 1: Determine machine learning category

Supervised learning	Unsupervised learning	Reinforcement learning
No	Yes	No

Unsupervised After a first requirements workshop with the project team and business stakeholders, the IT team proposed five potential solutions. In a second workshop, three were selected to be assessed in depth. One was a statistical approach and the other two were based on unsupervised machine learning. Based on the in-depth assessment, the business experts decided to implement one of the unsupervised learning algorithms.

5.3 Table 2: Potential showstoppers and promoters

Evaluation criterion	Weight	Machine learning business case viability						
		-3	-2	-1	0	+1	+2	+3
The Gartner analytic continuum	1					X		
Insufficient data volume	1							X
Insufficient data quality or availability	1							X
Unstable environment	2			X				
Regulatory barriers	1					X		
Lack of explanatory factors	0				X			
Strategic investment case	1				X			
Total (weighted sum)	n/a	+ 6						

Weights As a baseline, a weight of 1 was chosen for each criterion. The weights will then be changed accordingly if a criterion is deemed more important or less relevant than the 'standard'.

5.3.1 Assessing the evaluation criteria

The Gartner analytic continuum This is a use case in the 'predictive' category of the Gartner analytic continuum. Besides unsupervised machine learning, a statistical approach or sophisticated ruleset could also be applied. Hence, a rating of +1 is applied.

Insufficient data volume The data volume is not a problem for this project. First, the project is looking at internally generated data, i.e. there is no need to collect or buy additional data. Second, their financial transaction data is available in quite large and more than enough numbers. Hence, a rating of +3 is applied.

Insufficient data quality or availability The data quality of the feeds is good, since the financial transactions need to fulfill certain quality requirements in order for the transaction itself to be processed. Furthermore, outliers are exactly what we want the model to highlight, hence they shouldn't be removed.
The data availability is high, since all the data is available in electronic and structured form. Hence, a rating of +3 is applied.
Note: The data availability here is not to be confused with data sourcing, which certainly poses a bigger challenge for this project.

Unstable environment	<p>Financial transaction patterns can change over time in accordance with the client behavior. A client might change modes of payment, or shift payments from a competitor to this bank, make a big-ticket purchase or have a major event in his / her life which can all influence transaction patterns.</p> <p>It is important that a model can deal with those changing patterns to avoid over- or under-alerting potentially fraudulent transactions. Otherwise, it would require periodical re-baselining. Hence, this is deemed an important criterion and the weight is doubled.</p> <p>While the environment is rather unstable, there are unsupervised learning models which should be able to deal with this kind of variability. Hence, a rating of -1 is applied.</p>
Regulatory barriers	<p>This project does not have any (unusual) regulatory barriers. Quite the opposite, it is going to address an audit issue. Despite the audit point, risk and compliance assessed that the fraud detection processes do not pose a regulatory issue.</p> <p>Given that this is an audit point and not a regulatory issue, and updating the current fraud detection ruleset itself is only part of the overall project, a rating of +1 is applied.</p>
Lack of explanatory factors	<p>This criterion is not applicable since the project decided to use an unsupervised algorithm. Hence the weight and the rating were set to zero.</p>
Strategic investment case	<p>This bank already has a machine learning CoE. Hence, this is not an investment case. Since the overall project business case is favorable (see assessment tables below), this criterion was rated as 0.</p>

5.3.2 Assessing the table results

Evaluation Taking the weighted sums of the ratings, we get a total of +6, indicating a viable machine learning business case.

5.4 Table 3: Cost and benefit drivers

Evaluation criterion	Weight	Impact on costs					Expected benefits					
		-3	-2	-1	0	Estimate	0	+1	+2	+3	Estimate	
Data	0				X	n/a	X				n/a	
Technical complexity	1			X		n/a	X				n/a	
Project complexity	1		X			n/a	X				n/a	
Acceptable error margins	2	X				n/a			X		n/a	
Maturity of employed capabilities	1				X	n/a	X				n/a	
Data sourcing & response time	2	X				n/a			X		n/a	
Scalability	1			X		n/a	X				n/a	
Proof of reliability	1			X		n/a		X			n/a	
Explainability and interpretability	1				X	n/a		X			n/a	
Expected benefits	3				X	n/a				X	n/a	
Total (weighted sum)	n/a	- 17					n/a	+ 19				

Weights and estimates In order to demonstrate how these guidelines can be used in a qualitative manner, I am going to use the weights approach instead of the estimates. This is helpful in this context as it can be hard to estimate monetary benefits for closing an audit issue. As with the previous tables, weights are by default set to 1 and adjusted up or down for specific criteria.

5.4.1 Assessing the evaluation criteria

- Data Since data is not a specific issue on this project, I chose to weight and rate it with zero. Besides data sourcing, which is addressed in other criteria, no significant data-related cost is expected.
- Technical complexity Given the chosen algorithm, an existing CoE, as well as pre-existing machine learning infrastructure, the technical complexity is not that high for this organization. Relative to the overall budget, the cost impact is rated -1.
- Project complexity Since the project involves multiple applications, stretching across all the bank's divisions, several functions and has audit relevant productive outcomes, this criterion is rated -2.

Machine Learning - Evaluation guidelines for financial services business cases



Acceptable error margins	<p>In fraud detection, the error margins are very important, hence this criterion is double-weighted. The acceptable error margins have been subject to a hot debate. On one hand, a low alert threshold could yield tons of false negatives, while a high alert threshold would yield false positives – which is deemed the worse error type. Furthermore, as seen in the environment criterion, even if the perfect threshold is found, it may change over time.</p> <p>The end solution has to be a flexible threshold, which minimizes (ideally avoiding) error type I. Since the end users are also keen on a low as possible error type II, a cost impact rating of -3 is applied. On the other hand, we also apply a benefits rating of +2 since a low error type I prevents fraud cases, which can be expensive and reputation damaging. Also, a lower error type II reduces unnecessary business efforts.</p>
Maturity of employed capabilities	<p>The maturity of the employed algorithm is high (Level 1) and therefore this criterion is rated as neutral.</p>
Data sourcing & response time	<p>Since receiving the source data on time, as well as generating alerts and potentially blocking a transaction before it is executed are important requirements, the weight of this criterion is doubled.</p> <p>This project has to deal with high-volume, high-frequency and real-time data from multiple source systems. While the data is real-time, the response time doesn't have to be minimized like e.g. algorithmic high-frequency trading. Still, due to many different source systems, the cost impact is rated with -3.</p> <p>Being able to apply the fraud detection not only to overnight batch transactions but also in real-time is of course a substantial benefit which was rated with +2.</p>
Scalability	<p>Scalability is required across all the bank's division. However, since this is already addressed now at the start of the project, the cost impact can be curbed and is therefore rated with -1.</p>
Proof of reliability	<p>Proofing the reliability of the final model will be required to a certain degree to close the audit issue. However, we expect that the additional test efforts are not too strenuous, hence a cost impact of -1. Since this will allow to close the audit gap, the benefits rating is also deemed to be +1.</p>
Explainability and interpretability	<p>Explainability is not a hard requirement but of course helps to argue for closing the audit issue. Also, in case of false alerts produced, or alerts missed, an explanation will be expected. There is a small benefit rated with +1. Due to the chosen model, there is not really an additional cost, hence cost impact is rated neutral.</p>
Expected benefits	<p>The main benefit is of course closing the audit issue. While the fraud detection algorithm is not the only element which will allow to close the audit issue, it is nevertheless a required element. So, this is assigned a +8 rating.</p> <p>Another benefit is the prevention of formerly undetected fraud cases, which of course saves money and reputation. However, this is already considered in the error margin criterion. The project further expects to realize cost-savings through simplification of the cross-divisional application landscape & processes which is rated with +1.</p> <p>The overall score for +9 for expected benefits is represented in the table as +3 rating with triple weight but could have equally been represented as +1 rating with 9 times weight.</p>

5.4.2 Assessing the table results

Evaluation Taking the weighted sums of the cost and benefit ratings, we get a total of +2. This would normally indicate a business case with a slightly positive ROI, although ROI is not the right metric when talking about audit issues or regulatory compliance projects.

5.5 Table 4: Additional success factors

Evaluation criterion	Weight	Impact on business case						
		-3	-2	-1	0	+1	+2	+3
Time criticality	1				X			
Business acceptance & change	2			X				
Focus on smart data over big data	1					X		
Design actionable output	1				X			
Difficulty level of output prediction	1					X		
Identify and evaluate alternative options	1							X
Total (weighted sum)	n/a	+ 3						

Weights As a baseline, a weight of 1 was chosen for each criterion. The weights will then be changed accordingly if a criterion is deemed more important or less relevant than the 'standard'.

5.5.1 Assessing the evaluation criteria

Time criticality The project has an overall timeline which is ambitious but realistic. The chosen algorithm does not impact the timeline, and vice versa, the timeline does not restrict the algorithm choice. Hence, this criterion is rated neutral.

Business acceptance & change mgmt. While this project addresses an audit gap and should in fact reduce the overall financial risk for the company, it doesn't mean that business users happily agree to the project outcomes. While many changes will be introduced on the technology side through a central program, it still requires business user inputs and will ultimately change some of their processes. Due to the partial necessity of business cooperation, this criterion is double weighted. To realize the benefit of closing the audit issue, reasonable participation and adoption of the business users is required. This warrants to set aside some time budget to remind the business of well-known near-misses or actual fraud cases, and to establish a robust escalation path. However, the overall cost for this is rather low, hence rated with a -1.

Focus on smart data over big data	At this stage, the project team liaised with subject matter experts (SMEs) from the business side to define high-level requirements. Then, the IT team was briefed. Based on this, the IT team came up with a choice of algorithms, highlighting their advantages and disadvantages. The final choice was made by the business. So, domain experts were definitely involved and given the preferences of the IT team, we believe that this kind of 'smart data' approach helped to avoid going down some wrong paths. Hence a rating of +1 is applied here.
Design actionable output	One of the initial goals of this project is that transactions identified as 'potentially fraudulent' could be automatically blocked by the relevant system. It would then require a manual user review to either unblock or further investigate the case. This is a user action, enabled by the model output. Since this is already an expected benefit, the rating here is neutral.
Difficulty level of output prediction	The output to be predicted here is binary and therefore can be considered in the 'easy' level. Therefore, a slightly positive rating of +1 is applied.
Identify and evaluate alternative options	As mentioned before, after the project & business provided base requirements, IT came up with five options. Furthermore, there was also a baseline in form of the current state. I believe this constitutes a good exploration of the option space and therefore apply a rating of +3.

5.5.2 Assessing the table results

Evaluation Taking the weighted sums of the ratings, we get a total of +3, indicating that business case factors not directly related to costs and benefits also support the business case.

5.6 Overall assessment

Overall evaluation With every table yielding a positive result, the overall assessment is rather straightforward. This is a business case that, based on this evaluation, should be pursued.

6 Appendix

6.1 Common financial services use cases

Background	<p>In case a machine learning proposal / project suddenly comes your way, it might help to have at least a rough idea if it's something 'standard' or not. I have therefore put together a list of common machine learning use cases in financial services.</p> <p>In case your project proposal is on this list, you may not have to challenge the machine learning use case but rather ensure a lean business case. In case it's not on the list, it may well be worth asking if the use case has been applied in other companies before, and with what results.</p>
Fraud detection	<ul style="list-style-type: none">▪ Fraud detection & prevention (e.g. credit card transactions, payments, trades) and AML: Machine learning can e.g. identify standard patterns and highlight deviations. Besides monetary amounts, it can also look at where money was spent, consider failed log-in attempts etc.
Investment predictions	<ul style="list-style-type: none">▪ Investment predictions: Machine learning can learn from historical asset prices but also consider regulatory filings, consumer data, news releases and other non-standard data. There are even models which analyze the number of cars in parking lots of companies to infer how a company is doing. Or weather-forecast models being used for insurance-linked securities trading.▪ Trading: High-volume real-time data can be used for signal generation in high-frequency trading.
Bulk-searching files	<ul style="list-style-type: none">▪ Machine learning can search hundreds of thousands of documents, scans, pictures, text or audio files for certain information, like a specific legal clause.
Virtual assistants	<ul style="list-style-type: none">▪ Chatbots and robo-advisors can be enabled to learn instead of following pre-defined scripts.
Risk evaluation	<ul style="list-style-type: none">▪ Credit scoring, credit decision & insurance underwriting: Machine learning can support and streamline these processes especially with respect to risk evaluation.
Improve client experience	<ul style="list-style-type: none">▪ Client onboarding: Machine learning can help the client by pre-filling already existing information into the required forms and templates.▪ Personalization: Machine learning can create personalized investment recommendations.▪ Client retention: Machine learning can detect possible customer churn – however preventing it remains the duty of the client advisor.
Marketing	<ul style="list-style-type: none">▪ Sentiment analysis of a brand (company or product): Machine learning can do this automated by e.g. analyzing newspaper articles or social media.▪ Customer targeting: Machine learning may identify better ways to segment your prospective clients. For existing clients, it may identify cross- or up-selling potential.▪ Evaluate client feedback & complaints: Machine learning can do this automated and e.g. cluster the most common responses.

- Miscellaneous
- Machine learning can be used to optimize ATM allocation and cash replenishment strategies.
 - Deep learning could be used to generate analyst reports for securities traders.
 - Insurances can apply deep learning algorithms to assess potential damage claims in disaster areas.

6.2 Sources

- Brunton, Steve (2019). *Intro to Data Science*. YouTube playlist: <https://www.youtube.com/playlist?list=PLMrJAKhleNNQV7wi9r7Kut8liLFMWQOXn>
- D'Amour, A., Heller, K., Moldovan, D. et al (2020). *Underspecification in Machine Learning*. Journal article.
- Fraunhofer-Gesellschaft (2018). *Maschinelles Lernen: Eine Analyse zu Kompetenzen, Forschung und Anwendung*. Study paper.
- Gartner (2022). *Market Guide for AIOps Platforms*. Online article.
- Gartner (2010). *Big Data and Analytics Overview*. Presentation.
- Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep Learning*. MIT Press book.
- Ng, Andrew (2018). *Machine Learning Yearning*. Online book.
- Nguyen, Chi Nhan & Zeigermann, Oliver (2021). *Machine Learning kurz & gut*. 2. Auflage. Published book.
- Nielsen, M. (2019). *Neural Networks and Deep Learning*. Online book.
- Parisi, G., Kemker, R., Part, J. et al. (2019). *Continual lifelong learning with neural networks: A review*. Published article (Neural Networks, Vol. 113, pp. 54-71).
- Stanford Online. *CS229 Machine Learning Full Course taught by Andrew Ng | Autumn 2018*. YouTube: <https://www.youtube.com/playlist?list=PLoROMvodv4rMiGQp3WXShTMGgzqpfVfbU>

6.3 Abbreviations

AI	Artificial Intelligence
AML	Anti-money laundering
APM	Association for Project Management
ATM	Automated Teller Machine
BAU	Business as usual
Ctrl+F	The shortcut Ctrl+F opens the search (respective 'find') function in most Microsoft applications
CoE	Center of Excellence
FX	Forex or Foreign Exchange, refers to buying or selling international currencies
IT	Information Technology
kCHF	thousand Swiss Francs
KPI	Key Performance Indicator

Mgmt.	Management
ML	Machine Learning
NLP	Natural Language Processing
ms	milliseconds
NNM	New New Money
OCR	Optical Character Recognition
OpEx	Operational Expenditure
Ops	Operations
OTC	Over-the-Counter, referring to securities which are not traded via a centralized exchange
PDF	Portable Document Format, a widely used file format
PRINCE2	PRojects IN Controlled Environments (second edition)
RFP	Request for Proposal
ROA	Return on Assets
ROE	Return on Equity
ROI	Return on Investment
SME	Subject Matter Expert
SOX	Sarbanes–Oxley Act
UK	United Kingdom