# DATABOOK : a standardised framework for dynamic documentation of algorithm design during Data Science projects

Anna Nesvijevskaia

# DATABOOK: a standardised framework for dynamic documentation of algorithm design during Data Science projects

Anna Nesvijevskaia[1]

## Abstract

This paper proposes a standard documentary framework called *Databook* for Data Science projects. This proposal is the result of five years of action-research on multiple projects in several sectors of activity in France and a confrontation of standard theoretical processes of Data Science, such as CRISP_DM, with the reality of the field. The minimalist and flexible structure of the Databook prototype, described and illustrated in this paper, has revealed its operationality on more than a hundred projects and has been recognised by various stakeholders as an excellent facilitator of Human Data Mediation, especially for multi-skilled projects. Beyond its proven benefits for project efficiency, this framework, conceived as a frontier object, can be applied more broadly to data project portfolio management and data value, governance and quality. By surpassing the computational aspect of the models, the Databook is an answer to the issues of interpretability and auditability of algorithms.

## Keywords

Data Science, Artificial Intelligence, Documentation, Reproducibility, Algorithm Transparency, Project Process, FAIR, Human Data Mediation

## Introduction

The proliferation of Data Science projects has accelerated knowledge discovery and generated new algorithms for commercial use. It has also produced massive amounts of exploratory data and metadata. Yet exploration is still poorly equipped to understand the processed data in terms of meaning, utility and value. On one hand, recent Data Science platforms tend to structure only the technical aspects of the data engineering pipeline, data linkage and algorithmic libraries. This technicity erects a barrier to the understanding of the data by all project stakeholders, especially in complex multi-skilled projects. On the other hand, traditional Master Data Management tools, which handle this type of metadata on the key records of an organisation, are generally incomplete and unable to absorb all the data created during Data Science projects, including the final algorithmic model. The lack of standards for the capitalisation of this data leads to difficulties in replicating results, a lack of transparency and efficiency of the arbitrations made during these very dynamic projects, and a loss of resources during the data understanding and qualification phases of subsequent Data Science projects (portfolio management). These limitations are critical in the context of increasing European regulation and growing acculturation of business decision-makers to algorithms. Both trends require a shared and facilitated data understanding that goes beyond technical and mathematical measures.

This paper proposes a standard documentation framework, called *Databook*. It has been conceived for Data Science projects in which algorithms are designed. The emergence of the Databook is guided by (1) the theoretical and practical limitations of standard Data Science processes. This gap has been (2) compensated in the field by a Databook prototype with a unique structure. The prototype was (3) tested and confirmed as efficient for several purposes and stakeholders: this paper proposes its evolution to the first standard algorithm design documentation framework.

## 1. Standard process in Data Science projects: from theory to field reality

To begin with, we will consider the theoretical processes of Data Science projects and their results, the main limits of these processes identified in the field, such as the lack of documentation, and the first attempts to fill the documentation gaps in practice.

### 1.1 Overview of the standard process in Data Science projects

Data Science projects aim to build an algorithmic model for a specific practical purpose (a *usage*). The model consists of an input data, a finite sequence of well-defined operations, and an output en terms of analytical result. The choice of a model depends on the problem to be solved, such as a phenomenon prediction or its correlation to root causes. The solution is usually an articulation of several algorithms chosen among thousands of possibilities. The algorithms are applied to data selected for the project from an expanding number of available sources: the data are then assumed to contain in past observations an insightful signal that is key to the problem, and the algorithm is, therefore, a means of revealing this signal. The uncertainty of these projects is highly substantial because the presence and usefulness of the signal in the data must be explored, and sometimes the emergence of a signal predates to formulation of the need. As recent technological advances have had an impact on the entire data chain value (Bertino *et al.* 2011; Miller & Mork 2013), the cost of these exploration projects has reduced and opened up new horizons for possible usages in all sectors (Manyika *et al.* 2011; Mayer-Schönberger & Cukier 2013). The business needs covered by these projects are currently very diverse, most of them being assimilated to knowledge generation or decision-making acceleration. In both cases, the sense and value creation by the algorithmic model depends on a broader usage device (Brynjolfsson *et al.* 2011; Provost & Fawcett 2013) that includes a purpose, a context, a decision-making process, a user community, an interface or a workflow and many other elements that are impossible to standardise.

Despite this variety in terms of usages, algorithms and exploitable data, Data Science projects are composed of a similar sequence of activities described in the widespread use of Data Mining for Knowledge Discovery in Databases (Fayyad *et al.* 1996; Piatetsky-Shapiro 1994). Commercial actors, researchers and companies leading these projects have attempted to standardise these activities: the most successful attempt is the Cross-Industry Standard Process for Data Mining, or CRISP_DM (Chapman 1999; Shearer 2000; Wirth & Hipp 2000) that resulted from a convergence of reference processes and their confrontation in the field by a mixed consortium funded by the European Union. This process model breaks down the project life cycle into six phases: business understanding, data understanding, data preparation, modelling, evaluation and deployment. It captures the complexity of data exploration by identifying the main iterations between these phases and remains neutral in terms of usage, tools and data. Since the suspension of the consortium, several proposals to improve the standard process have remained pending: to expand the number of use cases, to map and describe the activities and their results in more detail or to link the process to different project management methods.

As the most stable and widely used process in Data Science (Camiciotti & Racca 2015; Provost & Fawcett 2013), CRISP_DM was defined as a reference in the course of an action-research which was conducted on seven different projects from 2014 to 2017 (Nesvijevskaia 2019). The objective of this thorough qualitative multiple cases study was to understand why Big Data, as a myth-bearing socio-technical phenomenon (Boyd & Crawford 2012) reflected in companies by the implementation of the

first Data Science projects, did not generate the expected value. The relevance of the CRISP_DM framework was confirmed, as were its expected limitations. The iterative nature of the main tasks was revealed as a regular and beneficial overlap between the six phases. The framework also found to be too focused on the algorithmic model, with a risk of uprooting the project results from the practitioner's activity (Nesvijevskaia 2017). It explains potential project failure in terms of result exploitation. Indeed, the process delays the anticipation of the usage, data inclusion/exclusion criteria but also the co-construction of the results restitution. This delay creates a risk of inadequate expectations, project costs drifts for production launch, errors in analytical strategy, but also a lack of capitalisation throughout the project. This confrontation between the reference standard process and the field leads to the building of a global Data Project device called Brizo_DS[2] which includes an adjusted CRISP_DM model. The reference outputs of each phase of the adjusted model are mapped to the process critical path and the process documentation (Figure 1).

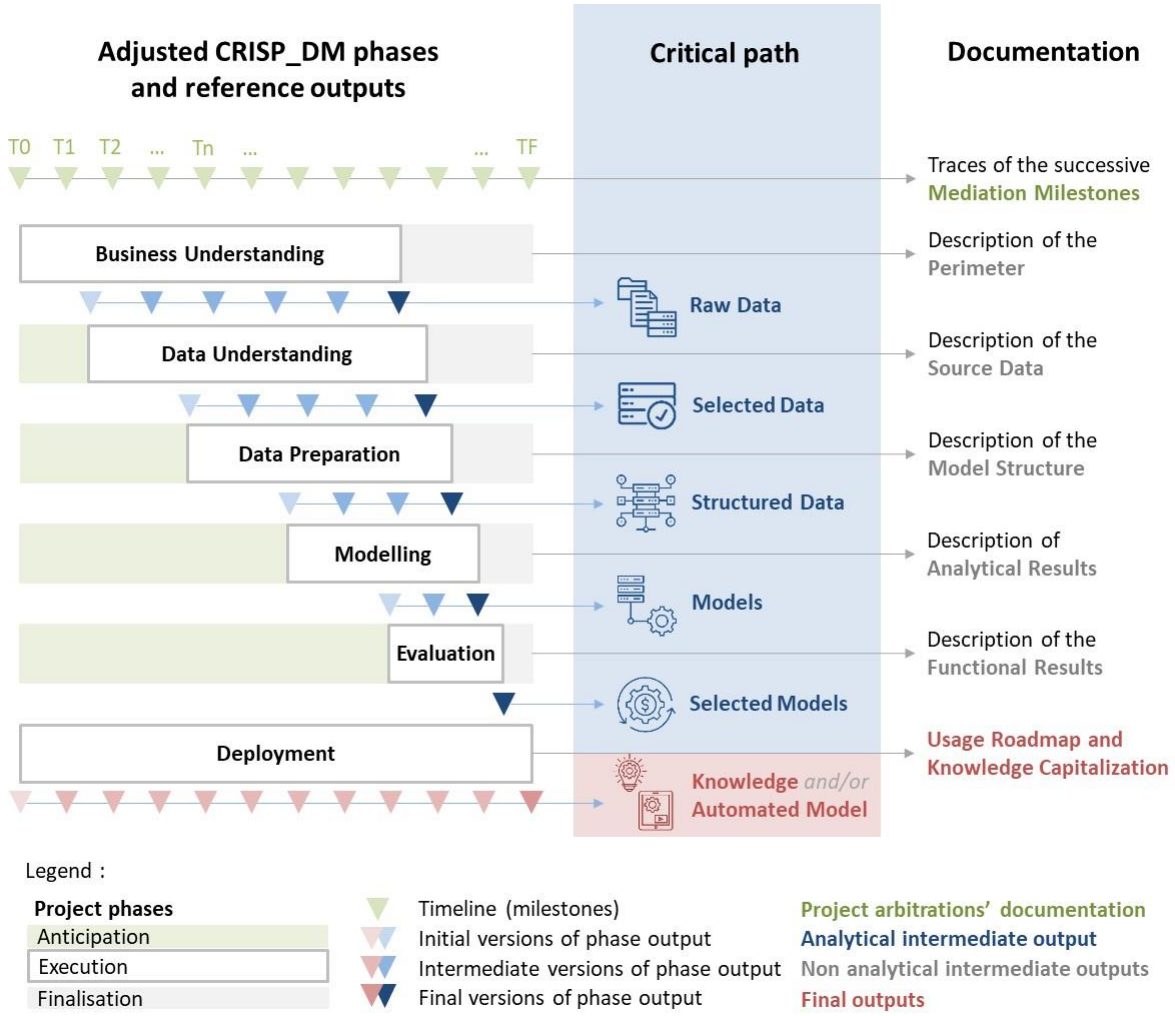



*Figure 1 - Output mapping of the adjusted CRISP_DM*

The mapping of the reference outputs in the figure above is based on the following principles. The *critical path* is composed of intermediate analytical outputs: *raw data* lead to *selected data* which are *structured* to feed algorithmic *models*, and the best models are *selected* to generate *knowledge* or decision-making. Some or all of the critical path outputs may be *automated*. These computational

objects are specific to Data Science projects as components of the finally used algorithm. They are materialized by the code of the algorithm and are accessible for the coding project team members. All intermediate outputs can be versioned: the first version allows the launch of the next phase of the project; the intermediate versioning explains the overlap between phases and the progressive optimisations of the algorithm during the project; and the final version corresponds to a component of the algorithm used for exploitation.

By isolation of this critical path, the *documentation* is composed of all the other outputs that can be shared between stakeholders in a tangible or intangible format. They can be generated before the execution of a phase (anticipation), during the phase or once the phase is finished. The numerous and non-mandatory possible documentation outputs (see the most common ones in Figure 8 in Appendices) can be classified into three main categories:

- *Critical Path analytical outputs documentation*: all the outputs in this category describe and qualify the intermediate analytical results and guide the convergence on the final optimum algorithm.
- *Usages*: these outputs describe the operational conditions for the final analytical result and knowledge activation (this activation can take place before the finalisation of the algorithm).
- *Mediation milestones*: all outputs in this category trace the arbitrations realized during the project and guide the project management.

The three categories above remain interdependent: the progressive design of analytical outputs feeds the project management; the project management makes decisions considering the value generation through usages; the usages emerge from analytical exploration and impose constraints and priorities. Once the outputs of each phase of the process are classified, their nature and production methods can be analysed by comparing practices and reviewing the state-of-the-art practices.

### 1.2 Main limits of the documentation outputs

The critical path has been broadly supported by the development of analytical tools (data engineering platforms, algorithmic libraries…) and the skills of freshly and progressively professionalised *Data Scientists* (Davenport & Patil 2012). It has therefore been increasingly productive. However, the actors implied in field projects can be more diverse and most of them are not supposed to open a Data Science application, read a line of code (even a well-commented code) or juggle technical and mathematical concepts.

In small-size projects, the most representative skills are usually divided into *business* skills and *data* skills, for instance when a Data Scientist works with a decision-maker: bridging the gap between these skills' carriers is still testified as insufficient and critical (Austin *et al.* 2021). In more complex projects observed in the field, more individuals are implied. Data skills can be carried for instance by machine learners, data engineers, data stewards or data analysts. Business skills are devised into strategic, analytical and operational. The last type of skills is often carried by users' representatives (for instance product owners) or knowledge managers, but they are very dependent on the expected usage. Some skills are dual, such as Business Intelligence skills. The most complex project teams are composed of members with mixed skills and with different levels of maturity: they require a strong mediation through project management (Nesvijevskaia 2019). Besides skills complexity, team members can be

confronted with difficulties related to data complexity (numerous sources, numerous extractions including erroneous ones, lack of meaning sharing by all actors, complex treatments, progressive identification of bias...). In the field, this complexity and high uncertainty require successive arbitrations implying a diversity of stakeholders. To achieve efficient arbitrations, heterogeneous stakeholders urgently need a common intelligible framework and semantics of raw and processed data at each stage of the project. However, reference documentation outputs are little mentioned in research work on data process, lacking anthropocentric anchoring.

The principle of the first Databook, as a dynamic documentation device, was guided by this urgent need and by a broader interdisciplinary approach to data quality. It had to take into account both computational (Berti-Equille 2012; Wang 1998) and cognitive (Arruabarrena *et al.* 2019; Broudoux & Scopsi 2011; Cottin & Nesme 2017; Odeh & Chartron 2016) aspects of transforming data into useful information by reducing uncertainties (Mayère 1990) in a given economic context (Doucet 2010). These historical approaches generally apply to the Master Data Management (Loshin 2010), where data governance issues are more largely focused: its objective is to 'increase business performance (by adjusting the value of the data) and reduce the costs associated with the processing and management of master data' (Mariko 2016). The inspiration also came from digital knowledge media engineering, seeking to establish standard attributes of knowledge elements (Zacklad *et al.* 2007). It also implied to follow the processes of capitalisation, sharing, knowledge creation, learning, selection and evaluation of useful information (Ermine 2003). However, the amount of data explored, created and discarded during time-limited Data Science projects did not allow for a full comprehensive data quality process, and the issues of a single exploratory project were not as significant to the investment as the meticulous processing of master data. The iterative exploration process required a flexible and dynamic data quality and knowledge sharing device that was difficult to transpose from MDM. It also had to be more practical for the project and the knowledge management needs common in consultancy practices. In opposition to the theoretical limitations and field pressure, a first Databook was imagined and tested in real-life situations.

## 2. The Databook: prototype structure

The Databook prototype is a generic documentation output specific to Data Science projects. It describes all the algorithm components and the decisions that occurred during the algorithm design. Hence, its structure reproduces the critical path outputs documentation and encapsulates the dynamic links between the algorithm, its usages and its design process punctuated by the mediation milestones. This structure is materialised in a single, common and shareable Excel file: each spreadsheet of this file represents a *module*. All the prototype modules are listed in Figure 2 and classified following the documentation categories listed in section 2. These modules can be supplemented with less specific documentation objects in different formats, such as PowerPoint reports, Data visualisation interfaces or tools required by the usage or by the project management.

| Databook Prototype Modules | | |
|---|---|---|
| | 0 | Databook Guide |
| Mediation Milestones | A | Project Roadmap |
| | B | Method of data inclusion/exclusion |
| | C | Exploration report |
| Critical Path analytical outputs documentation | 1 | Perimeter |
| | 2 | Source Data |
| | 3 | Model Structure |
| | 4 | Analytical Results |
| | 5 | Functional results |
| Usages | 6a | Usage Roadmap |
| | 6b | Expérience return |

*Figure 2 - Databook prototype modules*

The following sections present in detail the different Databook modules and their flexible building mechanism which relies on a clear distinction between *core structure* and *metadata structure*.

## 2.1 Core structure of the Databook

The core structure is the skeleton of the algorithm, usages and mediation milestones documentation. It is introduced in a guide (Module 0), which is a manual presenting the ten following modules grouped into three categories detailed in the following sections.

### 2.1.1 Documentation of the analytical outputs of the project

The structure of Modules from 1 to 5 replicates the critical path of intermediate analytical outputs of each phase of a standard Data Science process, excluding the deployment phase. An intermediate analytical output is defined as a *data object* (for example, a table) composed of *elements* (for example, variables in the table). The breakdown of data objects into more detailed elements remains specific to each project, but it always results in a structured list of homogeneous items. This list is usually presented like a hierarchical directory. Each data object or element in this list can then be completed with attributes, or *metadata*. These attributes result from the addition of several descriptive *criteria* that will drive the choice to keep or abandon an element for the following phase. This decision is a qualification traced through a *status* of the data object resulting from its judgement based on different criteria. These metadata (criteria and statuses) are also usually organised thematically or/and hierarchically: the choice of this structure remains specific to each project and will be presented in section 2.2.

The matrix representation of the structured list of homogeneous items associated with metadata fits perfectly formats such as Excel. For instance, the Module 2 (source data) aims the qualification of all the data that must be explored: an illustration of this module completed in a real-world project is presented in Appendices in Figure 9. Another illustration can be found in Figure 10 for the Module 3 (model structure), with the qualification of all new generated variables and their selection in a context of multiple algorithm development. The other 3 Modules follow the same matrix structure.

### 2.1.2 Documentation of usages and knowledge

The last deployment phase is often restricted in the literature to the usages directly aimed by the project. However, the Databook framework includes the documentation of both direct usages (technical and operational aspects) and knowledge generated throughout the project. It splits knowledge into two types: knowledge that can be potentially transformed into a business lever (indirect usage) and knowledge that can be useful for further Data Science projects (data project experience).

*Direct usages* are immediately operational levers which have been decided upon for deployment. Each direct usage is associated with deployment actions that can be described in terms of purpose, *modus operandi*, associated version of the solution to deploy, expected benefits, deadlines, responsibilities, key indicators to monitor and so on. In the case of an automatized algorithm, actions include the pipeline automatization tasks, and sometimes the interface development specifications. *Indirect usages* are potential levers with remaining uncertainties to investigate after the project. They result from knowledge that still requires concrete actions to be transformed into levers. Both types of usages require a usage roadmap (Module 6a): it contains a structured list of actions, usually broken down into a list of tasks and associated with metadata. As with the previous modules, the metadata includes the descriptive criteria and the status of each action, corresponding the decision to activate it or not. The matrix representation of this roadmap is very appropriate as a basis to feed other formats, more commonly presented to deciders (for instance, a report, a monitoring interface or the last version of the application). An illustration of this Module is presented in Appendices in Figure 11.

*Data project experience* covers all the qualitative feedback in the form of knowledge capitalisation useful for further data projects (Module 6b). Usually this experience feedback is tacit, intangible or orally shared, but it can also be documented, especially when the knowledge must be shared with stakeholders outside the project. For example, if team members judged that cleaning up the data in a particular table was not necessary for the project but had an intuition that it would be of great value to other projects or existing usages, this intuition can be capitalised upon. Another example can be a good coding practice capitalisation, or a business concept explored and finally judged as not of interest. This knowledge can potentially save significant time in future. Faced with the variety of possible knowledge that can arise from the experience of a project, the Databook prototype stops at a proposal to incrementally draw up a *list of insights* by application domain without seeking to structure the qualification of these ideas. In each project context, these ideas can then be shared with the appropriate stakeholders in the most suitable format.

### 2.1.3 Documentation of the milestones of the project

The milestones are usually the tip of the iceberg for the project management and provide essential elements for *arbitrations* throughout the project. As project decision facilitators, these milestones are usually more convenient to present with storytelling components, including texts, graphs and other project management best-in-class practices. However, they also necessarily include elements that must be fed with structured data and metadata issued from the other modules presented above.

This category includes three modules graded A, B and C. The Module A is a project roadmap with project advancement statistics based on statuses of elements qualified at each phase. It is illustrated in Appendices (see Figure 12). Each time the Databook is versioned, the Module A represents a

photography of the version. The Module B is a data inclusion/exclusion methodology synthetizing the structure of descriptive criteria and their expected impact on the qualification statuses. It represents the project decision rationale traced through metadata and results from a more complex mechanism described in section **Error! Reference source not found.**. The Module C refers to exploration reports: i t is very specific to each project. If the project has only one exploration report, it can be directly integrated in this module. However, usually a project generates several reports and each report is realized in its own format such as a Data Visualization or a presentation. In this case, the Module C lists the different reports, their versions, associated decision milestones (for instance, the date of a project committee) and key elements and decisions. This Module forms then a bridge between the decision milestones and earlier Databook versions as well as with other possible project documents.

### 2.1.4 Core structure synthesis

Each of the ten modules remains adaptable to the complexity of different projects thanks to a flexible database composed of custom lists, data objects, elements and associated metadata. The complete Databook core structure is presented below in Figure 3, with an illustration of the most common documented elements observed in the field for each module.

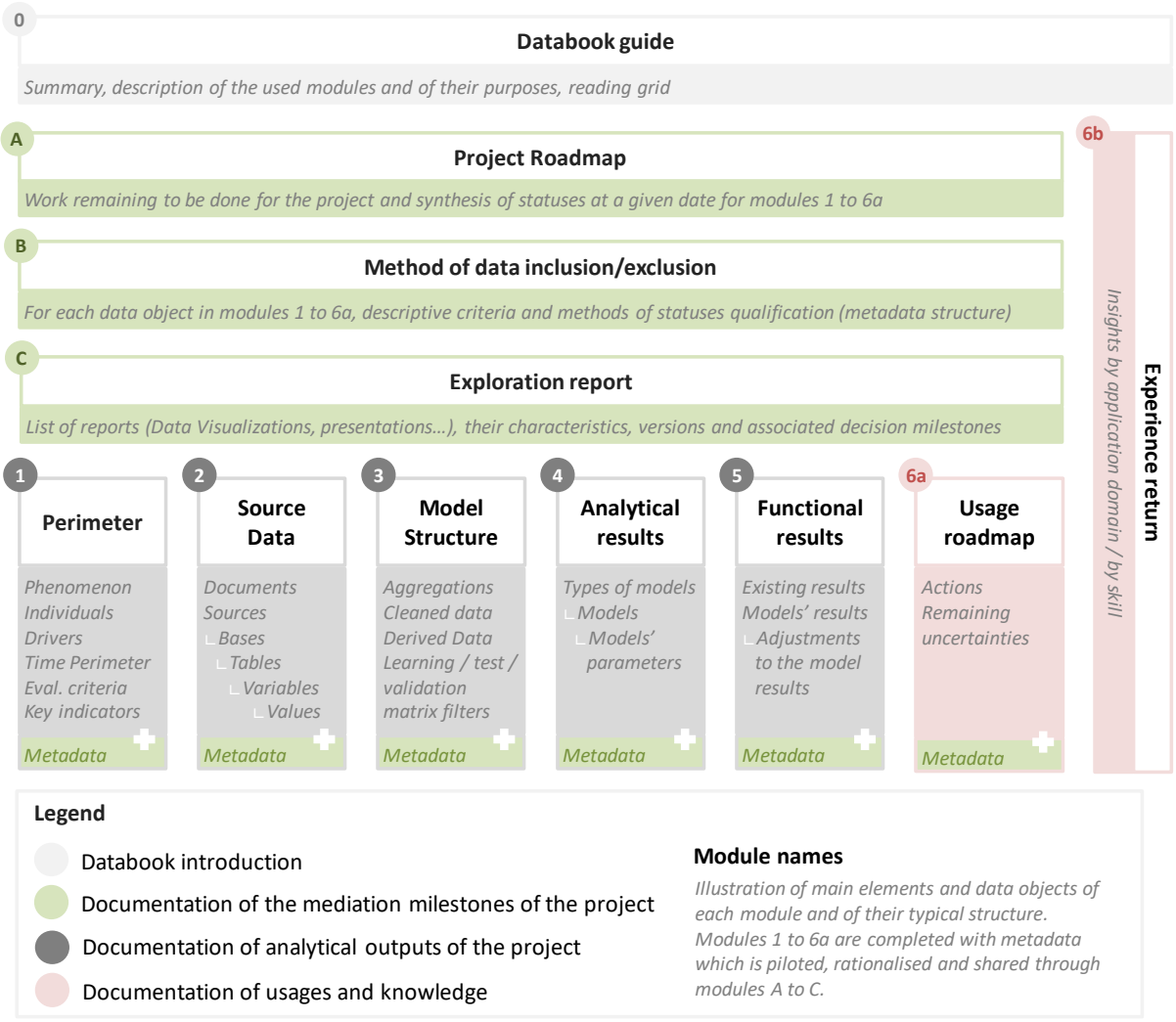



*Figure 3 - Databook core structure and its modules, illustrated with the most common elements and data objects*

## 2.2 Metadata structure

As presented above, each data object is described in terms of criteria and qualified with a status during the project. This documentation treatment is recorded through metadata. But, unlike the analytical treatments carried on the data throughout the critical path realization, the metadata treatment does not follow a sequential dynamic. Indeed, as seen in Figure 1, documentation can occur before the analytical work to anticipate it, during or after critical path completion. This dynamic is closely linked to the nature of the uncertainties reduced at each stage of the project: the skills required to anticipate the risks of each phase and to carry out the associated treatments are usually the same. The next sections reproduce the standard process and concentrate on the skills associated to each phase. It shows how those skills are implied in the documentation beyond their intervention in the critical path.

### 2.2.1 Business understanding

The data project must be anchored in a given business context in order to lead to a result that will be in line with the business strategy. This anchoring can be reflected in strategic criteria, business priorities and confidence in the relevance of each data object bearing real-life concepts in a business context. Business understanding metadata also includes all the regulatory constraints such as GDPR or discrimination rules that can lead to some data or model exclusion despite their statistical significance. The documentation in terms of business understanding requires the skills of *Strategic Management* and *Business Analysis*.

### 2.2.2 Data understanding

While source data understanding is part of the critical path, it does not stop here. All data objects produced during the project need semantics, units, names and other metadata that will make sense to all stakeholders in various communities. As observed in the field, this is one of the most used metadata in the Databook, facilitating co-construction and appropriation of all intermediate outputs. The level of detail can vary from a name of a data element to a definition or an in-depth explanation of its generation process. It must be aligned with the level of maturity of the stakeholders, their usual vocabulary, language, shortcuts and so on. If the Databook is to be shared outside the project team, these semantics can be completed with translations, comments and other facilitators. This semantic metadata can be structured and used as a dictionary or a repository, for instance in exploration reports or Data Visualisation, to complete technical structured nomenclatures of explored data with meaning. This documentation requires the skills of *Data Stewardship* and *Data Analysis*.

### 2.2.3 Data preparation

This documentation is predominantly technical and describes the data engineering issues in order to anticipate the exploration and the exploitation pipelines. The metadata for this purpose is composed of the function of data elements in the pipeline structure (keys, filters, place in query structures, filling controls, duplication controls…), of their formats and volumes with associated calculation time, and of other technical criteria. The technical documentation can lead to choose a given tool, language or even model or usage device and interface, and sometimes the exploration pipeline will differ from the exploitation pipeline. This documentation supports the anticipation of the usage deployment (controls, automations…) and requires the skills of *Data Engineering* and *Data Analysis*.

### 2.2.4 Modelling

Much more mathematical, this documentation links data objects to algorithmic models and evolves during the project as the model is anticipated, realised, calibrated, benchmarked and adapted to the usage. Modelling metadata corresponds to the analytical uncertainties and signal detection in the data. It includes metrics such as minimums, maximums, standard deviation of a given variable, modality distributions and a varied set of algorithm-specific metrics, such as parameters, hyperparameters or statistical evaluation criteria calculation. It also anticipates the re-learning mechanism and its monitoring if it is needed for the algorithm exploitation. This documentation requires the skills of *Machine Learning* and *Data Analysis*.

### 2.2.5 Evaluation

This documentation is paramount to understand the consistency of each data object in terms of its contribution to the key performance indicators of the expected usages. Naturally, this documentation concerns the translation of analytical results into business value: the statistical evaluation criteria (for example, the false positive rate for a churn prediction algorithm) must be translated into business evaluation criteria (for example, full-time equivalent or operational cost). This translation goes beyond semantics and includes the value calculation methods and its intelligible representation. Interpretability, rapidity, user appropriation or maintainability of a model can also be considered as performance indicators to judge a model for a given usage. This means that evaluation metadata can be both quantitative and qualitative. The best practice is to imagine the performance indicators first and then derive the statistical evaluation criteria from the business criteria, even for exploratory projects aimed at generating original knowledge. In this case, the initial performance indicators are specified as they are developed.

However, this type of documentation applies not only to analytical results but also to all the other data objects. For instance, interpretability can be judged for source data or newly generated variables in terms of consistency between the definition and the perceived meaning. Value calculation rules and orders of magnitude can also be anticipated since the beginning and progressively controlled. For instance, the control of the consistence of a customer database used in the project needs a comparison between the volume of the database lines and the number of customers usually measured by the company in existing reporting. These consistency controls are particularly useful when heterogeneous stakeholders from different parts of a company must make project decisions on a common objective basis. Consistency controls occur for intermediate and final data objects: they explain a significant number of iterations because they help to detect errors. They include not only obtained, but also expected metadata: for instance, a usage value can be judged through the delta versus an expected value. This documentation requires the skills of *Business Intelligence*, but also both *Data Analysis* and *Business Analysis*.

### 2.2.6 Deployment

This documentation corresponds to the usage anticipation throughout the project, whether they are defined from the start or gradually emerging. Indeed, this usage anticipation can lead to operational priorities or exclusions despite the business, technical or mathematical importance of certain data objects. For example, if the usage is a real-time decision-making, but a data source is collected only on a monthly basis, this data source may be eliminated from the project despite its strong predictive power. Another recurring example is the volume of explored data: big volume can be perfectly usable

for the exploration but inappropriate for usage exploitation. This qualification is also notably critical for exploitation of personal data. The usage anticipation is necessary to avoid the risk of producing interesting but unexploitable results.

Deployment documentation describes the operational constraints linked to the usage exploitation and strongly impacts the qualification of all the previous data objects. Indeed, their criteria must be adequately judged to determine the status. Operational criteria are very variable from a project to another. They depend on the stakeholders responsible for activating the project results and piloting the usage. Deployment requires the skills of *Product Ownership* for direct usages (this skill must be specified for each field of application) and/or *Knowledge Engineering* for indirect usages, usually completed with skills of *Business Analysis*.

### 2.2.7 Project management

As the analysis work progresses, data elements are treated at different phases of the critical path: these treatments are documented with associated metadata. This analytical work can be done phase by phase, but also iteratively or through phases' overlaps thanks to the versioning of intermediate analytic outputs, as described in Figure 1. A version then corresponds to a set of data elements that will evolve. For example, a Machine Learner can start working on a model with incomplete data, in order to test the first assumptions, without waiting for a complete dataset to be prepared by a Data Engineer, who is waiting for the last data extractions. In another situation, a business stakeholder may perceive a meaningful variable without knowing if this variable exists in a source. This variable will then be added to the Databook as being perceived as useful, but not directly confirmed as treatable. These advancement tactics cannot be followed if only completed analytical treatments are documented. However, they are perfectly followed when metadata are generated as soon as the treatment anticipation occurs. These advances are synthesised by a status of each data element in terms of qualification. The most common statuses are *to be started*, *in progress*, *included* or *excluded*, and can be adjusted for multiple uses in the same project (for example, a source can be *included* for algorithm A and *excluded* from algorithm B). The statuses are essential for sticking to the design dynamic and represent the pivot between phases: they vary with the versioning (Databook is then versioned in parallel) and can be supplemented with workload estimates or difficulties anticipated. This qualification thus requires the skills of *Project Management*, but also *Business Analysis* and *Data Analysis* to guarantee the understanding of each activity stakes and associated qualification methods.

### 2.2.8 Metadata structure synthesis

In the Databook, metadata can be generated independently form the analytical critical path as it aims both the anticipation and the reduction of different uncertainties of the project. Each phase of the project represents a type of uncertainty and involves a specific skill to reduce it. This skill must also be involved in the anticipation. This documentation mechanism generates rich criteria guiding the projects decisions traced through the statuses. Therefore, this framework is a boundary object than can be successfully adapted to different points of view for all the skills' carriers engaged in the project and robust enough to maintain identity between them (Star & Griesemer 1989). Thanks to this structure, skills' carriers participate using their documentation capacity in the main arbitrations leading to the construction of project results.

Given the diversity of possible metadata, Figure 4 reports a non-exhaustive list of the most commonly used criteria. This illustration represents the Databook modular core structure (vertical) enriched with typical metadata classified by type of uncertainty reduced for each phase of the project (horizontal). For each concrete project, this structure can be documented in the Module C, i.e. data inclusion/exclusion method. All metadata are then listed with associated modules and skills (usually, skills are represented by an individual skills' carrier) and for each criteria the status establishment method is mentioned. For instance, if the criteria 'personal data' is flagged as 'yes' for a variable, it will have to be flagged as 'excluded' in the status. This flexible mechanism linking the core structure and the metadata structure of the Databook results from an iterative confrontation between theorical and practical requirements.

| | | Business understanding | Data understanding | Data Preparation | Modelling | Evaluation | Deployment | Project management |
|---|---|---|---|---|---|---|---|---|
| 1 | Perimeter | Pertinence, business priority, consistence as business concepts | Sense of each key metric | Temporal perimeter, filters, keys… | Utility in the model construction | Translation of a key metric into data (calculation rules) | Importance in decision making, activability, historization possibilities… | Difficulties to reach exploitable data, status |
| 2 | Source Data | Confidence in the meaning, regulatory constraints… | Sense of each source data element, units, cognitive bias… | Accessibility, integrity, completeness… | Precision, bias, statistical metrics, nature of values… | Consistency between the perceived sense and the source data | Exploitability of the volumes, formats, techniques and resources needed for exploitation… | Difficulties to prepare the source data, status |
| 3 | Model Structure | Confidence in the construction process, regulatory constraints… | Sense of created data, units, nomenclatures, cognitive bias… | Description of treatment (keys, filtering, gross use, derivation, rules, matrix cutting…) | Models impacted, weight of signal for each data element… | Consistency between the perceived sense and the new data | Exploitability of created data, resources needed for deployment and maintenance… | Difficulties of the modeling, status |
| 4 | Analytical Results | Level of model and result generation process understanding | Sense of each model benchmarked (name, family…) | Calculation time and resources… | Statistical evaluation criteria : performance and confidence levels… | Consistency between the model process and its explanation | Frequency or thresholds of learning, maintainability… | Difficulties of result evaluation or monitoring (tools, competences…), status |
| 5 | Functional results | Business criteria evaluation, confidence in benefits potential… | Sense of the retreated results (name, nature, units…) and of evaluation metrics | Added indicators, format changes… | Retreatments issued from business appropriation (adjustments…) | Translation between business and statistical criteria (calculation rules, qualitative…) | Exploitation performance target and its confidence level… | Difficulties of deployment preparation of direct and indirect usages, status |
| 6 | Usages | Usage priorization, decision logs and criteria (benefits, uncertainties…) | Sense of each usage and their indicators | Nature of treatments of integration, automatization, controls, securisation… | Further model treatments needed for exploitation (self-learning parameters…) | Level of contribution of a usage to benefits… | Exploitation & maintenance modalities | Estimation of benefits and remaining uncertainties needing further investment, status |

*Figure 4 - Illustration of the most typical metadata used in the Databook modules*

## 3. The Databook: from a prototype to a standard documentation framework

In the next section, field feedback on the prototype is discussed in order to stabilise the Databook as a generic documentation framework and propose its main reading grids.

### 3.1 Benefits perception

The imagined prototype appeared in the field to be a complete, autonomous and dynamic object, logically linked to the stakeholders' documentation needs. Proposed as an Excel file with basic core and metadata structures, it was progressively filled up by the project teams on seven projects. The structure of the prototype appeared flexible enough to be adapted to meet the urgent needs and priorities of stakeholders, and if was successfully judged as operational. Besides its usability, the Databook prototype revealed its significant effectiveness in stimulating cooperation, enlightening the

arbitrations made during the project, and guiding the appropriation of the results by the stakeholders. Different beneficiaries highlighted through qualitative feedback that the device was used for two main purposes: project efficiency and data documentation efficiency.

The *project efficiency* was improved through the facilitation of the mediation milestones (meaning shearing, progress visualisation, decisions traceability…), the knowledge capitalisation useful for further data projects and the identification of indirect usages requiring further analytical investigation. The impact on the project results quality was also highlighted by the users and by business decision-makers: the traceability of all the analytical components of the algorithm and project decisions was perceived as a quality and solution auditability guarantee. Moreover, the recorded indirect usage ideas inspired not only further analytical iterations but also business offer and process evolution. Several data and business project team members also reported a change in posture: at the beginning of the project, they perceived filling in the Databook as an additional workload of little use, but as the project progressed, they realised that the device was indispensable and generated time savings at each iteration.

The *documentation efficiency* was appreciated not only by the team members but also stakeholders outside the project. For instance, data governance managers were very interested in semantic and other metadata generated during the project and the identification of referent data owners. Data protection officers kept and reused the personal data identification to control the discrimination drifts and the usage purposes. Financial managers were also very interested in the possibility to retrace the value generated by the usage and link it to the different data sources: this opened a new field for exploring patrimonial, finance and accounting concepts around data assets development. Finally, IT managers reused the Databook to size the technical resources required to explore, deploy and maintain other business applications. This feedback highlighted that the purpose of the Databook largely exceeded the project efficiency and aimed data project portfolio management and globally the data quality and value management.

### 3.2 Operational limitations

The prototype also showed its weaknesses, the main one being its Excel format which is not very practical for several modules. Indeed, exploration report, functional results and usage roadmap modules had to be completed with more specific formats such as Data Visualisation tools or PowerPoint reports. The analytical results module was also completed outside the Excel file when it was too specific to the benchmarked algorithms. The project roadmap module was simplified as much as possible in order to be coupled with more appropriate project management formats. If the prototype was to evolve to a more sophisticated tool, it would be necessary to handle other types of formats for these modules or favour the compatibility or interoperability with other dedicated tools.

The remaining modules have been adopted in Excel format and adapted to each project. For more complex projects that implied a production of several articulated algorithms, the prototype's modules have been multiplied to serve better the skeleton of the algorithm articulation. Most of the time, one or more modules was left empty: this selection revealed the adaptation to the specific needs and resources of each project. Finally, the module with the data exclusion/inclusion method, predefined with fixed metadata, was redefined orally by the stakeholders and applied mostly through column creation in modules 1 to 6a. Some mandatory regulatory criteria were dropped, such as the personal

data flag for anonymisation for projects without personal of data. Given the field feedback, the Databook structure is confirmed as flexible enough for different projects but attempts to make it more rigid (fixed metadata, mandatory modules…) are qualified as inoperative in practice. Module or metadata aborts were explained by the fact that time investment priorities were set at the small scale of each project, and rarely at the scale of a project portfolio or the company. They were also partly explained by organizational and human factors.

Indeed, as a collaborative tool for a multi-skilled Data Science project team, the Databook has raised several organisational issues. When a Data Science projects remains restricted to a small team, for example with one Data Scientist and one business decision-maker, the Data Scientist is often expected to produce all the documentation alone. Moreover, documentation production can be refused by some team members who do not see its benefits for their own technical tasks. Finally, Data Science is still a young profession and stakeholders often lack acculturation and experience: the anticipation of uncertainties is clearly difficult and time-consuming without experienced skill-carriers. In these circumstances, the definition of responsibilities can remain a weak point.

In theory, the operational application of the Databook requires a clear prior distinction between skills, individuals and responsibilities. Skills are needed to produce qualification metadata, as presented in section 0: they are essential in the Databook construction. Individuals can carry one or more skills, and each of their skill can be tainted with different maturity level. The maturity level is key for anticipation: an inexperienced team member is usually able to document his production only *a posteriori* or execute a qualification procedure only if it has been predefined by a more experienced skill carrier. Responsibilities are defined according the project specificities and individual skill range. Usually, this definition is realized by the Project Manager for the duration of the project. But the distinction of these three concepts and their articulation is often more confusing in practice. Surprisingly, the Databook appears as a good communication facilitator that can be used for responsibilities clarification.

### 3.3 Pilot evaluation

Since the prototype first tests, the Databook prototype was freely accessible to several Data Science teams in a leading French Data Science company called Quinten. Its appropriation continued on Data Science projects between 2017 and 2020, i.e. more than a hundred of projects in health, perfume, insurance, banking, media and industry sectors. Several completed Databooks have been reused from one project to another, mainly for projects for one given company, using the same data sources or with similar usages and data objects. The pilot phase main qualitative feedback confirms the precedent advantages and limitations: it is summarised in the Appendices (Figure 13). This confrontation between theory and various fields reality still must be considered as potentially biased by the Data Science practice of one company. Quinten is characterised by its own values, business offer and managerial practices. Most of the usages produced through the company's Artificial Intelligence projects are aimed at human users in highly regulated domains, thus transparency of the analytical work remains a priority.

The following proposal is an unprecedented attempt to standardise the most useful documentation principles and functionalities. The Databook, as a flexible framework should then be tested in other contexts.

### 3.4 The Databook as a standard documentation framework

The Databook is founded on the principle of distinction between the realization of analytical outputs on the critical path of algorithm design and the production of the documentation. Both dynamics imply similar skills but involve them at different stages and with different purposes. The Databook as a documentation framework provides an opportunity to share the story of how data is transformed into useful information during a collaborative Data Science project. It can be used as a dynamic device for capitalising on knowledge, a material object that helps to gradually retrace the memory of the project and to give transparency to the resulting algorithmic model. The Databook guarantees and respects by design the FAIR principles by guiding the construction of Findable, Accessible, Interoperable and Reusable data and metadata, as far as the business context allows the sharing of sensible data. For the Data Science projects, it provides the same advantages as the use of the FAIR principles in cross-institutional projects (Hansen et al. 2019): project planning, navigation through project changes, information sharing and data sharing outside the group. From a more operational point of view, the Databook structure, illustrated in appendices (Figures 14 to 24), provides three clear reading grids for each different purpose.

#### 3.4.1 Cross-skill data object qualification

The Databook can be used by all project actors to qualify one given data object and determine together its further treatment. It is the most basic use, achievable independently for each Module from 1 to 6. This cross-skill use promotes convergence towards the most relevant result and stimulates the productivity of the project team. The convergence is accelerated by more convenient representations of data, such as graphs, explanations and other reports listed in Module C. The convergence is also improved by Module B, representing the collective arbitration procedure. In parallel, the dynamics of the convergence are monitored in Module A through the statuses. In this situation, the project in less piloted through iterative or overlapped phases than by the progress of the qualification of one given data object. The Databook guide in Module 0 can facilitate this reading grid by presenting in priority core structure elements (horizontal in Figure 4).

For this application, the Databook should be read starting from the Module containing the data object and then viewing the project management modules. For example, Figure 5 illustrates how to use the Databook when the complete project team needs to qualify together the Analytical results.

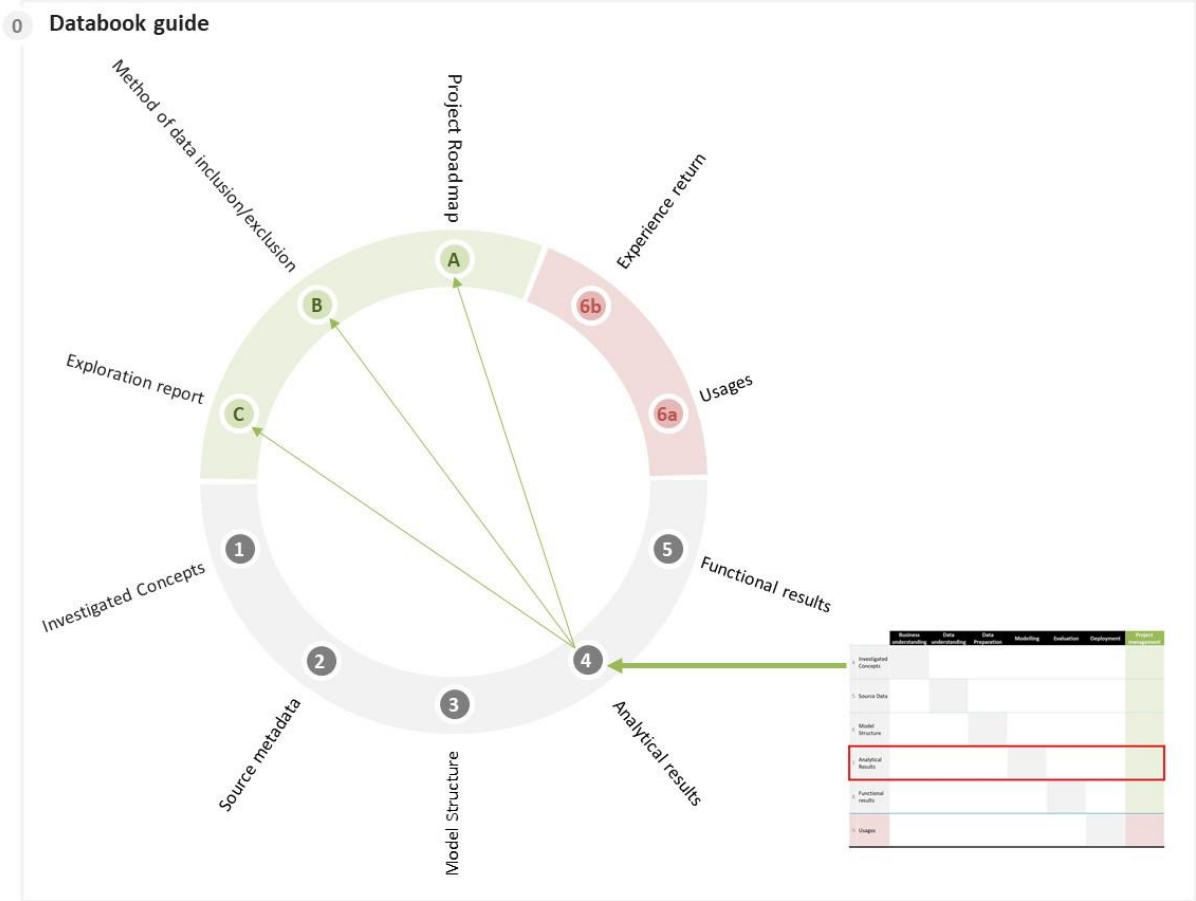



*Figure 5 - Databook reading grid for cross-skill qualification of a data object*

### 3.4.2 Skill capitalisation

The Databook can also be used by one given skill carrier who wishes to gain experience on all the data objects qualified and processed during the project. In this case, his attention will only focus on the qualification criteria that confer his skills on modules 1 to 6. For example, the Databook can be used by a Strategic Manager to understand how the components of the algorithm are related to the business priorities, by a Data steward who is interested in the semantics of all elements processed during the project or by a Data Engineer to optimise the exploration pipeline. The capitalisation is also very interesting for project management. Indeed, the difficulties solved throughout the project constitute a rise in maturity: the estimation of the benefits and remaining uncertainties on the usages at the end of the project can lead to a broader roadmap than a technical deployment of the result. All the knowledge capitalised can be qualitatively consolidated in Module 6b in the form of a return on experience from each of the skills' carriers involved. This capitalisation is essential to generate new ideas of usages and gain in productivity for further data projects (data projects portfolio management) requiring similar data objects or skills. The Databook guide in Module 0 can facilitate this reading grid by presenting in priority metadata structure elements (vertical in Figure 4).

For this application, the Databook should be read starting from each Module from 1 to 6 containing metadata of the skill of interest, and then list the generated knowledge in Module 6b, as presented in Figure 6.

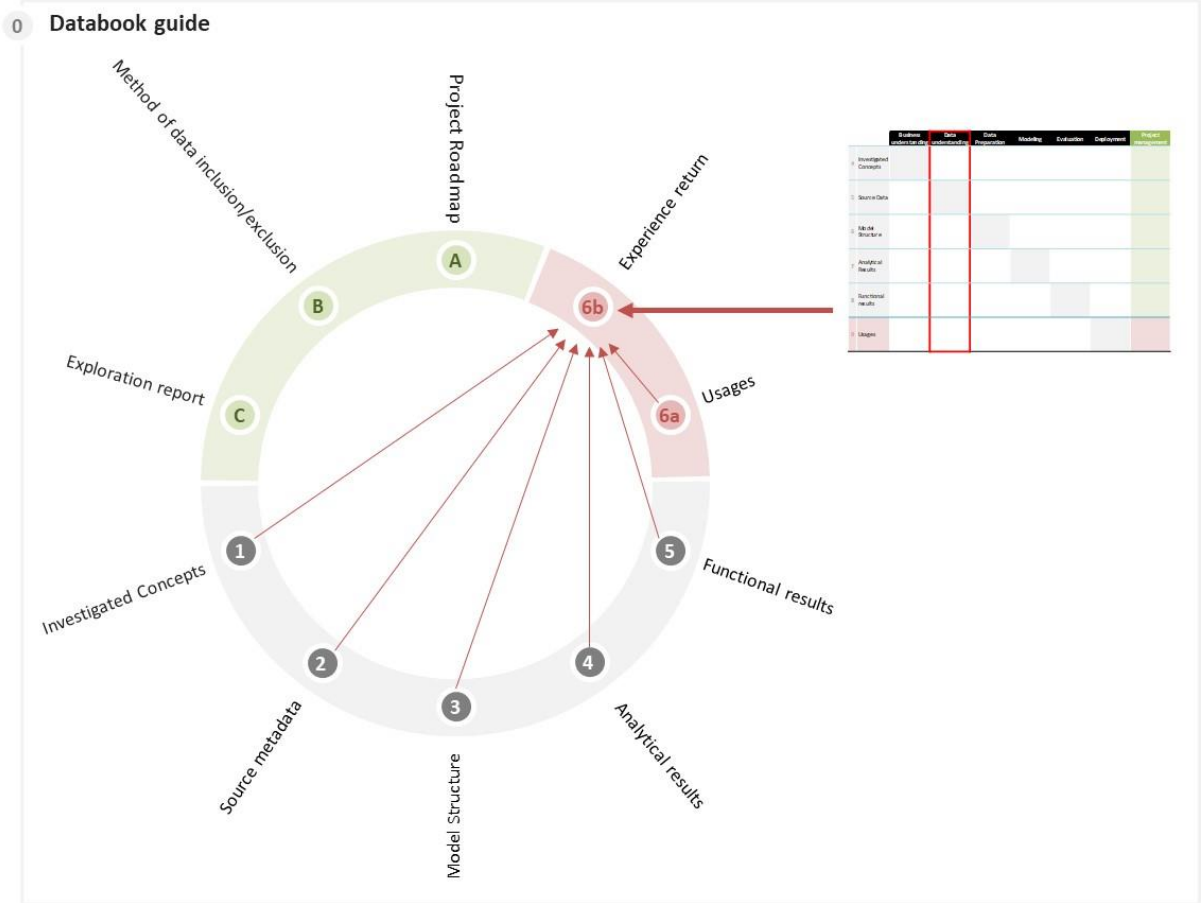



*Figure 6 - Databook reading grid for skill capitalisation from all data objects*

### 3.4.3 Final algorithm understanding

The Databook traces of all the data objects that compose the final algorithm. This trace is the documentation of the critical path: all the data elements notified as *included* are contained in the final versions of each intermediate output. This documentation is key for many purposes. First, it is a detailed specification for usage deployment. Then, it makes the algorithm auditable, including for external actors. And, finally, this traceability gives the possibility to propagate the value generated by the usage back on data sources: this inverse value cascade is hence an original tool for evaluating data assets from both patrimonial and operational views.

All these purposes follow the same reading grid illustrated in Figure 7. For example, within the framework of an audit of the final algorithm, the investigation will consist in going back from the operational usage to the mathematical algorithm, then to the data which feeds this algorithm, then to a set of source data. These source data will then point to key business concepts chosen in the construction of the algorithm. As the audit is interested only in used data elements included in the algorithm, his reading is focused on the diagonal in Figure 4 as soon as we consider that the final status of each type of data object is qualified by the skill carrier that produced the data object. If the auditor of the algorithm must go further to understand the design process, he may also be interested in project management metadata. This reading can be facilitated by the guide, especially by filtering the entire Databook only on elements with a status *included* or by zooming on more detailed project management metadata.

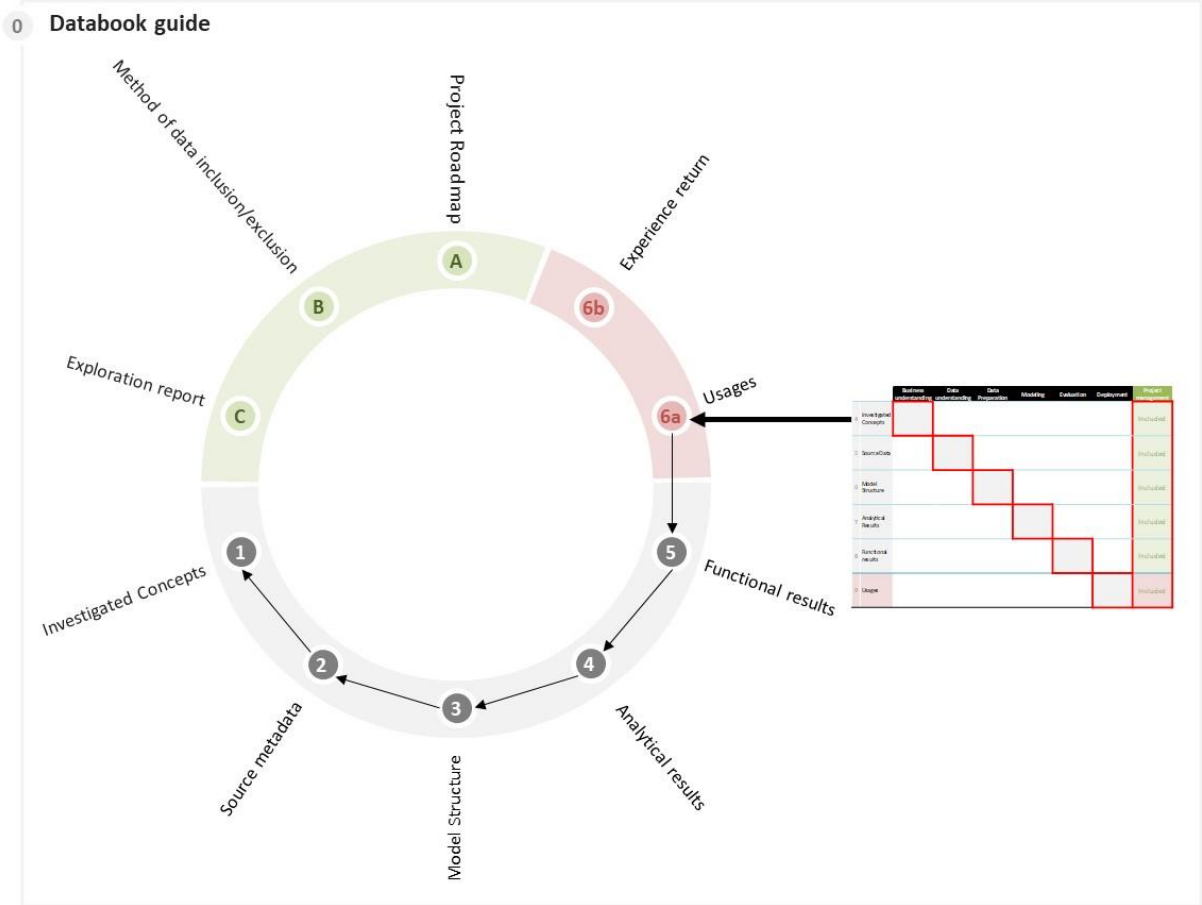



*Figure 7 - Databook reading grid for the final algorithm understanding*

These different reading grids, not exhaustively described above, can fit the priorities chosen for each project or purpose. This flexibility remains one of the advantages of such a documentation framework and justifies Databook definition as a new boundary object, like a portolan chart for navigating through the algorithm's metadata.

## Conclusion

The Databook emerges from the urgent documentation needs of data project stakeholders in the field and from interdisciplinary concepts inspiring the gap filling in the standard state-of-the-art Data Science process, CRISP_DM. Its structure is based on one main principle: in these exploratory algorithm design projects, each phase realization needs specific skills and all these skills are required to progressively adjust the entire process by producing dynamic documentation. Documentation is then the result of both anticipatory and informative qualification work, and the documentation process generates a faster convergence on the best project results. The Databook traces this dynamic and constitutes a boundary object for all the stakeholders. As the responsibilities and skills of a Data Scientist are still poorly and heterogeneously defined, Databook description sheds some light not only on the essential skills but also on their mobilisation mechanism. As a prototype, it is decanted and confirmed as a very efficient Human-Data Mediation facilitator. It can still be improved in terms of ergonomics, but its simple and flexible core structure completed with free metadata structure remains compatible with classical tools and independent of the data project purpose and pace. Outside of

projects, the Databook is useful for the development of a company's data assets and the governance of data quality.

Besides these theorical and operational considerations, one of the main benefits of the Databook remains its contribution to algorithm transparency in business companies. This lack of transparency is too often reduced to the algorithm learning process, especially for the deep learning. However, an algorithm is much more than that: the Databook can reveal all the objects that constitute it from end to end, but also the human choices than have driven its progressive design. The fear of this lack of transparency, crystalized after several scandals in the last years, lead to a search for a French and European position that has so far been unsuccessful, and to the affirmation of founding principles such as loyalty and vigilance by the CNIL in 2017 (Falque-Pierrotin *et al.* 2017). These principles are reflected in a set of recommendations, such as ethics training of all implied actors, the mediation between users to make algorithms more understandable, the subordination of algorithms to human freedom and to the general interest from the design phase, or the creation of a national algorithm audit platform. Capitalisation on French assets such as cultural values oriented towards people and ethics or the quality of training in engineering sciences and mathematics, is already mobilised, as presented by INRIA's annual report and its publications in 2017. While the contributions to the algorithm documentation framework remain limited and mainly oriented towards the control of external algorithms[3], this proposal offers the possibility for all algorithm designers to achieve transparency of their own algorithms.

## Acknowledgement

I would like to gratefully thank Professor Ghislaine Chartron for her determined and caring supervision during my thesis years and her valuable feedback on this paper. Many thanks to Quinten's team that made possible these Data Science projects' observations and the confrontation of the Databook prototype to real life. I am also very thankful to Catherine Lesperance for the careful and thorough revision of this article in English.

## References

Arruabarrena, B., Kembellec, G., & Chartron, G. (2019, March). *Data littératie & SHS : développer des compétences pour l'analyse des données*, Presented at the CODATA - Data Value Chain, Val d'Europe.

Austin, R. D., Joshi, M. P., Su, N., & Sundaram, A. K. (2021). Why So Many Data Science Projects Fail to Deliver. *MIT Sloan Management Review*, **Spring 2021**. Retrieved from https://sloanreview.mit.edu/article/why-so-many-data-science-projects-fail-to-deliver/

Berti-Equille, L. (2012). *La qualité et la gouvernance des données : Au service de la performance des entreprises*, Paris; Cachan: Hermes Science Publications.

Bertino, E., Bernstein, P., Agrawal, D., … Widom, J. (2011). Challenges and Opportunities with Big Data, Cyber Center Publications.

Boyd, D., & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, **15**(5), 662–679.

Broudoux, É., & Scopsi, C. (2011). Introduction. *Études de communication*, (36), 9–22.

Brynjolfsson, E., Hitt, L. M., & Kim, H. H. (2011). *Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?* (SSRN Scholarly Paper No. ID 1819486), Rochester, NY: Social Science Research Network.

Camiciotti, L., & Racca, C. (2015). *Creare valore con i Big Data. Gli strumenti, i processi, le applicazioni pratiche*, 1 edizione, Milano: Edizioni LSWR.

Chapman, P. (1999). *The CRISP-DM User Guide*, Presented at the Brussels SIG Meeting, NCR Systems Engineering Copenhagen.

Cottin, M., & Nesme, M.-F. (2017). La qualité : variations autour d'une notion essentielle, Quality: variations on an essential notion. *I2D – Information, données & documents*, **53**(4), 28–29.

Davenport, T. H., & Patil, D. J. (2012). Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review*, **90**(5), 70–76.

Doucet, C. (2010). *La qualité*, Paris: Presses Universitaires de France.

Ermine, J.-L. (2003). *La gestion des connaissances*, Hermes Lavoisier.

Falque-Pierrotin, I., Mahjoubi, M., & Villani, C. (2017). *Comment permettre à l'Homme de garder la main ? Rapport sur les enjeux éthiques des algorithmes et de l'intelligence artificielle*, CNIL. Retrieved from https://www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_garder_la_main_web.pdf

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, **39**(11), 27–34.

Hansen, Z. N., Kruse, F., & Thestrup, J. B. (2019). Managing data in cross-institutional projects. *IASSIST Quarterly*, **43**(3), 1–10.

Loshin, D. (2010). *Master Data Management*, Morgan Kaufmann.

Manyika, J., Chui, M., Brown, B., … Hung Byers, A. (2011). *Big data: The next frontier for innovation, competition, and productivity*, McKinsey Global Institute. Retrieved from https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation

Mariko, D. (2016). *Le Master Data Management (MDM) et la qualité des données de l'entreprise : synergies digitales et collaboratives*, INTD-CNAM.

Mayère, A. (1990). *Pour une Économie de L'Information*, C.N.R.S. Editions. doi: doi.org/10.3917/cnrs.mayer.1990.01

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution that Will Transform how We Live, Work, and Think*, Houghton Mifflin Harcourt.

Miller, H. G., & Mork, P. (2013). From Data to Decisions: A Value Chain for Big Data. *IT Professional*, **15**(1), 57–59.

Nesvijevskaia, A. (2017). Value creating through Data Science projects : insufficiency of the standart workflows.

Nesvijevskaia, A. (2019, October 18). *Phénomène Big Data en entreprise : processus projet, génération de valeur et Médiation Homme-Données* (thesis), Paris, CNAM. Retrieved from http://www.theses.fr/2019CNAM1247

Odeh, S., & Chartron, G. (2016). Acteurs et économie des métadonnées du livre en France : analyse et avenir. *Documentation et bibliothèques*, **62**(1), 21–32.

Piatetsky-Shapiro, G. (1994). An Overview of Knowledge Discovery in Databases: Recent Progress and Challenges. In *Rough Sets, Fuzzy Sets and Knowledge Discovery*, Springer, London, pp. 1–10.

Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, **1**(1), 51–59.

Shearer, C. (2000, Fall). The CRISP-DM model : the new blueprint for data mining. *Journal of Data Warehousing*, pp. 13–22, pp. 13–22.

Star, S. L., & Griesemer, J. R. (1989). Institutional Ecology, `Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science*, **19**(3), 387–420.

Wang, R. Y. (1998). A Product Perspective on Total Data Quality Management. *Commun. ACM*, **41**(2), 58–65.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, pp. 29–39.

Zacklad, M., Cahier, J.-P., Bénel, A., Zaher, L., Lejeune, C., & Zhou, C. (2007). Hypertopic: une métasémiotique et un protocole pour le Web socio-sémantique. In *Actes des 18eme journées francophones d'ingénierie des connaissances*, Francky Trichet, p. 13.

# Appendices

## 1. Reference outputs mapping

| Phases of adjusted CRISP_DM model | Reference outputs | Critical Path analytical outputs | Critical Path analytical outputs documentation | Mediation Milestones |
|---|---|---|---|---|
| Business Understanding | Terminology | | | Project Roadmap |
| | Project Plan | | | |
| | Background | | | |
| | Risks and Contingencies | | | |
| | Costs and Benefits | | | |
| | Business Objectives | | | |
| | Business Success Criteria | | | |
| | Assessment of Tools and Techniques (initial and final) | | | |
| | Requirements, Assumptions, and Constraints | | Perimeter description | |
| | Data Science Success Criteria | | | |
| | Data Science Goals | | | |
| | Inventory of Resources | Raw Data | | |
| Data Understanding | Data Collection Report | | Source data description | |
| | Data Description Report | | | |
| | Data Exploration Report | | | Exploration Report |
| | Data Quality Report | | | |
| | Data Set Description | | | |
| | Data Set | Selected Data | | |
| Data Preparation | Rationale for Inclusion / Exclusion | | | Data inc./exc. Method |
| | Data Cleaning Report | | Model Structure description | |
| | Data treatment description | | | |
| | Derivated Attributes | Structured Data | | |
| | Generated Records | | | |
| | Merged Data | | | |
| | Reformated Data | | | |
| Modelling | Modeling Technique | | Analytical results description | |
| | Modeling Assumptions | | | |
| | Test Design | | | |
| | Model Description | | | |
| | Analytical results Benchmark (model assessments results) | | | |
| | Parameter setting (initial and revised) | Models | | |
| | Model | | | |
| Evaluation | Assessment of Data Mining Results w.r.t. Business Success Criteria | | Functional results description | |
| | Approved models description | | | |
| | Review of Process | | | |
| | Approved Models | Selected Models | | |
| Deployment | List of Possible Actions & Decisions | | Usage Roadmap & Knowledge Capitalization | |
| | Restitution format | Knowledge | | |
| | Final Presentation | | | |
| | Deployment Plan | Automated Model | | |
| | Monitoring and Maintenance Plan | | | |
| | Experience Documentation | | | |

*Figure 8 - Reference outputs of the adjusted CRISP_DM model and their classification*

The Figure 8 presents a mapping between the main outputs of the adjusted CRISP_DM and the types of outputs presented in this paper. The CRISP_DM adjustment consists in rearranging its outputs in order to isolate for each phase its main intermediate analytical output (in blue), determining the dominant skill for this output and grouping the documentation outputs (in grey and red) around this dominant skill. The dominant skills are, in activity order: Strategy Management, Data Stewardship, Data Engineering, Machine Learning, Business Intelligence and Product Ownership/Knowledge Management. The documentation outputs can be produced collectively or with the help of Business Analysis and Data Analysis skills. Finally, the Project Management skill remain transversal to produce associated mediation milestones outputs (in green). Individual skills' carriers and project roles are not considered in this mapping.

## 2. Examples of Databook modules

In this appendix, four illustrations of Databook modules present its typical applications in Artificial Intelligence projects. These illustrations are extracted from complete project Databooks in Excel format and anonymised. For confidentiality reasons, communication of a complete Databook prototype with qualified data objects is avoided. However, further real-world illustrations and practical details emerging from the testing phase in France can be found in Appendix 11 of the multiple-case study (Nesvijevskaia, 2019, Appendix 11).

| Statut des Variables | |
|---|---|
| Exclue | Variables exclues de l'analyse |
| Incluse | Variables incluses dans l'analyse : voire nature d'usage |
| Nouvel extract attendu | Variables / tables nécessitant un extract complémentaire |
| ? | En cours d'analyse |

**Tableau des données reçues**

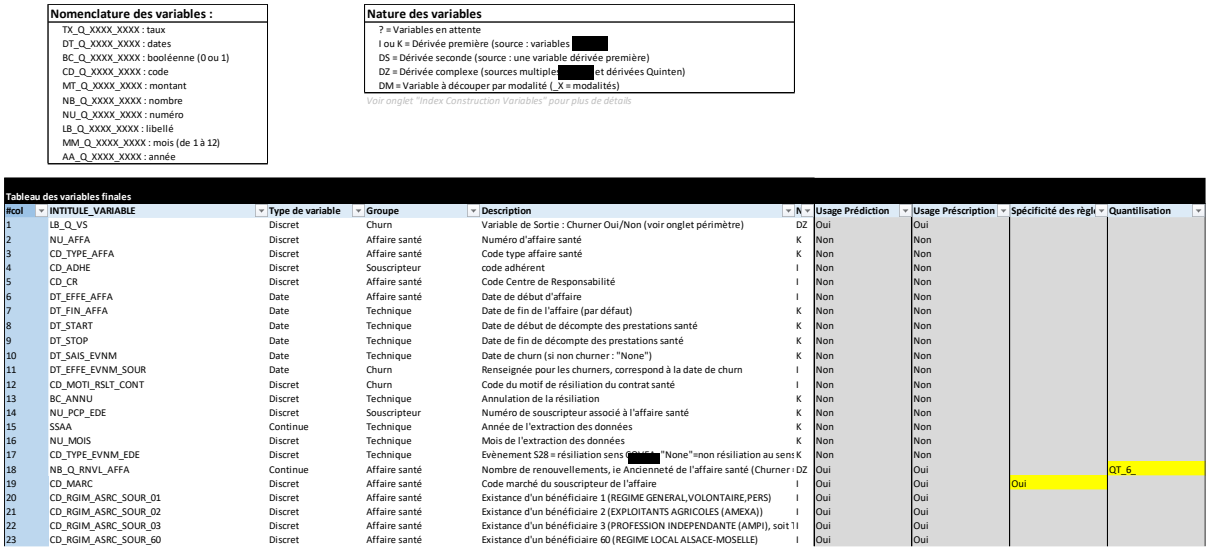
| Informations sur les fichiers reçus | | | Informations sur les variables reçues | | | | | | | | Nature de l'usage de la variable | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| N° de Lot | Table / File | Jointure(s) | Variable | Explication | TYPE de variable (AZ) (2 = Texte - 1 = Num.) | Priorité métier | DONNEE NOMINATIVE | Statut | test FORGE | Brute | Dérivée | Clé | Autre |
| 1 | sinmre15 | NOPOL | TOPCRAC | ? | ? | non | non | Exclue | Non - vide | | | | |
| 1 | sinmre15 | NOPOL | TOPASSTR | ? | ? | non | non | Exclue | Non - vide | | | | |
| 1 | sinmre15 | NOPOL | chargecie | Charge | ? | non | non | Exclue | OK | ? | | | |
| 1 | sinmre15 | NOPOL | regcie | Règlement | ? | non | non | Exclue | OK | ? | | | |
| 1 | sinmre15 | NOPOL | reccie | Recours encaissés | ? | non | non | Exclue | OK | ? | | | |
| 1 | sinmre15 | NOPOL | rapcie | Restant à payer | ? | non | non | Exclue | OK | ? | | | |
| 1 | ipfmre15 | NOPOL | DTMJNE | DATE DE MISE A JOUR DU SEGMENT (AQQQ) | ? | non | non | Exclue | OK | | | | |
| 1 | ipfmre15 | NOPOL | NOPOL | NUMERO DE POLICE (ANCIEN) X(14) | ? | | 1 non | Incluse | OK | x | | x | |
| 1 | ipfmre15 | NOPOL | NOINT | NUMERO DE L INTERMEDIAIRE | ? | | 1 non | Incluse | OK | x | | | |
| 1 | ipfmre15 | NOPOL | CDPOLE | CODE POLE | ? | | 1 non | Incluse | OK | x | | | |
| 1 | ipfmre15 | NOPOL | CMARCH | CODE MARCHE | ? | | 1 non | Incluse | OK | x | | | |
| 1 | ipfmre15 | NOPOL | CDREG | CODE REGION | ? | non | non | Exclue | OK | | | | |
| 1 | ipfmre15 | NOPOL | CDPROD | CODE PRODUIT | ? | | 1 non | Incluse | OK | | x | | |
| 1 | ipfmre15 | NOPOL | CSEGT | CODE SEGMENT | ? | | 1 non | Exclue | Non - Modalité Fixe | | | | |
| 1 | ipfmre15 | NOPOL | CSSEGT | CODE SOUS SEGMENT | ? | non | non | Exclue | OK | | | | |
| 1 | ipfmre15 | NOPOL | DTRESILP | DATE DE RESILIATION POLICE | ? | | 1 non | Incluse | OK | x | x | | |
| 1 | ipfmre15 | NOPOL | DTTRAMVT | DATE TRAITEMENT DERNIER MVT DU TRAITE | ? | non | non | Exclue | OK | | | | |
| 1 | ipfmre15 | NOPOL | NOAVEDER | DERNIER NUMERO D'AVENANT | ? | non | non | Exclue | OK | | | | |
| 1 | ipfmre15 | NOPOL | NOPOLORI | NO DE POLICE COMPAGNIE PRECEDENTE | ? | non | non | Exclue | OK | | | | |
| 1 | ipfmre15 | NOPOL | NOCIE | NUMERO DE COMPAGNIE | ? | non | non | Exclue | Non - Modalité Fixe | | | | |

*Figure 9 - Illustration of the Module 2 from a project on compliance*

The first illustration (see Figure 9 above) is extracted from a Databook adapted to the context of a French branch of leading insurance company which worked on the detection of contracts with non-compliance risks in order to optimise the control process. This module aims the qualification of variables used for the detection model (Module 5). The different components of this module include (from left to right):

- Technical information about the sources of variables needed by the Data Engineer: reception batch, name of the table and database join key for the table.
- Semantics of the variables: code and meaning of the variable (if the name does not exist in the database, or does not make sense, it is manually added in the Databook).
- Type of variable needed to choose the structuring methods (here, left unqualified).
- Business priority of the exploration of each variable (exclusion of a variables by a business decision-maker, based on his perception of the compliance control process).
- Flag of personal data to exclude (discussed with the DPO of the company).
- Status of the variable: the excluded variables are eliminated from the final algorithms.
- “Test Forge”: conclusion of a custom analytical qualification based on a mathematical method of elimination of variables with no signal or too much noise (realized by the Machine Learner).
- Usage of the variable: function of the variable in the learning matrix, qualified by the Data Engineer.

The columns colours represent the different skill-carriers that produced the qualification.



| Nomenclature des variables : |
|---|
| TX_Q_XXXX_XXXX : taux |
| DT_Q_XXXX_XXXX : dates |
| BC_Q_XXXX_XXXX : booléenne (0 ou 1) |
| CD_Q_XXXX_XXXX : code |
| MT_Q_XXXX_XXXX : montant |
| NB_Q_XXXX_XXXX : nombre |
| NU_Q_XXXX_XXXX : numéro |
| LB_Q_XXXX_XXXX : libellé |
| MM_Q_XXXX_XXXX : mois (de 1 à 12) |
| AA_Q_XXXX_XXXX : année |

| Nature des variables |
|---|
| ? = Variables en attente |
| I ou K = Dérivée première (source : variables [redacted] |
| DS = Dérivée seconde (source : une variable dérivée première) |
| DZ = Dérivée complexe (sources multiples [redacted] et dérivées Quinten) |
| DM = Variable à découper par modalité (_X = modalités) |

*Voir onglet "Index Construction Variables" pour plus de détails*

Tableau des variables finales

| #col | INTITULE_VARIABLE | Type de variable | Groupe | Description | N | Usage Prédiction | Usage Prescription | Spécificité des règl | Quantilisation |
|---|---|---|---|---|---|---|---|---|---|
| 1 | LB_Q_VS | Discret | Churn | Variable de Sortie : Churner Oui/Non (voir onglet périmètre) | DZ | Oui | Oui | | |
| 2 | NU_AFFA | Discret | Affaire santé | Numéro d'affaire santé | K | Non | Non | | |
| 3 | CD_TYPE_AFFA | Discret | Affaire santé | Code type affaire santé | K | Non | Non | | |
| 4 | CD_ADHE | Discret | Souscripteur | code adhérent | I | Non | Non | | |
| 5 | CD_CR | Discret | Affaire santé | Code Centre de Responsabilité | I | Non | Non | | |
| 6 | DT_EFFE_AFFA | Date | Affaire santé | Date de début d'affaire | I | Non | Non | | |
| 7 | DT_FIN_AFFA | Date | Technique | Date de fin de l'affaire (par défaut) | K | Non | Non | | |
| 8 | DT_START | Date | Technique | Date de début de décompte des prestations santé | K | Non | Non | | |
| 9 | DT_STOP | Date | Technique | Date de fin de décompte des prestations santé | K | Non | Non | | |
| 10 | DT_SAIS_EVNM | Date | Technique | Date de churn (si non churner : "None") | K | Non | Non | | |
| 11 | DT_EFFE_EVNM_SOUR | Date | Churn | Renseignée pour les churners, correspond à la date de churn | I | Non | Non | | |
| 12 | CD_MOTI_RSLT_CONT | Discret | Churn | Code du motif de résiliation du contrat santé | I | Non | Non | | |
| 13 | BC_ANNU | Discret | Technique | Annulation de la résiliation | K | Non | Non | | |
| 14 | NU_PCP_EDE | Discret | Souscripteur | Numéro de souscripteur associé à l'affaire santé | K | Non | Non | | |
| 15 | SSAA | Continue | Technique | Année de l'extraction des données | K | Non | Non | | |
| 16 | NU_MOIS | Discret | Technique | Mois de l'extraction des données | K | Non | Non | | |
| 17 | CD_TYPE_EVNM_EDE | Discret | Technique | Evènement S28 = résiliation sens [redacted] "None"=non résiliation au sens | K | Non | Non | | |
| 18 | NB_Q_RNVL_AFFA | Continue | Affaire santé | Nombre de renouvellements, ie Ancienneté de l'affaire santé (Churner : | DZ | Oui | Oui | | QT_6_ |
| 19 | CD_MARC | Discret | Affaire santé | Code marché du souscripteur de l'affaire | I | Oui | Oui | Oui | |
| 20 | CD_RGIM_ASRC_SOUR_01 | Discret | Affaire santé | Existance d'un bénéficiaire 1 (REGIME GENERAL, VOLONTAIRE,PERS) | I | Oui | Oui | | |
| 21 | CD_RGIM_ASRC_SOUR_02 | Discret | Affaire santé | Existance d'un bénéficiaire 2 (EXPLOITANTS AGRICOLES (AMEXA)) | I | Oui | Oui | | |
| 22 | CD_RGIM_ASRC_SOUR_03 | Discret | Affaire santé | Existance d'un bénéficiaire 3 (PROFESSION INDEPENDANTE (AMPI), soit l | I | Oui | Oui | | |
| 23 | CD_RGIM_ASRC_SOUR_60 | Discret | Affaire santé | Existance d'un bénéficiaire 60 (REGIME LOCAL ALSACE-MOSELLE) | I | Oui | Oui | | |

*Figure 10 - Illustration of the Module 3 from a project on health insurance churn*

The second illustration (see Figure 10 above) is extracted from a Databook of a health insurance churn project aiming the generation of two different algorithmic models (prescriptive profile generation and predictive scoring approach). In this project, more than 200 variables were collected and transformed into 500 new variables: these new variables are documented in Module 6. It is composed of (from left to right):

- Position of the variable in the new table, code and type of the variable (discrete or continuous): these criteria are critical for the Machine Learner to use the variables in a learning matrix. The nomenclatures for the variable codes have been defined specifically for the project in order to maintain homogeneity with existing nomenclature methods (Data analysis skills).
- Semantics of the new variables: all the variables are organised by groups with homogeneous meaning (customer characteristics, contract characteristics, trends in past claims…) and then described one by one. These semantics correspond to a new data dictionary.
- Type of structuring method to create the variable: closely linked to the Data Engineering pipeline, this qualification gives the possibility to see at a glance if the variable is identical to the source variable or if it is issued from a more complex treatment. The nomenclature of types of treatments have been defined specifically for the project context.
- The status of each variable is here split into two columns, each one corresponding to one of the two algorithms (predictive and prescriptive) used in the following step. Each status is here binary, representing the inclusion (yes) or exclusion (no) of the variable for each algorithm.
- The final columns correspond to two specific data treatments for the models: business rules association and mathematical quantiles generation. Each one is specifically documented in a complementary module.

This module was produced entirely by the Data Scientist and controlled by the business expert.

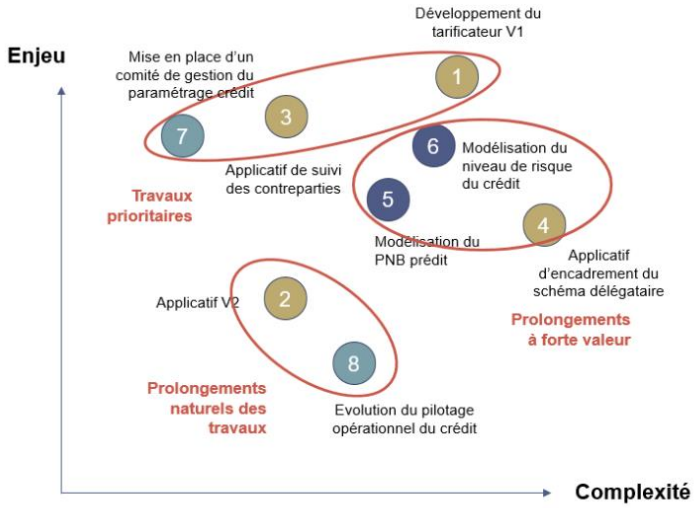



*Figure 11 - Illustration of a Usage Roadmap presented in Power Point format and based on the Module 6a*

The third illustration (see Figure 11 above) is an example of representation of the usage roadmap in a graphical Power Point format. This graph represents eight actions that need to be validated before deployment after the project. Each action is qualified in terms of:

- Number and name of each action
- Two types of categories of the actions: the colours represent the type of skills necessary for the deployment, and the red circles represent a qualitative priority judgement of the actions
- Estimated value creation (vertical axis) and complexity (horizontal axis)

This ergonomic representation is usually completed by a planning as soon as actions are validated. It is based on the Module 6a of the Databook (Usage Roadmap): the module is a matrix structure in Excel. Unsurprisingly, usually this Excel file contains not only the qualified actions, presented above, but also more detailed tasks, associated with constraints, responsibilities, dates and charge estimation (time and budget). The charge is usually qualified by the different skills' carriers of the project. The underlying Excel illustration remains confidential.

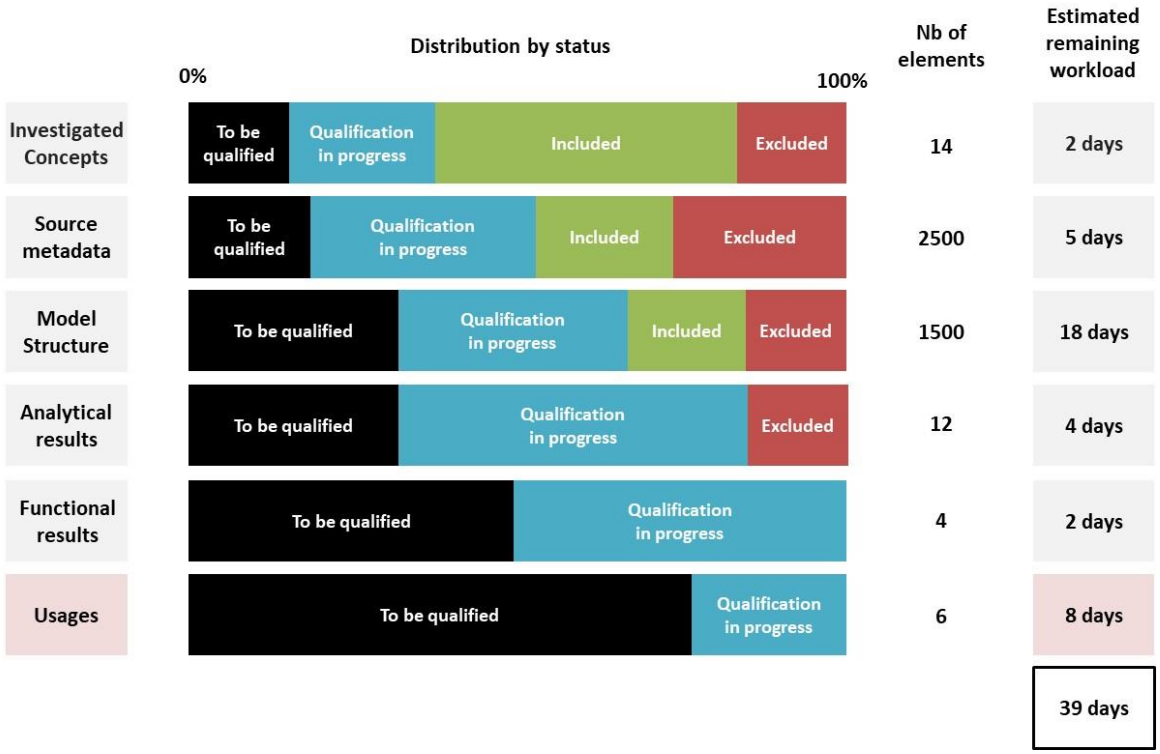



*Figure 12 - Illustration of a synthesis of statuses by module used for a given project milestone*

The last illustration (see Figure 12 above) is an example of synthesis of statuses issued from all the modules from 1 to 6a, presented in Module A (Project Roadmap). For each module 1 to 6a, data elements to be qualified are counted and distributed by type of status. For each type of status and module, a mean workload of qualification estimated: this gives the possibility to translate the remaining qualification workload in man-days. This charge anticipation method can be applied when all the elements and mean charges are correctly anticipated but also for more agile project management when elements emerge progressively and are associated with individual short-term charges. It is then compatible with typical backlog-burning monitoring tools.

## 3. Pilot feedback

| Databook Modules | | | Case 1 | Case 2 | Case 3 | Case 4 | Case 5 | Case 6 | Case 7 | Qualitative feedback from these 7 and other cases* |
|---|---|---|---|---|---|---|---|---|---|---|
| Mediation Milestones | 0 | Guide and Terminology | ♦ | ♦ | ♦ | ♦ | ♦ | ◆ | ♦ | The guide is necessary as a summary of structure of the modules; Terminology can be dispatched to modules 4 to 8 |
| | A | Project Roadmap | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | A minimal compatibility with usual project management tools is sufficient (qualification progress through statuses) |
| | B | Method of data inclusion/exclusion | ◆ | ◆ | ◆ | ◆ | ◆ | ◆ | ◆ | Too specific to each project to be a set of fixed parameters, the method is still needed to reflect the qualification process |
| | C | Exploration report | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | If the format is very usage specific, the module is still necessary as a link to the terminology for data presented in the report |
| Critical Path analytical outputs documentation | 1 | Investigated concepts | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | The format is difficult to fit to complex contexts, but has to structure the list of defined key metrics and concepts |
| | 2 | Source Metadata | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | As one of the most used modules, it has been completed by a hierarchy : sources, tables, variables, index of variable modalities |
| | 3 | Model Structure | ♦ | ♦ | ♦ | ♦ | ♦ | ♢ | ♦ | The module is most useful as the list of final variables in this format, and can be completed by nomenclatures and ER models |
| | 4 | Analytical Results | ♦ | ♦ | ♦ | ♦ | ♦ | ♢ | ♦ | Very algorithm specific, the module can be simplified to the evaluation metrics for benchmarked algorithms |
| | 5 | Functional results | ♦ | ♦ | ♦ | ♦ | ♦ | ♢ | ♦ | Very usage specific, the module can be simplified to the link between data and business evaluation criteria of the results |
| Usages | 6a | Usage Roadmap | ♦ | ♦ | ♦ | ♦ | ♦ | ♢ | ♦ | The module is a list of decisions of actions and remaining uncertainties, with expected investments and benefits |
| | 6b | Knowledge Capitalization | ♦ | ♦ | | ♦ | ♦ | ♦ | ♦ | Much more qualitative module, it can be progressively completed with experience returns throughout the project |
| **Case success (qualitative feedback)** | | | + | + | + | + | + | + | - | **The databook usage is globally correlated to the project success** |

**Legend**

- ♦ : used in the Excel Prototype
- ♦ : used in a different format
- ◆ : used orally
- ♦ : not used
- ♢ : phase not realized in the project

| | |
|---|---|
| Case 1 | Churn in health insurance |
| Case 2 | Turnover prediction in perfume industry |
| Case 3 | Risk prevention in health insurance |
| Case 4 | Compliance control in automobile insurance |
| Case 5 | Heavy property damage claims |
| Case 6 | Citrus price prediction in perfume industry (negative case) |
| Case 7 | Cross-selling in building insurance (negative case) |
| *Other cases | More than a hundred projects in health, banking, insurance, media and industry sectors by Data Scientists without supervision (independent and autonomous usage of the modules) |

*Figure 13 - Feedback on the implementation of the Databook in the field*

## 4. Databook: Excel format

The Figures 14 to 24 illustrate each Excel sheet of a Databook with documentation of the first iterations of an imaginary churn prediction project.

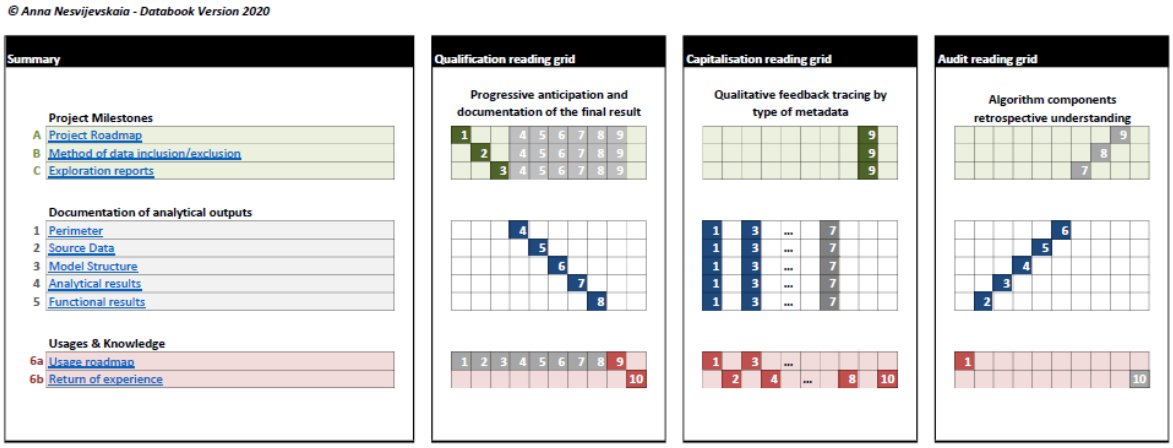
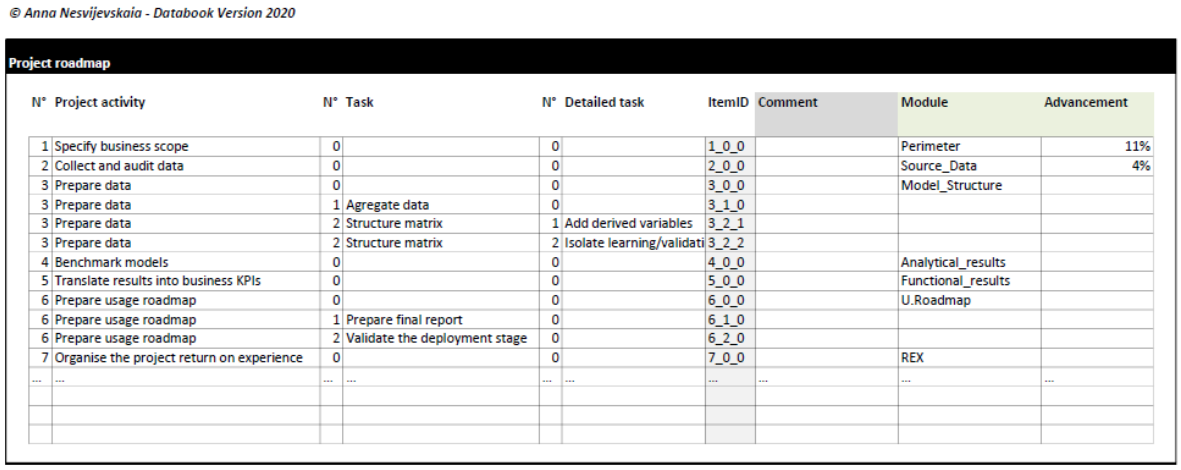



*Figure 14 – Module 0: Guide*




| N° | Project activity | N° | Task | N° | Detailed task | ItemID | Comment | Module | Advancement |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Specify business scope | 0 | | 0 | | 1_0_0 | | Perimeter | 11% |
| 2 | Collect and audit data | 0 | | 0 | | 2_0_0 | | Source_Data | 4% |
| 3 | Prepare data | 0 | | 0 | | 3_0_0 | | Model_Structure | |
| 3 | Prepare data | 1 | Agregate data | 0 | | 3_1_0 | | | |
| 3 | Prepare data | 2 | Structure matrix | 1 | Add derived variables | 3_2_1 | | | |
| 3 | Prepare data | 2 | Structure matrix | 2 | Isolate learning/validati | 3_2_2 | | | |
| 4 | Benchmark models | 0 | | 0 | | 4_0_0 | | Analytical_results | |
| 5 | Translate results into business KPIs | 0 | | 0 | | 5_0_0 | | Functional_results | |
| 6 | Prepare usage roadmap | 0 | | 0 | | 6_0_0 | | U.Roadmap | |
| 6 | Prepare usage roadmap | 1 | Prepare final report | 0 | | 6_1_0 | | | |
| 6 | Prepare usage roadmap | 2 | Validate the deployment stage | 0 | | 6_2_0 | | | |
| 7 | Organise the project return on experience | 0 | | 0 | | 7_0_0 | | REX | |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

*Figure 15 – Module A: Project Roadmap*

 Nesvijevskaia, Anna (2021) DATABOOK: a standardised framework for dynamic documentation of algorithm design during Data Science projects, IASSIST Quarterly 45(2), pp. 1-34. DOI: https://doi.org/10.29173/iq989

**Method of data inclusion / Exclusion**

| N° | Module | N° | Metadata Type | N° | Medatada Name | ItemID | Comment | Methods of statuses qualification |
|---|---|---|---|---|---|---|---|---|
| 1 | Perimeter | 1 | Business Understanding | 1 | Business concept description | 1_1_1 | | Include only clear and prioritaty concepts |
| 1 | Perimeter | 2 | Data Understandig | 2 | Data Source indentified | 1_2_2 | | If not identified, excusion |
| 1 | Perimeter | 3 | Data Preparation | 3 | Data preparation comment | 1_3_3 | | Exclude if the translation is impossible |
| 1 | Perimeter | 4 | Modelling | 4 | Impact on modellisation | 1_4_4 | | Exclude if the bias is expected as too high |
| 1 | Perimeter | 5 | Evaluation | 5 | Order of magnitude | 1_5_5 | | Anticipation prefered, but not needed |
| 1 | Perimeter | 6 | Usage Deployment | 6 | Operational comment | 1_6_6 | | Anticipation prefered, but not needed |
| 2 | Source_Data | 1 | Business Understanding | 1 | Business priority | 2_1_1 | | -- |
| 2 | Source_Data | 2 | Data Understandig | 2 | Description | 2_2_2 | | |
| 2 | Source_Data | 1 | Business Understanding | 3 | Compliance | 2_1_3 | | |
| 2 | Source_Data | 3 | Data Preparation | 4 | Volume | 2_3_4 | | |
| 2 | Source_Data | 3 | Data Preparation | 5 | Completeness | 2_3_5 | | |
| 2 | Source_Data | 3 | Data Preparation | 6 | Uniqueness | 2_3_6 | | |
| 2 | Source_Data | 5 | Evaluation | 7 | Accuracy | 2_5_7 | | |
| 2 | Source_Data | 6 | Usage Deployment | 8 | Operational comment | 2_6_8 | | |
| 3 | Model_Structure | 1 | Business Understanding | 1 | Business priority | 3_1_1 | | |
| 3 | Model_Structure | 2 | Data Understandig | 2 | Description | 3_2_2 | | |
| 3 | Model_Structure | 1 | Business Understanding | 3 | Business concept | 3_1_3 | | |
| 3 | Model_Structure | 4 | Data Understandig | 4 | Unit | 3_4_4 | | |
| 3 | Model_Structure | 3 | Data Preparation | 5 | Type of treatment | 3_3_5 | | |
| 3 | Model_Structure | 3 | Data Preparation | 6 | Volume | 3_3_6 | | |
| 3 | Model_Structure | 4 | Modelling | 7 | Base separation | 3_4_7 | | |
| 3 | Model_Structure | 5 | Evaluation | 8 | Order of magnitude | 3_5_8 | | |
| 3 | Model_Structure | 5 | Evaluation | 9 | Accuracy | 3_5_9 | | |
| 3 | Model_Structure | 6 | Usage Deployment | 10 | Operational comment | 3_6_10 | | |
| 4 | Analytical_results | 1 | Business Understanding | 1 | Business priority | 4_1_1 | | |
| 4 | Analytical_results | 2 | Data Understandig | 2 | Description | 4_2_2 | | |
| 4 | Analytical_results | 4 | Modelling | 3 | AUC | 4_4_3 | | |
| 4 | Analytical_results | 3 | Data Preparation | 4 | Volume | 4_3_4 | | |
| 4 | Analytical_results | 4 | Modelling | 5 | Lift | 4_4_5 | | |
| 4 | Analytical_results | 4 | Modelling | 6 | Coverage | 4_4_6 | | |
| 4 | Analytical_results | 5 | Evaluation | 7 | Model Interpretability | 4_5_7 | | |
| 4 | Analytical_results | 6 | Usage Deployment | 8 | Operational comment | 4_6_8 | | |
| 5 | Functional_results | 1 | Business Understanding | 1 | Business priority | 5_1_1 | | |
| 5 | Functional_results | 2 | Data Understandig | 2 | Description | 5_2_2 | | |
| 5 | Functional_results | 5 | Evaluation | 3 | Mean nb of calls per week | 5_5_3 | | |
| 5 | Functional_results | 5 | Evaluation | 4 | Target Precision | 5_5_4 | | |
| 5 | Functional_results | 5 | Evaluation | 5 | Time saving | 5_5_5 | | |
| 5 | Functional_results | 6 | Usage Deployment | 6 | Result Interpretability | 5_6_6 | | |
| 5 | Functional_results | 5 | Evaluation | 7 | Success rate | 5_5_7 | | |
| 5 | Functional_results | 6 | Usage Deployment | 8 | Pilot qualitative feedback | 5_6_8 | | |
| 5 | Functional_results | 5 | Evaluation | 9 | Turnover saved | 5_5_9 | | |
| 6 | U.Roadmap | 1 | Business Understanding | 1 | Business priority | 6_1_1 | | |
| 6 | U.Roadmap | 3 | Data Preparation | 2 | Data ressources need | 6_3_2 | | |
| 6 | U.Roadmap | 3 | Project Management | 3 | Time needed | 6_3_3 | | |
| 6 | U.Roadmap | 4 | Project Management | 4 | Risk | 6_4_4 | | |
| 6 | U.Roadmap | 5 | Project Management | 5 | Cost | 6_5_5 | | |
| 6 | U.Roadmap | 5 | Evaluation | 6 | Estimated benefits | 6_5_6 | | |
| 6 | U.Roadmap | 1 | Business Understanding | 7 | Remaining uncertainties | 6_1_7 | | |
| 7 | REX | 1 | Project Management | 1 | Interested stakeholders | 7_1_1 | | |
| 7 | REX | 1 | Business Understanding | 2 | Remaining uncertainties | 7_1_2 | | |
| -- | -- | -- | -- | -- | -- | -- | -- | -- |

*Figure 16 – Module B: Method of data inclusion/exclusion*

*© Anna Nesvijevskaia - Databook Version 2020*

**Exploration reports**

| N° | Document name | N° | Document part | ItemID | Version number | Version date | Document type | Comment | Document storage link | Milestone name | Milestone date | Milestone place | Participants of the milestone | Main decisions |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | IQ Quarterly Paper on Databook | 0 | | 1_0 | Final | 02/04/2021 | Paper in open access | Databook description | https://iassistquarterly.com/ | Conference IASSIST | 16/09/2021 | Gothenburg, Suede | Researchers | Use the Databook in Data Science Projects |
| 1 | IQ Quarterly Paper on Databook | 1 | Appendix Figure 9 | 1_1 | | | | Module 2 illustration | | | | | | Fill your own Module 2 |
| 1 | IQ Quarterly Paper on Databook | 2 | Appendix Figure 10 | 1_2 | | | | Module 3 illustration | | | | | | Fill your own Module 3 |
| 1 | IQ Quarterly Paper on Databook | 3 | Appendix Figure 11 | 1_3 | | | | Module 6a illustration | | | | | | Fill your own Module 6a |
| 2 | Big Data Phenomenon… | 0 | | 2_0 | Final | 2019 | Thesis PDF | | http://www.theses.fr/2019CNAM1247 | Defense of thesis | 18/10/2019 | Paris, France | Jury & fans | Write an paper on the Databook |
| 2 | Big Data Phenomenon… | 1 | Chapter 2.3.1 | 2_1 | | | | First Databook concepts | | | | | | Write an paper on the Databook |
| 3 | Project X - Databook V1 | 0 | | 3_0 | Initial | 01/01/2022 | Excel | Presentation of the structure | In my computer here : xxx | Project Committee N°1 | 05/01/2022 | Company X headquarters | Project Manager, Sponsor, XXX | Continue to use the Databook |
| 4 | Project X - Intermediary Exploration Report | 0 | | 4_0 | Final | 01/01/2022 | PPT | Fists exploration conclusions | Sent my mail : xxx | Project Committee N°1 | 05/01/2022 | Company X headquarters | Project Manager, Sponsor, XXX | Exclude all personal data |
| 5 | Project X - Dashboard | 0 | | 5_0 | Version 1 | 01/01/2022 | PowerBI | Details of the exploration | In my computer here : xxx | Project Committee N°2 | 01/02/2022 | Company X headquarters | Project Manager, Sponsor, XXX | Include semantics in the Dashboard |
| 5 | Project X - Dashboard | 1 | Graph1 | 5_1 | Version 1 | 01/01/2022 | PowerBI | name of Graph 1 | | Project Committee N°2 | 01/02/2022 | Company X headquarters | Project Manager, Sponsor, XXX | |
| 6 | Project X - Final Exploration Report | 0 | | 6_0 | Final | 01/02/2022 | PPT | All exploration conclusions | Sent my mail : xxx | Project Committee N°3 | 01/03/2022 | Company X headquarters | Project Manager, Sponsor, XXX | Deploy the results |
| 5 | Project X - Dashboard | 2 | Graph2 | 5_2 | Version 2 | 01/02/2022 | PowerBI | name of Graph 2 | | Project Committee N°3 | 01/03/2022 | Company X headquarters | Project Manager, Sponsor, XXX | Include semantics in the Dashboard |
| … | … | … | … | … | … | … | … | … | … | … | … | … | … | … |

*Figure 17 – Module C: Exploration reports*

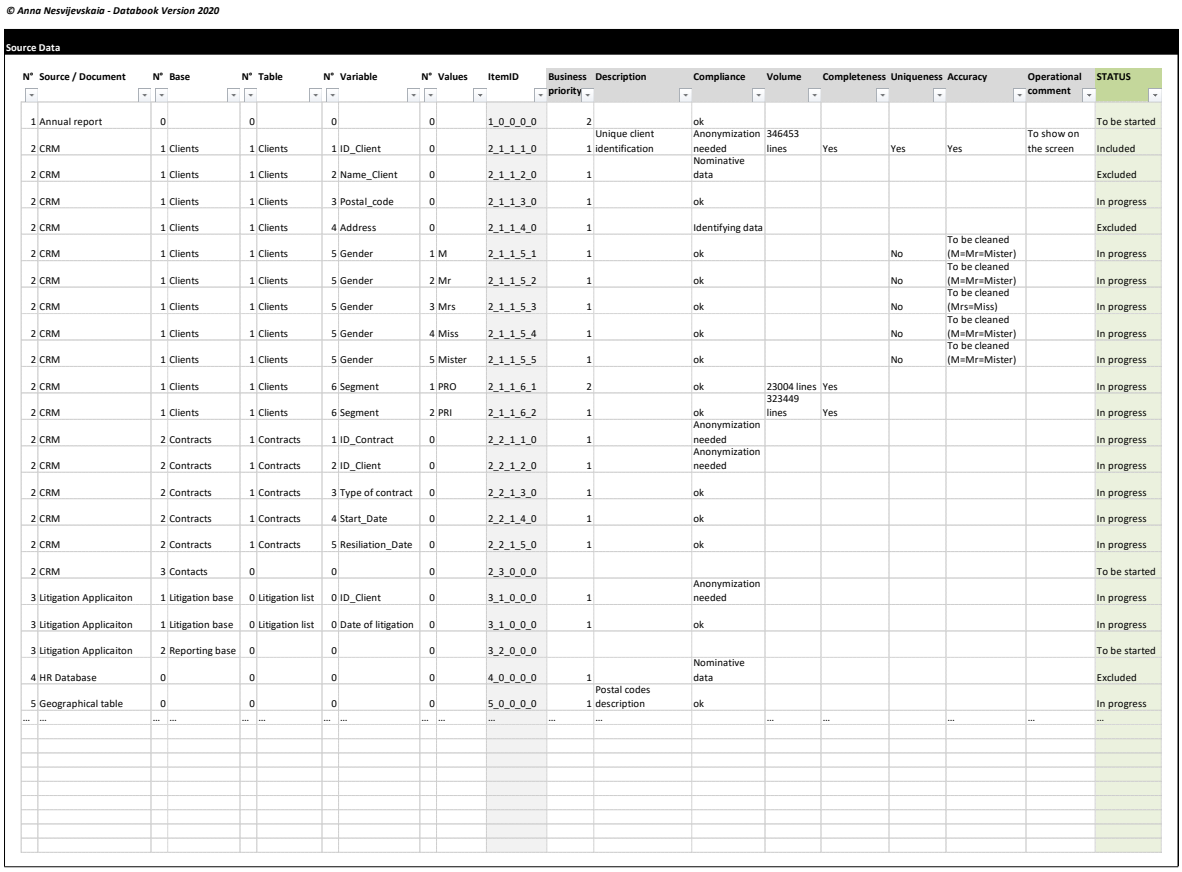
*© Anna Nesvijevskaia - Databook Version 2020*

**Exploration Perimeter**

| N° | Scope nature | N° | Business Concept | N° | Business Concept | ItemID | Business concept description | Data Source indentified | Data preparation comment | Impact on modellisation | Order of magnitude | Operational comment | STATUS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Phenomenon of interest | 1 | Active Churn | 0 | | 1_1_0 | Resiliation of all the contracts on a customer active demand | CRM | Date of the last resiliation demand | Active churn is the variable to predict | 15% per year | Retain active customers with high probability of churn | Included |
| 1 | Phenomenon of interest | 2 | Passive Churn | 1 | Non-conformity | 1_2_1 | Resiliation of all the contracts by the compagny | Litigation base | Identify customers with litigation | Exclude passive churn from active churn to predict | 2% per year | Never retain a "bad customer" | Included |
| 1 | Phenomenon of interest | 2 | Passive Churn | 2 | Passive Churn | 1_2_2 | Resiliation of all the contracts resulting from customer death | Impossible to source | Identify deceased customers | Exclude passive churn from active churn to predict | 1% per year | Avoid to call deceased customers | Excluded |
| 2 | Individuals (items to analyse) | 1 | Private clients | 0 | | 2_1_0 | Private Clients are prioritary to retain | CRM | Filter only on "PRI" customers | | 300000 private, 20000 professional | Concentrate the calls on BtoC vendors | In progress |
| 3 | Drivers | 1 | Client characteristics | 0 | | 3_1_0 | | CRM | | | | | To be started |
| 3 | Drivers | 2 | Contracts characteristics | 0 | | 3_2_0 | | CRM | | | | | To be started |
| 3 | Drivers | 3 | Historical commercial contacts | 0 | | 3_3_0 | | CRM | | | | | To be started |
| 4 | Time Perimeter | 1 | Prediction horizon | 0 | | 4_1_0 | | CRM | | | | | To be started |
| 4 | Time Perimeter | 2 | Historical clients | 0 | | 4_2_0 | All customers active in 2019 | CRM | | | | | In progress |
| 4 | Time Perimeter | 3 | Last contacts | 0 | | 4_3_0 | Consider only recent contacts (3 months) | CRM | Censure contacts on their date | | | | In progress |
| 5 | Eval. criteria | 1 | Target Precision | 0 | | 5_1_0 | | CRM | | | | | To be started |
| 5 | Eval. criteria | 2 | Target Interpretability | 0 | | 5_2_0 | | CRM | | | | | To be started |
| 5 | Eval. criteria | 3 | Time Saving for vendors | 0 | | 5_3_0 | | CRM | | | | | To be started |
| 6 | Key indicators | 1 | Turnover | 0 | | 6_1_0 | | CRM | | | | | To be started |
| 6 | Key indicators | 2 | FTE | 0 | | 6_2_0 | | Annual report | | | | | To be started |
| 6 | Key indicators | 3 | Salary costs | 0 | | 6_3_0 | | RH database | | | | | To be started |
| … | … | … | … | … | … | … | … | … | … | … | … | … | … |

*Figure 18 – Module 1: Perimeter*

 Nesvijevskaia, Anna (2021) DATABOOK: a standardised framework for dynamic documentation of algorithm design during Data Science projects, IASSIST Quarterly 45(2), pp. 1-34. DOI: https://doi.org/10.29173/iq989



© Anna Nesvijevskaia - Databook Version 2020

Source Data

| N° | Source / Document | N° | Base | N° | Table | N° | Variable | N° | Values | ItemID | Business priority | Description | Compliance | Volume | Completeness | Uniqueness | Accuracy | Operational comment | STATUS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Annual report | 0 | | 0 | | 0 | | 0 | | 1_0_0_0_0 | 2 | | ok | | | | | | To be started |
| 2 | CRM | 1 | Clients | 1 | Clients | 1 | ID_Client | 0 | | 2_1_1_1_0 | 1 | Unique client identification | Anonymization needed | 346453 lines | Yes | Yes | Yes | To show on the screen | Included |
| 2 | CRM | 1 | Clients | 1 | Clients | 2 | Name_Client | 0 | | 2_1_1_2_0 | 1 | | Nominative data | | | | | | Excluded |
| 2 | CRM | 1 | Clients | 1 | Clients | 3 | Postal_code | 0 | | 2_1_1_3_0 | 1 | | ok | | | | | | In progress |
| 2 | CRM | 1 | Clients | 1 | Clients | 4 | Address | 0 | | 2_1_1_4_0 | 1 | | Identifying data | | | | | | Excluded |
| 2 | CRM | 1 | Clients | 1 | Clients | 5 | Gender | 1 | M | 2_1_1_5_1 | 1 | | ok | | | No | To be cleaned (M=Mr=Mister) | | In progress |
| 2 | CRM | 1 | Clients | 1 | Clients | 5 | Gender | 2 | Mr | 2_1_1_5_2 | 1 | | ok | | | No | To be cleaned (M=Mr=Mister) | | In progress |
| 2 | CRM | 1 | Clients | 1 | Clients | 5 | Gender | 3 | Mrs | 2_1_1_5_3 | 1 | | ok | | | No | To be cleaned (Mrs=Miss) | | In progress |
| 2 | CRM | 1 | Clients | 1 | Clients | 5 | Gender | 4 | Miss | 2_1_1_5_4 | 1 | | ok | | | No | To be cleaned (M=Mr=Mister) | | In progress |
| 2 | CRM | 1 | Clients | 1 | Clients | 5 | Gender | 5 | Mister | 2_1_1_5_5 | 1 | | ok | | | No | To be cleaned (M=Mr=Mister) | | In progress |
| 2 | CRM | 1 | Clients | 1 | Clients | 6 | Segment | 1 | PRO | 2_1_1_6_1 | 2 | | ok | 23004 lines | Yes | | | | In progress |
| 2 | CRM | 1 | Clients | 1 | Clients | 6 | Segment | 2 | PRI | 2_1_1_6_2 | 1 | | ok | 323449 lines | Yes | | | | In progress |
| 2 | CRM | 2 | Contracts | 1 | Contracts | 1 | ID_Contract | 0 | | 2_2_1_1_0 | 1 | | Anonymization needed | | | | | | In progress |
| 2 | CRM | 2 | Contracts | 1 | Contracts | 2 | ID_Client | 0 | | 2_2_1_2_0 | 1 | | Anonymization needed | | | | | | In progress |
| 2 | CRM | 2 | Contracts | 1 | Contracts | 3 | Type of contract | 0 | | 2_2_1_3_0 | 1 | | ok | | | | | | In progress |
| 2 | CRM | 2 | Contracts | 1 | Contracts | 4 | Start_Date | 0 | | 2_2_1_4_0 | 1 | | ok | | | | | | In progress |
| 2 | CRM | 2 | Contracts | 1 | Contracts | 5 | Resiliation_Date | 0 | | 2_2_1_5_0 | 1 | | ok | | | | | | In progress |
| 2 | CRM | 3 | Contacts | 0 | | 0 | | 0 | | 2_3_0_0_0 | | | | | | | | | To be started |
| 3 | Litigation Applicaiton | 1 | Litigation base | 0 | Litigation list | 0 | ID_Client | 0 | | 3_1_0_0_0 | 1 | | Anonymization needed | | | | | | In progress |
| 3 | Litigation Applicaiton | 1 | Litigation base | 0 | Litigation list | 0 | Date of litigation | 0 | | 3_1_0_0_0 | 1 | | ok | | | | | | In progress |
| 3 | Litigation Applicaiton | 2 | Reporting base | 0 | | 0 | | 0 | | 3_2_0_0_0 | | | | | | | | | To be started |
| 4 | HR Database | 0 | | 0 | | 0 | | 0 | | 4_0_0_0_0 | 1 | | Nominative data | | | | | | Excluded |
| 5 | Geographical table | 0 | | 0 | | 0 | | 0 | | 5_0_0_0_0 | 1 | Postal codes description | ok | | | | | | In progress |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

*Figure 19 – Module 2: Source Data*

© Anna Nesvijevskaia - Databook Version 2020

Model structure

| N° | Agregated table | N° | Variable | N° | Values | ItemID | Business priority | Description | Business concept | Unit | Type of treatment | Volume | Base separation | Order of magnitude | Accuracy | Operational comment | STATUS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Main table | 0 | | 0 | | 1_0_0 | 1 | | | | | | | | | To run on a monthly basis | In progress |
| 1 | Main table | 1 | ID_Client | 0 | | 1_1_0 | 1 | Unique client identification number | Client | | Filtered on "PRI" and active contracts in time scope | 323449 lines | 60% learning 30% validation | | Yes | To show on the screen with associated client name | Included |
| 1 | Main table | 2 | Active_churn | 1 | YES | 1_2_1 | 1 | Resiliation of all the contracts on a customer active demand | Churn | | Flag: Resiliation date in time scope + client without litigation | 46900 lines | | 14,8% vs 15% | Yes | | In progress |
| 1 | Main table | 2 | Active_churn | 2 | NO | 1_2_2 | 1 | Active client or resiliation after litigation | Churn | | Flag: others | 276549 lines | | | | | In progress |
| 1 | Main table | 3 | Postal_code | 0 | | 1_3_0 | 1 | | Client characteristics | | Raw | | | | | | In progress |
| 1 | Main table | 4 | Department | 0 | | 1_4_0 | 1 | | Client characteristics | | Derivation | | | | | | In progress |
| 1 | Main table | 5 | Region | 0 | | 1_5_0 | 1 | | Client characteristics | | Derivation | | | | | | In progress |
| 1 | Main table | 6 | Border area | 0 | | 1_6_0 | 1 | | Client characteristics | | Derivation | | | | | | In progress |
| 1 | Main table | 7 | Gender | 1 | Male | 1_7_1 | 1 | | Client characteristics | nb | Cleaning | | | | | | In progress |
| 1 | Main table | 7 | Gender | 2 | Female | 1_7_2 | 1 | | Client characteristics | nb | Cleaning | | | | | | In progress |
| 1 | Main table | 8 | Contract_A | 0 | | 1_8_0 | 1 | | Contract characteristics | nb | Flag derivation | | | | | | In progress |
| 1 | Main table | 9 | Contract_B | 0 | | 1_9_0 | 1 | | Contract characteristics | | Flag derivation | | | | | | In progress |
| 1 | Main table | 10 | Contract_C | 0 | | 1_10_0 | 1 | | Contract characteristics | | Flag derivation | | | | | | In progress |
| 1 | Main table | 11 | Last_contract_purchase | 0 | | 1_11_0 | 1 | Number of months after the last substription | Contract characteristics | nb months | Date Derivation | | | | | | In progress |
| 1 | Main table | 12 | Client_Seniority | 0 | | 1_12_0 | 1 | Number of months after the first substription | Contract characteristics | nb months | Date Derivation | | | | | | In progress |
| 1 | Main table | 13 | Number_contacts | 0 | | 1_13_0 | 1 | | Contacts characteristics | nb | Aggregation | | | | | | To be started |
| 1 | Main table | 14 | Last_contact | 0 | | 1_14_0 | 1 | Number of days after the last contact | Contacts characteristics | nb days | Date Derivation | | | | | | To be started |
| 1 | Main table | 15 | Calls | 0 | | 1_15_0 | 2 | | Contacts characteristics | nb | Flag derivation | | | | | | To be started |
| 1 | Main table | 16 | Mails | 0 | | 1_16_0 | 2 | | Contacts characteristics | nb | Flag derivation | | | | | | To be started |
| 1 | Main table | 17 | Meetings | 0 | | 1_17_0 | 2 | | Contacts characteristics | nb | Flag derivation | | | | | | To be started |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

*Figure 20 – Module 3: Model Structure*

**Analytical results**

| N° | Type of model | N° | Model | N° | Model parameters | ItemID | Business priority | Description | AUC | Volume | Lift | Coverage | Model Interpretability | Operational comment | STATUS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Scoring | 1 | XGBoost | 0 | Default | 1_1_0 | 2 | See Final report | 0.7821 | 29481 | 2,5 | 69% | Difficult | Never show a score to a vendor | In progress |
| 1 | Scoring | 2 | Random Forest | 1 | Default | 1_2_1 | 2 | See Final report | 0.7048 | 32019 | 1,9 | 72% | Mean | Never show a score to a vendor | Excluded |
| 1 | Scoring | 2 | Random Forest | 2 | 4 trees, max depth 6 | 1_2_2 | 2 | See Final report | 0.7648 | 30245 | 2,4 | 70% | Mean | Never show a score to a vendor | In progress |
| 1 | Subgroup identification | 1 | Qfinder | 1 | Order 2, zscore, lift>1,5 | 1_1_1 | 1 | See Final report | NA | 30294 | 2,2 | 65% | Easy | Associate a lever to each profile | Excluded |
| 1 | Subgroup identification | 1 | Qfinder | 2 | Order 3, zscore, lift>2 | 1_1_2 | 1 | See Final report | NA | 14239 | 3,6 | 43% | Easy | Associate a lever to each profile | In progress |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

*Figure 21 – Module 4: Analytical results*



**Functional results**

| N° | Type of result | N° | Result | ItemID | Business priority | Description | Mean nb of calls per week | Target Precision | Time saving | Result Interpretability | Success rate | Pilot qualitative feedback | Turnover saved | STATUS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Existing | 0 | | 1_0 | 3 | Global retention plan | 35 | 18% | 0% | Yes | 25% | Vendors can not handle more than 35 calls per week | | In progress |
| 1 | Existing | 1 | Profile A | 1_1 | 3 | Recent clients (<2 months) | 20 | 15% | | Yes | 26% | | | In progress |
| 1 | Existing | 1 | Profile B | 1_1 | 3 | Parisian ancient clients | 15 | 22% | | Yes | 24% | | | In progress |
| 2 | Scoring | 2 | Score flag 1 | 2_2 | 2 | Clients with high churn risk (>75%) | 21 | 36% | | Difficult | 20% | Vendors do not know why thei have to call | | Excluded |
| 1 | Subgroup identification | 1 | Profile 1 | 1_1 | 1 | Recent customers (<4 months) with only 1 contract | 15 | 38% | | Yes | 35% | It works when vendors propose a satisfaction interview and cross-sell if satisfied | | In progress |
| 1 | Subgroup identification | 2 | Profile 2 | 1_2 | 1 | | | | | | | | | In progress |
| 1 | Subgroup identification | 3 | Profile 3 | 1_3 | 1 | | | | | | | | | In progress |
| 1 | Subgroup identification | 4 | Profile 4 | 1_4 | 1 | | | | | | | | | In progress |
| 1 | Subgroup identification | 5 | Profile 5 | 1_5 | 1 | | | | | | | | | In progress |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

*Figure 22 – Module 5: Functional results*



**Usage roadmap**

| N° | Action | N° | Task | ItemID | Business priority | Data ressources need | Time needed | Risk | Cost | Estimated benefits | Remaining uncertainties | STATUS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Analytic solution deployment | 0 | | 1_0 | 1 | See Databook final version | 10 days | Low | Low | High | | Delayed |
| 2 | Monitoring deployment | 1 | Add Profile monitoring | 2_1 | 1 | See Monitoring interface in PowerBI | 5 days | Low | Low | Medium | | Validated |
| 2 | Monitoring deployment | 2 | Add Vendor performance monitoring | 2_2 | 1 | See Monitoring interface in PowerBI | 7 days | Low | Low | High | Managerial levers ? | Validated |
| 3 | Maintainance anticipation | 1 | Once per year, look for new profiles | 3_1 | 1 | See Databook final version | 10 days | Medium | Low | Low | Churn stability | Delayed |
| 3 | Maintainance anticipation | 2 | Create a hotline for vendors | 3_2 | 1 | | 60 days / year | Low | High | Low | Train Marketing team ? | Excluded |
| 4 | Communication | 2 | Explain the profiles to vendors | 4_2 | 1 | | 30 days | Medium | Medium | High | | Delayed |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

*Figure 23 – Module 6a: Usage roadmap*

**Return of Experience**

| N° | Idea domain | N° | Idea | ItemID | Interested stakeholders | Remaining uncertainties | STATUS |
|---|---|---|---|---|---|---|---|
| 1 | Project Management | 1 | Better define who fill which item in the Databook | 1_1 | Data Scientists | Team organisation for the next project ? | To be shared |
| 2 | Data Analysis | 1 | Change the CRM interface in order to choose only between "Mr" & "Mrs" for each customer | 2_1 | IT / CRM team | Is it technically possible ? | To be shared |
| ... | ... | ... | ... | ... | ... | ... | ... |

*Figure 24 – Module 6b: Return of experience*

 Nesvijevskaia, Anna (2021) DATABOOK: a standardised framework for dynamic documentation of algorithm design during Data Science projects, IASSIST Quarterly 45(2), pp. 1-34. DOI: https://doi.org/10.29173/iq989

## Endnotes

[1] Anna Nesvijevskaia is Doctor of the Conservatoire National des Arts et Métiers in Science of Information and Communication and associate researcher at the laboratory DICEN Ile de France. She is also Partner at Quinten, expert firm in Artificial Intelligence, and can be reached by email: anna.nesvijevskaia@gmail.com (version: May 2021)

[2] Brizo_DS is a model of data project device fundamentally orientated towards value generation through the exploitation of usages, including knowledge capitalisation. It is intended to reduce the uncertainties inherent in these exploratory projects and is transferable to the scale of enterprise data project portfolio management. The device includes an adjusted CRISM_DM model, completed and reorganised in a Gantt chart to facilitate project management. Beyond the initials of the three reference indicators coordinating the trade-offs during the project between Benefits, Resources and Incertitudes, the name of the model is inspired by the Greek goddess Brizo, bearer of prophetic dreams and protector of sailors: the art of predicting the future through dreams is indeed a major asset for a Data Science project, by nature exploratory in an uncertain environment, and aiming to arrive 'safely', i.e. on a value-generating usage, whether anticipated or not.

[3] https://www.inria.fr/fr/pour-une-regulation-des-algorithmes