The BigMedilytics Blueprint: A Blueprint for Big Data and AI in Healthcare

Abstract
Setting up a Big Data or AI project and bringing it into production is always a challenging endeavor, starting from a business perspective, but also considering the technical or legal aspects. Following a proven structure, some guidelines, and recommendations can ease this process. Setting up such a project in the healthcare sector may introduce additional challenges, as it bears certain particularities, which need to be taken into account, such as patient consent for instance. This work presents the *BigMedilytics Blueprint*, a collection of guidelines and best-practices for Big Data and AI healthcare projects. Our work is based on the experience from twelve different BigMedilytics Big Data and AI pilots. Although the pilots varied widely with regards to use case, data used and medical problems to tackle, all pilots had to go through the same steps on an abstract level, such as defining a system architecture, (pre-)processing data or handling privacy issues. Putting this expert knowledge together with our experiences and recommendations, we created a Blueprint to support future developments in healthcare.

1. Introduction

BigMedilytics (Big Data for Medical Analytics) was a EU lighthouse project running from 2018-2021, addressing the application of Big Data and AI (artificial intelligence) techniques in different hospitals in Europe. The project was an innovation action, so the focus was primarily on applying already existing state-of-the-art Big Data and AI technologies to a variety of different use cases and domains in healthcare rather than inventing and researching new methods. The major goals of BigMedilytics was to show that state-of-the-art Big Data and AI technologies can significantly improve the productivity of the healthcare sector, by reducing costs, improving the quality through better patient outcomes and increasing access to healthcare facilities. To test this, the project was enrolled through twelve different pilots, covering three different themes, namely population health, chronic disease management and industrialization of healthcare services. All pilots addressed different use cases, different data sources and different problems and were rolled out in various countries across Europe. Based on the experience and lessons learned of our project, we present the *BigMedilytics Blueprint*, a collection of guidelines and best-practices to set-up a Big Data and AI project in the healthcare domain. The development of each pilot has been continuously monitored throughout the project by different work packages, involving technical, legal and business aspects.

2. BigMedilytics Blueprint

The ***BigMedilytics Blueprint*** is based on the experience and lessons learned from the different pilots and aligns them to an abstract level of common blocks, based on their similarities. Big Data and AI are both data driven techniques that have many aspects in common with data mining, so instead of creating a new process model , we build our blueprint upon the ***Cross-industry standard process for data mining***, also known as ***CRISP DM***. We use this well-established open standard process model, apply it to our healthcare domain, and incorporate our experiences and outcomes.

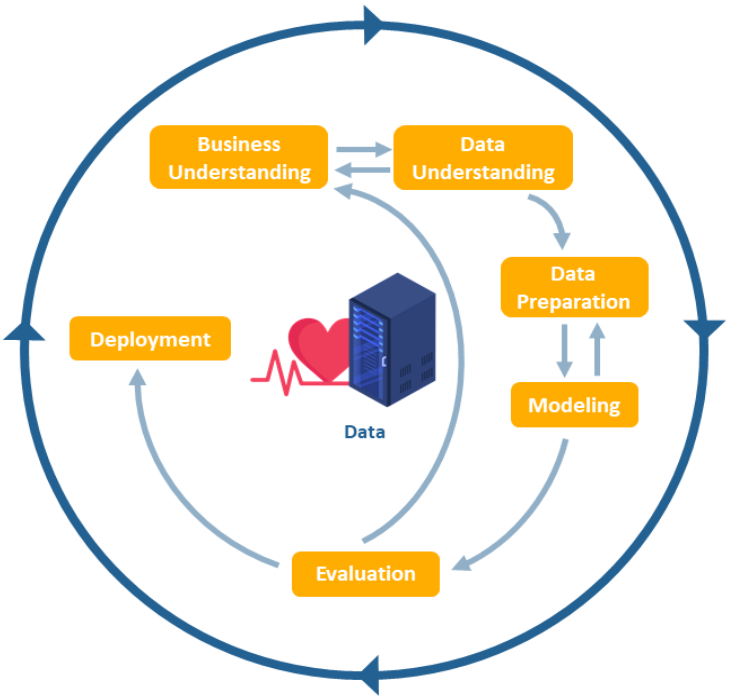

Figure 1: CRISP DM - Cross-industry standard process for data mining

The CRISP DM model is presented in Figure 1, and defines six basic phases (see Fig. 1) which will be described in the following sections. As the figure indicates, CRISP DM is not necessarily a fixed sequence of phases, as you can move back and forth if necessary. Moreover the arrows connecting the different phases are not the only way to move. The outer circle shows a clockwise movement and should highlight that the development of such a project is an iterative process. For instance, with the successful deployment and the finish of a project, previous problems, or new ideas could be then brought into a new use case, and the circle can restart. Also the general process may continue after the end of the project.

2.1 Business Understanding

An essential starting point in establishing a Big Data/AI project is a thorough understanding of the business, the business goal(s), assess current and target situation, define the technical task (Big Data/AI) and produce a project plan. Usually, this step results in a business model that describes how an organization creates, delivers, and captures value with the introduction of Big Data/AI technology.

Also with regards to business understanding the healthcare sector behaves slightly differently. In healthcare, not all stakeholders are “business” stakeholders, and a translation to our specific context is needed. Public, private for-profit, and private not-for-profit actors join forces to create value in the healthcare system. In addition, the value for patients is at least partly not only of monetary value but can be expressed in terms of quality of life, clinical outcomes, patient experience, and cost of treatment. Furthermore, value is also created for other stakeholders, such as healthcare professionals (better and faster decision-making, more efficient work processes), healthcare providers (higher productivity, better use of

resources), healthcare payers (better outcomes for lowest cost). Also value for the society can be discerned, e.g., a healthier population, increased labor productivity, lower health expenses (in total, or as percentage of the GDP).

In our project, all twelve pilots developed a business model to describe what value their BigData/AI innovation will create, which stakeholders, resources and partners are involved or affected and what activities are needed to create that value. Moreover, the process of business modeling helps to estimate which development costs and operational costs are related to the innovation, and how such costs may be covered in order to ensure that the innovation action does indeed create a positive value for the healthcare sector.

The classical “business model canvas”, a strategic management tool for developing and documenting new and existing business models, was as part of BigMedilytics adapted for the context of BigData/AI innovation in healthcare. The three main adaptations were:

- Acknowledge the multi-sided market in healthcare and acknowledge that these innovations create value not only for patients, but also for healthcare professionals, healthcare provider organizations, healthcare payers, and society at large.
- Acknowledge that “profit” is not the main driver for innovating in healthcare, but net positive value in terms of better outcomes and/or lower costs
- Acknowledge that in the Big Data/AI context, rules and regulations play a key role, and must be added as a separate payer in the business model canvas.

Business modeling is not a one-time activity, but often entails updates of the business model to finally meet the demands of all stakeholders involved in the activity. We recognized that it is rather a business modeling journey, moving from a business model of the pilot stage, to the stage of scaling up, and on to the phase of sustenance of the innovation.

In addition to that, compliance must be considered in this CRISP phase as well. The following requirements should be taken into account:

1. Integrate the Data Protection Officer and/or the Compliance Officer in the design team from early on.
2. Bear in mind that not only European legislation (GDPR) must be taken into account but also:
    - National laws regarding research in health.
    - Laws of other member states in trans-European projects.
    - Laws associated with international data transfers.
    - “Soft Laws” (such as Guidelines from EDPB, national data protection authorities, etc.).
3. In case of multi-partner projects, roles should be defined in terms of GDPR (controller-joint controller-processor) and data use or access (data provider-data consumer). These roles can then be used to define future agreements between the parties (joint-controller's agreement, processor's agreement, data sharing agreement).
4. Access to health system data requires in addition consideration of ethical procedures (research protocol, ethical protocol, informed (ethical) consent, ethics committee approval) as well as whether or not a clinical trial process of a medical device is necessary under Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices.

5. Integrate the recommendations of the EU High-Level Expert Group on Artificial Intelligence, the ethical principles of the OECD, and the ethical standards developed in the USA.

2.2 Data Understanding

The next phase is data understanding, which involves the sighting of the available data, as well as the data collection. The data has to be explored and examined, particular characteristics identified, and the quality of the data assessed. This analysis needs to ensure if the data actually can help to tackle the actual problem. If not, possibly the aims within the business understanding have to be amended or other data sources need to be found or incorporated.

Particularly in the medical domain this phase can be quite challenging: Getting access to data can be a time consuming step, for example in case the data scientist is not associated with the data provider, or is located in a different country. Bear also in mind, that access to data requires having been able to comply with regulations in a way that allows you to demonstrate:

- The legitimate origin of the data.
- The patient's consent, where necessary, and the guarantee of his or her rights.
- Plus the different requirements that each country impose on retrospective and prospective studies.

In addition, the data typically cannot be easily understood without medical (or additional) expertise from others. Therefore interdisciplinary work is essential, which typically requires extra time, as different stakeholders use different terminology and have a different understanding of the problem.

Furthermore, already in the phase of data understanding, the problem of 'anonymisation' emerges. The position of the data protection authorities of the EU in Working Party Opinion 5/2014 on anonymisation techniques is clear: irreversible anonymisation must be achieved. Afterwards, each data protection authority usually publishes its own guidelines.

The project has learned several lessons:

- De-identification and anonymisation are not synonymous.
- Resources must be foreseen to verify the risk of re-identification and to perform second anonymization when health system data have been "de-identified".
- A double layer of additional measures should be implemented:
    - Technical: a controlled platform environment with appropriate security measures should be designed. Among these, traceability of users is particularly relevant.
    - Legal: data sharing agreements, non-identification commitments of partners and/or users of the platform should be formalised.

2.3 Data Preparation

In the next phase the data has to be selected and integrated. Based on the previous analysis, data has to be cleaned and often converted into a different format, to be used in the next stage. However, although data preparation as well as data understanding might sound trivial, these steps often take the most amount of time within the overall project.

Often, in the healthcare domain, the creation and selection of digital data, such as electronic medical records, have grown over time, and so the quality, even of a single data source, may vary over time. However, missing and/or wrong information in the data is a very common phenomenon and needs to be dealt with. Moreover, possible errors and inconsistencies might be not directly obvious for an outsider, which highlights the need for interdisciplinary work on data cleansing.

The limitation to work with the data only on-site, within a secure environment of the data provider, may also create additional challenges which need to be taken under consideration (e.g. limited user rights, old infrastructure).

2.4 Modelling

The next phase in CRISP-DM describes the modelling phase. In this phase it needs to be decided, which AI algorithms to use and which test procedure to follow. For example, we may decide to train a classifier based on the mathematical concept of linear regression and use a 10-fold cross validation training strategy, i.e. we will perform the fitting procedure ten times each time randomly selected a training set of 90% of the data, keeping 10% of the data for validation.

Regarding Big Data we also need to decide if we follow a batch (the data is collected and stored first and then in one or more batches analysed) or streaming approach (data is generated and analysed continuously). In this respect and especially together with AI, a combination of both approaches is possible. For instance, a model may be trained in a batch, but used with streaming data.

Since CRISP-DM focus on Data Mining there are certain aspects of modelling that from a software engineering part need to be taken into account with respect to modelling but that are not addressed by CRISP-DM, such as modelling of the software architecture or domain modelling that may need to be considered in a healthcare project.

According to the experience within our twelve pilots a large range of different lessons learned have been made. One of the lessons learned is that due to the rather specialized data also the evaluation of the result of model training can only be interpreted successfully by a team of both medical experts and data scientists.

Modeling involves the application of the data protection by design principle and usually starts with the data protection impact assessment from which the risks to be eliminated, mitigated and reduced will be derived. Lessons learned from the project in this area include:

- The emergence of lists of criteria and methodologies of data protection authorities that are similar, but not exactly the same. For example, the CNIL offers a multilingual tool with a very open methodology while the Spanish AEPD defines a comprehensive checklist of controls with a complementary guide of controls for AI.
- It is essential to train the whole staff.
- To take into account the requirements of the Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices.
- Decisions must be taken on the development of an ethical impact analysis on AI.

2.5 Evaluation

In this phase an evaluation is performed if the previously defined business success criteria (e.g. KPIs) are fulfilled and supported by applying Big Data and AI technology. If this is not the case, a thorough analysis of all previous phases is required. The reasons for not reaching the final goals may be manifold. For example, not enough data is available for training a model successfully, too much missing data, or relevant data has not been collected, etc. When all fails it may even be necessary to amend the business success criteria in the Business Understanding phase. Moreover, this phase will review the overall work (e.g., were all steps properly executed? Summarize findings etc), and determine the next steps.

To measure the impact of Big data innovations, business and research projects frequently rely on key performance indicators (KPIs). In BigMedilytics the three core KPIs are: improving the quality of the healthcare system by improving the effectiveness, providing more people access to the healthcare system but at the same time reducing the costs by applying Big Data and AI technology. But which KPIs capture the impact of Big data innovations on healthcare in general? In the BigMedilytics project we learned the following: First order KPIs outline how Big Data innovations change the information provided. Second order KPIs shed light on how Big Data innovations change the decision-making process. Third order KPIs capture the perceived usefulness of the data and in how far value is attributed to the information. And fourth order KPIs reveal whether and in how far Big Data innovations might affect patient experience, population health, costs, and professional satisfaction in the longer run.

But KPIs might not only differ in their order effects but also in their function. Some KPIs can have a temporary function and can be revoked and revised as one sees fit. Core KPIs remain relevant over time. Consequently, using KPIs is an iterative, recursive process of moving back-and-forth between finding out which indicators are feasible, acceptable, measurable, and informative.

While assessing how Big Data affects long-term health outcomes, one needs to keep in mind that long-term outcomes are dependent on a sequence of decisions and exogenous factors. In how far changes in long-term outcomes can causally be attributed to Big data innovations is therefore dependent on how rigorously one can establish the counterfactual scenario of what would have happened if the Big data innovation had not been developed and implemented.

For evaluating Big Data and AI in healthcare we also should keep in mind that the impact of applying the trained models for healthcare often only can be validated with new patients, i.e. over a longer period of time.

2.6 Deployment

The final phase defines the deployment of the model. This includes planning, role-out, monitoring and maintenance. Planning of the deployment should already be considered in the business understanding phase, since the deployment of the Big Data and AI technology may create costs, such as additional hardware, for example, buying a cluster in for handling Big Data or the purchase of a fast computer with GPU support for training AI models. In addition, organisational structures may need to be changed or adapted, for example, if

telemedicine data is automatically monitored, it requires a team of medical experts to check and decide if further actions need to be taken in case the system triggers an alarm.
Already in the phase, an ethical and legal governance model must be implemented that should be able to:

- Ensure that the AI system respects the values of human rights, human centric-approach and human oversight, explainability and fairness.
- Ensure transparency at two levels:
  - Internal:
    - by clearly identifying and notifying the roles and responsibilities of users, particularly with regard to those uses involving the adaptation of decision-making processes subject to the risk of bias.
    - ensuring the involvement of users in continuous improvement;
  - External:
    - Providing adequate information to patients
    - Designing dialogue methodologies with all relevant stakeholders
- Regularly audit the system from a legal and security point of view.
- Maintains adequate incident management procedures, particularly those relating to security breaches and those identifying reliability issues in AI results.

Governance models may involve, depending on the characteristics of the entity:
- Adopting and implementing ethical codes.
- Promoting and adopting the codes of conduct and/or certifications provided for in the GDPR.
- Defining governance bodies for the systems.

In this phase also the way how information is presented and reported is also of importance, this includes also such as the development of a graphical user interface (GUI). Additionally, planning the deployment should also include a fall-back strategy, in case the new technology is not working as expected and interferes with the day-to-day operation. This applies in particular to disruptive technologies, such as asset management technologies that replace older techniques.

The actual deployment of the developed Big Data/AI technology (the role-out) should first happen after the model at the IT infrastructure using the model and attached components such as a newly developed GUI has been thoroughly tested and evaluated. In contrast to most companies at which a role-out of a new system can take place after a (partial-) shutdown of some services, in hospitals usually the the IT-system has to be running all the time and must only be interrupted for a short period and therefore should not be interfered by erroneous systems. In case a backup twin-system is available it would be a good idea to test the system there first.

The team involved in using the new system should be informed about the role-out timely.
Also the outcomes of the model need to be communicated with the decision makers, which were particularly included in the business understanding phase.

Monitoring is necessary in order to see how the newly integrated Big Data/AI technology behaves over time. With regards to Big Data an aspect to monitor may for example be how fast the assigned storage is filled. With regards to AI models an aspect to monitor can be on

how the prediction or classification based on the new model behaves over time with new and thereby unseen data. In a hospital this data is usually based on patients (vitals, lab values etc.) so monitoring may be necessary over a longer period of time because only a few new patients are being hospitalized every day.

Maintenance may mean that a trained AI model is outdated and needs to be replaced either because the algorithm calculating the model has been significantly improved or due to a larger sum of new data collected a new training, testing and deployment of a model makes sense. With regards to Big Data the hardware infrastructure may be evaluated and depending on the result a Big Data cluster needs more compute power.