7 Stages of Machine Learning — A Framework

Introduction

Today, most businesses generate a lot of data, and Machine Learning (ML) is the tool to make use of it and unlock significant value for the business.

However, every company and project is different, so people working on data science problems need to have a strong understanding of all important steps for successful Machine Learning. Only then, it is possible to identify and build end-to-end ML solutions that can actually make a business impact.

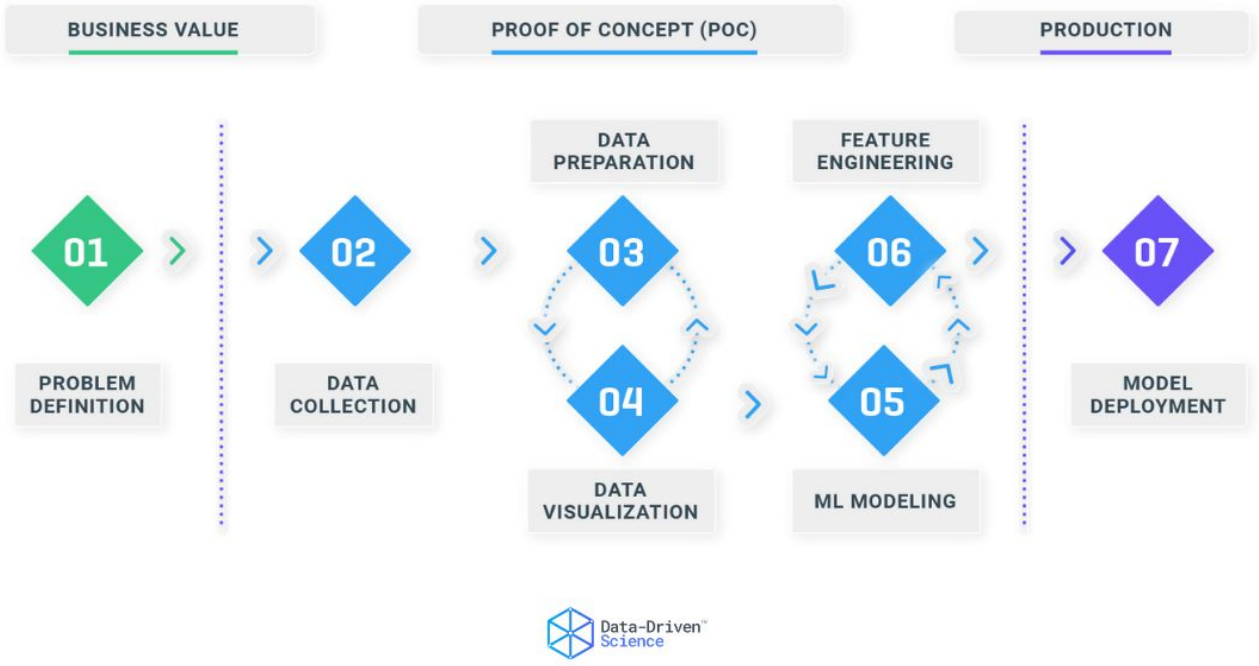
In this blog post, we walk through our Machine Learning framework that will provide a clear and effective structure for any ML project. It will help solving complex problems by having a simple step-by-step “recipe”.

The goal of the 7 Stages framework is to break down all necessary tasks in Machine Learning and organize them in a logical way. At the end, the framework acts as a general process that can be universally applied to any project independently of industry and type of business.

Data-Driven Science (DDS) will also use that framework for its upcoming comprehensive Machine Learning online course published on Udemy — stay tuned. We will go deep into each stage and give you everything that is needed to complete Data Science projects successfully.

The 7 Stages of Machine Learning:

1. Problem Definition
2. Data Collection
3. Data Preparation
4. Data Visualization
5. ML Modeling
6. Feature Engineering
7. Model Deployment



These 7 stages are the key steps in our framework. We have categorized them additionally into groups to get a better understanding of the larger picture.

The stages are grouped into 3 phases:

1. Business Value
2. Proof of Concept (POC)
3. Production

Phase 1 — Business Value

It is absolutely crucial to adopt a business mindset when thinking about a problem that should be solved with Machine Learning — defining customer benefits and creating business impact is top priority. Domain expertise and knowledge is also essential as the true power of data can only be harnessed if the domain is well known and understood.

Phase 2 — Proof of Concept (POC)

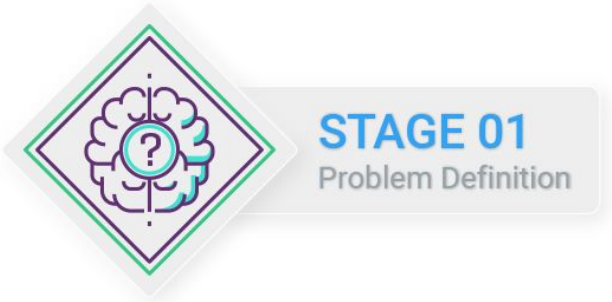
Proof of Concept (POC) is the most comprehensive part of our framework. From Data Collection to Feature Engineering, 5 stages of our ML framework are included here. Core of any POC to test an idea in terms of its feasibility and value to the business. Also, questions around performance and evaluation metrics are answered in that phase. Only a strong POC that delivers business value and is feasible allows one putting the ML Model into production.

Phase 3 — Production

In the third phase, one is taking the ML model and scaling it. The goal is to integrate Machine Learning into a business process solving a problem with a superior solution compared to, for example, traditional programming. The process of taking a trained ML model and making its predictions available to users or other systems is known as model deployment. Lastly, it is also essential to iterate on the ML model over time to improve it.

7 Stages of Machine Learning

1. Problem Definition



The first stage in the DDS Machine Learning Framework is to define and understand the problem that someone is going to solve. Start by analyzing the goals and the *why* behind a particular problem statement. Understand the power of data and how one can use it to make a change and drive results. And asking the right questions is always a great start.

Few possible questions:

- What is the business?
- Why does the problem need to be solved?
- Is a traditional solution available to solve the problem?
- If probabilistic in nature, then does available data allow to model it?
- What is a measurable business goal?

2. Data Collection

Once the goal is clearly defined, one has to start getting the data that is needed from various available data sources.

At this stage, some of the questions worth considering are:

- What data do I need for my project?
- Where is that data available?

- How can I obtain it?
- What is the most efficient way to store and access all of it?

There are many different ways to collect data that is used for Machine Learning. For example, focus groups, interviews, surveys, and internal usage & user data. Also, public data can be another source and is usually free. These include research and trade associations such as banks, publicly-traded corporations, and others. If data isn't publicly available, one could also use web scraping to get it (however, there are some legal restrictions).

3. Data Preparation

The third stage is the most time-consuming and labor-intensive. Data Preparation can take up to 70% and sometimes even 90% of the overall project time. But what is the purpose of this stage?

Well, the type and quality of data that is used in a Machine Learning model affects the output considerably. In Data Preparation one explores, pre-processes, conditions, and

transforms data prior to modeling and analysis. It is absolutely essential to understand the data, learn about it, and become familiar before moving on to the next stage.

Some of the steps involved in this stage are:

- Data Filtering
- Data Validation & Cleansing
- Data Formatting
- Data Aggregation & Reconciliation

4. Data Visualization

Data Visualization is used to perform Exploratory Data Analysis (EDA). When one is dealing with large volumes of data, building graphs is the best way to explore and communicate findings. Visualization is an incredibly helpful tool to identify patterns and trends in data, which leads to clearer understanding and reveals important insights. Data Visualization also helps for faster decision making through the graphical illustration.

Here are some common ways of visualization:

- Area Chart
- Bar Chart
- Box-and-whisker Plots
- Bubble Cloud
- Dot Distribution Map
- Heat Map
- Histogram
- Network Diagram
- Word Cloud

5. ML Modeling

Finally, this is where *'the magic happens'*. Machine Learning is finding patterns in data, and one can perform either supervised or unsupervised learning. ML tasks include regression, classification, forecasting, and clustering.

In this stage of the process one has to apply mathematical, computer science, and business knowledge to train a Machine Learning algorithm that will make predictions

based on the provided data. It is a crucial step that will determine the quality and accuracy of future predictions in new situations. Additionally, ML algorithms help to identify key features with high predictive value.

6. Feature Engineering

Machine Learning algorithms learn recurring patterns from data. Carefully engineered features are a robust representation of those patterns.

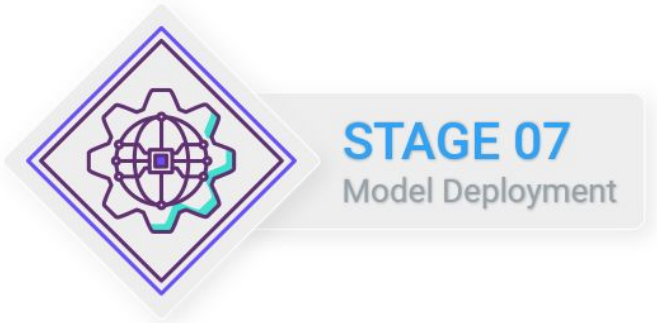
Feature Engineering is a process to achieve a set of features by performing mathematical, statistical, and heuristic procedures. It is a collection of methods for identifying an optimal set of inputs to the Machine Learning algorithm. Feature Engineering is extremely important because well-engineered features make learning possible with simple models.

Following are the characteristics of good features:

- Represents data in an unambiguous way
- Ability to captures linear and non-linear relationships among data points

- Capable of capturing the precise meaning of input data
- Capturing contextual details

7. Model Deployment



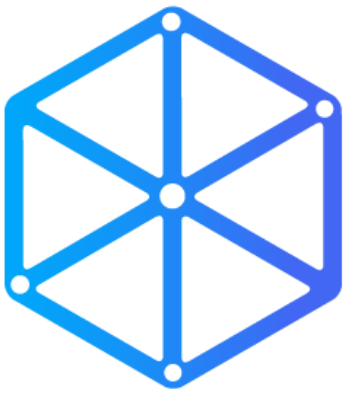
The last stage is about putting a Machine Learning model into a production environment to make data-driven decisions in a more automated way. Robustness, compatibility, and scalability are important factors that should be tested and evaluated before deploying a model. There are various ways such as Platform as a Service (PaaS) or Infrastructure as a Service (IaaS). For containerized applications, one can use container orchestration platforms such as Kubernetes to rapidly scale the number of containers as demand shifts.

Another important part of the last stage is iteration and interpretation. It is critical to constantly optimize the model and pressure test the results. At the end, Machine Learning has to provide value to the business and make a positive impact. Therefore, monitoring the model in production is key.

Conclusion

This was an overview about '**The 7 Stages of Machine Learning**' — a framework that helps to structure the typical process of a ML project. The idea is to equip practitioners with a template that can be universally applied and simplifies the process from idea to implementation.

"Being a data scientist is not only about data crunching. It's about understanding the business challenge, creating some valuable actionable insights to the data, and communicating their findings to the business." — Jean-Paul Isson



www.datadrivenscience.com