# Machine Learning Engineer Assignment

ZDP-AI Team *
Data & Analytics | ZEISS Digital Partners
Carl Zeiss AG

2023–06–30

[*] All cases and data in this assignment are entirely fictional.

# 1 Introduction

The ZEISS Digital Partners unit based in Munich, Oberkochen and Jena is realizing cutting-edge customer-centric digital solutions jointly with the ZEISS business units. We are a very young and diverse growing team of Data Scientists, Machine Learning Engineers and MLOps Engineers. The team is being built to support strategic business units of ZEISS in their journey towards digitalization. We from ZEISS Digital Partners believe that one can see Data Science as software engineering with a data component. We focus on building our solutions in cloud working with both structured and unstructured data using Azure cloud services, terraform and Python Machine Learning eco system. We strive to have standards for open-source code and to master our tools. The aim of the following test is to assess some of the areas of expertise we think are necessary for a Machine Learning Engineer at ZEISS Digital Partners.

### How to take this test?

This test consists of two parts:

- Part 1: Productionise a simple Data Science case
- Part 2: Cloud architecture

Part 1 is a task more related to the role of an Machine Learning Engineer, Part 2 is more related to the role of an MLOps Engineer. Pick at least one of the tasks and provide a solution to it, dependent on which position you are applying for. Feel free to also provide solutions for the other task. Please spend no more than **five (5)** hours on the assignment. Provide all writing, scripts, code, scans of written work you deem to be important for us to evaluate your solution. Please send solution files back by email or with another file sharing protocol. We understand that questions are open and time is limited, thus, try to organize your answers and keep in mind that we value quality over quantity.

## Note

Here are a few things to consider while answering our questions:

- Machine Learning Engineering:
  1. What is the business case and KPIs to optimize?
  2. What kind of data is available?
  3. What is to be modeled and predicted?
  4. Are labels needed?
  5. What is the machine learning objective?
  6. Code structure?
  7. What would you do with the Jupyter notebooks?
  8. Usage of quickly changing and incompatible external libraries (Pandas, Tensorflow)
  9. What types of open source packages might you take advantage of?
  10. Which code quality standards would you enforce?
  11. Monitoring result quality and managing machine learning models
  12. CI/CD processes/pipelines
  13. Testing. Unit testing.
  14. Throughput performance bottlenecks
  15. Logging in code
- MLOps Engineering:
  1. Operations
  2. Failure Redundancy
  3. Scaling
  4. Privacy Concerns
  5. User Access and Security

## 2 Assignment

### Part 1. Simple Data Science case

This task is more related to the role of an a Machine Learning Engineer . Thousands of our clients around the world use our hardware devices to conduct fundamental research. Our team needs to reduce the downtime and optimize the periodic maintenance operations for these devices with the help of machine learning algorithms. These devices are collecting log data about their internal state of operation, which we analyse with the help of machine learning.
In this part you are given a simple example Data Science use case: classification based on sensor data. Imagine a Data Scientist came up with a proof of concept in a notebook and her solution seem to solve the business problem at hand. You as a Machine Learning Engineer are joining the project at that stage to help productionise this use case and create cloud infrastructure.
You can assume that

- The log data are being written periodically to a text file on a computer connected to each device.
- Data are being collected and saved in the Data Lake (e.g. Azure Blob Storage, Amazon S3, Google Cloud Storage) by another team for you. You don't have to think about direct connections and protocols to those devices.
- You have all necessary authorisations and permissions to do what you want
- Data Science code is committed to git (say Azure DevOps) and you are given access to the repository
- The infrastructure should have at least 2 environments for development and production

Note, that the given code is intentionally kept simple and trivial, as we would like you to focus on maintenance, software and Machine Learning Engineering side of things, not the model itself. You are, however, encouraged to make additional and more ambitious assumptions in your answers, such as: more data, more complex Data Science algorithm, etc. ... For this example, we chose sensor data because it is close to the real problems that our team solves at ZEISS.
In **data_science_problem** folder you will find:

- model.ipynb (Data Science code) [exist also as model.html]
- requirements.txt (Python environment)
- data/historical_sensor_data.csv (sample historical data)
- data/real_sensor_data.csv (sample inference data)

### How to run this code?

Assuming that you have pip and Python $3.8$ on your system or virtual environment:

```
#!/bin/bash
python -m pip install --upgrade pip
pip install -r requirements.txt
# install kernel, that we named ds
python -m ipykernel install --user --name ds --display-name "ds"
# run jupyter server locally and navigate to the notebook
jupyter notebook
```

*If you don't want to run the code you can explore model.html in the browser of your choice.*

### Questions

Taking the provided use case, prepare a production-ready code example, specifically concentrating on covering the following aspects:

- How would you assess if a code is *ready* for production? What do you think about the given code in model.ipynb?
- Would you change anything in the code? How and why? (Provide code examples)
- How would you rearrange code for production system? (Provide code examples)
- Would you add anything to the code? (Provide code examples)
- Discuss training and inference data provided.
- How would you organize the codebase management. E.g. how do you deliver code to the production environment? (make any assumption about code management system, that you like e.g. GitHub, Bitbucket, GitLab, Azure Repos)

### Optional part

- Would you consider any alternatives to the used algorithm and if so, why?
- Discuss pros and cons of your chosen algorithms as well as evaluation metrics.
- Discuss Big O notation of the algorithms, its algorithmic and memory complexity.
- List libraries in various programming languages that you know or have used that implement those algorithms.
- Would your setup change if these were Databricks notebooks with Spark compute clusters behind them, written in PySpark?

## Part 2. Cloud architecture

This task is more related to the role of an MLOps Engineer.
For this part, we continue working on the sensor data. Without loss of generalization, let's assume the code consists of two models, each having a notebook with data preprocessing and a notebook with training and inference code. One model is classifying devices into faulty and non-faulty, and another detects anomalies in sensor data. Training data are available in .parquet format files. We have 1000 devices sending data from 100 sensors each every day. Log files contain sensor data for every minute. The customers would like to see the results on a webhosted dashboard and be somehow alerted about anomalies.
Sketch an architecture diagram and list cloud services and resources suitable for this case. You are free to make assumption about cloud providers. Consider having multiple Machine Learning models working simultaneously. Think of training, evaluation, and monitoring to be implemented.

### Questions

- How would a cloud architecture look like and why?
- What tools and implementation approach would you take?
- Which pipelines would you build there? How?
- Which assumptions have you made?
- How can code management be done?
- Discuss maintenance aspects (How to ensure the system will work for 1 year? 10 years?)

### Optional part

- How would you integrate the near real time model updates (online training) with the once-a-day slow full-update (batch training)?
- Discuss organisational aspects and challenges of a multi team setup.
- How would your cloud architecture change if some data were external and collected outside the ZEISS organisation?

**Good luck! We are looking forward to your answers.**