DataScientest

This document explains what is expected of you during the realization of your project. You will have three main renderings:

- **report 1** : introduce, describe, visualize and manipulate your dataset;
- **report 2** : modeling and solving the problem using Machine Learning techniques;
- **final report** : made up of the first two reports, it embodies your vision and your work in relation to your project by integrating a conclusion and an opening.

This methodology is therefore intended to advise you in the production of renderings, but does not represent specific instructions that you must follow. You are in control of your work, your analyses, as well as the format that your various reports will take.

Report 1: exploration, data visualization and data pre-processing report

Introduction to the project

Context

- Context of the project's integration into your business.
- From a technical point of view.
- From an economic point of view.
- From a scientific point of view.

Objectives

- What are the main objectives to be achieved? Describe in a few lines.
- For each member of the group, specify the level of expertise around the problem addressed?
- Have you contacted business experts to refine the problem and the underlying models? If yes, detail the contribution of these interactions.
- (Are you aware of a similar project within your company, or in your entourage? What is its progress? How has it helped you in the realization of your project? How does your project contribute to improving it?).

DataScientest

Understanding and manipulation of data

Framework

- Which set(s) of data(s) did you use to achieve the objectives of your project?
- Are these data freely available? If not, who owns the data?
- Describe the volume of your dataset?

Relevance

- Which variables seem most relevant to you with regard to your objectives?
- What is the target variable?
- What features of your dataset can you highlight?
- Are you limited by some of your data?

Pre-processing and feature engineering

- Did you have to clean and process the data? If yes, describe your treatment process.
- Did you have to carry out normalization/standardization type transformations of your data? If yes, why?
- Are you considering dimension reduction techniques in the modeling part? If yes, why?

Visualizations and Statistics

- Have you identified relationships between different variables? Between explanatory variables? and between your explanatory variables and the target(s)?
- Describe the distribution of these data, distribution, outliers.. (pre/post processing if necessary)
- Present the statistical analyzes used to confirm the information present on the graphs.
- Draw conclusions from the elements noted above allowing them to project themselves into the modeling part

Assessment methods:
Reconstituted professional situation: from a set of company data, the candidate must implement various pre-processing and data augmentation to make them usable through machine learning techniques.

DataScientest

Report 2: modeling report

Stages of the project

Classification of the problem

- What kind of machine learning problem is your project like? (classification, regression, clustering, etc)
- What task does your project relate to? (fraud detection, facial recognition, sentiment analysis, etc)?
- What is the main performance metric used to compare your models? Why this one?
- Did you use other qualitative or quantitative performance metrics? If yes, detail it.

Model choice and optimization

- What algorithms have you tried?
- Describe which one(s) you selected and why?
- Did you use parameter optimization techniques such as Grid Search and Cross Validation?
- Have you tested advanced models? Bagging, Boosting, Deep Learning... Why?

Interpretation of results

- Have you analyzed the errors in your model?
- Did this contribute to his improvement? If yes, describe.
- Have you used interpretability techniques such as SHAP, LIME, Skater... (Grad-CAM for Deep Learning...)
- What has (or not) generated a significant improvement in your performance?

Assessment methods:
Professional scenario: based on a proposed solution, the candidate will have to produce a summary report including: the explanation of the choices of AI

solutions implemented, the interpretation of the results, the evaluation of the reliability of the algorithms and an optimization proposal.

Final report :

Conclusion drawn

Difficulties encountered during the project

- What was the main scientific obstacle encountered during this project?
- For each of the following points, if you encountered difficulties, detail how they slowed you down in setting up your project.
- Forecast: tasks that took longer than expected, etc.
- Datasets: acquisition, volumetry, processing, aggregation, etc.
- Technical/theoretical skills: timing of skill acquisition, skill not offered in training, etc.
- Relevance: of the approach, model, data, etc.
- IT: storage power, computational power, etc.
- Other

Report

- Detail what was your main contribution to achieving the project's goals.
- Have you changed the model since the last iteration? If yes, provide details.
- Present the results obtained and compare them to the benchmark
- For each of the project's goals, detail how they were achieved or not.
- If they have been reached, in which process(es) can your model fit? Detail.

Continuation of the project

- What avenues for improvement do you suggest to increase the performance of your model?
- How has your project contributed to an increase in scientific knowledge?

Bibliography

- What bibliographical elements (research articles, blog, books, etc.) did you rely on to carry out your project?

Appendices

- Gantt diagram.
- Description of code files.