MLOps project

The aim of the project is to simulate the real world process of deploying machine learning models, using the concepts that we have discussed during the classes.

Project Deliverables

Report with maximum of 6 pages:

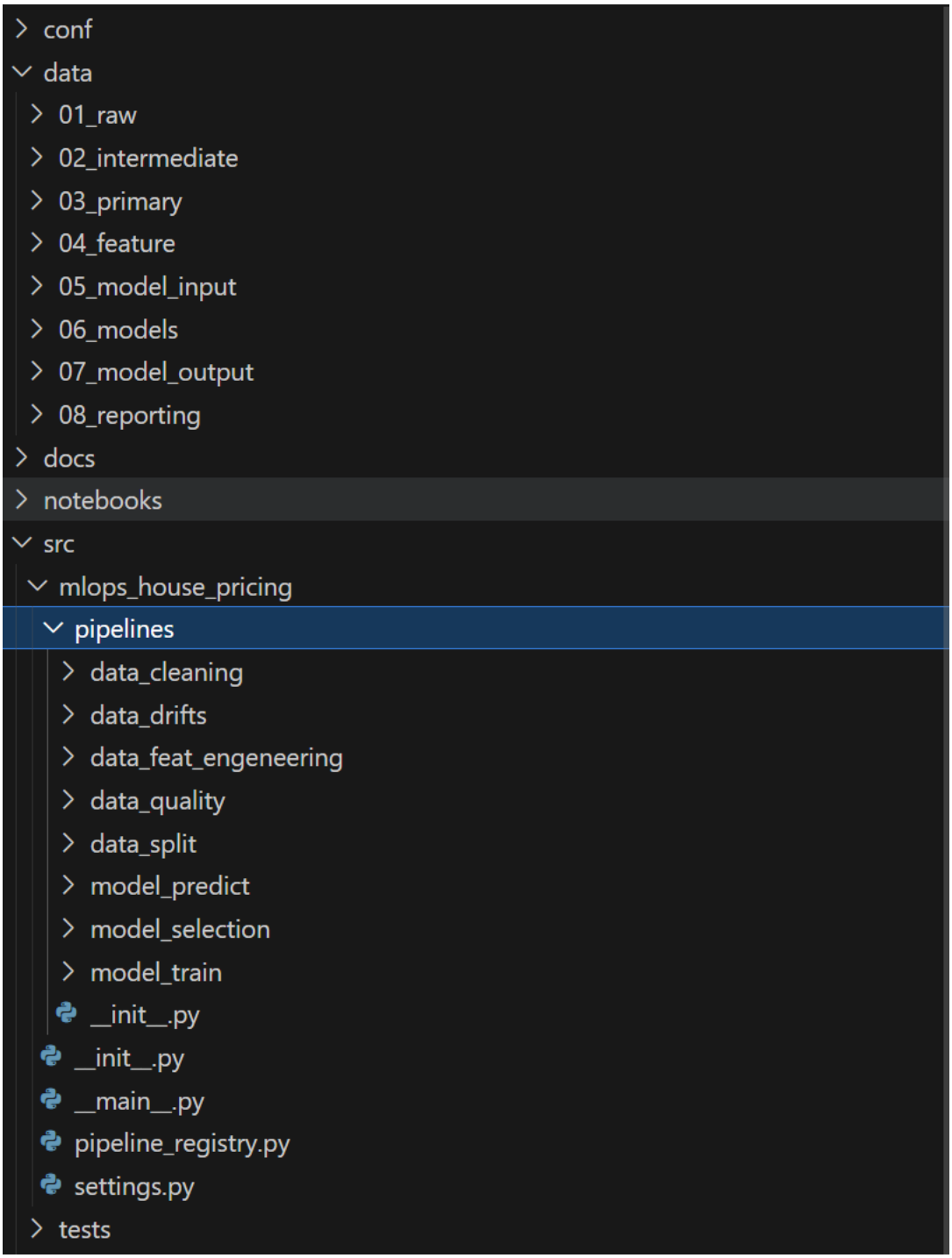
- Explain why you choose that data and what you try to achieve. Define **your success metrics.**
- Project planning: how you organized and scheduled the different steps (you can be inspired by sprints in the agile methodology).
- **Results and conclusions** from data exploration and data modelling (plots, feature importance, explainability).
- Since this is a proof of concept, discuss how this would be implemented in production and what are the **advantages of the technologies used, risks and possible mitigations**, e.g. "we are using only Pandas, so if there the amount of data scales up, our pipeline will not be efficient. We propose more x weeks to build in Spark, as a mitigation solution."
- **List of the packages and versions** used for the project.

Code for generating your pipeline:

- Preference to use **Kedro organization and modular code**, also for orchestration. You can keep your exploration notebooks in your appropriate folder. **(Classes Week 3 and Week 4)**
- Try to include the following components in your pipeline:
    1. **Unit data tests and feature store**: you can use one of the tools from the class or your own solutions, but it is important to have several asserts for the data quality. (**Class Week 1**)
    2. **MLflow** for **experimentation and model versioning**. (**Class Week 2**)
    3. Save model main metrics and explainability (SHAP) and include in the reports some explanations about them.
    4. Model serving (**Class Week 4/5).**
    5. **Data drift evaluation**: if you build a pipeline to test a sample of data of your strongest model, include this component as well. You can play with your sample if you want to generate drift or see how the metrics would change if drift happened. (**Class Week 6**)
    6. Try to **build tests** for your relevant functions and pipelines. **(Classes Week 3 and Week 4)**

In the end, everyone should be able to run their pipeline and to produce the same results. Projects will be graded based on the **quality of the report, code and creativity shown** for using the technologies.

A real use case organized and ready to be used as in MLOps environments with separated pipeline orchestration.



Each component is a pipeline that can run in full sequential sequence:

- e.g. data quality -> data cleaning -> data_feat_engineering -> ... model_train -> data drift
- Or we can choose just to run separated pipeline, e.g. after the model being in production only runs the data quality part or the data drift part.

You should send the report, zip of the code with a sample of data to run or a Git link.