kThe following is a slightly more concise list of the requirements for Assignment 1. This document is intended to be a sort of “notepad” and task list where we can claim sections of work/elaborate on choices/ and generally collaborate on this assignment.

*I created this little chart thing for us to document what we have done so far just so we don’t have any crossover*

|  |  |  |
| --- | --- | --- |
| **Output** | **Team Member** | **Finished?** |
| *What was done* | *Who did it* | *Is it done?* |

*I will also be porting this over to our github, because why not use the Project page*. *The document is a nice to have for now*

**Section 1: Overview**

1.1 Project scope and Objectives

The output of this section must be in a minimum 300 word writeup that covers the following elements

1. Project objective (I assume what our ML app analyzes)
2. Framework used
3. Initial data we chose, source, what it is, why
4. The model we initially used, why we used it
5. Document 3rd party packages (whoever does this, consult Matt to verify package list)

|  |  |  |
| --- | --- | --- |
| **Output** | **Team Member** | **Finished?** |
| Submission doc v1 | Henry | In Progress |

1.2 Selection of Data:

This should be documented in the previous section’s writeup, and it should be in the README as well (I assume). We must highlight

* Choice of data, and justification (why we use this data for this project instead of other)
* Source of data
* Preprocessing steps
* Do we have backup data

|  |  |  |
| --- | --- | --- |
| **Output** | **Team Member** | **Finished?** |
| We selected a diabetes dataset in kaggle https://www.kaggle.com/datasets/prosperchuks/health-dataset | Matt | Yes |

1.3 Model Architecture

This, again, should be documented in both the writeup and in the README. Document:

* What model architecture(s) we will use
* Describe what we chose, what the parameters are, and in theory what should happen

|  |  |  |
| --- | --- | --- |
| **Output** | **Team Member** | **Finished?** |
| We are using Gradient Boosting via XGBoost  https://xgboost.readthedocs.io/en/stable/ | Matt | Model decision is finished |

1.4

What tools are we using? Should be in write up and ReadMe I think?

|  |  |  |
| --- | --- | --- |
| **Output** | **Team Member** | **Finished?** |
| Hydra, Make, Poetry | Matt | Other tools may be added when necessary |

**2: Code Organization and Setup**

2.1 Code Repository Setup

- set up repo with Cookiecutter

|  |  |  |
| --- | --- | --- |
| **Output** | **Team Member** | **Finished?** |
| Code Repository is created | Henry | Yes |

2.2 Environment Set up

- set up Py environment

- Document dependencies within requirements.txt

- Google Colab

|  |  |  |
| --- | --- | --- |
| **Output** | **Team Member** | **Finished?** |
| Poetry environment is set up. I believe as we add packages with poetry, it automatically adds to the reqs file. Matt set up Colab connection | Henry | Yes |

**3. Version Control**

3.1 Use of Git

- branching strategy(ies)

- rules

|  |  |  |
| --- | --- | --- |
| **Output** | **Team Member** | **Finished?** |
| Basic branching rules have been created. Will likely be iterated upon | Henry | Yes |

3.2 Best Practices for Team Collaboration

Nothing really to assign here, maybe repository owner (Henry) will set up a rule that automatically assigns PRs to the other two members if possible?

|  |  |  |
| --- | --- | --- |
| **Output** | **Team Member** | **Finished?** |
| PR review rule | Henry | yes |

**4: Data Handling**

It seems he wants the Data and Training to be in their own document for submission. Whoever takes this on should have their own writeup

4.1 Data Preparation

- automation of data cleaning, normalization, and augmentation using scripts

*I know we have the scripts created, but I am not sure if we can count this as automated... should there be a git workflow made or is this for later?*

|  |  |  |
| --- | --- | --- |
| **Output** | **Team Member** | **Finished?** |
| We have a python file called preprocess that contains two functions for data prep – the first function splits data into training and testing sets and the other function normalizes the data. That’s all we are doing for prep – the data was pretty clean from kaggle  https://github.com/MLOps-Team-Spring-2024/Team-Project/blob/main/mlops\_team\_project/src/preprocess/preprocess.py | Matt | Yes |

4.2 Data Documentation

- provide detailed documentation of the data preparation process, include any transformations applied to dataset

*Data should have it’s own documentation*

|  |  |  |
| --- | --- | --- |
| **Output** | **Team Member** | **Finished?** |
| We have a python file called preprocess that contains two functions for data prep – the first function splits data into training and testing sets and the other function normalizes the data. That’s all we are doing for prep – the data was pretty clean from kaggle  https://github.com/MLOps-Team-Spring-2024/Team-Project/blob/main/mlops\_team\_project/src/preprocess/preprocess.py | Matt | Needs to be added into the README |

**5: Model Training**

5.1 Training Infrastructure

- configure training environment to use available resources

- if using colab ensure understanding of GPU capabilities

- document set up and running so team members can follow (Data Documentation)

|  |  |  |
| --- | --- | --- |
| **Output** | **Team Member** | **Finished?** |
| There’s a notebook to run the model and a python module.  The notebook can be started via poetry – I will update make to run it from there as well  <https://github.com/MLOps-Team-Spring-2024/Team-Project/blob/main/notebooks/1_modeling.ipynb>  https://github.com/MLOps-Team-Spring-2024/Team-Project/blob/main/mlops\_team\_project/models/cli.py | Matt | Yes |

5.2 Initial model training and evaluation

- train on selected dataset

- establish a baseline performance

- iteratively improve the model through tuning and further training

- utilize cross-validation and other statistical techniques to validate

- training **must** be well documented: *specify config, process, and evaluation metrics (Data Documentation)*

|  |  |  |
| --- | --- | --- |
| **Output** | **Team Member** | **Finished?** |
| I have a baseline model and exp1 model. They only tune 1 hyper parameter but I think that’s ok for this project since that’s not the focus. I also added cross validation. | Matt | Yes |

**6: Documentation and Reporting**

6.1 README

This must contain

* Overview of project and objectives
* Instructions for setting up environment
* Steps to replicate the training and evaluation
* List of dependencies and how to install (should really automate tbh)
* Contributions of each team member (found on this doc I guess)

|  |  |  |
| --- | --- | --- |
| **Output** | **Team Member** | **Finished?** |
| Will have a PR on github for this | Allan | Yes, need to update Steps to replicate the training and evaluation |

* Architectural overview

|  |  |  |
| --- | --- | --- |
| **Output** | **Team Member** | **Finished?** |
| Created an overview on draw.io. Added that into github | Allan | Yes. Continuously evolving |

6.2 Code Documentation

- Docstrings for all functions and classes using consistent style

- inline comments for complex logic

- ruff and mypy for styling and type checking

- makefile

*I think this is just for each person, make sure you document your code. Ruff and MyPy should be confirmed that they work, I believe they are already a part of the project?*

|  |  |  |
| --- | --- | --- |
| **Output** | **Team Member** | **Finished?** |
| Github actions and Makefile commands | Henry deBuchananne | Yes, ruff and MyPy have github actions that run on PR open and Push |

TURN IN:  
- GitHub repo with ReadMe and Source Code  
- Documentation of Data Prep and Model Training  
- Brief report on findings, challenges, and areas for improvement  
- One team member submits team list, GitHub repo, and I assume the report to D2L  
  
  
Based on the "Related to Future Parts" section, i think the next part is going to be the CI/CD section, and the final part is the active deployment of the model in some way. May even have to have a proper UI?  
  
We can provide more detail on the sections that must be written out based on the checklist he provided at the end of the assignment page. He does point out that not all points in that checklist will be met, but if you find anything relevant to what you are working on, may as well check it off of the list.