AIMLOps-B4 M5 Week-2 Lab - Questions and Answers

Question 1: Is there any tool to create these file structures automatically?

Answer 1: "Cookiecutter" is a tool that will help in developing a flexible, and reasonably standardized project structure.
http://cookiecutter-data-science.drivendata.org/

Question 2: Please share other tools that help us move from Collab to py code base

Answer 2: Link shared by mentor: https://nbdev.fast.ai/

Question 3: Explain the BaseEstimator and TransformerMixin once more on how it works.

Answer 3: In short, these classes are used to make a new child class compatible with the Scikit-learn pipeline.

Base estimator: It is the base class for all estimators in scikit-learn. It allows the custom class to have the ability to set params, get params, and allows working with GridSearch. Also helps our custom class to be easily integrated with the Pipeline function.

Transformer MixIn: a fit_transform method that delegates to fit and transform you defined in the class.

Question 4: Can we have a detailed document of the code base structure like it is getting described in this walkthrough.

Answer 4: Refer to M5-W1-AST.pdf, & M5-W2-AST2.pdf shared, they include the information about the files structure and purpose.

Question 5: In our cohort(in future classes) will we be covering pytorch pipeline prod deployments too ?

Answer 5: What we are discussing today is model or library agnositic. We can apply the same principles to Pytorch today. We try to give those examples as we go ahead with labs in coming classes.

Question 6: In the mini project, the need to drop columns. What is the right place to implement it? I had built that also as a class and added to pipeline - Is that correct approach?

Answer 6: Instead of Pipeline, its better to include in data preprocessing or feature engineering steps.
If you are creating different features using that column and old column is not useful after feature engineering then you can drop that column in tranform method of your respective class.

Question 7: Does the pipeline run in sync way?
If yes how to make it run in parallel/async way

Answer 7: So the sklearn pipeline will run sequentially. Supposing you have a lot of processing files and files that to run apart from the ML model, to make them async one need to take a different approach like pipeline orchestration tools (airflow/prefect/mage)

Question 8: Can we define the mappings for all the columns and use that in the pipeline rather than specifying for each column feature

Answer 8: It is advised to keep mappings and encoding definitions in the config.yaml as it gives you more control over your data variables. so whenever you have additional data variables coming in for a data field, you can just update the yaml to capture additional records. Makes it more dynamic.

Question 9: Once build, so anyone can install this package by doing a pip install is it?

Answer 9: Yes, you can share the .whl file from which anyone can install the package by doing "pip install /path/to/.whl".
You can also push this to PyPI(Python Package Index) site to make it available publically, and then anyone can install it directly from PyPI using pip(Package Installer for Python) without you explicitely sharing the .whl file with them.

Question 10: Can we have tmultiple environments in the same project, one for developing the ML application and other for testing?

Answer 10: Yes, technically you can.

Practically, it is an overkill and might actually be hampering in nature. Scenario: something passes in your test environment, but doesn’t for your dev.