Machine Learning Operations (MLOps): Machine Learning from R&D to Production

Mid-Term Exercise

In groups of two, please **submit a client charter paper (the template of which can be found in: https://github.com/Azure/Azure-TDSP-ProjectTemplate/blob/master/Docs/Project/Charter.md) for the final project**. **The maximum document length should be 5 pages, font size 12 and default margins. The file format must be PDF (to avoid compatibility issues). Please send it to my email by the abovementioned deadline.**

The project definition is:

"You are given the bank marketing dataset (which can be downloaded from: https://archive.ics.uci.edu/ml/datasets/bank+marketing) and the German credit risk dataset (which can be downloaded from: https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)). You should use the following simple XGBoost models on each as your baseline: https://www.kaggle.com/kevalm/xgboost-implementation-on-bank-marketing-dataset (Version 3) and https://www.kaggle.com/hendraherviawan/predicting-german-credit-default (Version 4; The baseline model for this dataset is 'XGBoost 1b: Unbalance Dataset (ROC_AUC:0.79)', which appears in: https://www.kaggle.com/hendraherviawan/predicting-german-credit-default?scriptVersionId=5889660&cellId=26), respectively.

The goal of the project is to design and implement a basic ML pipeline for each model (you can, but don't have to, use Scikit-Learn pipelines, as shown in: https://nabeelvalley.netlify.app/docs/data-science-with-python/xgboost-and-pipelines/) that would contain an **automatic step** that would improve the overall performance of the model **in any method that does NOT involving changing the model's type or hyperparameters (although retraining on different data IS allowed) and is NOT dataset-specific** (although it might support only datasets with tabular data only, that is, you don't have to support, e.g., textual or image data).

Example: **"We will add a feature selection phase that will only use the top 90% features with the highest Shapley values, in an attempt to improve the AuROC metric of the baseline model."**

The improvement is to be measured by the metrics you should define and specify in the Metric section of the charter document (e.g., accuracy, precision, recall, a specific fairness metric, etc.).

Like in a real industry ML project, in order to implement the automatic step, you are welcome to use any free or open-source tool we saw during class, you are familiar with, or you find on the internet. Proprietary or private tools are not allowed. A list of suggested tools is specified below, but any other tool can be used instead."

You will get your graded exercise back with comments, as well as a message that the project was approved, or need additional clarification.

You SHOULD fix the design document based on these comments.

However, you need to submit this revised version only with the final project submission.

Resubmission BEFORE THAT is only required if the project is not approved.

You can submit before the deadline: The earlier you submit your design – the earlier you get an approval and can start working on the project!

When designing the final project in this exercise, please consider that the final project will be graded by these criteria:

1. A working (as in: not crashing and producing an output in less than one hour) code.
2. Automatic (as opposed to a manual/human-in-the-loop) improvement step.
3. You will successfully explain your implementation and understand the theory behind it. The theory should be specified in the Scope and Architecture sections in the charter document.
4. The step is generic and isn't dataset specific. In order to demonstrate that, you should use the exact same automatic step in the pipelines of both the bank marketing dataset and of the German credit risk dataset.
5. The relative performance improvement size (from the baseline model to the model including the automatic step), measured by the metric you determined and specified in the character document.
6. Non-trivial an interesting (theory-wise) implementation.

Examples of possible tools/methods to integrate/implement

1. https://towardsdatascience.com/how-to-find-weaknesses-in-your-machine-learning-models-ae8bd18880a3 - A BIG bonus will be given to the student group who would implement and use this method.
2. https://aif360.mybluemix.net/
3. https://fairlearn.org/
4. https://github.com/ydataai/ydata-synthetic
5. https://ai.googleblog.com/2021/10/baselines-for-uncertainty-and.html - A bonus will be given to the student group who would use this method.
6. https://blogs.oracle.com/ai-and-datascience/post/macest-release - A bonus will be given to the student group who would use this method.
7. https://research.ibm.com/blog/uncertainty-quantification-360 - A bonus will be given to the student group who would use this library.
8. https://towardsdatascience.com/how-can-i-measure-data-quality-9d31acfeb969
9. https://concept-drift.fastforwardlabs.com/ - A bonus will be given to the student group who would implement and use this method.
10. https://medium.com/a3data/data-quality-with-hermione-46233529517b
11. https://towardsdatascience.com/great-expectations-always-know-what-to-expect-from-your-data-51214866c24
12. https://adversarial-robustness-toolbox.org/
13. https://www.tensorflow.org/tfx/guide/tfdv
14. https://www.tensorflow.org/tfx/model_analysis/get_started
15. https://feature-engine.readthedocs.io/en/1.1.x/
16. https://syntheticdata.community/