Case Study Description

A company developed a product which is available in numerous markets. The product batch may be released to a market only if it meets strict quality criteria. Product quality is assessed based on costly and time-consuming lab measurements.

The company would like to automatize and accelerate the product release process; thus, it would like to have a tool indicating the key product quality indicators at a batch level. Ideally, the automatized batch release process requires understanding and knowledge of the key product quality indicators value determined by the at-line measurements (see attached dataset.parquet). If not feasible or the available data does not allow to make such predictions, an indication if a batch meets releasing criteria should suffice. A warning flag shall be raised for a batch if key product indicators exceeds a predefined value.

To develop such solution, the company provided 238 product batches data that have been produced already. The dataset includes both categorical and numerical predictors along with the key product quality indicator values i.e. the "Target". A threshold of 1.3 is used to classify a batch as meeting / not meeting release criteria. Batches with Target variable below the threshold are considered as successes and the ones above the threshold as failures.

Your job is to create (within one week):

- a model that predicts (or at least classifies) the success of a batch and
- an associated presentation you will be presenting (in English) that is going to be pitched in-front of the business owner.

Data Sources

Use the following file for the analysis: dataset.parquet

Objectives

1. Prepare and evaluate a predictive model for the Target variable. Feel free to use any modelling technique suitable for the data provided. Provide us with your complete code in Python – so we can reproduce all the steps (can be in a form of Jupyter Notebook/Lab, or .py file(s)). Ensure sufficient code quality and comments are within code helping reviewers to navigate through it.
2. Create a presentation (10 minutes + 5 minutes for Q&A) with your results for a managerial audience with some technical knowledge using necessary visualizations. Your presentation must include which predictors were selected and how your model performed. Add any recommendations and next steps if you had to continue the analysis. While discussing the technical part, justify and explain modelling techniques you used, evaluation metrics and assumptions underlying the preprocessing phase.
3. Be ready to discuss your code in depth with other data scientists after the business presentation is over.