Programming assignment #4

Course: CHE1148H - Data Process Analytics

Coding options

If you encounter issues with Databricks and feel more comfortable with working on .ipynb files for both sections, it's fine. You may have to reproduce the best Random Forest model in Python. Databricks files will also be accepted.

1 Data quality

You will generate some data quality insights for two models we have worked in the past:

1. DFT error estimation [0]
2. campaign response from transaction data

The first model is based on a regularized Ridge algorithm and the code is given. Use the L2 model with $\alpha = 2$. Also, notice that the features are already imputed with zero values when a subgroup is absent. In your analysis, you need to calculate the completeness as the percentage of non-zero elements per column.

For the second model, please use the **best Random Forest model** you created in the previous assignment. We don't need the perfect model (meaning don't spend more time on fine-tuning it), as long as it does not overfit.

Deliverables:

Create the completeness-feature importance graph for each model and comment on the results and trends you observe. Make sure that appropriate scaling of the axis is used for the best visualization and interpretation (e.g. loglog or semi-log axis). Also, if you want to improve the two models what kind of features would you design or incorporate in your data? **2 notebooks: 1 graph for each model and interpretation in Markdown language comments.**

[0] From Bhattacharjee et al (2021) Regularized machine learning on molecular graph model explains systematic error in DFT enthalpies, 11-14372 Scientific Reports

2 Data and model drift

Your team has built a number of models over the years and now there is concern about their quality and possible drift. You will build a monitoring tool that captures model quality and drift. Use all data prior to Dec-2013 (inclusively) to train a model and calculate the baseline distributions you need for the first point of the dashboard based on this timeframe.

Then, you put this model in production and you calculate the scores for all clients after Jan-2014 (inclusively). You are asked to capture the drift of the model scores and the top 5 features of the model once it goes in production on Jan-2014.

Deliverables:

Create a dashboard that includes the Jensen-Shannon divergence metric with the training data as baseline distributions and monthly values after Jan-2014 for the following variables:

1. the monthly scores
2. the monthly top 5 features of the model

1 notebook: 1 graph with scores and top 5 features drift; comments and interpretation of the results.