STAT3612 Statistical Machine Learning: Group Project Call for Proposal

October 20, 2020

Submit via Moodle before 11:59pm October 30, 2020

You are required to form small groups (with 4 or 5 members each group) to carry out a statistical machine learning project. The theme of this year is **interpretable machine learning (IML)** with applications to a real data case study in the banking industry. For an IML model, both "prediction accuracy" and "model explainability" are equally important (each accounting for 30% in the final grade). The remaining 40% of grade will be about story-telling, based on both your oral presentation and final written report.

The data includes about 10,400 anonymized Home Equity Line of Credit (HELOC) loans[1], together with 23 raw features. You may obtain the dataset from here, together with the data description file at here. Note that in the second Excel file the monotonicity constraints are included in the data dictionary, which are based on the prior knowledge about the input-output relationships. You are required to build two sets of machine learning models, one without the monotonicity constraints and the other one with the monotonicity constraints.

Each team has to be formed with strictly 4 to 5 members (only with exceptions subject to approval). Each team will receive a unique Group after collection of all the proposals. Such group ID will be used as the random seed for splitting data into training (80%) and testing (20%) sets. You can build your final IML models based on the training data only. There is no restriction in the choice of feature engineering techniques or machine learning algorithms. The final model evaluations will be based on three aspects:

1. Prediction performance evaluated on the testing data (30%);
2. Model interpretability (both global and local) for the final model (30%);
3. Story-telling based on your oral presentation and final written report (40%).

As long as your team is formed, you can launch the project immediately, till December 1 (the day of oral presentations). Before oral presentations, you will need to submit your final model in the Python notebook format with adequate description, so that your results can be reproduced by tutors. After December 1, you will also have 5 days to revise your Python notebook before submitting it as the final project report (due date: December 6).

[1] We acknowledge these data are obtained from FICO, and it is purely for our academic purpose.