STAT3612 Statistical Machine Learning: Group Project Call for Proposal

October 11, 2019

Submit via Moodle before 11:59pm October 25, 2019

You are required to form small groups to carry out a statistical machine learning project. The theme of this semester is **interpretable machine learning (IML)** with applications to a real data case study. For an IML model, both prediction accuracy and model explainability are equally important. You will need to take both objectives into account when building models in this project.

The data includes about 10,400 anonymized Home Equity Line of Credit (HELOC) loans, together with 23 attributes. You may obtain the dataset and data descriptions from here and 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 feature-target effect. We acknowledge these data are obtained from FICO, and it is purely for our academic purpose.

Each team has to be formed with four or three members, and a unique Group ID will be assigned after collection of all the proposals. Such group ID will then used as the random seed for splitting data into training (80%) and testing (20%) sets. Build your IML models based only on the training data. 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 accuracy as tested on the remaining 20% dataset;
2. Global interpretability about the model effects;
3. Local interpretability about individual predictions (Good/Bad cases from testing data).

The project can be started immediately, till November 25 (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 November 25, you will also have one week time to improve your Python notebook before submitting it as the final project report.