STAT3612 Group Project: Data Science Challenge in Predicting International Mathematics Proficiency

April, 2018

Problem description:

This project is to build a statistical machine learning model for predicting international mathematics proficiency for eighth grade students among six countries/regions, namely Hong Kong, Taiwan, Singapore, Japan, United Kingdom and United States. The training data `x_train.csv` used for this data science challenge is an assembled version of the TIMSS 2015 database, consisting of 28,063 students with the following variables:

- `StudentID`: Student ID
- `Gender`: Male/Female
- `Region`: HKG, TWN, SGP, JAP, ENG, USA
- `InMotif_1-3`: intrinsic motivation variables 1-3
- `ExMotif_1-3`: extrinsic motivation variables 1-3
- `SelfEff_1-3`: self-efficacy variables 1-3
- `Teacher_1-3`: math teacher effect variables 1-3
- `NumBook, NumDevice`: number of books and number of digital devices at home
- `EdFather, EdMother`: education level of father, educaton level of mother

Our aim is to predict the label `FlagAIB` that indicates whether the student performance in mathematical assessment is better than the Advanced International Benchmark. The labels for training data is provided by `y_train.csv`.

The model performance of classification will be assessed by the area under the ROC curve (AUC) based on a test data `x_test.csv` with 7,018 students. You are required to submit the probability of `FlagAIB` per student from the test data, in the same format as `y_train.csv`.

Timeline:

- **Data release:** April 4 (Download from the course website: http://stat3612.saas.hku.hk)
- **Contest period:** April 4 – 30 (One and only one submission every 3 days)
- **Presentation:** Week 1 of May (Oral presentation with peer assessments)
- **Final report:** due May 6 (Written report with detailed methodology and results)

Remarks:

1. R/Python programming is required.
2. There is no restriction on the use of statistical machine learning algorithms.
3. All the members in the same team share the same score (30% of final grade).