IE30301-Final Project Guidelines

Final Project Description

There are two types of data sets for the final project, which are sourced from Kaggle. However, you should use the dataset provided by us, as the dataset was partially modified.

- Tasks (choose one)

1. Regression
   - Dataset: "Rossmann Store Sales"
   - Description: See variable description file
2. Classification
   - Dataset: "Titanic dataset generated with CTGAN."
   - Description: See variable description file

The description for each dataset is below. Please read the instruction carefully. If you have questions regarding the final project, please use the discussion board. TA will not reply to emails regarding projects to avoid any possible information disparity between students. Please submit your Assignment to the blackboard.

✓ **Due Date:** 2020.05.30(Sunday) 10 pm Korea Standard Time.
✓ **We will not accept late work!!**
✓ **Submission format:** Please submit a .zip file containing the below **two** files. Follow the naming rules as stated.

your_student_ID_FinalProject_your NAME.zip

your_student_ID_FinalReport_yourNAME.pdf

your_student_ID_FinalProject_yourNAME.ipynb

Ex. 20205318_FinalReport _HongGilDong.pdf

Coding Policy

✓ You can refer to external references for this project. But **you must cite** all the materials and websites on your report.
✓ You can also try unsupervised learning techniques you've learned in the class (e.g., PCA, k-means clustering) for richer analysis.
✓ Don't use deep learning models or Boosting models such as Xgboost, Gradient Boost, Catboost, etc.

Evaluation Policy (100 Points)

- The logical process of problem definition and hypothesis construction (including feature engineering) **(20 points)**
- Exploratory data analysis **(20 points)**
  - ✓ Show at least three plots or tables related to problems and your hypotheses
  - ✓ Provide corresponding explanations/interpretations
- Correctness of the process of data munging and data pre-processing
- Usage of appropriate models **(15 points)**
  - ✓ Use at least three different models for the task
- Usage of appropriate evaluation measures with useful visualizations **(20 points)** (e.g., ROC curve, k-fold CV result with an error bar, etc.)
  - ✓ Compare the performance of models
- Thorough post-analysis of results (including implications) **(25points)**
  - ✓ What insights were you able to gain from this data?