Data Mining Project

Steel Plate Defect Prediction

——Multi Binary Classification

Task Introduction

- Task: Multiclass Classification
  - Predict the probability of various defects on steel plates given multiple factors.
- Key points:
  - Feature Analysis, Classification
- Project URL:
  - https://www.kaggle.com/t/c67f14743e83489f8f9dfa20780df769

Dataset & Data Format

- Dataset
  - The dataset (both train and test) was generated from a deep learning model trained on the Steel Plates Faults dataset from UCI.
  - Training: 28 features and 7 target labels (Total 19218 records)
  - Tasting: 28 features (Total 32032-19218 = 12814 records)
  - Label: 7 class(Pastry, Z_Scratch, K_Scatch, Stains, Dirtiness, Bumps, Other_Faults)

Dataset & Data Format

- The Submission file should contain a header and have the following format:

```
id,Pastry,Z_Scratch,K_Scatch,Stains,Dirtiness,Bumps,Other_Faults
19219,0.5,0.5,0.5,0.5,0.5,0.5,0.5
19220,0.5,0.5,0.5,0.5,0.5,0.5,0.5
19221,0.5,0.5,0.5,0.5,0.5,0.5,0.5
etc.
```

Evaluation

- Submissions are evaluated using area under the ROC curve using the predicted probabilities and the ground truth targets.
- To calculate the final score, AUC is calculated for each of the 7 defect categories and then averaged.
- In other words, the score is the average of the individual AUC of each predicted column.

TODO List

- The Kaggle competition
  - Submission Deadline: 1 July (Monday of 18th week)
  - Leaderboard release: 3 July.
    - The detailed results will be public available in QQ Group
- Two Reports
  - HW1: Data Analysis
    - Including data preprocessing, feature correlation analysis, feature distribution analysis, et al.
    - Deadline: 1 June
  - HW2: Final Summary
    - Including model design, training strategy, results analysis, et al.
    - Deadline: 1 July