Binary Classification of Imbalanced data from Bosch Production Line

P-1: Xi Yang Liang Dong Weijie Zhou Yeojin Kim

1 Data Set

The data from an ongoing Kaggle competition: Bosch Production Line Performance (`https://www.kaggle.com/c/bosch-production-line-performance/data`). The data represents measurements of parts as they move through Bosch's production lines and it is a binary classification problem. The raw data has three types of feature: numerical, categorical, and date.

2 Project Idea

This project has two challenges: 1) The size of raw data is large(15.4GB) and cannot directly fit into a single machine. We will do data exploration and feature reduction to fit important features into a single memory with 16GB memory. 2) The raw data is highly imbalanced (6879:1176868). We will utilize different methods such as upsampling, downsampling and SMOTE to do the sampling.

Following the above two procedures, we will then train the machine learning models with gradient boosting tree, and then compare its performance with other speeding up gradient boosting tree models, including Xgboost, FastBDT and SAS Viya.

3 Software

Python open source libraries: pandas, scikit-learn, matplotlib, python API for SAS Viya (SAS cloud platform)

Gradient boosting tree libraries: xgboost and FastBDT, gradient boosting tree in SAS Viya

4 Papers to Read

[1] Tianqi Chen, & Carlos Guestrin. (2016) XGBoost: A Scalable Tree Boosting System. *KDD.*

[2] Keck T. (2016) A speed-optimized and cache-friendly implementation of stochastic gradient-boosted decision trees for multivariate classification *arXiv preprint arXiv*: 1609.06119.

5 Teammate and Work Division

Xi & Yeojin: Data exploration and feature reduction, sampling

Weijie & Liang: Tree building, hyperparameter optimization

6 Midterm Milestone

Do feature engineering and sampling of raw data, generate submission in Kaggle by SAS Viya platform, single machine version xgboost and FastBDT.