Student id: 20123133
Name: Hui Kwat Kong
Itsc: kkhuiaa

readme

All the scripts (including __init__.py) and trainFeatures.csv, testFeatures.csv, trainLabels.csv are assumed to be run in the current working directory.

The feature generating is divided into two parts: numeric and categorical variables.

For numeric variables:

1. Fill the numeric missing value (assuming there will be missing value in unseen dataset)
2. Apply PCA for feature generation and reduction (selected as 95% of variance)
3. Normalize the data again for accelerating the model, since XGboost requires gradient descend.

For categorical variables:
It is not good to transform the columns as dummy variables, since it will produce sparse data and curse of dimensionality. Since the target is binary, we can rank the categorical column to ordinal variable with its mean response rate ranking.

Model is built on XGboost. Different hyperparameter are selected by GridsearchCV with 10-folds of cross validation.
Final best train score is 0.8369423597821433 and best test score is 0.8348591652285823.
The two scores difference is small that the model is stable.