The data for this competition represents measurements of parts as they move through Bosch's production lines. Each part has a unique Id. The goal is to predict which parts will fail quality control (represented by a 'Response' = 1).
The dataset contains an extremely large number of anonymized features. Features are named according to a convention that tells you the production line, the station on the line, and a feature number. E.g. L3_S36_F3939 is a feature measured on line 3, station 36, and is feature number 3939.
On account of the large size of the dataset, we have separated the files by the type of feature they contain: numerical, categorical, and finally, a file with date features. The date features provide a timestamp for when each measurement was taken. Each date column ends in a number that corresponds to the previous feature number. E.g. the value of L0_S0_D1 is the time at which L0_S0_F0 was taken.
In addition to being one of the largest datasets (in terms of number of features) ever hosted on Kaggle, the ground truth for this competition is highly imbalanced. Together, these two attributes are expected to make this a challenging problem.

File descriptions

- **train_numeric.csv** - the training set numeric features (this file contains the 'Response' variable)
- **test_numeric.csv** - the test set numeric features (you must predict the 'Response' for these Ids)
- **train_categorical.csv** - the training set categorical features
- **test_categorical.csv** - the test set categorical features
- **train_date.csv** - the training set date features
- **test_date.csv** - the test set date features
- **sample_submission.csv** - a sample submission file in the correct format