You are asked to build the most accurate model you can to predict target column for data\_test.csv. The metric to reflect accuracy can be defined by yourself.

'id': id column for data\_train, data\_test, respectively

'num\*': numerical features

'der\*': derived features from other features

'cat\*': categorical features

'target': target column, only exists in data\_train. it is binary.

There are potentially missing values in each column.

The goal is to predict 'target' column for data\_test.csv.

The solution should have a result csv file with two columns:

1. 'id': the id column from data\_test.csv

2. 'target': the predicted probability of target being 1

The corresponding code to reproduce the result csv file should be included as well.

Solution:

* I have come up with two classification model for the data-set because the target have binary values deciding whether you can classify as ‘0’ or ‘1’:

1. Random Forest (0.963522877901). . .
2. Stochastic Gradient Descent (0.963480930888)

* Both have same approximately same prediction values.
* Data\_Science\_Challenge\_RandomForest
  + Result.csv
* Data\_Science\_Challenge\_SGDClassifier
  + Result\_new\_1.csv