Documentation of AIML 2017 Project:
Default Prediction

B03901191 Fu-Hsuan Liu (劉馥瑄)

December 21, 2017

Problem 1.

簡述此次作業主要使用的演算法，並說明為何使用此演算法:

In this project, I used an algorithm which is a kind of gradient boosting on decision trees, that is, CatBoost, since it's efficient for classification problem and it's one of the state-of-the-art in machine learning.

Problem 2.

呈上題，簡述此演算法比較重要的參數，並說明如何選取適合的參數值

```
model = CatBoostClassifier (iterations=1000, depth=5, learning_rate=0.05,
loss_function='Logloss', logging_level=‘Verbose')
```

Several parameters in the algorithm are very crucial and with different meaning. The more iterations or depth is set, the more training time and less loss (,which is defined as ‘learn' in training result), and learning_rate is something like differential, the distance for each step forward.

After several trying, I found that the default is most suitable and close to baseline. Hence my code was turn out to be:

```
model = CatBoostClassifier()
```

Problem 3.

簡述對資料的輸入特徵 (features) **的處理方式**

Actually I've tried many ways to pre-process the training and testing data, however, after I pre-processed them, the result (or score) was not improved a lot, even not improved. Below are my tries:

a. LIMIT_BAL: One-hot encoding,
categorized them into several groups based on their value (e.g. 7~10 groups).

b. PAY_[1~6]: turn -2, -1 into 0.

c. MARRIAGE: remove this feature. Since I think that this feature is unrelated.

d. BILL_AMT[1~6], PAY_AMT[1~6]: Normalized

Problem 4.

若程式中有使用較特殊的套件，請敘述其名稱及版本，並簡述為何使用此套件

f*rom CatBoost import CatBoostClassifier* is the specific module I used in this project, the version and related documentation are linked below: https://tech.yandex.com/catboost/doc/dg/concepts/python-quickstart-docpage/#classification