# CMPSC497 Fall 2018 Programming Assignment 2.

**Assigned:** Wednesday, October 10, 2018

**Due:** Monday, October 29, 2018 (by midnight, submit a package of codes in .py and a report .pdf via Canvas)

**Maximum:** 100 point

**Note:** This assignment is to be done by an individual student, no team work allowed.

---

Based on the result of data preprocessing (in Programming Assignment 1), you are asked to solve a classification problem, the decision making of clients of bank marking [1], and perform a number of tasks in order to evaluate the performance of models under different settings. The goal of this assignment is to use the (preprocessed) data to train classifiers for prediction and to exercise on various issues usually arising in training a classifier in order to find the best model.

## Dataset

- Bank Marketing Dataset [1] (https://archive.ics.uci.edu/ml/datasets/Bank+Marketing). Please use Column 1 to 20 as input features and Column 21 as the output class.

## Classification Experiment

- Task 1. In this task, you will train a logistic regression classifier on Bank Marketing Dataset to predict whether a client will subscribe a term deposit.
    - There are 10 numerical features and 10 categorical features. Please train **Logistic Regression Model 1** based on normalized numerical features and one-hot encoded categorical features, and train **Logistic Regression Model 2** based on unnormalized numerical feature and one-hot encoded categorical feature.
    - Please use 5-Fold cross-validation for experiments. (See textbook and https://en.wikipedia.org/wiki/Cross-validation_(statistics))

  Please summarize the definitions and mathematical formulae of **confusion matrix**, **precision metric**, **recall metric**, **f-measure metric,** and **accuracy metric.** Please compare the performance of **Logistic Regression Model 1** and **Logistic Regression Model 2** in terms of these four metrics.

## Imbalanced Issue

- Task 2. Note that data imbalance exists in this dataset. Please explain why we want to avoid imbalance issue in training classifiers? Briefly summarize at least 3 methods deal with data imbalance issue. Generate new datasets by either downsampling or upsampling and repeat the steps in Task 1 to compare the performance of generated datasets with the original dataset. Note that you only need to perform sampling algorithms on training set. Please explain why.

## Feature Selection

- Task 3. Please summary the reason why we perform feature selection? Please perform feature selection based on the correlation results in Assignment 1 (using chi-square for categorical data and mutual information for numerical data). Generate partial datasets by only using top k (k =1, 3, 5) most correlated categorical features and numerical features for model training (i.e., k categorical features + k numerical features). Follow the setup in Task 1 to compare the performance of partial datasets with the original dataset.
- For those who are not confident with their results from Assignment 1, we provide the top k most relevant features extracted using scikit-learn packet. https://psu.box.com/s/kxp9mc1t8gn288dwh0rkkp2v247u7cbi.
  Data are stored in numpy array. You can load them using python-pickle packages.

```
import pickle
with open('/Users/CongWeilin/Downloads/feature_selection/feature_k=5.pkl','rb') as f:
    data = pickle.load(f)
print(data.keys())
print(data['numerical'].shape, data['categorical'].shape)
```

```
dict_keys(['numerical', 'categorical'])
(41188, 5) (41188, 28)
```

## Model Comparison

- Task 4. In addition to logistic regression classification, Decision tree and Multi-layer perceptron neural network are also widely used for classification. Please follow the setup in Task 1 to compare the performance of these two models with Logistic Regression Model on both balanced and imbalanced datasets.

## Submission

- Please submit a report in pdf and a code file in .py.

## Packages

- sklearn (http://scikit-learn.org/). A machine learning framework in Python

## Reference

[1] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014