# Competition 1

Deadline: [Upload Prediction at Website] **2019/04/19** (Friday) 23:59

Deadline: [Upload Report & Code at Moodle] **2019/04/21** (Sunday) 23:59

Competition Website: http://140.116.52.202:5566

## Task: Churn Prediction of Bank Customers

Welcome to the first competition of Introduction to Data Science. This CP1 could be your first ever data science competition! ☺ Through the competition and teamwork, you are expected to practice the usage of scikit-learn classification package. Pursuing high ranks by competing with each other will also make you learn more than what we taught in the class. In other words, you can learn by yourself to find any advanced and state-of-the-art classification methods and packages, and apply that methods in this competition. The more you explore about the classification in Google, the more you will learn. We sincerely hope you enjoy the competition process, and learn more beyond about classification using Python scikit-learn package.

| | RowNumber | CustomerId | Surname | CreditScore | Geography | Gender | Age | Tenure | Balance | NumOfProducts | HasCrCard | IsActiveMember | EstimatedSalary | Exited |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 551 | 15806307 | S2336 | 720 | S2 | Male | 38 | 5 | 114051.97 | 2 | 0 | 1 | 107577.29 | 0 |
| 1 | 6897 | 15709621 | S1500 | 682 | S0 | Female | 54 | 4 | 62397.41 | 1 | 1 | 0 | 113088.6 | 1 |
| 2 | 4588 | 15619340 | S1865 | 672 | S0 | Female | 31 | 5 | 119903.67 | 1 | 1 | 1 | 132925.17 | 0 |
| 3 | 291 | 15620746 | S1672 | 592 | S2 | Female | 40 | 4 | 104257.86 | 1 | 1 | 0 | 110857.33 | 0 |
| 4 | 1673 | 15646372 | S2532 | 753 | S2 | Male | 42 | 5 | 120387.73 | 1 | 0 | 1 | 126378.57 | 0 |
| 5 | 648 | 15649129 | S1548 | 575 | S0 | Female | 42 | 5 | 104472.9 | 1 | 1 | 1 | 71641.38 | 0 |
| 6 | 6113 | 15729557 | S37 | 572 | S0 | Male | 37 | 6 | 135715.66 | 1 | 1 | 0 | 115928.95 | 0 |
| 7 | 8957 | 15579112 | S750 | 753 | S1 | Female | 34 | 6 | 124281.61 | 1 | 1 | 0 | 89136.06 | 0 |
| 8 | 1678 | 15680895 | S2079 | 546 | S0 | Male | 46 | 3 | 62397.41 | 2 | 1 | 1 | 79809.09 | 0 |
| 9 | 5202 | 15580935 | S252 | 657 | S0 | Male | 45 | 4 | 141238.54 | 2 | 0 | 0 | 95281.51 | 0 |
| 10 | 4868 | 15738150 | S1717 | 617 | S2 | Male | 35 | 4 | 62397.41 | 2 | 1 | 1 | 132607.99 | 0 |

Figure 1: A snapshot of training data. The column “Exited” is the target to be predicted.

In this competition, your task is **churn prediction of bank customers**. Given hundreds of financial behavioral data of bank customers, the goal is to predict whether a customer will eventually quit the bank, i.e., no longer having transactions in the bank in the future. We provide you a training dataset (**train.csv**) and a test dataset (**test.csv**). A snapshot of training data is presented in Figure 1. There are 13 features whose meanings are correspondingly described in Table 1, and **the prediction target is the “Exited” column**. In the test data, we provide you only the features and the “Exited” column is hidden for you to predict. To upload your prediction results, you have to exactly follow the format of the sample file (**sample_upload.csv**) to generate the predicted “exited” in the binary form (1: exited, 0: not exited). Your ranking in the leaderboard in the competition website will be generated and updated based on your latest and best prediction record.

Your prediction results will be evaluated by three metrics: Accuracy, Precision, and F-Score. You may want to understand the details of these three evaluation metrics. We give the interpretation of them in the following Table 2. It is very important to notice that in grading your performance of this

competition based on the ranking, three evaluation metrics have different weights: Accuracy (30%), Precision (30%), and F-Score (40%). Giving F-Score the highest weight means that we hope your prediction is able to not only accurately predict "0" and "1", but also find all of the exited customers.

Table 1: The list of features and their meanings.

| Feature Name | Meaning |
|---|---|
| RowNumber | ID of each bank's customer record |
| CustomerId | ID of each bank's customer |
| Surname | Anonymized surname of the customer |
| CreditScore | Credit score of the customer. Higher credit score means better banking behaviors |
| Geography | Anonymized zip code of the customer |
| Age | Age of the customer |
| Tenure | Number of tenures (不動產) of the customer |
| Balance | Amount of money the customer has in a bank account |
| NumOfProducts | Number of financial products of the customer |
| HasCrCard | Does the customer have credit card? (1: True, 0: False) |
| IsActiveMember | Whether the customer has frequent transactions in the bank? (1: True, 0: False) |
| EstimatedSalary | Estimated and perturbed salary money of the customer |
| Exited | Whether the customer does eventually leave (exit) the bank? (1: True, 0: False) |

Table 2: The interpretation of three evaluation metrics.

| Metrics | Definition |
|---|---|
| Accuracy | $\frac{\text{測試資料中，正確分類為 1 的筆數} + \text{正確分類為 0 的筆數}}{\text{所有測試資料的筆數}}$ |
| Precision | $\frac{\text{測試資料中，正確分類為 1 的筆數}}{\text{測試資料中，所有分類為 1 的筆數}}$ |
| Recall | $\frac{\text{測試資料中，正確分類為 1 的筆數}}{\text{測試資料中，所有答案為 1 的筆數}}$ |
| F Score | $\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$<br>The F score is the harmonic average of the precision and recall, where an F score reaches its best value at 1 (perfect precision and recall) and worst at 0. (Wikipedia) |

In order to encourage and excite your participation of this competition and respect the competition results, we create the following rules that you are required to absolutely follow.

(1) Each team has at most **20 times per day** for uploading the prediction results.
(2) You can use any machine learning packages, do not limit by scikit-learn we taught in the class. We strongly encourage you to learn by yourself to explore more advanced ML packages. We also very believe that those teams ranked at top positions have used advanced ML packages.
(3) Your final score of this competition will be graded by three parts
    (a) The **ranking in the leaderboard**
    (b) The **quality of your report** that describes the details of your prediction (see below for details)
    (c) The **clearness of your code** in either .ipynb or .py files.
(4) We will provide **awards to the final top-3** teams in both certifications and gifts.
(5) The top-3 teams need to deliver a **15-minute presentation** to explain their methods on **4/24**

In writing the report, you should to include the following items. Please submit your report in PDF.

(1) **小組各成員的姓名、系級與學號**。
(2) **競賽敘述與目標**：簡述競賽是要做什麼。
(3) **資料前處理**：若有對資料進行某些前處理 (如過濾、取樣)，請說明。
(4) **特徵處理與分析**：若有對特徵進行某些處理或分析，請說明。
(5) **預測訓練模型**：採用什麼分類模型進行訓練與預測，以及最終選定的參數是什麼；若有對既有模型進行修改或調整，請說明。若你嘗試過多種方法、多個分類模型，也請詳細說明。

(6) 預測結果分析：針對不同分類模型、不同特徵、不同參數、不同評估指標(Accuracy、Precision 與 F Score)的預測結果之圖形或表格呈現；若有記錄上傳到競賽網站的預測結果，可整理呈現在報告中；若有自行切自己的訓練與測試資料，也可以回報自行測試的預測結果。此外，也可以分析討論哪些模型與哪些特徵會產生較高與較低的效果，最終歸納出對於特徵、分類模型選擇的建議。得明確指出最終上傳的結果是哪一組分類模型與其參數。

(7) 感想與心得：經過這次競賽，每一位組員必須寫下一段至少 200 字的心得感想，其中可包含(但不限)譬如：從競賽學到什麼、自學到什麼、花最多時間的地方、最困難在哪(或是太簡單)、給競賽的建議、給網站的建議等。

## Important Notes

This competition is for each **team**. You are asked to write comments or markdown sections to describe the meaning of each part of your codes in either “cp1.ipynb” or “cp1.py”.

## How to Submit Your Homework?

**Submission in NCKU Moodle.** Before submitting your report and code, please **zip** your code files (.py or .ipynb) and the report in PDF into a zip file, and name the zip file as "StudentID_cp1.zip" or "StudentID_cp1.rar". For example, if StudentIDs of your team members are H12345678 and H87654321, then name your file as "H12345678_H87654321_cp**1**.zip". Then submit your zipped file using NCKU Moodle platform http://moodle.ncku.edu.tw .

## Have Questions about This Homework?

Please feel free to visit TAs, and ask/discuss any questions in their office hours. We will be more than happy to help you.