Cloud Computing and Big Data Analytics

HW5: Large-Scale with PySpark

TA: Wei-Lun Tseng (Eric) (曾偉倫)
Email: eric840610.ee02@g2.nctu.edu.tw

Table of Contents

- Requirement and Environment Setting
  - Requirement
  - Environment Setting
- Problem Description
  - Customer Churn Prediction
  - Dataset
  - Problem Description
- Grading Policy
- Requirement and Notification
- Deadline

Requirement

- Google Colab with PySpark ML lib
- Use jupyter notebook template in HW5.zip

Customer Churn Prediction

Customer churn occurs when customers or subscribers stop doing business with a company or service.

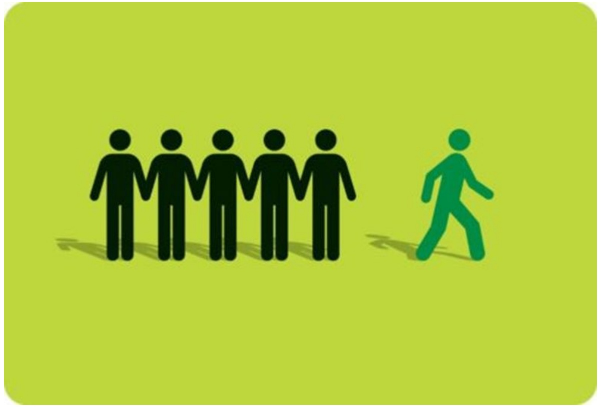


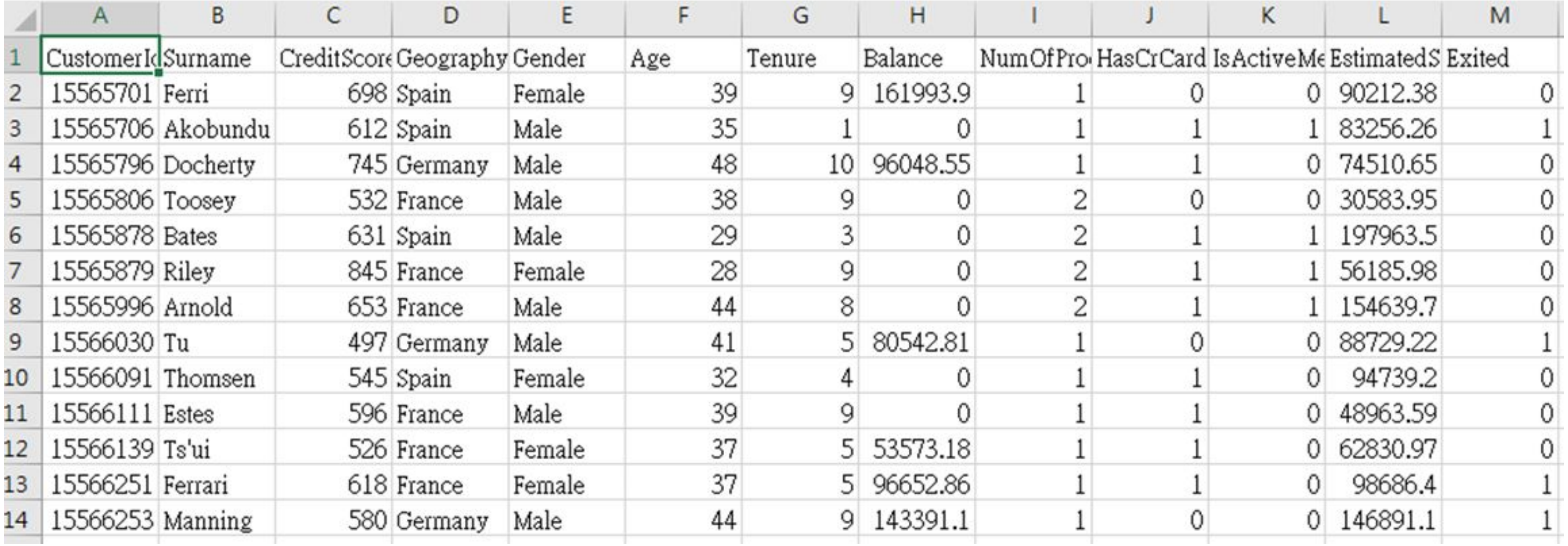
Photo credited CUSTOMER **BLISS**

Dataset Description

- Bank Customer Churn
- Dataset Format: csv
- There are the following information in public.csv:

CustomerId,Surname,CreditScore,Geography,Gender,Age,Tenure,Balance,
NumOfProducts,HasCrCard,IsActiveMember,EstimatedSalary,Exited

public.csv



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	CustomerI	Surname	CreditScor	Geography	Gender	Age	Tenure	Balance	NumOfPro	HasCrCard	IsActiveMe	EstimatedS	Exited
2	15565701	Ferri	698	Spain	Female	39	9	161993.9	1	0	0	90212.38	0
3	15565706	Akobundu	612	Spain	Male	35	1	0	1	1	1	83256.26	1
4	15565796	Docherty	745	Germany	Male	48	10	96048.55	1	1	0	74510.65	0
5	15565806	Toosey	532	France	Male	38	9	0	2	0	0	30583.95	0
6	15565878	Bates	631	Spain	Male	29	3	0	2	1	1	197963.5	0
7	15565879	Riley	845	France	Female	28	9	0	2	1	1	56185.98	0
8	15565996	Arnold	653	France	Male	44	8	0	2	1	1	154639.7	0
9	15566030	Tu	497	Germany	Male	41	5	80542.81	1	0	0	88729.22	1
10	15566091	Thomsen	545	Spain	Female	32	4	0	1	1	0	94739.2	0
11	15566111	Estes	596	France	Male	39	9	0	1	1	0	48963.59	0
12	15566139	Ts'ui	526	France	Female	37	5	53573.18	1	1	0	62830.97	0
13	15566251	Ferrari	618	France	Female	37	5	96652.86	1	1	0	98686.4	1
14	15566253	Manning	580	Germany	Male	44	9	143391.1	1	0	0	146891.1	1

Problem Description

- Predict customers exit (1) or not(0) exited.
- This is a binary prediction result.
- You need to use ‘public.csv’ to build PySpark ML model.
- TA will load hidden dataset to do evaluataion.

- Please show your output as the following type:

```
CustomerID,Exited
12313123,0
32121311,0
…
```

Grading Policy

- Total score: 100
- If your result over baseline, your score is more than 70.
- The Top-10% students get 100, Top-30% students get 90, and so on.
- Baseline: f1 score ≥ 0.72

	Top-10%	Top-30%	Top-50%	Over baseline
score	100	90	80	70

Requirement and Reminder

- Use template Jupyter Notebook file to do this homework.
- TA will use public dataset to validate your model, then load private dataset and use your model to predict the result with Jupyter Notebook.

- If your output format is wrong, your score will have some discount (score*0.8).

Deadline

- Submission Deadline: before 2022/06/12   23:59 (on E3)
- Submission File: CCBDA-HW5-[Student_ID].ipynb
  - E3: jupyter notebook (only one file)
  - Remember write your student ID to rename the file.

- If you have any question, feel free to send email to contact TA.
  - TA: Wei-Lun Tseng (Eric) (曾偉倫)
  - Email: eric840610.ee02@g2.nctu.edu.tw

Useful Resource

- (English)Tutorial: Build a machine learning app with Apache Spark MLlib - Azure Synapse Analytics | Microsoft Docs
- (Traditional Chinese)教學課程：使用 Apache Spark MLlib 建置機器學習應用程式 - Azure Synapse Analytics | Microsoft Docs