

# 實作機器學習

Kaggle 數據資料實作

**Kristen Chan / Ning Chen**

2019.02.16

# Agenda

- Kaggle
- Microsoft Azure Machine Learning
- Python



\* kaggle

# Hello !

I am Kristen Chan

Data Scientist

E-Commerce / Telecom

R-Ladies Taipei

Co-Organizer [link](#)



# Hello !

I am Ning Chen

Data Scientist

Media/Social Good

Data for Social Good

PM/Consultant [link](#)

R-Ladies Taipei

Co-Founder [link](#)



# Kaggle Introduction

你知道 **kaggle** 嗎？



# 什麼是 Kaggle ?

- Website : <https://www.kaggle.com>
- 資料科學和機器學習競賽平台
- 目前已累積超過 50 萬名、遍布超過 194 個國家的註冊用戶
- 涵蓋電腦科學、電腦視覺、生物、醫藥
- Kaggle 排行榜更成為業界找尋人才的指標

The Kaggle logo is displayed in a large, lowercase, blue sans-serif font. The letter 'k' is slightly taller than the other letters. A small trademark symbol (TM) is located at the top right corner of the letter 'g'.

kaggle™

# 開始前，你需要先...

- Sign in

Kaggle is the place to do data science projects

[See how it works](#)



Register with just one click:

We won't share anything without your permission

[Sign up with Google](#)

[Sign up with Facebook](#)

Manually create an account:

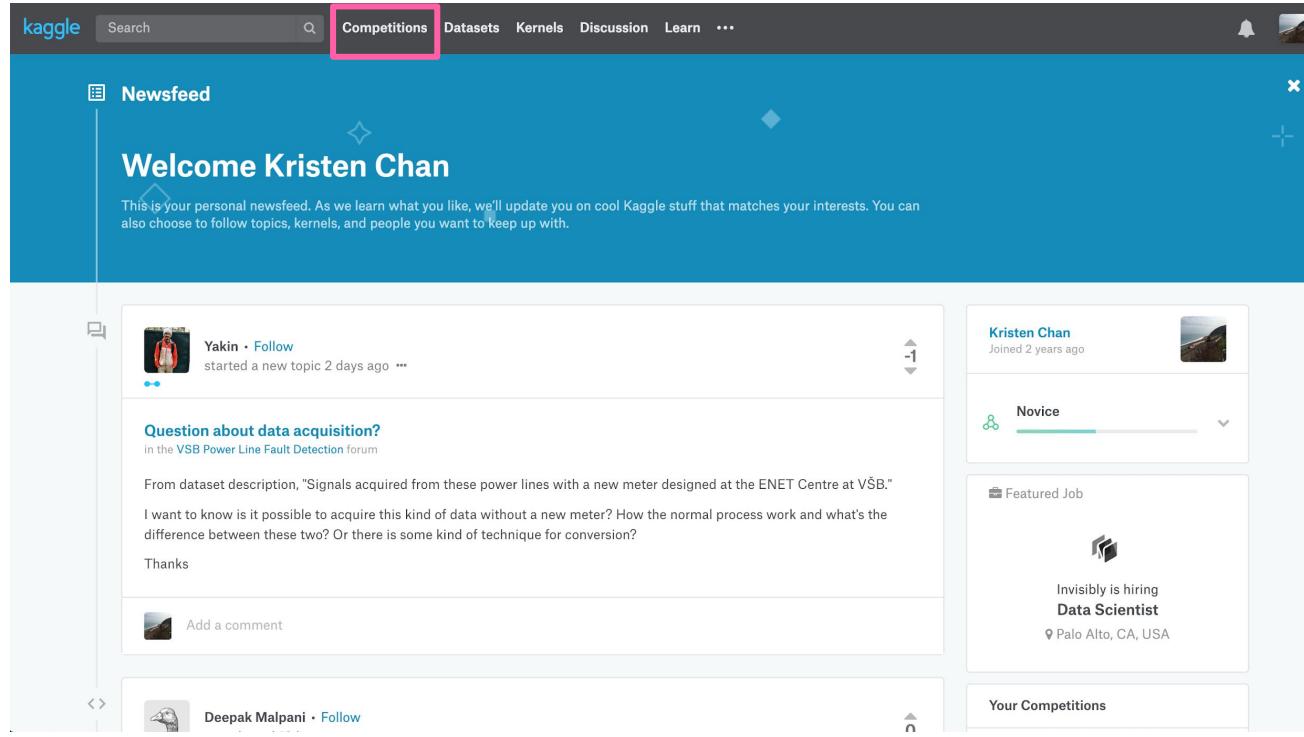
Email

Password

[Register](#)

# 準備開始...

- 找到 Competitions



The screenshot shows the Kaggle homepage. At the top, there is a navigation bar with links for 'Search', 'Competitions' (which is highlighted with a pink box), 'Datasets', 'Kernels', 'Discussion', 'Learn', and '...'. Below the navigation bar is a 'Newsfeed' section with a welcome message for 'Kristen Chan': 'Welcome Kristen Chan'. It says, 'This is your personal newsfeed. As we learn what you like, we'll update you on cool Kaggle stuff that matches your interests. You can also choose to follow topics, kernels, and people you want to keep up with.' On the left side of the newsfeed, there is a post from 'Yakin' with the title 'Question about data acquisition?' and a link to 'the VSB Power Line Fault Detection forum'. The post includes a snippet of text: 'From dataset description, "Signals acquired from these power lines with a new meter designed at the ENET Centre at VŠB." I want to know is it possible to acquire this kind of data without a new meter? How the normal process work and what's the difference between these two? Or there is some kind of technique for conversion?' Below this post is a comment input field with the placeholder 'Add a comment'. At the bottom of the newsfeed, there is another post from 'Deepak Malpani' with the title 'Your Competitions'.

Yakin · Follow  
started a new topic 2 days ago ...

**Question about data acquisition?**  
in the VSB Power Line Fault Detection forum

From dataset description, "Signals acquired from these power lines with a new meter designed at the ENET Centre at VŠB." I want to know is it possible to acquire this kind of data without a new meter? How the normal process work and what's the difference between these two? Or there is some kind of technique for conversion?

Thanks

Add a comment

Deepak Malpani · Follow

**Kristen Chan**  
Joined 2 years ago

Novice

**Featured Job**

Invisibly is hiring  
**Data Scientist**  
Palo Alto, CA, USA

Your Competitions

Kristen Chan 9

# 找一個想要比的競賽

- Titanic: Machine Learning from Disaster

The screenshot shows the Kaggle Competitions page. A pink box highlights the 'Titanic: Machine Learning from Disaster' competition, which is labeled 'Entered Competition'. An arrow points from a pink box labeled '經典' (Classic) to this highlighted section. Below it, another pink box highlights the 'Two Sigma: Using News to Predict Stock Movements' competition, labeled 'Active Competition'. A blue box at the bottom left contains the text '目前有 18 個正在進行中的比賽' (Currently there are 18 active competitions).

Competitions

Documentation InClass

General InClass

Sort by Grouped

All Categories Search competitions

1 Entered Competition

Titanic: Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

Getting Started · Ongoing · tutorial, tabular data, binary classification

Knowledge 10,049 teams

18 Active Competitions

Two Sigma: Using News to Predict Stock Movements

Performance

\$100,000 2,895 teams

to go · news agencies, time series, finance, money

LANL Earthquake Prediction

Can you predict upcoming laboratory earthquakes?

Research · 4 months to go · earth sciences, physics, signal processing

\$50,000 1,073 teams

Elo Merchant Category Recommendation

Help understand customer loyalty

\$50,000 3,727 teams

目前有 18 個正在進行中的比賽

KristenChian 10

# 獎牌資格



## Competition Medals

Competition medals are awarded for top competition results. The number of medals awarded per competition varies depending on the size of the competition. Note that InClass, playground, and getting started competitions do not award medals.

	0-99 Teams	100-249 Teams	250-999 Teams	1000+ Teams
Bronze	Top 40%	Top 40%	Top 100	Top 10%
Silver	Top 20%	Top 20%	Top 50	Top 5%
Gold	Top 10%	Top 10	Top 10 + 0.2%*	Top 10 + 0.2%*

\* (Top 10 + 0.2%) means that an extra gold medal will be awarded for every 500 additional teams in the competition. For example, a competition with 500 teams will award gold medals to the top 11 teams and a competition with 5000 teams will award gold medals to the top 20 teams.

# Titanic Data Sets

# Titanic : Machine Learning from Disaster

The screenshot shows the Kaggle competition page for 'Titanic: Machine Learning from Disaster'. At the top, there's a banner with the competition name and a small icon. Below the banner, the title 'Titanic: Machine Learning from Disaster' is displayed in bold. A sub-headline says 'Start here! Predict survival on the Titanic and get familiar with ML basics'. The Kaggle logo is present with the text 'Kaggle · 10,049 teams · Ongoing'. Below this, there's a navigation bar with links: Overview (underlined), Data, Kernels, Discussion, Leaderboard, Rules, Team, My Submissions, and Submit Predictions. The 'Submit Predictions' button is highlighted in blue.

Overview

Description

**Start here if...**

Evaluation

You're new to data science and machine learning, or looking for a simple intro to the Kaggle prediction competitions.

Tutorials

Frequently Asked Questions

**Competition Description**

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

In this challenge, we ask you to complete the analysis of what sorts of people were likely to survive. In

KristenChau 13

# All Data Sets

資料集	資料維度	欄位說明		
A. training.csv	891 X 12	1. PassengerId	乘客 ID	
B. testing.csv	418 X 11	2. Survived	是否存活	0 = No, 1 = Yes
C. gender_submission.csv	418 X 2	3. Pclass	船票的等級	1 = 1st, 2 = 2nd, 3 = 3rd
		4. Name	乘客姓名	
		5. Sex	性別	
		6. Age	年齡	
		7. SibSp	兄弟姊妹/配偶數	
		8. Parch	父母/小孩數	
		9. Ticket	船票編號	
		10. Fare	船票價格	
		11. Cabin	船艙號碼	
		12. Embarked	登岸的港口	C = Cherbourg Q = Queenstown S = Southampton

# All Data Sets

表示社會經濟地位 : 1st = Upper,  
2nd = Middle,  
3rd = Lower

資料集
A. training.csv
B. testing.csv
C. gender_submission.csv

資料維度
891 X 12
418 X 11

1. PassengerId	乘客 ID	
2. Survived	是否存活	0 = No, 1 = Yes
3. Pclass	船票的等級	1 = 1st, 2 = 2nd, 3 = 3rd
4. Name	乘客姓名	
5. Sex	性別	
6. Age	年齡	
7. SibSp	兄弟姊妹/配偶數	
8. Parch	父母/小孩數	
9. Ticket	船票編號	
10. Fare	船票價格	
11. Cabin	船艙號碼	
12. Embarked	登岸的港口	C = Cherbourg Q = Queenstown S = Southampton

# All Data Sets

資料集
A. training.csv
B. testing.csv
C. gender_submission.csv

資料維度
891 X 12
418 X 11
418 X 2

小於 1 歲：會以小數表示  
若看到 xx.5 的年齡表示是估計的

1. PassengerId	乘客 ID	
2. Survived	是否存活	0 = No, 1 = Yes
3. Pclass	船票的等級	1 = 1st, 2 = 2nd, 3 = 3rd
4. Name	乘客姓名	
5. Sex	性別	
6. Age	年齡	
7. SibSp	兄弟姊妹/配偶數	
8. Parch	父母/小孩數	
9. Ticket	船票編號	
10. Fare	船票價格	
11. Cabin	船艙號碼	
12. Embarked	登岸的港口	C = Cherbourg Q = Queenstown S = Southampton

# All Data Sets

資料集
A. training.csv
B. testing.csv
C. gender_submission.csv

資料維度
891 X 12
418 X 11

Sibling = 兄弟, 姐妹, 繼兄弟, 繼姐妹  
Spouse = 丈夫, 妻子(不包括情婦和未婚夫)

1. PassengerId	乘客 ID	
2. Survived	是否存活	0 = No, 1 = Yes
3. Pclass	船票的等級	1 = 1st, 2 = 2nd, 3 = 3rd
4. Name	乘客姓名	
5. Sex	性別	
6. Age	年齡	
7. SibSp	兄弟姊妹/配偶數	
8. Parch	父母/小孩數	
9. Ticket	船票編號	
10. Fare	船票價格	
11. Cabin	船艙號碼	
12. Embarked	登岸的港口	C = Cherbourg Q = Queenstown S = Southampton

# All Data Sets

資料集
A. training.csv
B. testing.csv
C. gender_submission.csv

資料維度
891 X 12
418 X 11

Parent = 父親, 母親  
Child = 兒子, 女兒, 繼兒子, 繼女兒  
[Note] 有些小孩只有保母陪同, 所以 Parch = 0

1. PassengerId	乘客 ID	
2. Survived	是否存活	0 = No, 1 = Yes
3. Pclass	船票的等級	1 = 1st, 2 = 2nd, 3 = 3rd
4. Name	乘客姓名	
5. Sex	性別	
6. Age	年齡	
7. SibSp	兄弟姊妹/配偶數	
8. Parch	父母/小孩數	
9. Ticket	船票編號	
10. Fare	船票價格	
11. Cabin	船艙號碼	
12. Embarked	登岸的港口	C = Cherbourg Q = Queenstown S = Southampton

# All Data Sets

資料集
A. training.csv
B. testing.csv
C. gender_submission.csv

資料維度
891 X 12
418 X 11

Cherbourg : 瑟堡, 法國西北屬重要軍港和商  
Queenstown : 科芙, 愛爾蘭  
Southampton : 南安普敦, 英國南方 ← Titanic 出航

1. PassengerId	乘客 ID	
2. Survived	是否存活	0 = No, 1 = Yes
3. Pclass	船票的等級	1 = 1st, 2 = 2nd, 3 = 3rd
4. Name	乘客姓名	
5. Sex	性別	
6. Age	年齡	
7. SibSp	兄弟姊妹/配偶數	
8. Parch	父母/小孩數	
9. Ticket	船票編號	
10. Fare	船票價格	
11. Cabin	船艙號碼	
12. Embarked	登岸的港口	C = Cherbourg Q = Queenstown S = Southampton

# All Data Sets

資料集	資料維度	
A. training.csv	891 X 12	
B. testing.csv	418 X 11	
C. gender_submission.csv	418 X 2	

1. PassengerId 乘客 ID  
2. Pclass 船票的等級 1 = 1st, 2 = 2nd, 3 = 3rd  
3. Name 乘客姓名  
4. Sex 性別  
5. Age 年齡  
6. SibSp 兄弟姊妹/配偶數  
7. Parch 父母/小孩數  
8. Ticket 船票編號  
9. Fare 船票價格  
10. Cabin 船艙號碼  
11. Embarked 登岸的港口 C = Cherbourg  
Q = Queenstown  
S = Southampton

# All Data Sets

資料集

A. training.csv

資料維度

891 X 12

B. testing.csv

418 X 11

C. gender\_submission.csv

最後上傳的格式

418 X 2

1. PassengerId 乘客 ID  
2. Survived 是否存活 0 = No, 1 = Yes

# Azure ML Studio

# 什麼是 Microsoft Azure Machine Learning

- 不用安裝
- 不用寫程式
  - 資料清洗
  - 機器學習演算法
- 也可以支援 Python 或 R 加強運算
- 可佈署成 Web Services, 分享給他人使用

# Microsoft Azure Machine Learning Studio

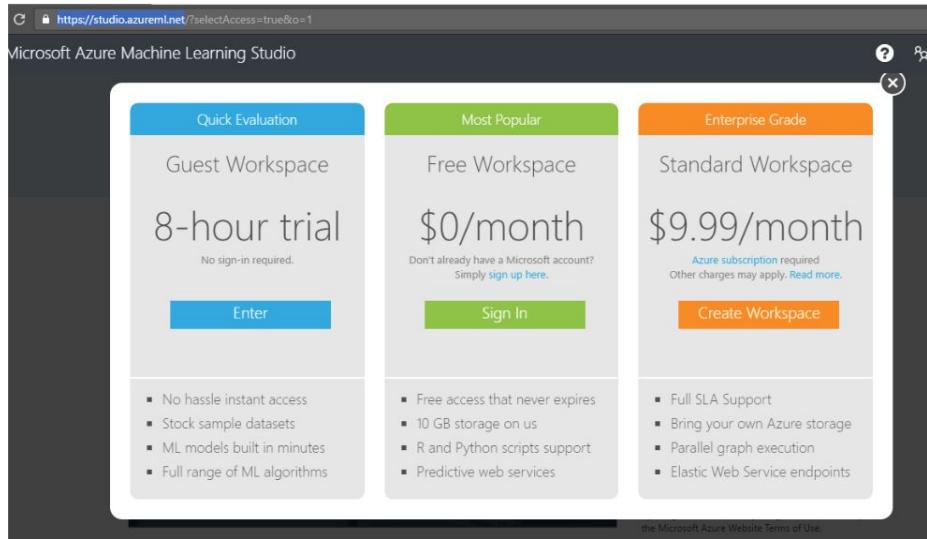
- Website : <https://studio.azureml.net/>

The screenshot shows the homepage of the Azure Machine Learning Studio. At the top, there's a navigation bar with the Microsoft logo, a search icon, and user profile icons. Below the header, a large banner features a blue 3D geometric shape and the text "Azure Machine Learning service". A yellow "New!" badge is in the top right corner of the banner. Below the banner, a call-to-action button says "Try it today!". To the right of the banner, there's a "Welcome to Azure Machine Learning" message, a "Sign In" button, and links for "Pricing & FAQ" and "Terms of Use". At the bottom of the page, there's a section for "Announcements NEW!" with three items: "Azure Machine Learning Studio R Runtime Upgrade" (aired August 31, 2017), "Mining Campaign Funds" (aired August 03, 2017), and "Inside the Data Science VM" (aired June 21, 2016).



# Microsoft Azure Machine Learning Studio

- 使用 Microsoft Account 登入, 選擇 Free Workspace
- 不登入使用 Guest Workspace



# 建立 Dataset

Microsoft Azure Machine Learning Studio

Kristen-Free-Workspace    

datasets

MY DATASETS SAMPLES

NAME	SUBMITTED BY	DESCRIPTION	DATA TYPE	CREATED	SIZE	PROJECT	SEARCH
No datasets found							

1 新增

PROJECTS EXPERIMENTS WEB SERVICES NOTEBOOKS DATASETS TRAINED MODELS SETTINGS

NEW DOWNLOAD DELETE OPEN IN NOTEBOOK GENERATE DATA ACCESS ADD TO PROJECT



# 建立 Dataset

The screenshot shows the Microsoft Azure Machine Learning Studio interface. At the top, there's a navigation bar with the title "Microsoft Azure Machine Learning Studio", a workspace dropdown "Kristen-Free-Workspace", and user icons. Below the navigation bar, the main area is titled "datasets". It features a "NEW" section with several options: "DATASET" (which is highlighted with a pink border), "MODULE", "PROJECT PREVIEW", "EXPERIMENT", and "NOTEBOOK PREVIEW". A sub-section titled "Upload a new dataset from a local file" contains a "FROM LOCAL FILE" button. On the left side of the main area, there are two tabs: "MY DATASETS" and "SAMPLES". A large pink circle with the number "2" is overlaid on the "DATASET" button, with the text "選擇上傳檔案" (Select Upload File) written next to it.

Microsoft Azure Machine Learning Studio

Kristen-Free-Workspace

PROJECTS datasets

MY DATASETS SAMPLES

NEW

DATASET FROM LOCAL FILE

Upload a new dataset from a local file

2 選擇上傳檔案

MODULE

PROJECT PREVIEW

EXPERIMENT

NOTEBOOK PREVIEW

KristenChian 27

# 建立 Dataset

The screenshot shows the Microsoft Azure Machine Learning Studio interface. On the left, there's a sidebar with icons for Projects, Experiments, Web Services, Notebooks, Datasets (which is selected), Trained Models, and Settings. The main area is titled 'datasets' and shows a table with columns: NAME, SUBMITTED BY, DESCRIPTION, DATA TYPE, CREATED, SIZE, and PROJECT. A search bar is at the top right of the table. Below the table, it says 'No datasets found'. A modal window titled 'Upload a new dataset' is open in the center. It contains fields for selecting a file ('SELECT THE DATA TO UPLOAD:'), marking it as a new version of an existing dataset ('This is the new version of an existing dataset'), entering a name ('ENTER A NAME FOR THE NEW DATASET:'), selecting a type ('SELECT A TYPE FOR THE NEW DATASET:'), providing an optional description ('PROVIDE AN OPTIONAL DESCRIPTION:'), and a checkmark button at the bottom right. A pink circle with the number '3' is overlaid on the 'SELECT THE DATA TO UPLOAD:' field.

Microsoft Azure Machine Learning Studio

Kristen-Free-Workspace

PROJECTS

EXPERIMENTS

WEB SERVICES

NOTEBOOKS

DATASETS

TRAINED MODELS

SETTINGS

NEW

datasets

MY DATASETS SAMPLES

NAME SUBMITTED BY DESCRIPTION DATA TYPE CREATED SIZE PROJECT

No datasets found

Upload a new dataset

SELECT THE DATA TO UPLOAD:  
[選擇檔案] Titanic\_Train.csv

This is the new version of an existing dataset

ENTER A NAME FOR THE NEW DATASET:  
Titanic\_Train.csv

SELECT A TYPE FOR THE NEW DATASET:  
Generic CSV File with a header (.csv)

PROVIDE AN OPTIONAL DESCRIPTION:  
Kaggle Titanic

3 找到對應的檔案上傳

DOWNLOAD DELETE OPEN IN NOTEBOOK GENERATE DATA ACCESS ADD TO PROJECT

KristenChian 28

# 建立 Dataset

Microsoft Azure Machine Learning Studio    Kristen-Free-Workspace ▾    ?    🚧    😊    🚙

PROJECTS    EXPERIMENTS    WEB SERVICES    NOTEBOOKS    DATASETS    TRAINED MODELS    SETTINGS

## datasets

MY DATASETS    SAMPLES

NAME	SUBMITTED BY	DESCRIPTION	DATA TYPE	CREATED	SIZE	PROJECT	🔍
Titanic_Test.csv	sinue625	Kaggle Titanic	GenericCSV	2/12/2019 4:08:50 PM	27.96 KB	None	🔍
Titanic_Train.csv	sinue625	Kaggle Titanic Train	GenericCSV	2/12/2019 4:06:41 PM	59.76 KB	None	🔍

上傳 Train.csv & Test.csv

NEW    DOWNLOAD    DELETE    OPEN IN NOTEBOOK    GENERATE DATA ACCESS    ADD TO PROJECT



# 建立 Experiment

Microsoft Azure Machine Learning Studio

Kristen-Free-Workspace    

experiments

MY EXPERIMENTS SAMPLES

LAST EDITED  PROJECT 

No experiments found

0 items selected

 1

 + NEW

 DELETE  ADD TO PROJECT  COPY TO WORKSPACE

KristenChian 30

# 建立 Experiment

Microsoft Azure Machine Learning Studio

Kristen-Free-Workspace

experiments

PROJECTS

EXPERIMENTS

NEW

DATASET

MODULE

PROJECT PREVIEW

EXPERIMENT

NOTEBOOK PREVIEW

2 Blank Experiment

Microsoft Samples

VIEW MORE IN GALLERY

Sample 1: Download dataset from UCI: Adult 2 class dataset

Sample 2: Dataset Processing and Analysis: Auto Imports Regression

Sample 3: Cross Validation for Binary Classification: Adult

Sample 4: Cross Validation for Regression: Auto Imports

Sample 5: Train, Test, Evaluate for Binary Classification: Adult

Sample 6: Train, Test, Evaluate for Regression: Auto Imports Dataset

Sample 7: Train, Test, Evaluate for Multiclass Classification: Letter

Sample 8: Apply SQL transformation

Sample 9: Split, partition and sample system

Anomaly Detection: Credit Risk

Binary Classification:

Binary Classification:

Binary Classification:

Binary Classification:

Binary Classification:

Binary Classification:

KristenChian 31

# 介面介紹

可以改標題名稱

The screenshot shows the Microsoft Azure Machine Learning Studio interface. At the top, there's a navigation bar with the title "Microsoft Azure Machine Learning Studio" and a user profile "Kristen-Free-Workspace". On the left, a sidebar labeled "A" contains a search bar and a list of modules: Saved Datasets, Data Format Conversions, Data Input and Output, Data Transformation, Feature Selection, Machine Learning, OpenCV Library Modules, Python Language Modules, R Language Modules, Statistical Functions, Text Analytics, Time Series, Web Service, and Deprecated. A green callout box highlights the title "Kaggle Titanic Easy" at the top of the workspace. The main workspace is titled "[Canvas]" and has a subtitle "編輯實驗 並放上你需要的 Module". It features a "Drag Items Here" placeholder and a "Mini Map" window. The right side of the interface includes sections for "Properties", "Experiment Properties" (Status Code: InDraft), "Summary" (with a text input field), "Description" (with a text input field), and "Quick Help". At the bottom, there are buttons for "RUN HISTORY", "SAVE", "DISCARD CHANGES", "RUN", "SET UP WEB SERVICE", and "PUBLISH TO GALLERY".

# 介面介紹

B Microsoft Azure Machine Learning Studio

Kristen-Free-Workspace ? ☺ 🚧

In draft Properties Project

Experiment Properties STATUS CODE InDraft

Summary Enter a few sentences describing your experiment (up to 140 characters).

Description Enter the detailed description for your experiment.

Quick Help

Google Titanic Easy [Module]

執行各種操作 To create your experiment, drag and drop datasets and modules here

Drag Items Here

a. 可從上方搜尋欄找 Module  
b. 可直接拖曳要使用到 Module 到 Canvas 中

Mini Map

RUN HISTORY SAVE AS DISCARD CHANGES SET UP WEB SERVICE PUBLISH TO GALLERY

The screenshot shows the Azure Machine Learning Studio interface. On the left is a sidebar with a 'Saved Datasets' section and a 'Modules' section containing categories like Data Input and Output, Data Transformation, Feature Selection, Machine Learning, etc. The main canvas area has a green header 'Google Titanic Easy [Module]' with a sub-instruction 'To create your experiment, drag and drop datasets and modules here'. Below this is a 'Drag Items Here' placeholder. A dashed line connects the 'Data Input and Output' module in the sidebar to this placeholder. At the bottom of the canvas is a 'Mini Map' window showing the current experiment structure. Along the bottom edge are several toolbars: RUN HISTORY, SAVE AS, DISCARD CHANGES, SET UP WEB SERVICE, and PUBLISH TO GALLERY. The top right corner features a user profile for 'Kristen-Free-Workspace' with icons for help, user, smiley face, and a gear.

Kristen Chian 33

# 介面介紹

Kaggle Titanic Easy

Drag Items Here

**[Properties]**

當點選 Canvas 上的 Module 時, Properties 會出現對應的參數供調整

**[Note] Quick Help**  
可以幫助你更深入了解個參數意義

The screenshot shows the Microsoft Azure Machine Learning Studio interface. On the left is a sidebar with various icons and a list of modules: Saved Datasets, Data Format Conversions, Data Input and Output, Data Transformation, Feature Selection, Machine Learning, OpenCV Library Modules, Python Language Modules, R Language Modules, Statistical Functions, Text Analytics, Time Series, Web Service, and Deprecated. The main area is titled "Kaggle Titanic Easy" and contains a "Mini Map" visualization. A callout box highlights the "Properties" panel on the right, which is titled "[Properties]" and displays "In draft". It includes sections for Experiment Properties (Status Code: InDraft), Summary (a text input field), Description (another text input field), and a "Quick Help" section. A large callout box in the center-right area explains that when a module is selected on the canvas, its properties will appear in the properties panel, and it also mentions the [Note] Quick Help feature. The bottom navigation bar includes buttons for NEW, RUN HISTORY, SAVE, DISCARD CHANGES, RUN, SET UP WEB SERVICE, and PUBLISH TO GALLERY.

Kristen-Free-Wo...@live.com

C



Properties Project

Experiment Properties

STATUS CODE InDraft

Summary

Enter a few sentences describing your experiment (up to 140 characters).

Description

Enter the detailed description for your experiment.

Quick Help

Kristen Chian 34



RUN HISTORY



SAVE



DISCARD CHANGES



RUN



SET UP WEB SERVICE



PUBLISH TO GALLERY

# 介面介紹

Microsoft Azure Machine Learning Studio

Kaggle Titanic Easy

In draft

Properties Project

Experiment Properties

Status Code InDraft

Summary

Description

Quick Help

NEW

D

RUN HISTORY SAVE SAVE AS DISCARD CHANGES RUN SET UP WEB SERVICE PUBLISH TO GALLERY

To create your experiment, drag and drop datasets and modules here

Drag Items Here

Mini Map

【控制執行】

儲存、執行、發布實驗...

Kristen Clark 35

# All Data Sets

剛剛Kaggle 的資料集



資料集	資料維度		
A. training.csv	891 X 12	1. PassengerId	乘客 ID
B. testing.csv	418 X 11	2. Survived	是否存活 0 = No, 1 = Yes
C. gender_submission.csv	418 X 2	3. Pclass	船票的等級 1 = 1st, 2 = 2nd, 3 = 3rd

- 4. Name 乘客姓名
- 5. Sex 性別
- 6. Age 年齡
- 7. SibSp 兄弟姊妹/配偶數
- 8. Parch 父母/小孩數
- 9. Ticket 船票編號
- 10. Fare 船票價格
- 11. Cabin 船艙號碼
- 12. Embarked 登岸的港口
  - C = Cherbourg
  - Q = Queenstown
  - S = Southampton

# [Method 1] Access Data

Microsoft Azure Machine Learning Studio   Kristen-Free-Workspace ▾ ? 🌐 😊 🚙

Search experiment items  Drag Items Here

Saved Datasets

- My Datasets
  - Titanic\_Test.csv
  - Titanic\_Train.csv**
- Samples
- Data Format Conversions
- Data Input and Output
- Data Transformation
- Feature Selection
- Machine Learning
- OpenCV Library Modules
- Python Language Modules
- R Language Modules
- Statistical Functions
- Text Analytics
- Time Series

匯入資料  
→ 從 Module 選擇, Saved Datasets 中的 [Titanic\_Train.csv]

Summary

Enter a few sentences describing your experiment (up to 140 characters).

Description

Enter the detailed description for your experiment.

Quick Help

Mini Map

Run History Save Discard Changes Run Set Up Web Service Publish to Gallery

Kristen Chian 37

# [Method 1] Access Data

資料設定完成後，檢查資料輸入狀況

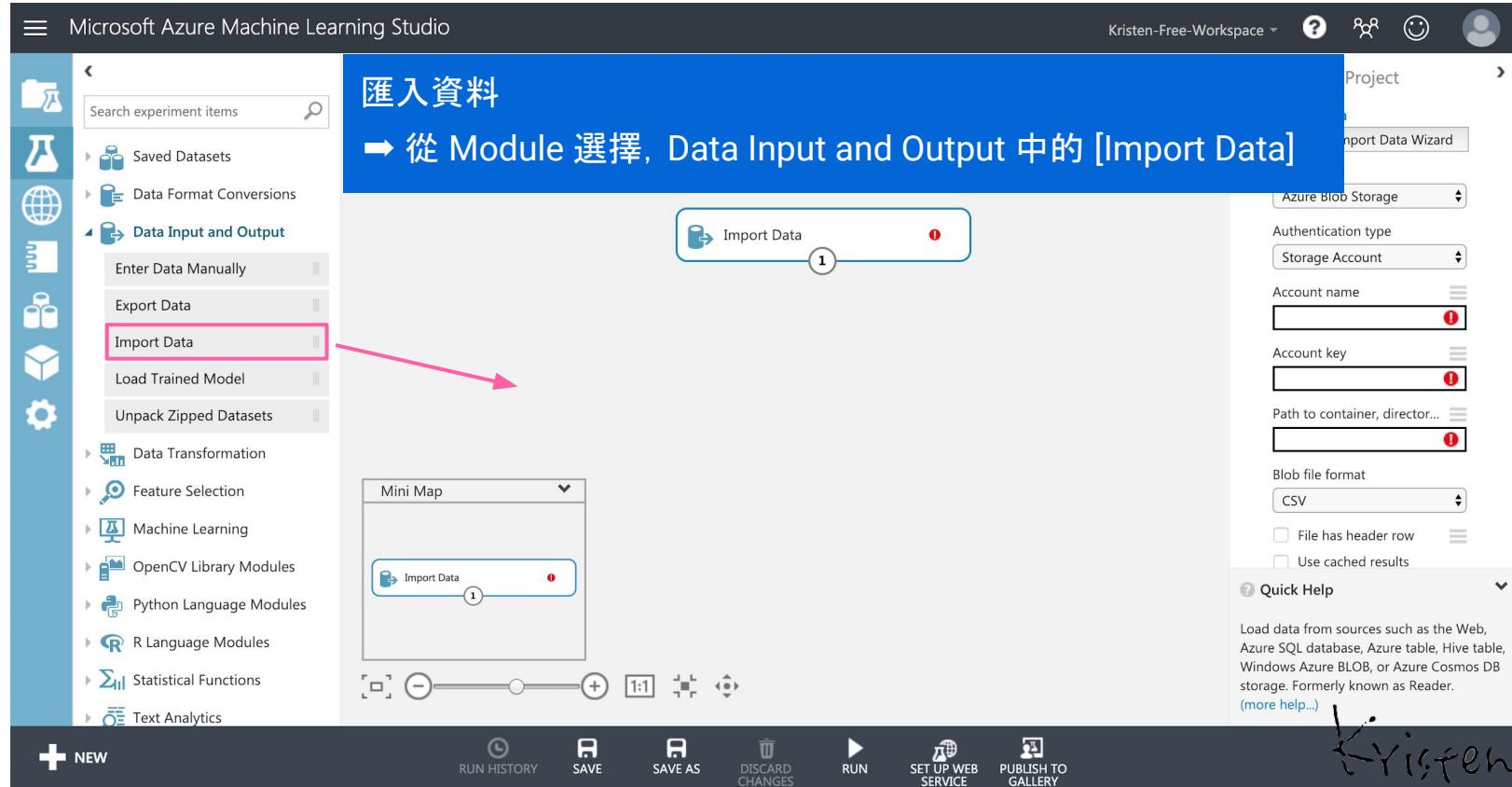
1. [Run]
2. Datasets 右鍵 → 選擇 [Visualize]

The screenshot shows the Azure ML Studio interface. On the left, there's a sidebar with various icons and a list of datasets and input/output options. In the center, a 'Mini Map' shows the 'Titanic\_Train.csv' dataset. A context menu is open over the dataset icon in the mini map, with the 'Visualize' option highlighted. The 'Properties' pane on the right displays the following information for 'Titanic\_Train.csv':

Submitted By	sinue625
Size	59.8 KB
Format	GenericCSV
Created On	2/12/2019 ...

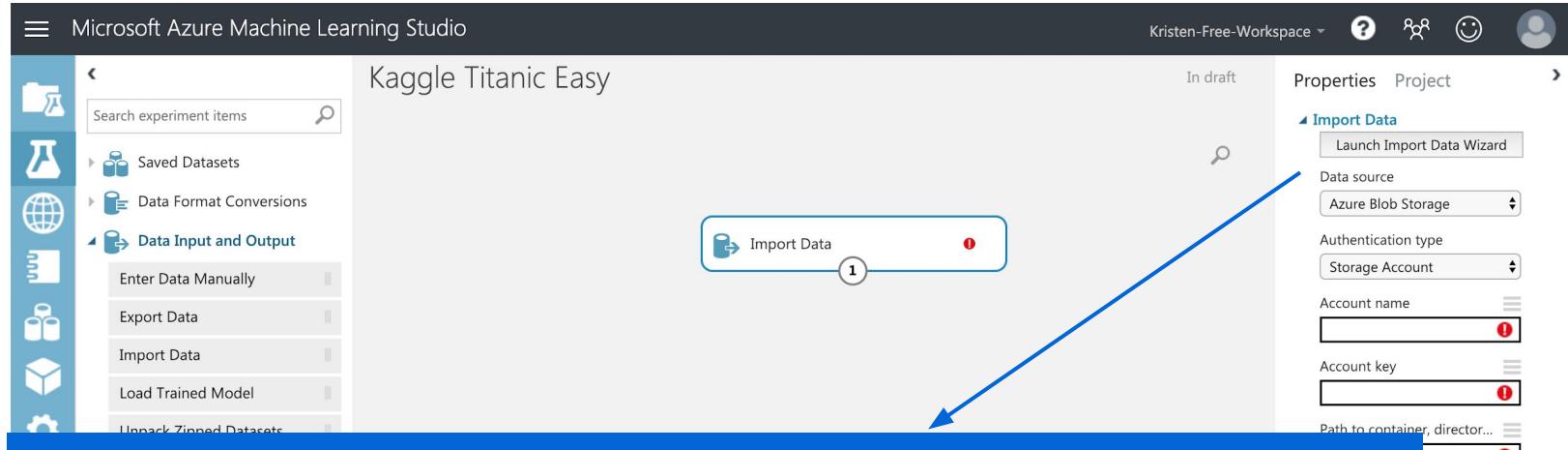
At the bottom, there are navigation buttons for Run History, Save, Publish to Gallery, etc.

# [Method 2] Access Data



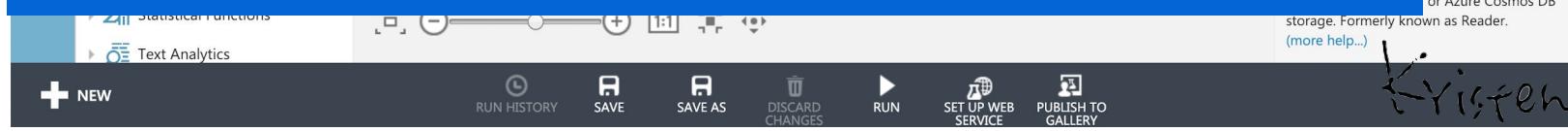
KrisenCheck 39

# [Method 2] Access Data



## Properties 設定

1. Data Source 選擇 : [Web URL via Http]
2. Data Source URL 輸入
3. CSV or TSV has header : 若資料含有 Header 就要打勾
4. Use cached results : 打勾表示把資料 cached 起來就不用每次執行實驗都重抓資料



Kristen Chian 40

[https://raw.githubusercontent.com/radestaipei/Azureml-shiny-app/master/Shiny\\_Titanic/Titanic\\_train.csv](https://raw.githubusercontent.com/radestaipei/Azureml-shiny-app/master/Shiny_Titanic/Titanic_train.csv)

# [Method 2] Access Data

The screenshot illustrates the process of importing data from a web URL in Azure ML Studio. It is divided into three main sections:

- Step 1: Launch Import Data Wizard** (Top Right): A button labeled "Import Data" with a magnifying glass icon is highlighted with a pink circle and arrow. The "Properties" panel on the right shows the "Data source" set to "Web URL via HTTP".
- Step 2: Choose data source** (Center): A modal window titled "IMPORT DATA" lists various data sources. The "Web URL via HTTP" option is highlighted with a pink box and a pink circle with the number "2". Other options include Hive Query, Azure SQL Database, Azure Table, Azure Blob Storage, Data Feed Provider, On-Premises SQL Database (Preview Feature), and Azure DocumentDB.
- Step 3: Connect to Web URL via HTTP** (Left): The main workspace shows the "Import Data" step. The "Data source URL" field contains the URL "https://raw.githubusercontent.com/radestaipei/Azureml-shiny-app/master/Shiny\_Titanic/Titanic\_train.csv". The "Data format" dropdown is set to "CSV". A checkbox "CSV or TSV has header row" is checked and highlighted with a pink box and a pink circle with the number "3". A pink box with the text "資料有 Header 就要打勾" (Check if the data has a header) is overlaid on this section. The "RUN HISTORY" tab at the bottom is also highlighted with a pink box and a pink circle with the number "3".

A large pink arrow points from Step 1 down to Step 2, and another pink arrow points from Step 2 down to Step 3. A dashed blue arrow points from the URL in Step 3 up towards the top center of the slide.

Kiyffen Chian 41

# [Method 2] Access Data

資料設定完成後，檢查資料輸入狀況

## 1. [Run]

The screenshot shows the Azure ML Studio interface. On the left, a sidebar menu is open under the 'Data Input and Output' section, with 'Import Data' selected. In the center, a 'Mini Map' window displays a single step labeled 'Import Data'. At the bottom right of the map, there is a pink circle with the number '1'. Below the map, a toolbar has several buttons, with the 'Run' button highlighted in pink and also having a pink circle with the number '1' above it. The main workspace on the right is titled 'Import Data' and shows configuration options for importing data from a Web URL via HTTP. The URL is set to 'https://raw.githubusercontent.com'. The 'Data format' is set to 'CSV'. A checked checkbox indicates 'CSV or TSV has header'. The status bar at the bottom shows 'Kristen-Free-Workspace' and 'In draft'.

In draft

Draft saved at 下午1:35:46

Properties Project

Import Data

Launch Import Data Wizard

Data source

Web URL via HTTP

Data source URL

https://raw.githubusercontent.com

Data format

CSV

CSV or TSV has header

Use cached results

Quick Help

Load data from sources such as the Web, Azure SQL database, Azure table, Hive table, Windows Azure BLOB, or Azure Cosmos DB storage. Formerly known as Reader.  
(more help...)

NEW

RUN HISTORY

SAVE

DISCARD CHANGES

RUN

SET UP WEB SERVICE

PUBLISH TO GALLERY

Kristen Chian 42

# [Method 2] Access Data

資料設定完成後，檢查資料輸入狀況

1. [Run]
2. Import Data 右鍵 → Results dataset 選擇 [Visualize]

The screenshot shows the Azure ML Studio interface. On the left, there is a 'Mini Map' showing the flow of the experiment. A blue rounded rectangle highlights the 'Import Data' step. A context menu is open over this step, with a pink circle labeled '1' pointing to the 'Results dataset' option. Another pink circle labeled '2' points to the 'Visualize' option in the same menu. To the right of the mini map, the main workspace shows the 'Import Data' step connected to other components. On the far right, the 'Properties' pane is open, showing details for the 'Import Data' step, including the data source URL (<https://raw.githubusercontent.com>), data format (CSV), and status (Finished). Below the properties pane, a 'Quick Help' section provides information about loading data from various sources.

Properties Project

Import Data

- Launch Import Data Wizard
- Data source
- Web URL via HTTP
- Data source URL
- https://raw.githubusercontent.com
- Data format
- CSV
- CSV or TSV has header
- Use cached results
- START TIME 9/3/2018 1...
- END TIME 9/3/2018 1...
- ELAPSED TIME 0:00:09.990
- STATUS CODE Finished
- STATUS DETAILS None

View output log

Import Data

1

2

Results dataset

- Download
- Save as Dataset
- Save as Trained Model
- Save as Transform
- Visualize
- Generate Data Access Code...
- Open in a new Notebook

Mini Map

Import Data

Quick Help

Load data from sources such as the Web, Azure SQL database, Azure table, Hive table, Windows Azure BLOB, or Azure DocumentDB storage. Formerly known as Reader. (more help)

# Access Data

Microsoft Azure Machine Learning Studio

Kaggle Titanic Easy

Kaggle Titanic Easy > Import Data > Results dataset

rows 891 col 8

Note 總共有 891 筆資料

檢查 Import 進來的資料有沒有問題

	Survived	PassengerClass	Gender	Age	SiblingSpouse	ParentChild	FarePrice	PortEr
0	3	male	22	1	0	7.25	S	
1	1	female	38	1	0	71.2833	C	
1	3	female	26	0	0	7.925	S	
1	1	female	35	1	0	53.1	S	
0	3	male	35	0	0	8.05	S	
0	3	male	0	0	0	8.4583	Q	
0	1	male	54	0	0	51.8625	S	
0	3	male	2	3	1	21.075	S	
1	3	female	27	0	2	11.1333	S	
1	2	female	14	1	0	30.0708	C	
1	3	female	4	1	1	16.7	S	
1	1	female	58	0	0	26.55	S	
0	2	male	20	0	0	0.25	S	

To view, select a column in the table.

Statistics

Visualizations

NEW RUN HISTORY SAVE AS DISCARD CHANGES RUN SET UP WEB SERVICE PUBLISH TO GALLERY

Kristen Chian 44

# 篩選資料

## 篩選欄位

→ 從 Module 選擇, Data Transformation 中 Manipulation 的 [Select Columns in Dataset]

The screenshot shows the Azure ML Studio interface. On the left, there is a vertical toolbar with icons for New, Run History, Save, Discard Changes, Set Up Web Service, and Publish to Gallery. Below the toolbar is a navigation bar with 'NEW', 'RUN HISTORY', 'SAVE', 'DISCARD CHANGES', 'RUN', 'SET UP WEB SERVICE', and 'PUBLISH TO GALLERY'. The main workspace contains a flowchart. At the top of the flowchart is an 'Import Data' module. Below it is a 'Select Columns in Dataset' module, which is highlighted with a red box and has a pink arrow pointing from the left sidebar towards it. To the right of the main workspace is a 'Properties' panel. In the 'Properties' panel, under 'Select Columns in Dataset', there is a section titled 'Selected columns' with the sub-section 'Selected columns' highlighted in a red box. A tooltip for 'Selected columns' says: 'Launch the selector tool to make a selection'. Another tooltip for 'Launch column selector' says: 'Selects columns to include or exclude from a dataset in an operation. Formerly known as Project Columns.' At the bottom right of the image, there is a signature that reads 'Kristen Chian 45'.

The screenshot shows the 'Properties' panel in the Azure ML Studio interface. At the top, it says 'Kristen-Free-Workspace'. There are tabs for 'Properties' and 'Project'. Under 'Properties', there is a section for 'Select Columns in Dataset'. It includes a 'Selected columns' section with a tooltip: 'Launch the selector tool to make a selection'. Below that is a 'Launch column selector' button with a tooltip: 'Selects columns to include or exclude from a dataset in an operation. Formerly known as Project Columns.' At the bottom, there is a 'Quick Help' section with a tooltip: 'Selects columns to include or exclude from a dataset in an operation. Formerly known as Project Columns.' A 'more help...' link is also present. The bottom right corner of the image has a signature that reads 'Kristen Chian 45'.

# 篩選資料

Microsoft Azure Machine Learning Studio

Kaggle Titanic Easy

In draft

Draft saved at: 午後1:39:45

Properties Project

◀ Select Columns in Dataset

Select columns

Selected columns: Launch the selector tool to make a selection

Launch column selector

Import Data → Select Columns in Dataset

按著連接點拖到連接處

Note 因為還沒設定 Properties

Mini Map

Import Data → Select Columns in Dataset

Quick Help

Selects columns to include or exclude from a dataset in an operation. Formerly known as Project Columns.

(more help...)

NEW

RUN HISTORY

SAVE

SAVE AS

DISCARD CHANGES

RUN

SET UP WEB SERVICE

PUBLISH TO GALLERY

按著連接點拖到連接處

Note 因為還沒設定 Properties

# 篩選資料

Microsoft Azure Machine Learning Studio   Kristen-Free-Workspace   Properties Project

Kaggle Titanic Easy   In draft   Draft saved at 下午1:40:54

Search experiment items   Data Format Conversions   Data Input and Output   Data Transformation   Filter   Learning with Counts   Manipulation   Add Columns   Add Rows   Apply SQL Transformation...   Clean Missing Data   Convert to Indicator Value...   Edit Metadata   Group Categorical Values   Join Data   Remove Duplicate Rows   Select Columns in Dataset...   Select Columns Transform...

BY NAME   WITH RULES   BY NAME   WITH RULES

Available Columns   Selected Columns

PassengerId  
Name  
Sex  
SibSp  
Parch  
Ticket  
Fare  
Cabin  
Embarked

Survived  
Pclass  
Age

要的選過來

1 Selected columns: Use the selector tool to make a selection   Launch column selector

2 Select columns

3 選完記得打勾

Quick Help: Selects columns to include or exclude from a dataset in an operation. Formerly known as Project Columns. (more help...)

RUN HISTORY   SAVE   DISCARD CHANGES   RUN   SET UP WEB SERVICE   PUBLISH TO GALLERY

NEW   Kristen Chaw 47

The screenshot shows the 'Select Columns in Dataset' tool in the Azure ML Studio. The sidebar on the left has 'Manipulation' selected. The main area shows a 'Select columns' dialog with 'BY NAME' selected. A large pink box covers the list of available columns, with the text '要的選過來' (Select what you need) overlaid. A callout box labeled '1' points to the 'Selected columns:' dropdown. A callout box labeled '2' points to the 'Select columns' button. A callout box labeled '3' points to a checked checkbox in the bottom right corner, with the text '選完記得打勾' (Don't forget to check after selecting) overlaid. The status bar at the bottom right says 'Kristen Chaw 47'.

# 清理資料

Microsoft Azure Machine Learning Studio

篩選欄位

→ 從 Module 選擇, Data Transformation 中 Manipulation 的 [Clean Missing Data]

Import Data

Select Columns in Dataset

Clean Missing Data

1 2

Minimum missing value range: 0

Maximum missing value range: 1

Cleaning mode: Custom substitution value

Replacement value: 0

Generate missing values

Quick Help: Specifies how to handle the values missing from a dataset (more help...)

Kristen Chan 48

# [Method 1] 清理資料 -- 整筆刪掉

Kaggle Titanic Easy

In draft

Draft saved at 下午1:23:23

Titanic\_Train.csv → Select Columns in Dataset → Clean Missing Data

Properties Project

Clean Missing Data

Selected columns: All columns

Launch column selector

Minimum missing value ratio: 0

Replace using MICE  
Custom substitution value  
Replace with mean  
Replace with median  
Replace with mode  
**Remove entire row** (highlighted)

Replace entire column  
Replace using Probabilistic PCA

Mini Map

處理 Missing 方法

在 Cleaning mode 中選 [Remove entire row] : 當其中一個欄位出現遺失值時，整筆(row)刪掉

Quick Help

Specifies how to handle the values missing from a dataset (more help...)

NEW RUN HISTORY SAVE AS DISCARD SET UP WEB PUBLISH TO

## 處理 Missing 方法

在 Cleaning mode 中選  
[Remove entire row] : 當其中一個欄位出現遺失  
值時，整筆(row)刪掉

Kirilen Chata 49

# [Method 1] 清理資料 -- 整筆刪掉

Kaggle Titanic Easy

Kaggle Titanic Easy > Clean Missing Data > Cleaned dataset

ROWS: 714

Note 剩下有 714 筆資料

	Survived	Pclass	Age
A	0	3	22
A	1	1	38
A	1	3	26
A	1	1	35
A	0	3	35
A	0	1	54
C	0	3	2
C	1	3	27
E	1	2	14
G	1	3	4
J	1	1	58
R	0	3	20
R	0	3	20

To view, select a column in the table.

Statistics

Visualizations

missing

Run History

Save

Save As

Discard Changes

Run

Set Up Web Service

Publish to Gallery

Kristen Chian 50

# [Method 2] 清理資料 -- 用中位數取代

Microsoft Azure Machine Learning Studio

Kaggle Titanic Easy

Kristen-Free-Workspace

Properties Project

Finished running ✓

**Clean Missing Data**

Columns to be cleaned  
Selected columns: All columns

Launch column selector

Minimum missing value r... 0

Maximum missing value r... 0

Replace using MICE  
Custom substitution value  
Replace with mean  
**Replace with median** (highlighted)

Replace with mode  
Remove entire row  
Remove entire column  
Replace using Probabilistic PCA

Generate missing val...

Mini Map

START TIME 2/14/2019 ...  
END TIME 2/14/2019 ...  
Quick Help

**處理 Missing 方法**  
在 Cleaning mode 中選 [Replace with median] : 當數值欄位出現遺失值時, 用中位數取代整筆

```

graph TD
    A[Titanic_Train.csv] --> B[Select Columns in Dataset]
    B --> C[Clean Missing Data]
    style C fill:#0078d4,color:#fff
    style C stroke:#0078d4
    style C stroke-width:2px
    C -- "1" --> C -- "2" --> C
  
```

Kristen Chan 51

# [Method 2] 清理資料 -- 用中位數取代

Kaggle Titanic Easy

Finished running ✓ Properties Project

ROWS 891 Note 維持 891 筆資料

Kaggle Titanic Easy > Clean Missing Data > Cleaned dataset

Survived Pclass Age

檢查資料處理狀況

1. Clean Missing Data 右鍵 → Cleaned dataset 選擇 [Visualize]

	1	3	26
A	1	1	35
A	0	3	35
A	0	3	28
C	0	1	54
C	0	3	2
E	1	3	27
G	1	2	14
J	1	3	4
R	1	1	58

To view, select a column in the table.

missing

Select Columns in Dataset

NEW RUN HISTORY SAVE SAVE AS DISCARD CHANGES RUN SET UP WEB SERVICE PUBLISH TO GALLERY

KristenChian 52

# Azure ML Studio with Python Notebook

# 建立一個 Notebook

Microsoft Azure Machine Learning Studio

Kristen-Free-Workspace

PROJECTS EXPERIMENTS WEB SERVICES NOTEBOOKS DATASETS TRAINED MODELS SETTINGS

notebooks preview

NAME	LANGUAGE	LAST MODIFIED	PROJECT	...
No notebooks found				

**NEW**

DELETE RENAME ADD TO PROJECT

# 建立一個 Python3 Notebook

Microsoft Azure Machine Learning Studio

Kristen-Free-Workspace

notebooks [preview](#)

NAME LANGUAGE LAST MODIFIED PROJECT

NEW

FROM LOCAL FILE Upload a new notebook from a local file

Search notebooks

Microsoft Samples

[VIEW MORE IN GALLERY](#)

Python 3 Blank Notebook

Python 2 Blank Notebook

R Blank Notebook

Tutorial on Azure Machine Learning Notebook

[Tutorial](#) [Jupyter Notebook](#)

[Azure ML](#) [Web Service](#) [Linear](#)

Access Azure ML Experiment Data

Variable Selection in Azure ML Jupyter Notebook

GBM in Azure ML Jupyter Notebook

Evaluating Multiple Models

KristenChian 55

The screenshot shows the Microsoft Azure Machine Learning Studio interface. On the left, there's a sidebar with icons for Projects, Dataset, Module, Project (Preview), Experiment, and Notebook (Preview). The 'Notebook (Preview)' icon is highlighted with a pink border. The main area is titled 'notebooks' with a 'preview' link. It has columns for NAME, LANGUAGE, LAST MODIFIED, and PROJECT. A button 'Upload a new notebook from a local file' is present. Below this, there's a search bar labeled 'Search notebooks'. A section titled 'Microsoft Samples' shows cards for 'Python 3 Blank Notebook', 'Python 2 Blank Notebook', and 'R Blank Notebook'. To the right of these is a card for a 'Tutorial on Azure Machine Learning Notebook' featuring a woman in a lab coat and a flask. At the bottom, there are links for 'Access Azure ML Experiment Data', 'Variable Selection in Azure ML Jupyter Notebook', 'GBM in Azure ML Jupyter Notebook', and 'Evaluating Multiple Models'.

# 建立一個 Python3 Notebook

The screenshot shows the Microsoft Azure Machine Learning Studio interface. On the left, there is a sidebar with icons for PROJECTS, EXPERIMENTS, WEB SERVICES, NOTEBOOKS (which is selected), DATASETS, TRAINED MODELS, and SETTINGS. The main area is titled "notebooks" and shows a table with columns: NAME, LANGUAGE, LAST MODIFIED, and PROJECT. A modal window titled "Name Notebook" is open in the center, containing a "NOTEBOOK NAME" input field with the value "Kaggle Titanic". A blue arrow points from the text "幫 Notebook 取一個名字" (Help Notebook get a name) to the input field. At the bottom of the modal is a checkmark icon.

Microsoft Azure Machine Learning Studio

Kristen-Free-Workspace

notebooks preview

NAME LANGUAGE LAST MODIFIED PROJECT

PROJECTS EXPERIMENTS WEB SERVICES NOTEBOOKS DATASETS TRAINED MODELS SETTINGS

+

NEW

DELETE RENAME ADD TO PROJECT

Name Notebook

NOTEBOOK NAME

Kaggle Titanic

✓

幫 Notebook 取一個名字

# 建立一個 Python3 Notebook

The screenshot shows the Microsoft Azure Machine Learning Studio interface. On the left, there is a sidebar with various project management options: PROJECTS, EXPERIMENTS, WEB SERVICES, NOTEBOOKS (which is currently selected), DATASETS, TRAINED MODELS, and SETTINGS. At the bottom of the sidebar is a 'NEW' button. The main area displays a table titled 'notebooks' with one entry: 'Kaggle\_Titanic'. The table has columns for NAME, LANGUAGE, LAST MODIFIED, and PROJECT. The 'NAME' column shows 'Kaggle\_Titanic', 'LANGUAGE' shows 'Python 3', 'LAST MODIFIED' shows '2/12/2019 4:22:55 PM', and 'PROJECT' shows 'None'. Below the table are three buttons: DELETE, RENAME, and ADD TO PROJECT. The top right corner of the screen shows the workspace name 'Kristen-Free-Workspace' and some user icons.

NAME	LANGUAGE	LAST MODIFIED	PROJECT
Kaggle_Titanic	Python 3	2/12/2019 4:22:55 PM	None

**Actions:** DELETE, RENAME, ADD TO PROJECT

# 匯入 Package

In [1] :

```
# Warning 不顯示
import warnings
warnings.filterwarnings('ignore')
warnings.filterwarnings('ignore', category=DeprecationWarning)
```

In [2] :

```
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
import seaborn as sns

# 圖在 notebook 中顯示
%matplotlib inline
```

# 匯入 Data

In [3] :

```
from azureml import Workspace

ws = Workspace()
titanic_train = ws.datasets['Titanic_Train.csv']
titanic_test = ws.datasets['Titanic_Test.csv']

data_train = titanic_train.to_dataframe()
data_test = titanic_test.to_dataframe()
```

從 Workspace 中的資料集

# Exploratory Data Analysis

# 處理遺失值

In [4] : data\_train.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
PassengerId    891 non-null int64
Survived        891 non-null int64
Pclass          891 non-null int64
Name            891 non-null object
Sex             891 non-null object
Age             714 non-null float64
SibSp           891 non-null int64
Parch           891 non-null int64
Ticket          891 non-null object
Fare            891 non-null float64
Cabin           204 non-null object
Embarked         889 non-null object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.6+ KB
```

177 個遺失值

用年齡中位數來做插補

In [5] : # Age : null values with the median age  
data\_train['Age'] = data\_train['Age'].fillna(data\_train['Age'].median())

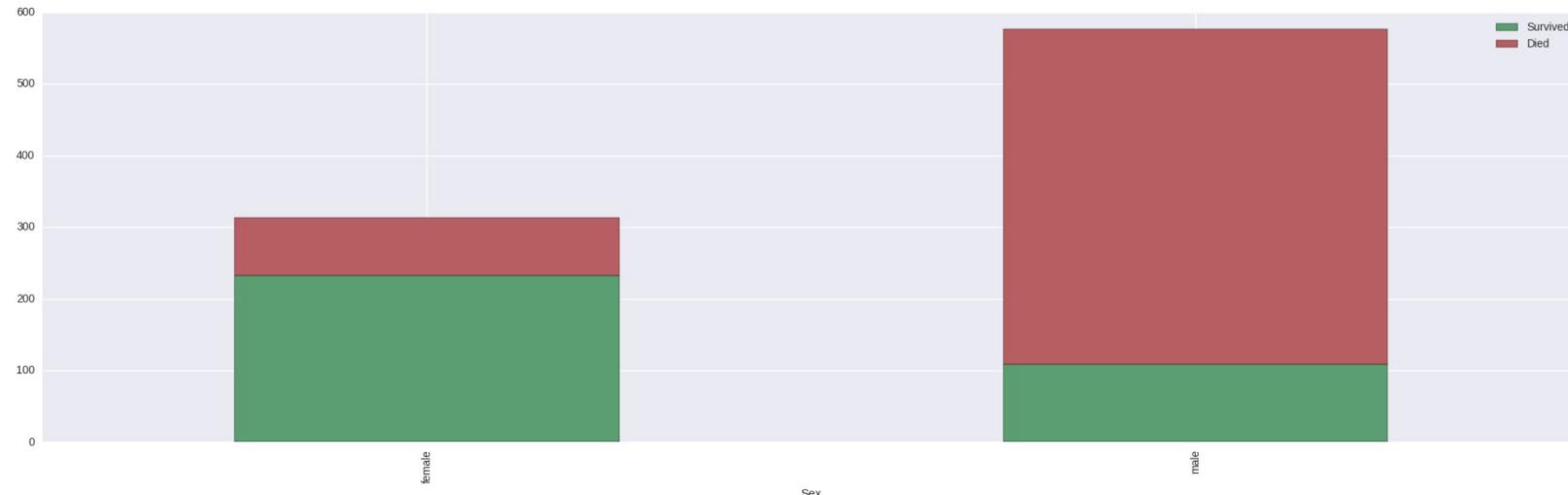
# Sex VS. Survived (count)

In [6] :

```
# 為了畫圖用建立的變數  
data_train['Died'] = 1 - data_train['Survived']
```

In [7] :

```
data_train.groupby('Sex').agg('sum')[['Survived', 'Died']].plot(kind='bar', figsize=(25, 7),  
stacked=True, colors=['g', 'r']);
```



# Sex VS. Survived (ratio)

In [8] :

```
data_train.groupby('Sex').agg('mean')[['Survived', 'Died']].plot(kind='bar', figsize=(25, 7),  
stacked=True, colors=['g', 'r']);
```



# Age VS. Sex VS. Survived

In [9] :

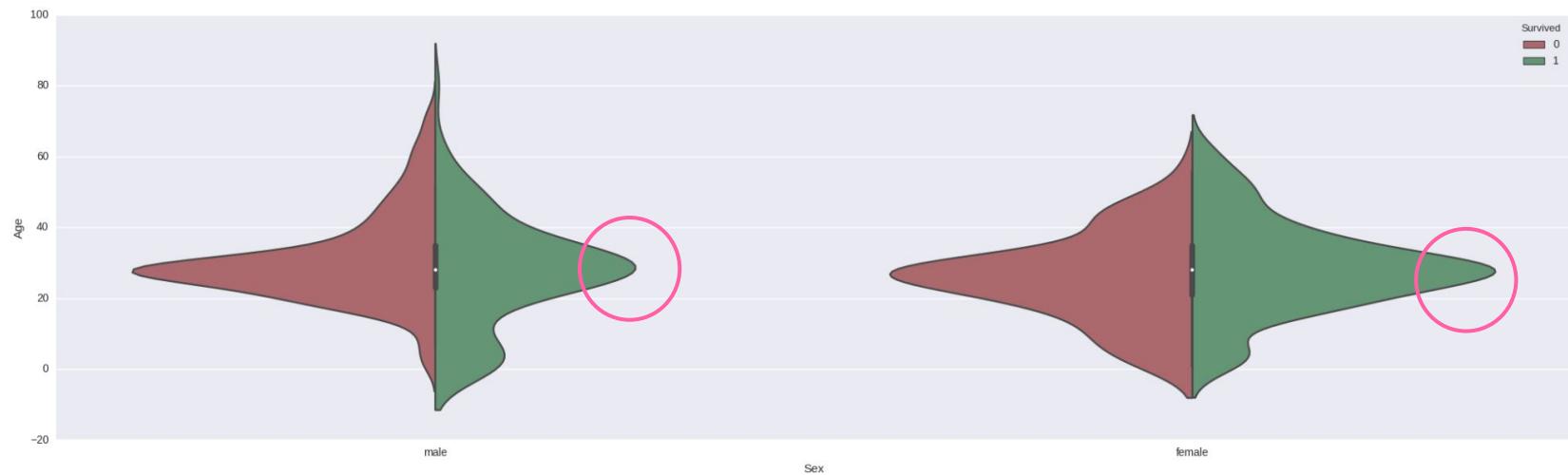
```
fig = plt.figure(figsize = (25, 7))
sns.violinplot(x = 'Sex', y = 'Age',
                hue = 'Survived', data = data_train,
                split = True,
                palette = {0: "r", 1: "g"})
);
```



# Age VS. Sex VS. Survived

In [9] :

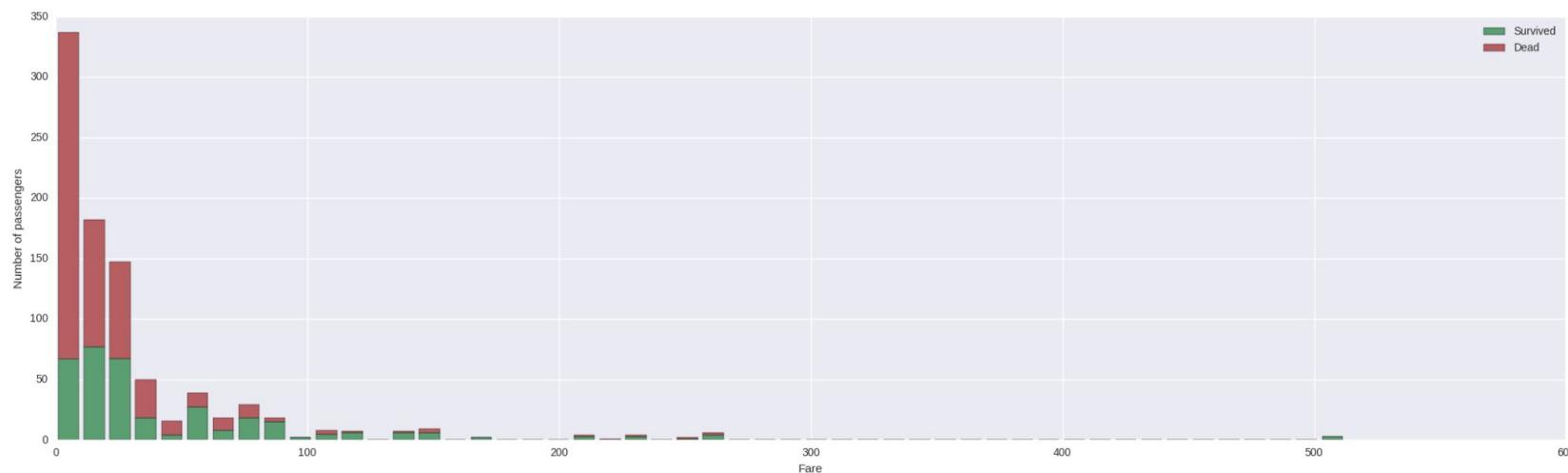
```
fig = plt.figure(figsize = (25, 7))
sns.violinplot(x = 'Sex', y = 'Age',
                hue = 'Survived', data = data_train,
                split = True,
                palette = {0: "r", 1: "g"}
);
```



# Fare VS. Survived

In [10]:

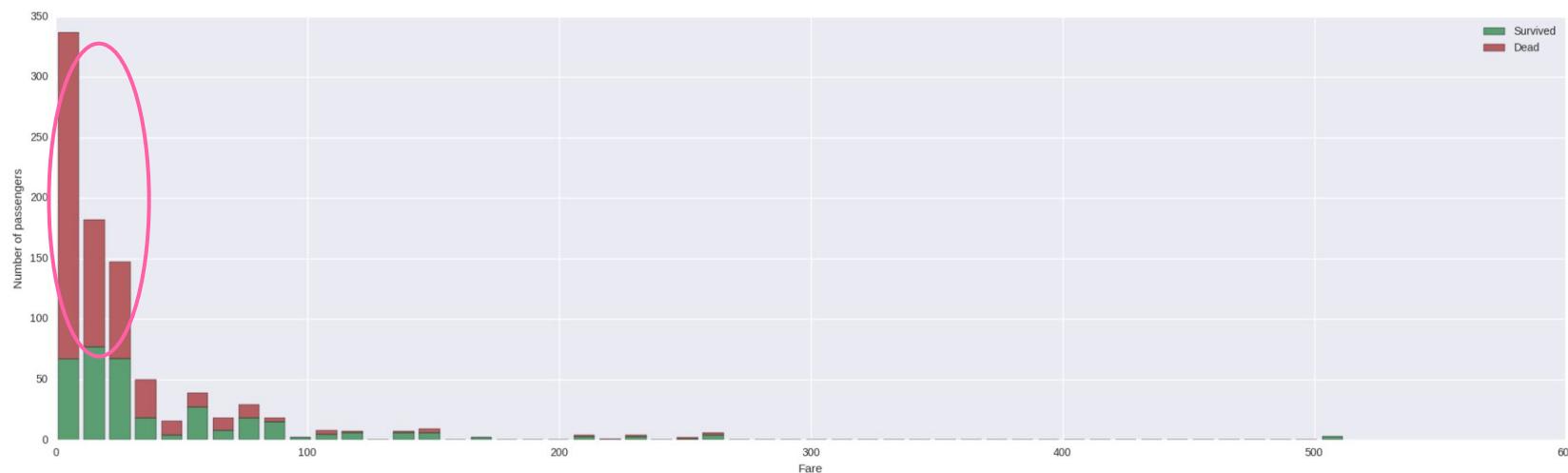
```
figure = plt.figure(figsize = (25, 7))
plt.hist([data_train[data_train['Survived'] == 1]['Fare'], data_train[data_train['Survived'] == 0]['Fare']],
         stacked = True, color = ['g','r'],
         bins = 50, label = ['Survived', 'Dead'])
plt.xlabel('Fare')
plt.ylabel('Number of passengers')
plt.legend();
```



# Fare VS. Survived

In [10]:

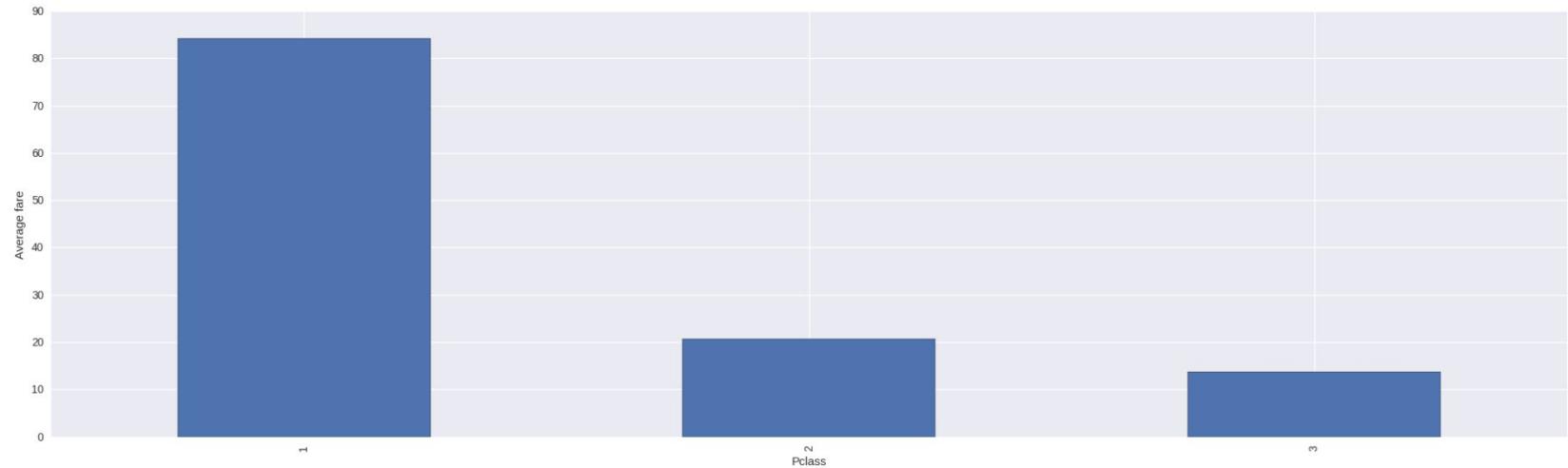
```
figure = plt.figure(figsize = (25, 7))
plt.hist([data_train[data_train['Survived'] == 1]['Fare'], data_train[data_train['Survived'] == 0]['Fare']],
         stacked = True, color = ['g','r'],
         bins = 50, label = ['Survived', 'Dead'])
plt.xlabel('Fare')
plt.ylabel('Number of passengers')
plt.legend();
```



# Fare VS. Survived

In [11]:

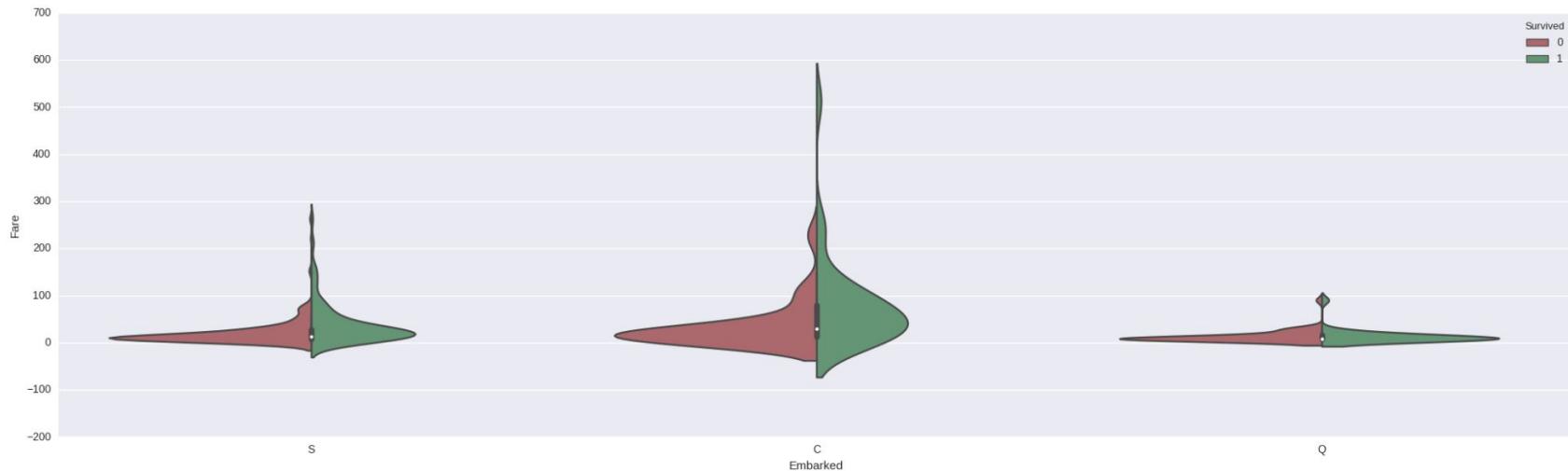
```
ax = plt.subplot()
ax.set_ylabel('Average fare')
data_train.groupby('Pclass').mean()['Fare'].plot(kind = 'bar', figsize = (25, 7), ax = ax);
```



# Fare VS. Survived

In [12] :

```
fig = plt.figure(figsize = (25, 7))
sns.violinplot(x = 'Embarked', y = 'Fare', hue = 'Survived', data = data_train, split = True,
                 palette = {0: "r", 1: "g"});
```



# Embarked VS. Fare VS. Survived

In [12]:

```
fig = plt.figure(figsize = (25, 7))
sns.violinplot(x = 'Embarked', y = 'Fare', hue = 'Survived', data = data_train, split = True,
                 palette = {0: "r", 1: "g"});
```



# Feature Engineering

# 合併訓練和測試樣本

```
In [13]: data_train = titanic_train.to_dataframe()  
data_test = titanic_test.to_dataframe()  
  
In [14]: targets = data_train.Survived  
  
In [15]: data_train.drop(['Survived'], 1, inplace = True)  
  
        data_combine = data_train.append(data_test)  
        data_combine.reset_index(inplace = True)  
        data_combine.drop(['index', 'PassengerId'], inplace = True, axis = 1)  
  
In [16]: data_combine.info()
```

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 1309 entries, 0 to 1308 891 + 418 = 1309  
Data columns (total 11 columns):  
Age          1223 non-null float64  
Cabin         295 non-null object  
Died          891 non-null float64  
Embarked      1307 non-null object  
Fare          1308 non-null float64  
Name          1309 non-null object  
Parch         1309 non-null int64  
Pclass        1309 non-null int64  
Sex           1309 non-null object  
SibSp         1309 non-null int64  
Ticket        1309 non-null object  
dtypes: float64(3), int64(3), object(5)  
memory usage: 112.6+ KB
```

# Passenger Name

Name
Braund, Mr. Owen Harris
Cumings, Mrs. John Bradley (Florence Briggs Th...)
Heikkinen, Miss. Laina
Futrelle, Mrs. Jacques Heath (Lily May Peel)
Allen, Mr. William Henry
Moran, Mr. James
McCarthy, Mr. Timothy J
Palsson, Master. Gosta Leonard
Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)
Nasser, Mrs. Nicholas (Adele Achem)

In [17] :

```
titles = set()
for name in data_combine['Name']:
    titles.add(name.split(',') [1].split('. ') [0].strip())

print(titles)

{'Capt',
 'Col',
 'Don',
 'Dona',
 'Dr',
 'Jonkheer',
 'Lady',
 'Major',
 'Master',
 'Miss',
 'Mlle',
 'Mme',
 'Mr',
 'Mrs',
 'Ms',
 'Rev',
 'Sir',
 'the Countess'}
```

# Passenger Name

```
In [18]: Title_Dictionary = {  
    "Capt" : "Officer",  
    "Col" : "Officer",  
    "Major" : "Officer",  
    "Jonkheer" : "Royalty",  
    "Don" : "Royalty",  
    "Dona" : "Royalty",  
    "Sir" : "Royalty",  
    "Dr" : "Officer",  
    "Rev" : "Officer",  
    "the Countess" : "Royalty",  
    "Mme" : "Mrs",  
    "Mlle" : "Miss",  
    "Ms" : "Mrs",  
    "Mr" : "Mr",  
    "Mrs" : "Mrs",  
    "Miss" : "Miss",  
    "Master" : "Master",  
    "Lady" : "Royalty"  
}
```



- ◆ Officer
- ◆ Royalty
- ◆ Mr
- ◆ Mrs
- ◆ Miss
- ◆ Master

```
In [19]: # Split Name  
data_combine['Title'] = data_combine['Name'].map(lambda name:name.split(',') [1].split('.')[0].strip())  
  
# Mapping new Title  
data_combine['Title'] = data_combine.Title.map(Title_Dictionary)
```

Kristen Chan

# Passenger Name

```
In [20]: # Encoding in dummy variable
dummy_title = pd.get_dummies(data_combine['Title'], prefix = 'Title')
data_combine = pd.concat([data_combine, dummy_title], axis = 1)
```

```
In [21]: data_combine.head()
```

Dummy Variable

```
Out[21]:
```

	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Title	Title_Master	Title_Miss	Title_Mr	Title_Mrs	Title_Officer	Title_Royalty
0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S	Mr	0	0	1	0	0	0
1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C	Mrs	0	0	0	1	0	0
2	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2, 3101282	7.9250	NaN	S	Miss	0	1	0	0	0	0
3	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	Mrs	0	0	0	1	0	0
4	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S	Mr	0	0	1	0	0	0

# Passenger Age

```
In [22]: print(data_combine['Age'].isnull().sum())
```

263

train + test 的 missing 共有 263

```
In [23]: grouped_train = data_combine.iloc[:891].groupby(['Sex', 'Pclass', 'Title'])
grouped_train_median = grouped_train.median()
grouped_train_median = grouped_train_median.reset_index()[['Sex', 'Pclass', 'Title', 'Age']]
print(grouped_train_median)
```

	Sex	Pclass	Title	Age
0	female	1	Miss	30.0
1	female	1	Mrs	40.0
2	female	1	Officer	49.0
3	female	1	Royalty	40.5
4	female	2	Miss	24.0
5	female	2	Mrs	31.5
6	female	3	Miss	18.0
7	female	3	Mrs	31.0
8	male	1	Master	4.0
9	male	1	Mr	40.0
10	male	1	Officer	51.0
11	male	1	Royalty	40.0
12	male	2	Master	1.0
13	male	2	Mr	31.0
14	male	2	Officer	46.5
15	male	3	Master	4.0
16	male	3	Mr	26.0

依照 Sex, Pclass, Title 分組計算年齡的中位數

# Passenger Age

```
In [22]: print(data_combine['Age'].isnull().sum())
```

```
263
```

```
In [23]: grouped_train = data_combine.iloc[:891].groupby(['Sex', 'Pclass', 'Title'])
grouped_train_median = grouped_train.median()
grouped_train_median = grouped_train_median.reset_index()[['Sex', 'Pclass', 'Title', 'Age']]
print(grouped_train_median)
```

	Sex	Pclass	Title	Age
0	female	1	Miss	30.0
1	female	1	Mrs	40.0
2	female	1	Officer	49.0
3	female	1	Royalty	40.5
4	female	2	Miss	24.0
5	female	2	Mrs	31.5
6	female	3	Miss	18.0
7	female	3	Mrs	31.0
8	male	1	Master	4.0
9	male	1	Mr	40.0
10	male	1	Officer	51.0
11	male	1	Royalty	40.0
12	male	2	Master	1.0
13	male	2	Mr	31.0
14	male	2	Officer	46.5
15	male	3	Master	4.0
16	male	3	Mr	26.0

使用 Training Data

依照 Sex, Pclass, Title 分組計算年齡的中位數

# Passenger Age

```
In [24]: def fill_age(data):
    fill_id = ( (grouped_train_median['Sex'] == data['Sex']) &
                (grouped_train_median['Title'] == data['Title']) &
                (grouped_train_median['Pclass'] == data['Pclass']) )
    return grouped_train_median[fill_id]['Age'].values[0]

In [25]: data_combine['Age'] = data_combine.apply(lambda data: fill_age(data)
                                                if np.isnan(data['Age']) else data['Age'], axis=1)
```

遇到遺失值, 去找對應的組別補上值

# Fare

```
In [26]: print(data_combine['Fare'].isnull().sum())
```

```
1
```

```
In [27]: data_combine.Fare.fillna(combined.iloc[:891].Fare.mean(), inplace = True)
```



補平均票價

# Embarked

```
In [28]: print(data_combine['Embarked'].isnull().sum())
```

```
2
```

```
In [29]: data_combine.loc[:891, 'Embarked'].value_counts()
```

```
Out [29]: S    644  
C    168  
Q     78  
Name: Embarked, dtype: int64
```

```
In [30]: data_combine.Embarked.fillna('S', inplace = True)
```

```
In [31]: # Encoding in dummy variable  
dummy_embarked = pd.get_dummies(data_combine['Embarked'], prefix = 'Embarked')  
data_combine = pd.concat([data_combine, dummy_embarked], axis = 1)
```



- 補出現最多的港口
- Add Dummy Variable

# Cabin

```
In [32]: print(data_combine['Cabin'].isnull().sum())
```

```
1014
```

```
In [33]: cabins = set()
for c in data_combine.iloc[:891]['Cabin']:
    cabins.add(c)

print(cabins)
```

```
{nan, 'A32', 'E36', 'A16', 'D47', 'B49', 'F2', 'B4', 'D56', 'E50', 'A14', 'D45', 'D9', 'F G63', 'C110', 'C99', 'B96 B
98', 'C62 C64', 'D26', 'C86', 'B41', 'C54', 'C22 C26', 'C124', 'C128', 'B28', 'C82', 'B82 B84', 'F33', 'C106', 'B39',
'B5', 'D15', 'E17', 'B86', 'B37', 'D11', 'D28', 'A23', 'G6', 'E34', 'C47', 'B42', 'B71', 'D6', 'B77', 'C103', 'B35',
'D33', 'C101', 'B30', 'D7', 'A19', 'C45', 'B3', 'C49', 'F4', 'B18', 'B73', 'D19', 'C50', 'C65', 'B58 B60', 'C93', 'F3
8', 'D37', 'E67', 'C148', 'E58', 'B80', 'B102', 'A7', 'E25', 'C85', 'C87', 'C123', 'E40', 'D30', 'A26', 'C52', 'D36',
'D10 D12', 'C90', 'E68', 'B22', 'B94', 'E63', 'E38', 'A36', 'B57 B59 B63 B66', 'E44', 'A20', 'A34', 'E12', 'C2', 'D4
8', 'E33', 'C30', 'B51 B53 B55', 'C118', 'D49', 'C78', 'D', 'C111', 'C83', 'E46', 'C95', 'F G73', 'A5', 'E8', 'C70',
'C23 C25 C27', 'C125', 'E24', 'B20', 'B69', 'A24', 'E77', 'A6', 'C126', 'B38', 'C68', 'E31', 'C91', 'A10', 'E49', 'D2
1', 'E10', 'D46', 'C32', 'D50', 'C7', 'E121', 'B78', 'C92', 'E101', 'A31', 'B101', 'B79', 'B50', 'F E69', 'D20', 'T',
'C104', 'D17', 'B19', 'D35', 'C46'}
```

所有艙號

# Cabin

```
In [34]: # Missing : M
data_combine.Cabin.fillna('M', inplace = True)

# Get Each Cabin First letter
data_combine['Cabin'] = data_combine['Cabin'].map(lambda l: l[0])
```

```
In [35]: # Encoding in dummy variable
dummy_cabin = pd.get_dummies(data_combine['Cabin'], prefix = 'Cabin')
data_combine = pd.concat([data_combine, dummy_cabin], axis = 1)
```



- Missing 的用 M 來代表 Missing
- 取 Cabin 的第一個英文字
- Add Dummy Variable

# Passenger Sex

```
In [36]: data_combine['Sex'] = data_combine['Sex'].map({'male':1, 'female':0})
```



Female : 0  
Male : 1

# Pclass

```
In [37]: # Encoding in dummy variable  
dummy_pclass = pd.get_dummies(data_combine['Pclass'], prefix = 'Pclass')  
data_combine = pd.concat([data_combine, dummy_pclass], axis = 1 )
```



Add Dummy Variable

# Ticket

Ticket	
0	A/5 21171
1	PC 17599
2	STON/O2. 3101282
3	113803
4	373450
5	330877
6	17463
7	349909
8	347742
9	237736



In [38] :

```
def cleanTicket(ticket):
    ticket = ticket.replace('.', ' ')
    ticket = ticket.replace('/', ' ')
    ticket = ticket.split()
    ticket = map(lambda t : t.strip(), ticket)
    ticket = list(filter(lambda t : not t.isdigit(), ticket))
    if len(ticket) > 0:
        return ticket[0]
    else:
        return 'Null'
```

In [39] :

```
tickets = set()
for t in data_combine['Ticket']:
    tickets.add(cleanTicket(t))

print(tickets)
```

```
{'Fa', 'SCA3', 'WEP', 'SCParis', 'AS', 'CASOTON', 'SCOW', 'SOTONOQ', 'SOC', 'STONO', 'Null', 'CA', 'C', 'SCPARIS', 'SP',
'STONO2', 'PPP', 'STONOQ', 'LINE', 'FCC', 'SWPP', 'FC', 'SOTONO2', 'LP', 'PC', 'A', 'AQ3', 'PP', 'SCA4', 'AQ4',
'SCAH', 'SOPP', 'A5', 'A4', 'SOP', 'WC', 'SC'}
```

取出票號前的英文, 若沒有則顯示 Null

# Ticket

```
In [40]: data_combine['Ticket'] = data_combine['Ticket'].map(cleanTicket)
```

```
In [41]: # Encoding in dummy variable
dummy_tickets = pd.get_dummies(data_combine['Ticket'], prefix = 'Ticket')
data_combine = pd.concat([data_combine, dummy_tickets], axis = 1 )
```



Add Dummy Variable

# Family

In [42] :

```
data_combine['Family_size'] = data_combine['Parch'] + data_combine['SibSp'] + 1

# Introducing other features based on the family size
data_combine['Single_family'] = data_combine['Family_size'].map(lambda s: 1 if s == 1 else 0)
data_combine['Small_family'] = data_combine['Family_size'].map(lambda s: 1 if 2 <= s <= 4 else 0)
data_combine['Big_family'] = data_combine['Family_size'].map(lambda s: 1 if 5 <= s else 0)
```



## 新增變數

- Family Size : 兄弟姊妹/配偶 + 父母/小孩
- Single Family : Family Size = 1 ← 獨自一人
- Small Family : Family Size = 2~4 ← 小家庭
- Big Family : Family Size >= 5 ← 大家庭

# Final Data

```
In [43]: data_combine.drop(['Name', 'Title', 'Embarked', 'Cabin', 'Pclass', 'Ticket'], axis = 1, inplace = True)
```

```
In [44]: data_combine.shape
```

```
Out[44]: (1309, 67)
```

```
In [45]: data_combine.head()
```

```
Out[45]:
```

	Sex	Age	SibSp	Parch	Fare	Title_Master	Title_Miss	Title_Mr	Title_Mrs	Title_Officer	...	Ticket_STONO	Ticket_STONO2	Ticket_STONOQ	Ticket_SW
0	1	22.0	1	0	7.2500	0	0	1	0	0	...	0	0	0	0
1	0	38.0	1	0	71.2833	0	0	0	1	0	...	0	0	0	0
2	0	26.0	0	0	7.9250	0	1	0	0	0	...	0	1	0	0
3	0	35.0	1	0	53.1000	0	0	0	1	0	...	0	0	0	0
4	1	35.0	0	0	8.0500	0	0	1	0	0	...	0	0	0	0

5 rows × 67 columns

# Azure ML Studio

# Execute Python Script

# 介紹 Execute Python Script

Microsoft Azure Machine Learning Studio    Kristen-Free-Workspace ▾ ? ☰ ☺ ☰ ☰

Kaggle Titanic Advance    In draft    Properties    Project

Search experiment items

→ 從 Module 選擇, Python Language Modules 中 [Excute Python Script]

Saved Datasets  
Data Format Conversi  
Data Input and Output  
Data Transformation  
Feature Selection  
Machine Learning  
OpenCV Library Modules  
**Python Language Modules**  
Execute Python Script  
R Language Modules  
Statistical Functions  
Text Analytics  
Time Series  
Web Service  
Deprecated

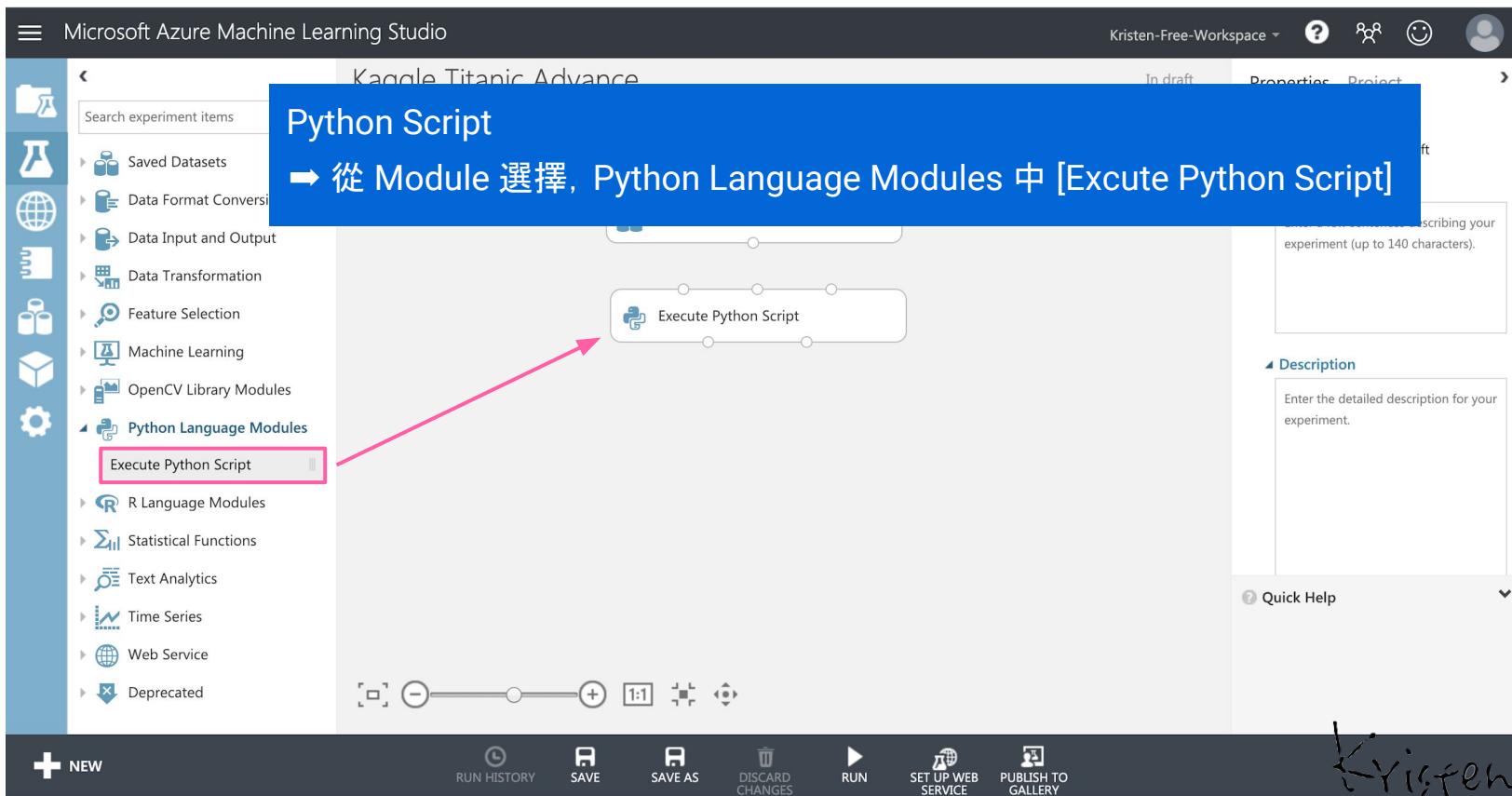
Execute Python Script

scribing your experiment (up to 140 characters).

Description  
Enter the detailed description for your experiment.

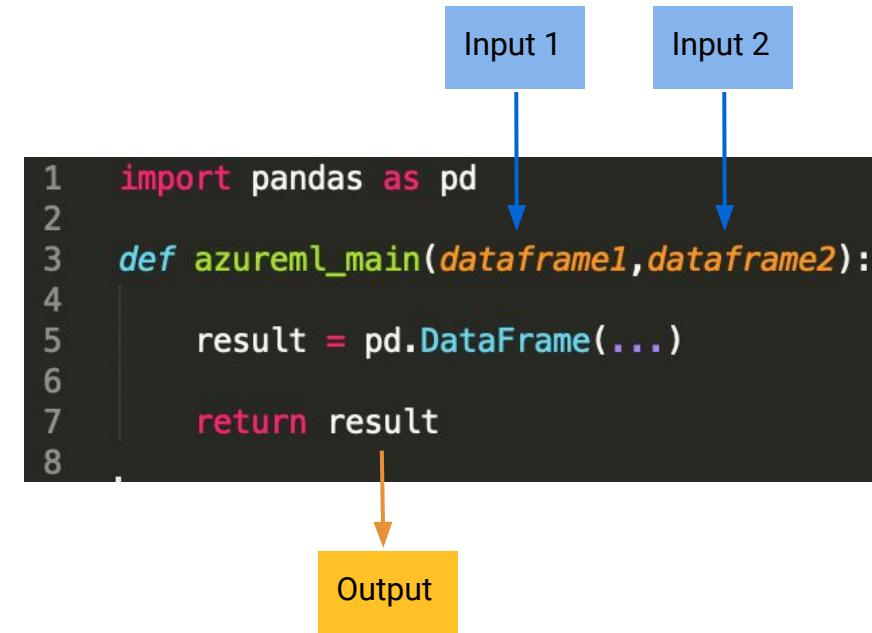
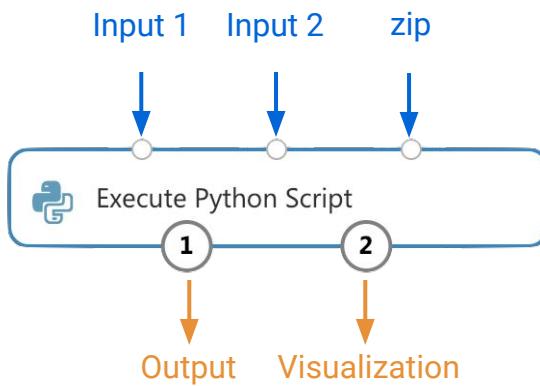
Quick Help

NEW    RUN HISTORY    SAVE    SAVE AS    DISCARD CHANGES    RUN    SET UP WEB SERVICE    PUBLISH TO GALLERY



Kristen Chian 90

# 介紹 Excute Python Script



# Python Script -- Titanic Feature Engineering

Search or jump to... Pull requests Issues Marketplace Explore

[https://github.com/kristenchan/Sharing/blob/master/Kaggle\\_Titanic/Titanic\\_FeatureEngineering.py](https://github.com/kristenchan/Sharing/blob/master/Kaggle_Titanic/Titanic_FeatureEngineering.py) Fork 0

Issues 0 Pull requests 0 Projects 0 Wiki Insights Settings

Branch: master Sharing / Kaggle\_Titanic / Titanic\_FeatureEngineering.py Find file Copy path

kristenchan Kaggle Titanic Feature Engineering 53c658b a minute ago

1 contributor

108 lines (90 sloc) | 3.91 KB Raw Blame History

```
1 #---- Import Package ----
2 import numpy as np
3 import pandas as pd
4
5 #---- Call Azure ML ----
6 def azureml_main(dataframe1):
7
8     data_final = dataframe1
9
10    #-- Passenger Name --
11    Title_Dictionary = {
12        "Capt": "Officer",
13        "Col": "Officer",
14        "Major": "Officer",
15        "Jonkheer": "Royalty",
16        "Don": "Royalty",
17        "Dona": "Royalty",
18        "Sir": "Royalty",
19        "Dr": "Officer",
20        "Rev": "Officer",
21        "the Countess": "Royalty",
22        "Mme": "Mrs",
```



KristenChan 92

# Excute Python Script

Microsoft Azure Machine Learning Studio

Kaggle Titanic Advance

In draft

Draft saved at 上午2:47:38

Properties Project

Experiment Properties

- START TIME 2/14/2019 ...
- END TIME 2/14/2019 ...
- STATUS CODE InDraft
- STATUS DETAILS None

Summary

Enter a brief title describing your experiment (up to 140 characters).

Description

Enter the detailed description for your experiment.

Quick Help

按著連接點拖到(Input1)連接處

```
graph TD; Titanic_Train[Titanic_Train.csv] --> Execute_Python[Execute Python Script]
```

NEW

RUN HISTORY SAVE SAVE AS DISCARD CHANGES RUN SET UP WEB SERVICE PUBLISH TO GALLERY

Kristen Chian 93

# Excute Python Script

Kaggle Titanic Advance

In draft

Saving...

Properties Project

Execute Python Script

Python script

```
1 #---- Import Package ----
2 import numpy as np
3 import pandas as pd
4 |
5 #---- Call Azure ML ----
6 def azureml_main(dataframe1):
7
8     data_final = dataframe1
9
10    #-- Passenger Name --
11    Title_Dictionary = {
12        "Capt": "Officer",
13        "Col": "Officer",
14        "Major": "Officer",
15        "Jonkheer": "Royalty",
16        "Don": "Royalty",
17        "Dona": "Royalty",
18        "Sir": "Royalty",
19        "Dr": "Officer",
20        "Rev": "Officer",
21        "the Countess": "Royalty",
```

Quick Help

Executes a Python script from an Azure Machine Learning experiment.

1. 選 Execute Python Script 把剛剛的 Code 貼過
2. [Run]
3. 在 [Execute Python Script] ① Results dataset 點右鍵 → 選擇 [Visualize]

Kristen Chian 94

# Excute Python Script

Microsoft Azure Machine Learning Studio

Kaggle Titanic Advance

Finished running ✓ Properties Project

rows 891 columns 62

view as

Survived	Sex	Age	SibSp	Parch	Fare	Title_Master	Title_Miss	Title_I
0	1	22	1	0	7.25	0	0	1
1	0	38	1	0	71.2833	0	0	0
1	0	26	0	0	7.925	0	1	0
1	0	35	1	0	53.1	0	0	0
0	1	35	0	0	8.05	0	0	1
0	1	26	0	0	8.4583	0	0	1
0	1	54	0	0	51.8625	0	0	1
0	1	2	3	1	21.075	1	0	0
1	0	27	0	2	11.1333	0	0	0
1	0	14	1	0	30.0708	0	0	0
1	0	4	1	1	16.7	0	1	0
1	0	58	0	0	26.55	0	1	0
0	1	20	0	0	2.075	0	0	1

To view, select a column in the table.

Statistics

Visualizations

NEW RUN HISTORY SAVE SAVE AS DISCARD CHANGES RUN SET UP WEB SERVICE PUBLISH TO GALLERY

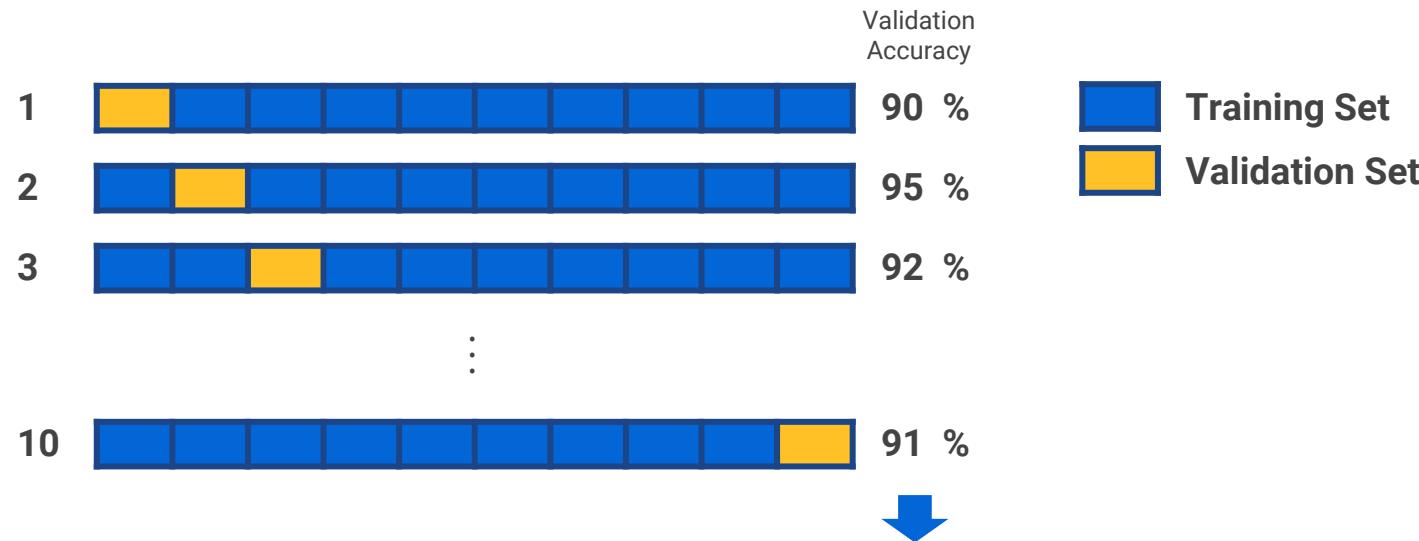
Kristen Chian 95

# Modeling

# 交叉驗證 Cross-Validation

- 避免因為選用某一特定的訓練和測試資料 產生偏差
- 交叉驗證做法：
  - Non-Exhaustive Methods
    - Holdout Method
    - K-Folder cross-validation
    - Stratified K-Fold Cross Validation
  - Exhaustive Methods
    - Leave-p-out cross-validation
    - Leave-one-out cross-validation

# 交叉驗證 Cross-Validation -- K-Folder cross-validation



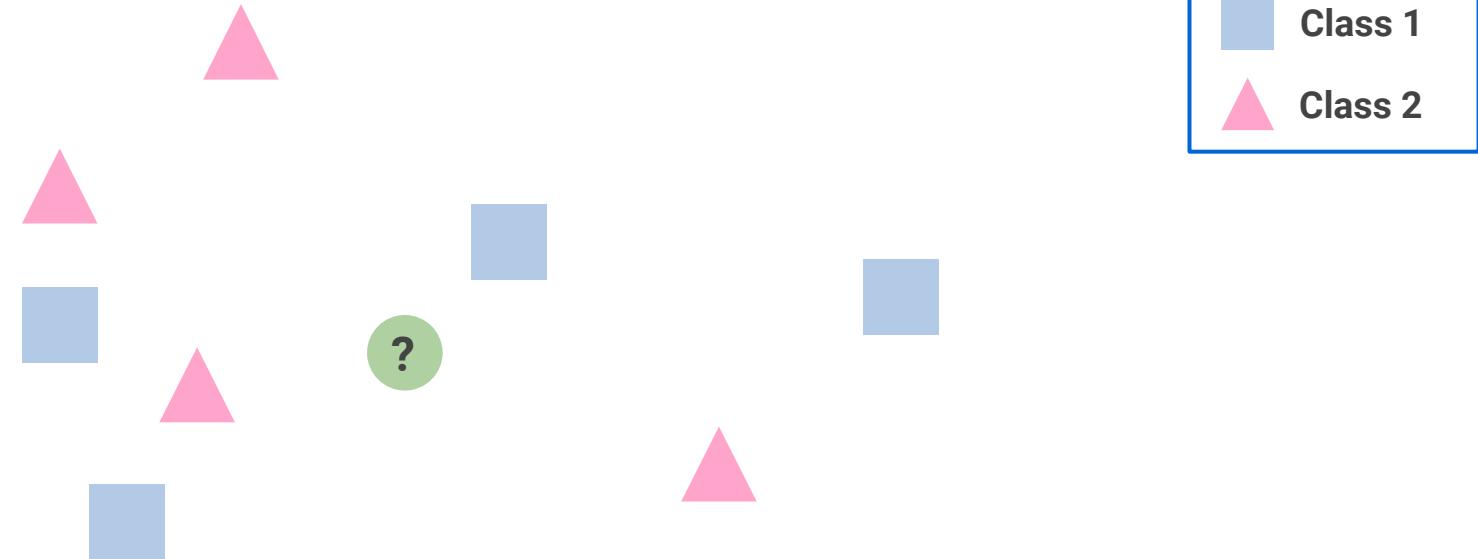
Final Accuracy = Average(1,2,3,...10)

# Model

- Logistic Regression
- KNN
- Decision Tree
- Random Forests
- Gradient Boosted Trees
- SVM

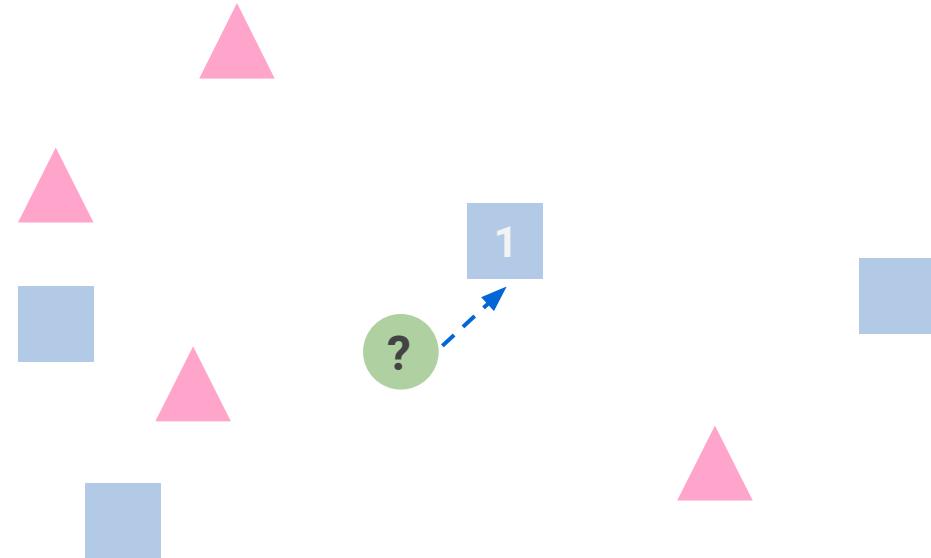
# Model -- KNN

- K Nearest Neighbor



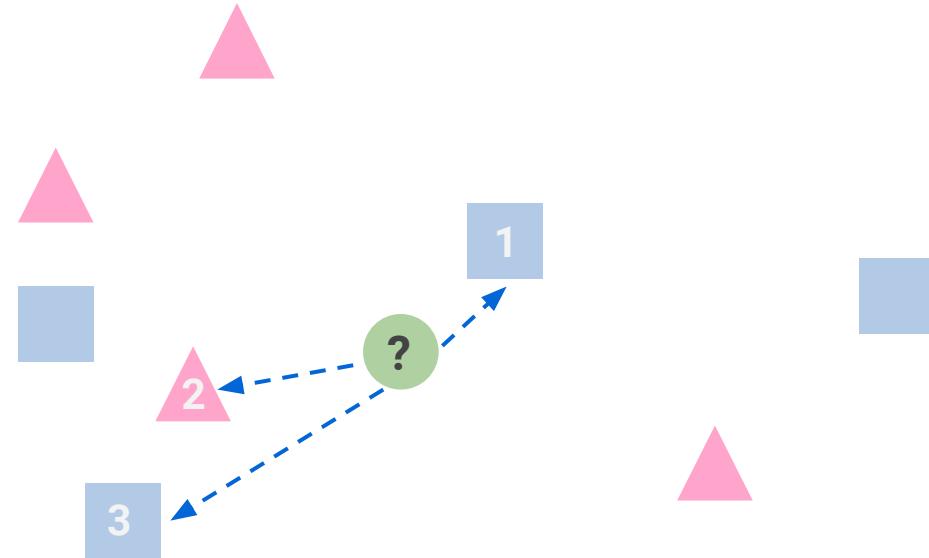
# Model -- KNN

- K Nearest Neighbor

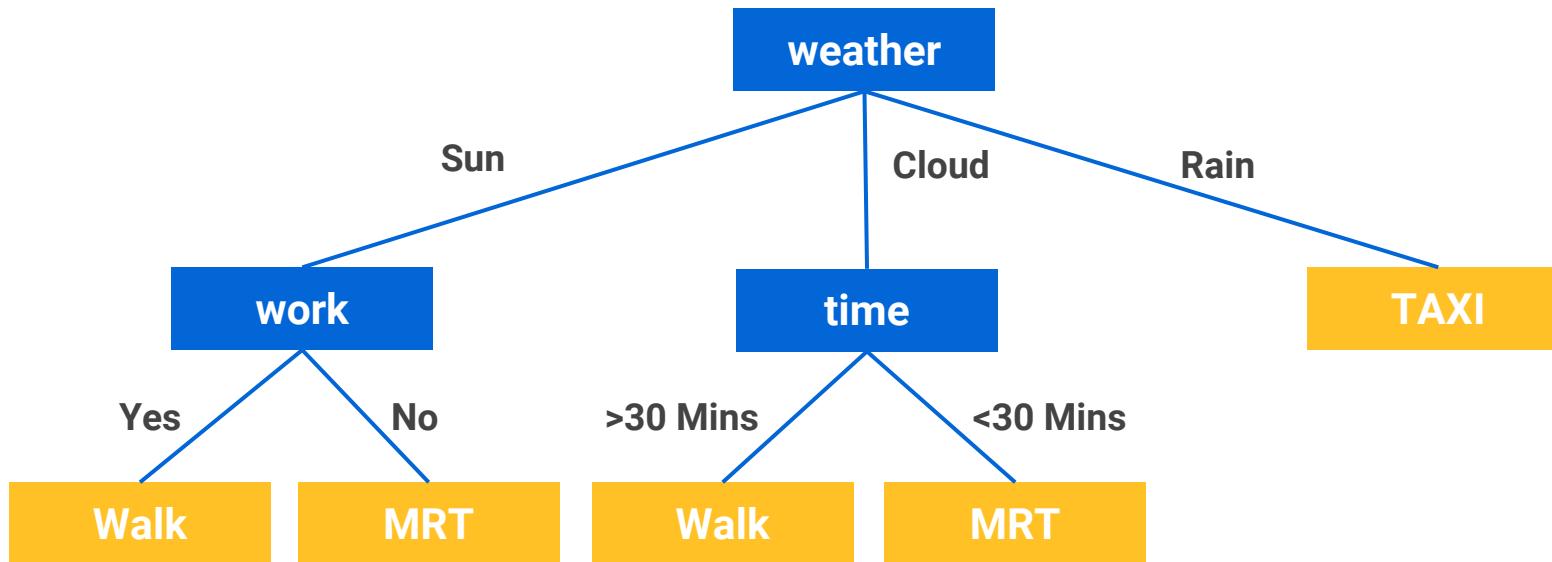


# Model -- KNN

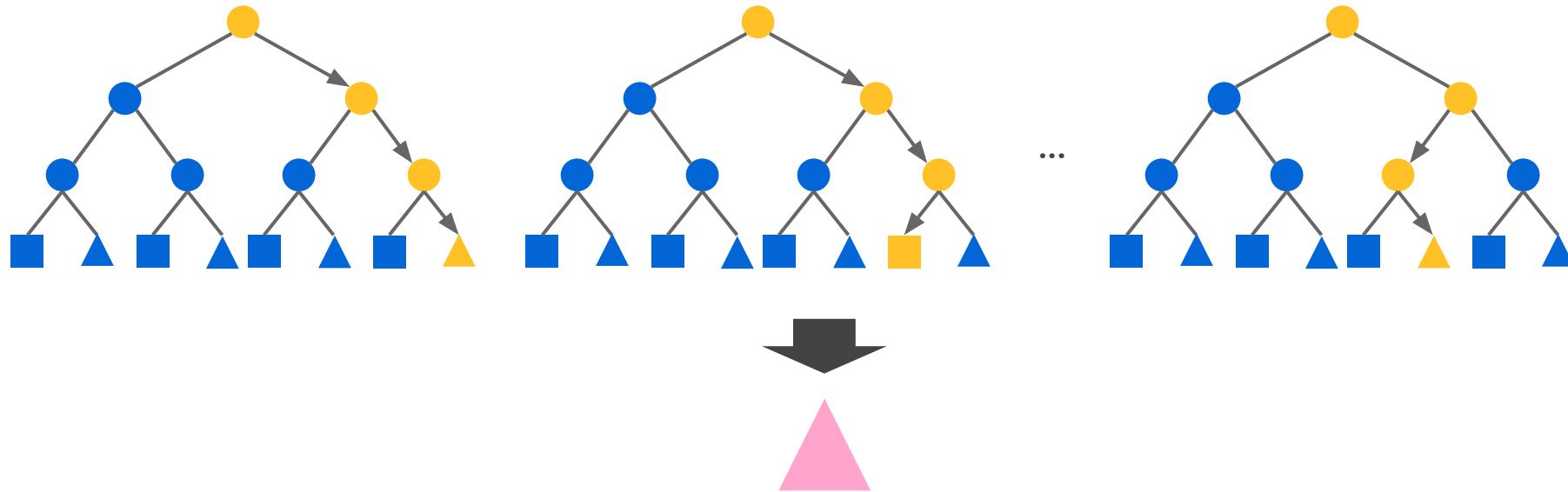
- K Nearest Neighbor



# Model -- Decision Tree

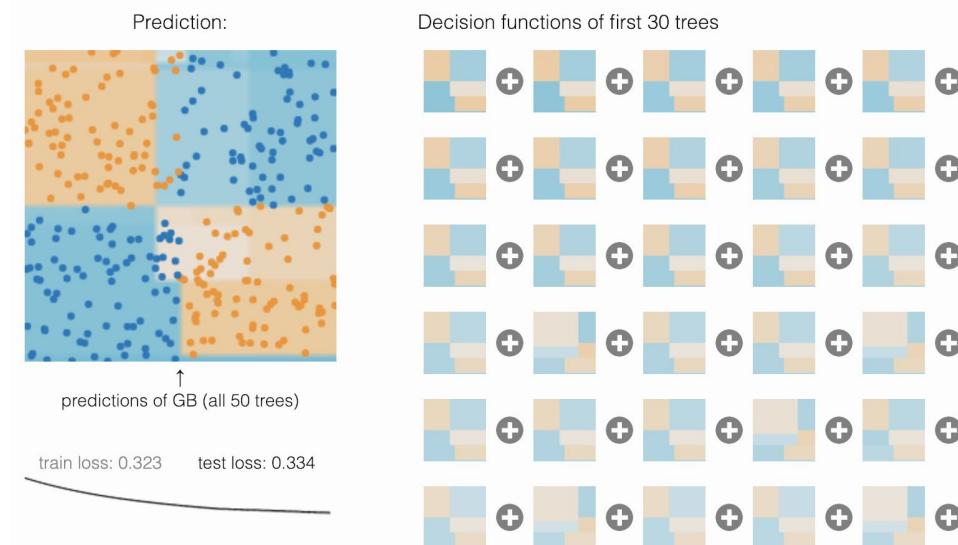


# Model -- Random Forests



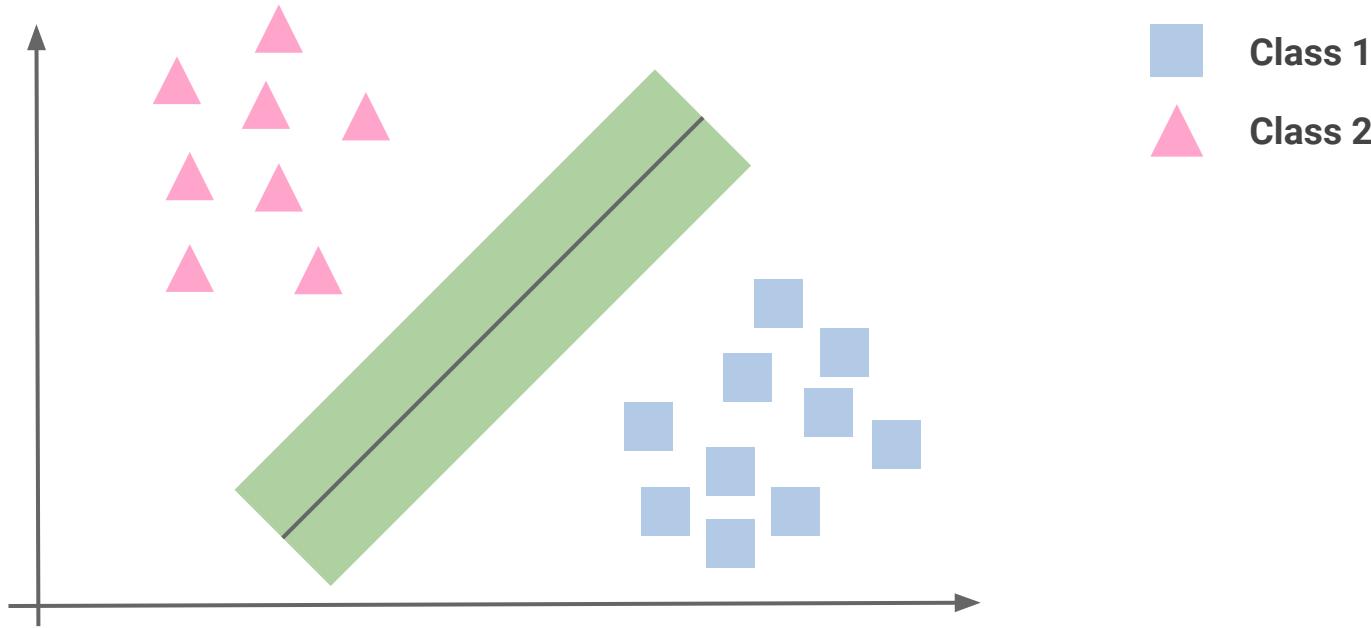
# Model -- Gradient Boosted Trees

- 什麼是 Boosting
  - 三個臭皮匠勝過一個諸葛亮
  - Weak Model
    - 複雜度低
    - 訓練的成本低
    - 不容易 Overfitting
- Gradient Boosting
  - 可以用在許多不同的 Loss Function



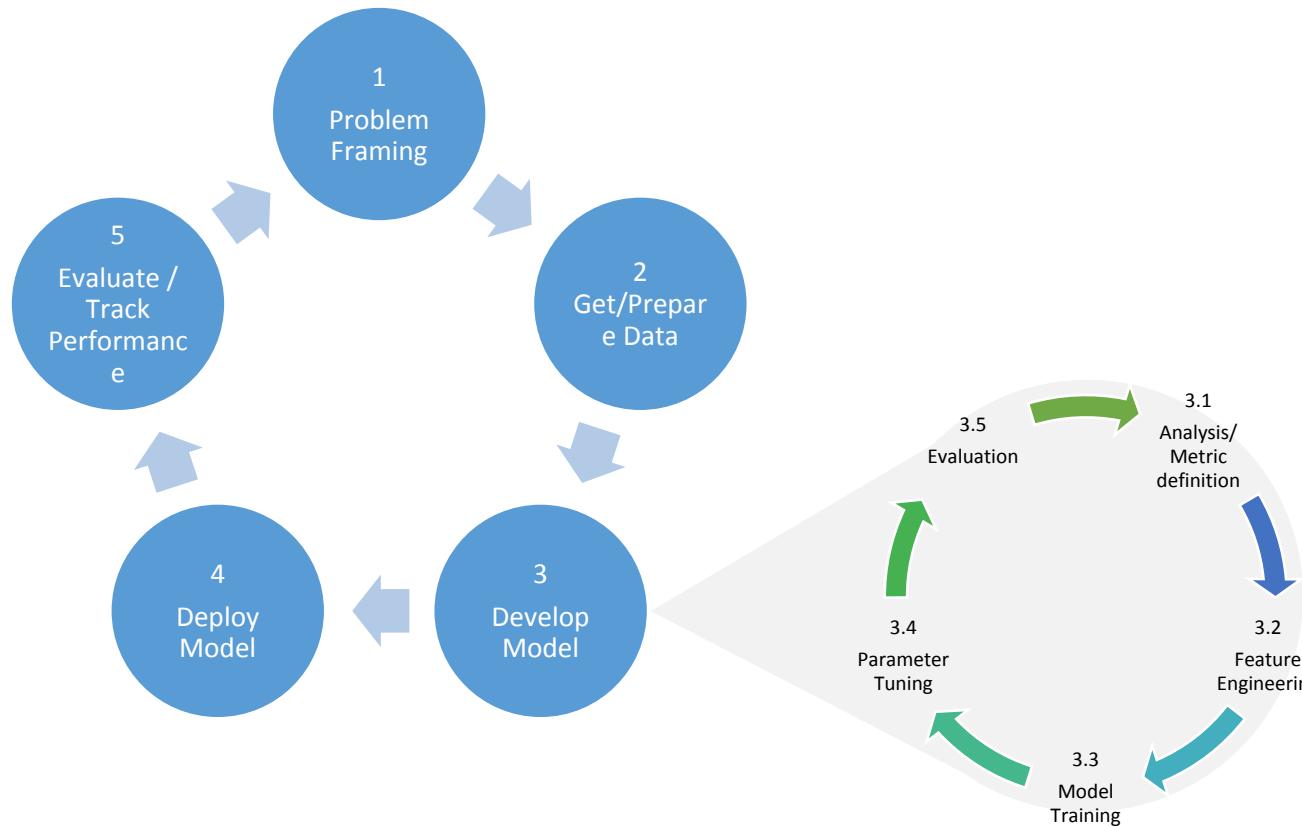
# Model -- SVM

- Support Vector Machine



# Azure ML Studio Advance

# Steps to Build a Machine Learning Solution



# Steps to Build a Machine Learning Solution

- Create a model
  - Step 1: Get data
  - Step 2: Prepare the data
  - Step 3: Define features
- Train the model
  - Step 4: Choose and apply a learning algorithm
- Score and test the model
  - Step 5: Predict

# Steps to Build a Machine Learning Solution

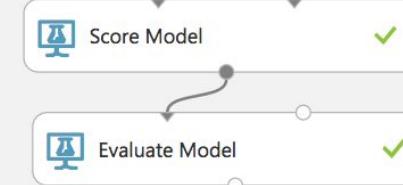
Create a model



Train the model

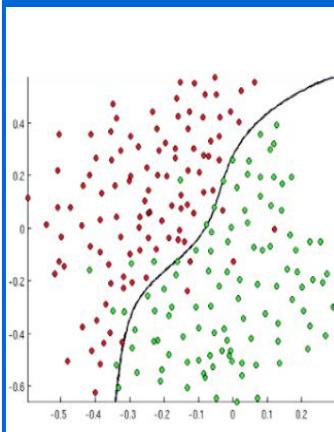


Score and test  
the model

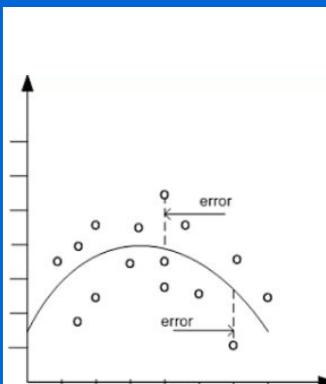


# Azure ML 主要演算法

Classification  
(分類)



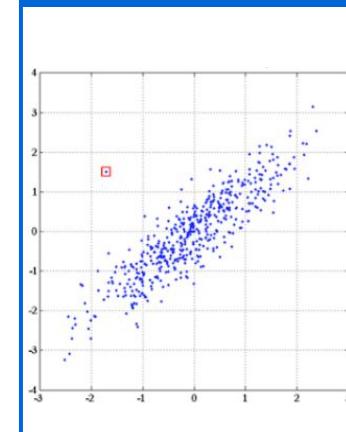
Regression  
(迴歸)



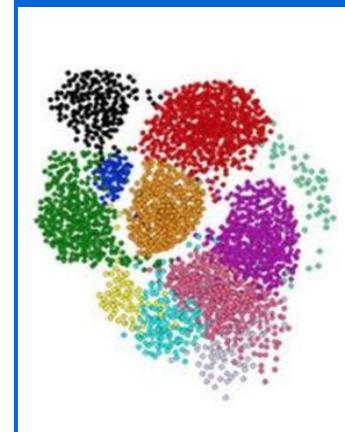
Recommenders  
(推薦)



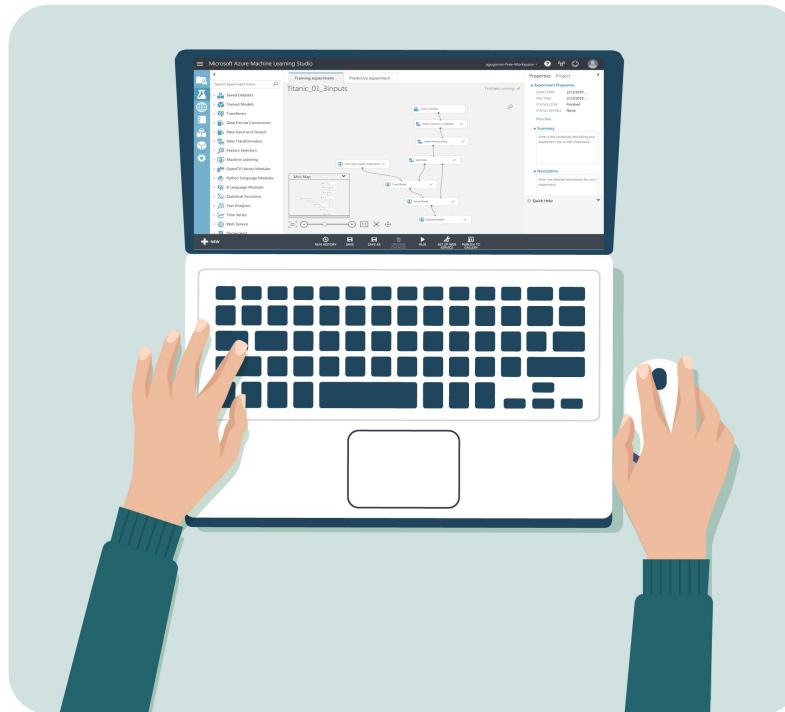
Anomaly Detection  
(異常偵測)



Clustering  
(集群)

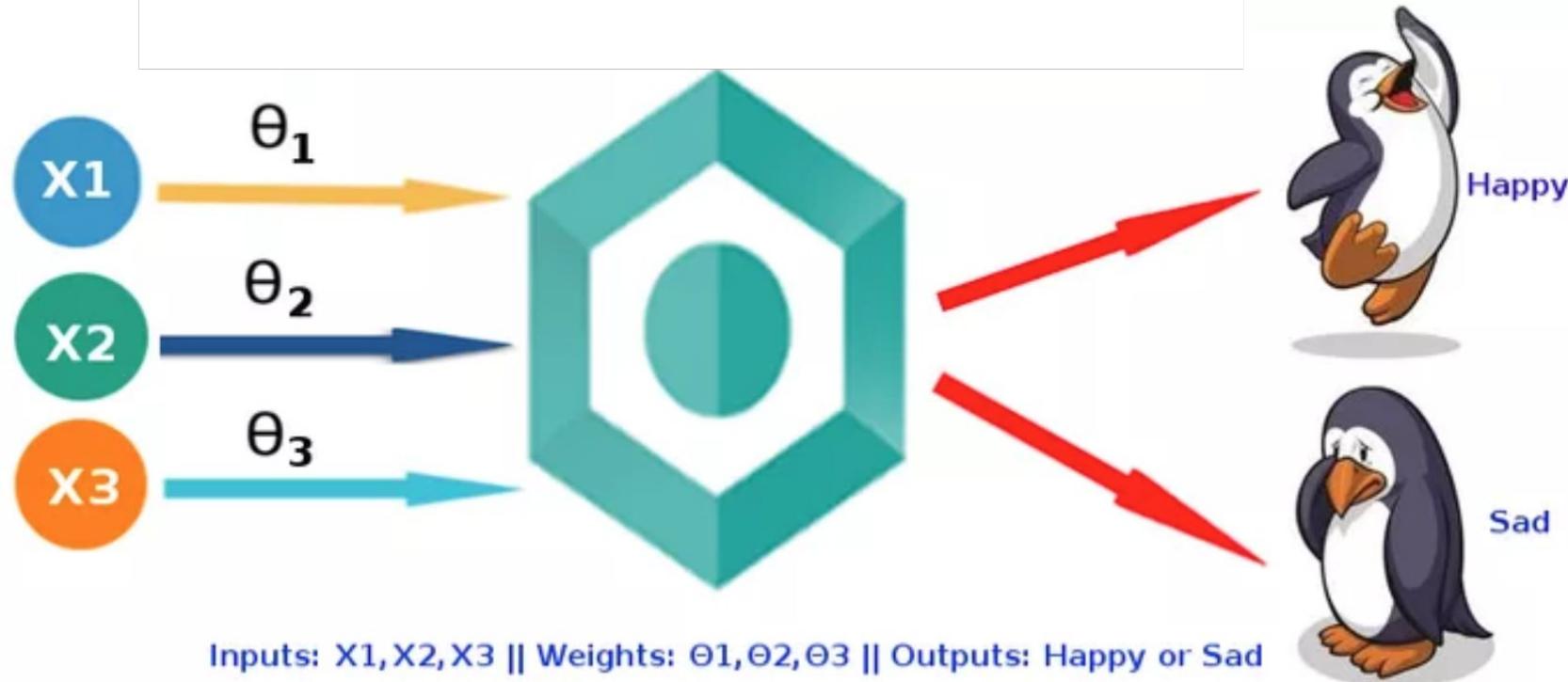


# Steps to Build a Machine Learning Solution



## Demo Time !!

# Logistic Regression Model



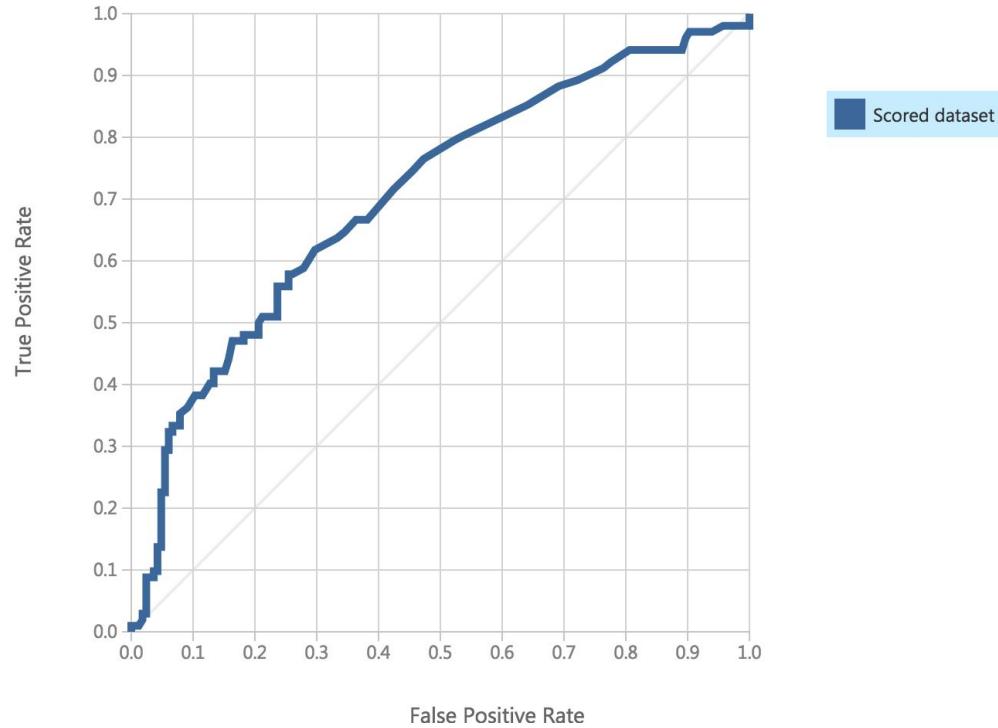
# Score Model

Titanic Survival Predication - OneModelLo... ➤ Score Model ➤ Scored dataset

rows	columns	ex	Age	SibSp	Parch	Fare	Embarked	Scored Labels	Scored Probabilities
267	10								
		male	33	0	0	7.775	S	0	0.346324
		female	38	0	0	227.525	C	1	0.665548
		male	52	0	0	30.5	S	0	0.363519
		male	32	0	0	7.925	S	0	0.347288
		male	55.5	0	0	8.05	S	0	0.329883
		male	28	0	0	7.225	C	0	0.349352
		female	2	0	1	10.4625	S	0	0.400525
		female	28	1	0	24.15	Q	0	0.353884
		male	14	5	2	46.9	S	0	0.371498
		male	32	0	0	56.4958	S	0	0.417075

# Evaluate Model

- ROC (Receiver Operating Characteristic Curve)



# Evaluate Model

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
<b>10</b>	<b>92</b>	<b>0.633</b>	<b>0.625</b>	<b>0.5</b>	<b>0.707</b>
False Positive	True Negative	Recall	F1 Score		
<b>6</b>	<b>159</b>	<b>0.098</b>	<b>0.169</b>		

# Evaluate Model

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	1	0	0.004	0.622	0.019	1.000	0.010	0.620	1.000	0.000
(0.800,0.900]	0	0	0.004	0.622	0.019	1.000	0.010	0.620	1.000	0.000
(0.700,0.800]	0	2	0.011	0.614	0.019	0.333	0.010	0.617	0.988	0.000
(0.600,0.700]	2	2	0.026	0.614	0.055	0.429	0.029	0.619	0.976	0.000
(0.500,0.600]	7	2	0.060	0.633	0.169	0.625	0.098	0.633	0.964	0.001
(0.400,0.500]	29	11	0.210	0.700	0.494	0.696	0.382	0.701	0.897	0.021
(0.300,0.400]	63	148	1.000	0.382	0.553	0.382	1.000	1.000	0.000	0.707
(0.200,0.300]	0	0	1.000	0.382	0.553	0.382	1.000	1.000	0.000	0.707
(0.100,0.200]	0	0	1.000	0.382	0.553	0.382	1.000	1.000	0.000	0.707
(0.000,0.100]	0	0	1.000	0.382	0.553	0.382	1.000	1.000	0.000	0.707

# Evaluate Model

Threshold 區間	TP區間內 增加	FP區間內 增加	$\frac{(TP+FP)}{All}$						$\frac{TN}{(FN+TN)}$	$\frac{TN}{(FP+TN)}$
Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	1	0	0.004	0.622	0.019	1.000	0.010	0.620	1.000	0.000
(0.800,0.900]	0	0	0.004	0.622	0.019	1.000	0.010	0.620	1.000	0.000
(0.700,0.800]	0	2	0.011	0.614	0.019	0.333	0.010	0.617	0.988	0.000
(0.600,0.700]	2	2	0.026	0.614	0.055	0.429	0.029	0.619	0.976	0.000
(0.500,0.600]	7	2	0.060	0.633	0.169	0.625	0.098	0.633	0.964	0.001
(0.400,0.500]	29	11	0.210	0.700	0.494	0.696	0.382	0.701	0.897	0.021
(0.300,0.400]	63	148	1.000	0.382	0.553	0.382	1.000	1.000	0.000	0.707
(0.200,0.300]	0	0	1.000	0.382	0.553	0.382	1.000	1.000	0.000	0.707
(0.100,0.200]	0	0	1.000	0.382	0.553	0.382	1.000	1.000	0.000	0.707
(0.000,0.100]	0	0	1.000	0.382	0.553	0.382	1.000	1.000	0.000	0.707

# Web Service

Microsoft Azure Machine Learning Studio

Training experiment Predictive experiment

Titanic\_01\_3inputs

Finished running ✓

Properties Project

Experiment Properties

- START TIME 2/13/2019 ...
- END TIME 2/13/2019 ...
- STATUS CODE Finished
- STATUS DETAILS None

Prior Run

Summary

Description

Quick Help

[SET UP WEB SERVICE]

Mini Map

Titanic Dataset → Select Columns in Dataset → Clean Missing Data → Split Data → Two-Class Logistic Regression → Train Model → Score Model → Predictive Web Service [Recommended] Retraining Web Service

RUN HISTORY SAVE AS DISCARD CHANGES RUN SET UP WEB SERVICE PUBLISH TO GALLERY

# Web Service

## 挑戰一 [predictive exp.]

DASHBOARD CONFIGURATION

General [New Web Services Experience preview](#)

Published experiment

[View snapshot](#) [View latest](#)

Description

No description provided for this web service.

API key

VufamCtbA4MsEdlQHDCg74TivzzU/c0e5aW9Fz7zFZNF7EEvO2nGpaXf8qn2npHPSdSPE6HM22Hk0SEDD2KTww==

Default Endpoint

API HELP PAGE	TEST	APPS	LAST UPDATED	
REQUEST/RESPONSE	<a href="#">Test</a> <a href="#">Test preview</a>	Excel 2013 or later    Excel 2010 or earlier workbook	4/20/2017 5:00:29 PM	
BATCH EXECUTION	<a href="#">Test</a> <a href="#">Test preview</a>	Excel 2013 or later workbook	4/20/2017 5:00:29 PM	

# Advanced Version -- Python

Microsoft Azure Machine Learning Studio

Kaggle Titanic Advance

In draft

Draft saved at 上午2:47:38

Properties Project

Experiment Properties

- START TIME 2/14/2019 ...
- END TIME 2/14/2019 ...
- STATUS CODE InDraft
- STATUS DETAILS None

Summary

Description

Quick Help

Execute Python Script

Titanic\_Train.csv

Execute Python Script

Diagram:

```
graph TD; A[Titanic_Train.csv] --> B[Execute Python Script]; B --> C[Execute Python Script]
```

Diagram description: The diagram shows a flow from a dataset node 'Titanic\_Train.csv' to two parallel execution nodes 'Execute Python Script'. A blue arrow points from the bottom of the second 'Execute Python Script' node towards the bottom of the slide.

[https://github.com/kristenchan/Sharing/blob/master/Kaggle\\_Titanic/Titanic\\_FeatureEngineering.py](https://github.com/kristenchan/Sharing/blob/master/Kaggle_Titanic/Titanic_FeatureEngineering.py)

Kristen Chan 121

# Advanced Version -- Python

篩選欄位

→ 從 Module 選擇, Data Transformation 中 Manipulation 的 [Select Columns in Dataset]

Select columns

AVAILABLE COLUMNS

BY NAME WITH RULES

PassengerId  
Pclass  
Name  
Ticket  
Cabin  
Embarked  
Title

Selected Columns

Survived  
Sex  
Age  
Title\_Master  
Title\_Miss  
Title\_Mr  
Title\_Mrs  
Title\_Officer  
Title\_Royal  
Embarked\_C  
Embarked\_Q  
Embarked\_S  
Cabin\_A  
Cabin\_B  
Cabin\_C

Leave

7 columns

1. PassengerId 5. Embarked  
2. Pclass 6. Cabin  
3. Name 7. Ticket  
4. Title

剩下的移到右邊

Quick Help

Selects columns to include or exclude from a dataset in an operation. Formerly known as project Columns.

KristenChen 122

# Advanced Version -- Python

Microsoft Azure Machine Learning Studio   Kristen-Free-Workspace

Kaggle Titanic Advance   Finished running ✓

Properties Project

Experiment Properties

- START TIME 2/25/2019 ...
- END TIME 2/25/2019 ...
- STATUS CODE Finished
- STATUS DATA... None

Prior Run

Summary

Enter a few sentences describing your experiment (up to 140 characters).

Description

Enter the detailed description for your experiment.

Quick Help

Search experiment items

Saved Datasets

- My Datasets
  - Titanic Dataset
  - Titanic\_Test.csv
  - Titanic\_Train.csv
- Samples
  - Adult Census Income Bi...
  - Airport Codes Dataset
  - Automobile price data (...)
  - Bike Rental UCI dataset
  - Bill Gates RGB Image
  - Blood donation data
  - Book Reviews from Am...
  - Breast cancer data
  - Breast Cancer Features
  - Breast Cancer Info
  - CRM Appetency Labels ...
  - CRM Churn Labels Shared
  - CRM Dataset Shared
  - CRM Upselling Labels S...

清理資料

Mini Map

Titanic\_Train.csv

Execute Python Script

Select Columns in Dataset

Two-Class Logistic Regression

Split Data

Train Model

Score Model

Evaluate Model

RUN HISTORY   SAVE   SAVE AS   DISCARD CHANGES   RUN   SET UP WEB SERVICE   PUBLISH TO GALLERY

```
graph TD; A[Titanic_Train.csv] --> B[Execute Python Script]; B --> C[Select Columns in Dataset]; C --> D[Two-Class Logistic Regression]; D --> E[Split Data]; E --> F[Train Model]; F --> G[Score Model]; G --> H[Evaluate Model]
```

KrisjenChen 123

# Advanced Version -- Python

Microsoft Azure Machine Learning Studio    Kristen-Free-Workspace ▾ ? 🔍 🌐 🎉 🚀

Kaggle Titanic Advance

Search experiment items

Experiment Properties

START TIME 2/26/2019 ...  
END TIME 2/26/2019 ...  
STATUS CODE Finished  
STATUS DETAILS None

Prior Run

Summary

Enter a few sentences describing your experiment (up to 140 characters).

Description

Enter the detailed description for your experiment.

Quick Help

Public Web Service

Titanic\_Train.csv

Execute Python Script

Select Columns in Dataset

Two-Class Logistic Regression

Split Data

Train Model

Score Model

Evaluate Model

Predictive Web Service [Recommended]

Retraining Web Service

RUN HISTORY SAVE AS DISCARD CHANGES RUN SET UP WEB SERVICE PUBLISH TO GALLERY

NEW

```
graph TD; A[Titanic_Train.csv] --> B[Execute Python Script]; B --> C[Select Columns in Dataset]; C --> D[Two-Class Logistic Regression]; D --> E[Split Data]; E --> F[Train Model]; F --> G[Score Model]; G --> H[Evaluate Model];
```

KrisjenChen 124

# Advanced Version -- Python

Microsoft Azure Machine Learning Studio   Kristen-Free-Workspace ▾ ? 🔍 😊 🚙

### [Web Service] Step 1

Kaggle Titanic Advance [Predictive Exp.]   In draft   Draft saved at 上午1:24:31

```
graph TD; A[Titanic_Train.csv] --> B[Execute Python Script]; A --> C[Select Columns in Dataset]; B --> D[Score Model]; C --> D; D --> E[Web service output]; D -- feedback --> F[Kaggle Titanic Advance [train...]]
```

Properties Project ▾

**Experiment Properties**

- START TIME -
- END TIME -
- STATUS CODE InDraft
- STATUS DETAILS None

**Summary**  
Enter a few sentences describing your experiment (up to 140 characters).

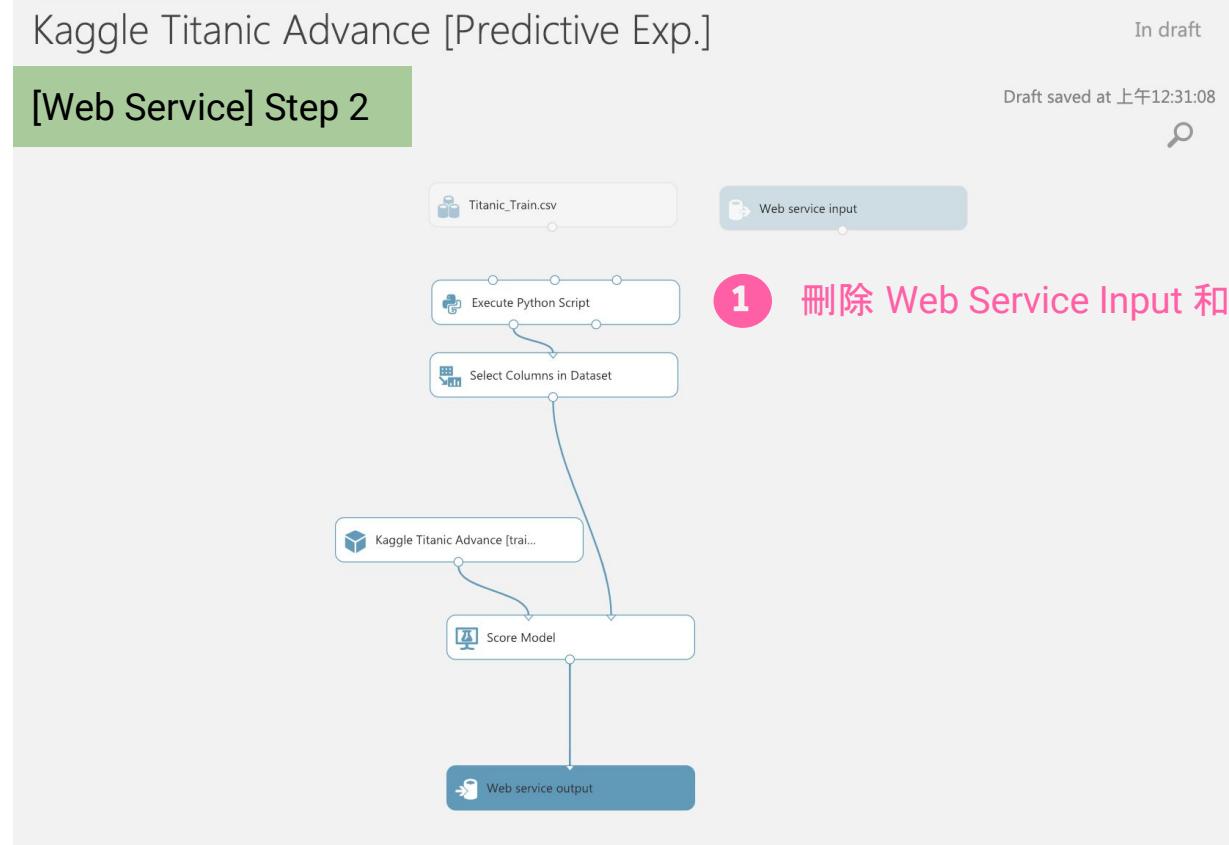
**Description**  
Enter the detailed description for your experiment.

Quick Help

NEW RUN HISTORY SAVE SAVE AS DISCARD CHANGES RUN DEPLOY WEB SERVICE PUBLISH TO GALLERY

KrisenChen 125

# Advanced Version -- Python



# Advanced Version -- Python

Kaggle Titanic Advance [Predictive Exp.] In draft Saving...

[Web Service] Step 2

```
graph TD; A[Titanic_Train.csv] --> B[Select Columns in Dataset]; C[Kaggle Titanic Advance [trai...]] --> D[Select Columns in Dataset]; B --> E[Execute Python Script]; D --> E; E --> F[Score Model]; F --> G[Web service output]; B --> H[Web service input]
```

2 新增一個篩選欄位接上 Titanic\_Train

# Advanced Version -- Python

Microsoft Azure Machine Learning Studio

[Web Service] Step 2 Training experiment Predictive experiment Kaggle Titanic Advance [Predictive Exp.] In draft Draft saved at: 上午9:21:34

Properties Project Select Columns in Dataset

Select columns

Available columns: PassengerId, Survived

Selected columns: Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked

3 選擇 Python Script 中會處理到的變數

[Note]

- 這裡選擇欄位是給 Web Service Input 要丟進來的
- PassengerId 對於 Model 來說沒有意義，所以不用選擇
- 然後因為我們的 Web Service 是要預測 Survived，所以當然不能把 Survived 選過來

RUN HISTORY SAVE SAVE AS DISCARD CHANGES RUN DEPLOY WEB SERVICE PUBLISH TO GALLERY

Yifan Chen 128

# Advanced Version -- Python



# Advanced Version -- Python

Microsoft Azure Machine Learning Studio   Kristen-Free-Workspace

[Web Service] Step 2   Training experiment   Predictive experiment   Kaggle Titanic Advance [Predictive Exp.]   Failed 2/26/2019 2:35:46 AM

Saved Datasets   My Datasets   Titanic Dataset, Titanic\_Test.csv, Titanic\_Train.csv   Samples   Adult Census Income Bi..., Airport Codes Dataset, Automobile price data ..., Bike Rental UCI dataset, Bill Gates RGB Image, Blood donation data, Book Reviews from Ama..., Breast cancer data, Breast Cancer Features, Breast Cancer Info, CRM Appetency Labels ..., CRM Churn Labels Shared

Titanic\_Train.csv → Select Columns in Dataset → Web service input

Kaggle Titanic Advance [train... → Execute Python Script → Select Columns in Dataset → Score Model → Web service output

Properties   Project   Select Columns in Dataset   Selected columns: Column names: Survived,Sex,Age,Title,Maste...

START TIME: 2/26/2019 ...  
END TIME: 2/26/2019 ...  
ELAPSED TIME: 0:00:26.262  
STATUS CODE: Failed  
STATUS DETAILS: requestId = 1550cf1287...; errorComp...; taskStatusC...; {"Exception": {"ErrorId": "... 0001: Column with name or index \"Survived\" not found"})}Error.

Quick Help: Selects columns to include or exclude from a dataset in an operation. Formerly known as Project Columns. (more help...)

NEW   RUN HISTORY   SAVE   SAVE AS   DISCARD CHANGES   RUN   DEPLOY WEB SERVICE   PUBLISH TO GALLERY

5 將原本的 Select Columns in Dataset 中的 Survived 刪除

KristenChian 130

# Advanced Version -- Python

Microsoft Azure Machine Learning Studio

Kristen-Free-Workspace

[Web Service] Step 2

Predictive experiment

Kaggle Titanic Advance [Predictive Expt]

Select columns

BY NAME

WITH RULES

Ticket\_C ✕ Ticket\_CA ✕ Ticket\_CASOTON ✕  
Ticket\_FC ✕ Ticket\_FCC ✕ Ticket\_Fa ✕  
Ticket\_LINE ✕ Ticket\_Null ✕ Ticket\_PC ✕  
Ticket\_PP ✕ Ticket\_PPP ✕ Ticket\_SC ✕  
Ticket\_SCA4 ✕ Ticket\_SCAH ✕ Ticket\_SCOW ✕  
Ticket\_SCPARIS ✕ Ticket\_SCParis ✕ Ticket\_SOC ✕  
Ticket\_SOP ✕ Ticket\_SOPP ✕ Ticket\_SOTONO2 ✕  
Ticket\_SOTONOQ ✕ Ticket\_SP ✕ Ticket\_STONO ✕  
Ticket\_STONO2 ✕ Ticket\_SWPP ✕ Ticket\_WC ✕  
Ticket\_WEP ✕ Family\_size ✕ Single\_family ✕  
Small\_family ✕ Big\_family ✕ Fare ✕ SibSp ✕  
Parch ✕ Survived ✕

Failed 2/26/2019 2:35:46 AM ✕

Properties Project

▪ Select Columns in Dataset

Select columns

BY NAME

WITH RULES

Ticket\_C ✕ Ticket\_CA ✕ Ticket\_CASOTON ✕  
Ticket\_FC ✕ Ticket\_FCC ✕ Ticket\_Fa ✕  
Ticket\_LINE ✕ Ticket\_Null ✕ Ticket\_PC ✕  
Ticket\_PP ✕ Ticket\_PPP ✕ Ticket\_SC ✕  
Ticket\_SCA4 ✕ Ticket\_SCAH ✕ Ticket\_SCOW ✕  
Ticket\_SCPARIS ✕ Ticket\_SCParis ✕ Ticket\_SOC ✕  
Ticket\_SOP ✕ Ticket\_SOPP ✕ Ticket\_SOTONO2 ✕  
Ticket\_SOTONOQ ✕ Ticket\_SP ✕ Ticket\_STONO ✕  
Ticket\_STONO2 ✕ Ticket\_SWPP ✕ Ticket\_WC ✕  
Ticket\_WEP ✕ Family\_size ✕ Single\_family ✕  
Small\_family ✕ Big\_family ✕ Fare ✕ SibSp ✕  
Parch ✕

not found"})Error.

Quick Help

Selects columns to include or exclude from a dataset in an operation. Formerly known as Project Columns.  
(more help...)

+

NEW

RUN HISTORY

SAVE

SAVE AS

DISCARD CHANGES

RUN

DEPLOY WEB SERVICE

PUBLISH TO GALLERY

KristenChian 131

The image shows a comparison between two ways of selecting columns in Azure ML Studio. The left dialog, titled 'Select columns', has 'WITH RULES' selected, listing columns like Ticket\_C, Ticket\_CA, Ticket\_CASOTON, etc., each preceded by a small 'x' icon. The right dialog, also titled 'Select columns', has 'BY NAME' selected, listing the same columns but preceded by a small checkmark icon. A pink arrow points from the left dialog to the right one. In the background, there's a 'Training experiment' tab, a 'Predictive experiment' tab, and a message 'Failed 2/26/2019 2:35:46 AM'. The bottom navigation bar includes 'RUN HISTORY', 'SAVE', 'SAVE AS', 'DISCARD CHANGES', 'RUN', 'DEPLOY WEB SERVICE', and 'PUBLISH TO GALLERY'.

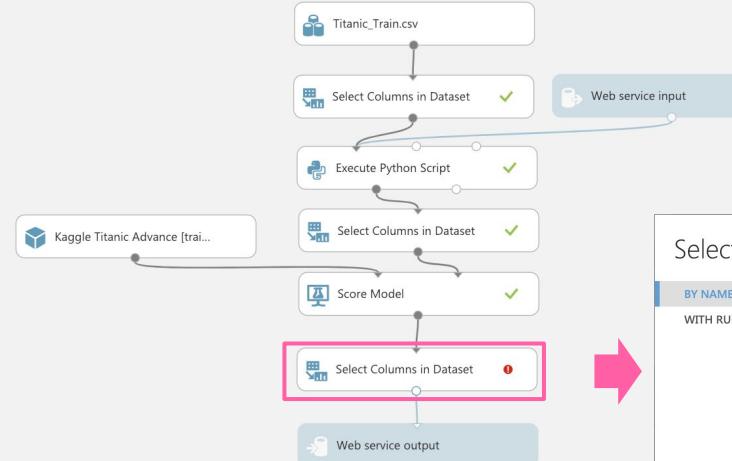
# Advanced Version -- Python

Kaggle Titanic Advance [Predictive Exp.]

[Web Service] Step 2

In draft

Saving...



Select columns

AVAILABLE COLUMNS	SELECTED COLUMNS
All Types <input type="button" value="search columns"/> <input type="button" value=""/>	All Types <input type="button" value="search columns"/> <input type="button" value=""/>
Sex	<input checked="" type="checkbox" value="Scored Labels"/>
Age	
SibSp	
Parch	
Fare	
Title_Master	
Title_Miss	
Title_Mr	
Title_Mrs	
Title_Officer	
Title_Royalty	
Embarked_C	
Embarked_Q	
Embarked_S	
Cabin_A	

62 columns available      1 columns selected

- 6 新增一個篩選欄位，並選擇 Scored Label  
Output 僅需呈現 Scored Label



# Advanced Version -- Python

Microsoft Azure Machine Learning Studio   Kristen-Free-Workspace ▾ ? ☺ 🔍

[Web Service] Step 2   Training experiment   Predictive experiment   Finished running ✓

Kaggle Titanic Advance [Predictive Exp.]

Titanic\_Dataset  
Titanic\_Test.csv  
Titanic\_Train.csv

Adult Census Income Bi...  
Airport Codes Dataset  
Automobile price data (...  
Bike Rental UCI dataset  
Bill Gates RGB Image  
Blood donation data  
Book Reviews from Ama...  
Breast cancer data  
Breast Cancer Features  
Breast Cancer Info  
CRM Appetency Labels ...  
CRM Churn Labels Shared

Titanic\_Train.csv  
Select Columns in Dataset  
Execute Python Script  
Kaggle Titanic Advance [trai...  
Select Columns in Dataset  
Score Model  
Select Columns in Dataset  
Web service input  
Web service output

Properties   Project   ▶

Experiment Properties

START TIME 2/26/2019 ...  
END TIME 2/26/2019 ...  
STATUS CODE Finished  
STATUS DETAILS None

Go to web service  
Prior Run

Summary

Enter a few sentences describing your experiment (up to 140 characters).

Description

Enter the detailed description for your experiment.

Quick Help

Run 執行

```
graph TD; A[Titanic_Train.csv] --> B[Select Columns in Dataset]; B --> C[Execute Python Script]; C --> D[Kaggle Titanic Advance [trai...]]; D --> E[Select Columns in Dataset]; E --> F[Score Model]; F --> G[Select Columns in Dataset]; G --> H[Web service output];
```

KristenChian 133

# Advanced Version -- Python

Microsoft Azure Machine Learning Studio   Kristen-Free-Workspace ▾ ? ☺ 🔍

[Web Service] Step 3   Training experiment   Predictive experiment   Finished running ✓

Kaggle Titanic Advance [Predictive Exp.]

Titanic\_Dataset  
Titanic\_Test.csv  
Titanic\_Train.csv

Adult Census Income Bi...  
Airport Codes Dataset  
Automobile price data (...  
Bike Rental UCI dataset  
Bill Gates RGB Image  
Blood donation data  
Book Reviews from Ama...  
Breast cancer data  
Breast Cancer Features  
Breast Cancer Info  
CRM Appetency Labels ...  
CRM Churn Labels Shared

Titanic\_Train.csv  
Select Columns in Dataset  
Execute Python Script  
Kaggle Titanic Advance [trai...  
Select Columns in Dataset  
Score Model  
Select Columns in Dataset  
Web service input  
Web service output

Properties   Project   ▶ Experiment Properties   START TIME 2/26/2019 ...  
END TIME 2/26/2019 ...  
STATUS CODE Finished  
STATUS DETAILS None  
Go to web service  
Prior Run  
Summary  
Enter a few sentences describing your experiment (up to 140 characters).  
Description  
Enter the detailed description for your experiment.  
Quick Help

Deploy Web Service

```
graph TD; A[Titanic_Train.csv] --> B[Select Columns in Dataset]; B --> C[Execute Python Script]; C --> D[Score Model]; D --> E[Select Columns in Dataset]; E --> F[Web service output];
```

KrisjenChia 134

# Advanced Version -- Python

Microsoft Azure Machine Learning Studio

Kristen-Free-Workspace    

kaggle titanic advance [predictive exp.]

DASHBOARD CONFIGURATION

General [New Web Services Experience preview](#)

Published experiment

[View snapshot](#) [View latest](#)

Description

No description provided for this web service.

API key

Olj7

Default Endpoint

API HELP PAGE TEST APPS LAST UPDATED  

	REQUEST/RESPONSE	BATCH EXECUTION	TEST	APPS	LAST UPDATED
			<a href="#">Test preview</a>	<a href="#">Excel 2013 or later</a>   <a href="#">Excel 2010 or earlier workbook</a>	2/26/2019 2:43:22 AM
			<a href="#">Test preview</a>	<a href="#">Excel 2013 or later workbook</a>	2/26/2019 2:43:22 AM

 NEW  DELETE

[Web Service] Step 4



135

# Advanced Version -- Python

kaggle titanic advance [predictive exp.]

DASHBOARD CONFIGURATION

General New Web Services Experience [preview](#)

Published experiment

[View snapshot](#) [View latest](#)

Description

No description provided for this web service.

API key

Olj7

Default Endpoint

API HELP PAGE TEST APPS LAST UPDATED

REQUEST/RESPONSE TEST [Test preview](#)

BATCH EXECUTION TEST [Test preview](#)

選 Test 做測試

[Web Service] Step 4

Kristen-Free-Workspace ? ☺ ☺ ☺

+ NEW DELETE

# Advanced Version -- Python

Enter data to predict

測試

PCLASS

1

PARCH

1

NAME

Ostby, Miss. Helene Ragnhild

TICKET

113509

SEX

female

FARE

61.9792

AGE

22

CABIN

B36

SIBSP

0

EMBARKED

C

# Advanced Version -- Python

Microsoft Azure Machine Learning Studio    Kristen-Free-Workspace ▾    ?    🔍    😊    🚙

kaggle titanic advance [predictive exp.]

DASHBOARD    CONFIGURATION

General    New Web Services Experience [preview](#)

Published experiment

[View snapshot](#)    [View latest](#)

Description

No description provided for this web service.

API key

JqO1bPzuYd8xK2Fpj0WyNGe39/4IMcb2tmtBbvfUvI4uM895uQGBePNHiiC5JtINS1ePjSHY+C0W9Y4JD0Eyeg==

Default Endpoint

API HELP PAGE    TEST    APPS    LAST UPDATED

REQUEST/RESPONSE    Test [preview](#)

BATCH EXECUTION    Test [preview](#)

	Excel 2013 or later	Excel 2010 or earlier workbook	LAST UPDATED
<a href="#">Excel 2013 or later workbook</a>			2/27/2019 1:35:03 AM
<a href="#">Excel 2013 or later workbook</a>			2/27/2019 1:35:03 AM

成功 !!

✓ 'Kaggle Titanic Advance [Predictive Exp.]' test returned ["1"]...

DETAILS    i    CLOSE    X

NEW    DELETE

KrisfenChian 138

# Note

# Reference

- Kaggle Titanic
  - [https://github.com/ahmedbesbes/How-to-score-0.8134-in-Titanic-Kaggle-Challenge/blob/master/article\\_1.ipynb](https://github.com/ahmedbesbes/How-to-score-0.8134-in-Titanic-Kaggle-Challenge/blob/master/article_1.ipynb)
- Cross Validation
  - <https://blog.contactsunny.com/data-science/different-types-of-validations-in-machine-learning-cross-validation>
- Azure ML Execute Python Script
  - <https://docs.microsoft.com/zh-tw/azure/machine-learning/studio/execute-python-scripts>

# Jupyter Notebook

- 執行
  - 執行此 cell 內容 : Ctrl + Enter
  - 執行此 cell 內容 : Shift + Enter
  - 執行此 cell 內容 : Alt + Enter , Option + Enter
- 儲存 : Ctrl + S , Command + S
- 編輯模式(綠色)
  - 修改 cell 內容
  - 啟動編輯 : Enter
- 指令模式(藍色)
  - 使用快捷鍵執行指令
  - 啟動指令 : Esc

# Jupyter Notebook

- 執行
  - 執行此 cell 內容 : Ctrl + Enter
  - 執行此 cell 內容 : Shift + Enter
  - 執行此 cell 內容 : Alt + Enter , Option + Enter
- 儲存 : Ctrl + S , Command + S
- 編輯模式(綠色)
  - 修改 cell 內容
  - 啟動編輯 : Enter
- 指令模式(藍色)
  - 使用快捷鍵執行指令
  - 啟動指令 : Esc

# Jupyter Notebook

- 常用快速鍵(無大小寫之分)
  - 上方新增一個 cell:**A**
  - 下方新增一個 cell:**B**
  - 刪除一個 cell:**D**
  - 複製一個 cell:**C**
  - 貼上一個 cell:**V**
  - 啟動 Code 模式:**Y** ( 程式碼撰寫及執行程式 )
  - 啟動 Markdown 模式:**M** ( 可以做筆記; [Markdown語法介紹](#))